TECHNISCHE
UNIVERSITÄT
DARMSTADT

AUTOMATIC SHORT ANSWER GRADING *using* NEURAL MODELS

Examining Adversarial Robustness and Elaborated Feedback Generation

Vom Fachbereich Elektrotechnik und Informationstechnik
der Technischen Universität Darmstadt
zur Erlangung des akademischen Grades eines
Doktor-Ingenieurs (Dr.-Ing.)
genehmigte Dissertation

von

ANNA MARIE FILIGHERA, M.SC.

Vorsitz: Prof. Dr. rer. nat. Florian Steinke
Referent: Prof. Dr.-Ing. Ralf Steinmetz
Korreferent: Prof. Dr.-Ing. Ulrik Schroeder

Tag der Einreichung: 12. Mai 2023
Tag der Disputation: 13. Juli 2023

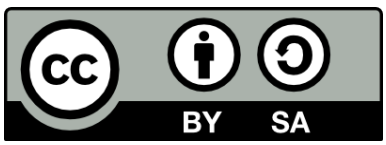
Darmstadt 2023

Anna Marie Filighera, M.Sc.: *Automatic Short Answer Grading Using Neural Models.*
Examining Adversarial Robustness and Elaborated Feedback Generation
Technical University of Darmstadt, Darmstadt

Jahr der Veröffentlichung der Dissertation auf TUprints: 2023
Tag der mündlichen Prüfung: 13. Juli 2023

Dieses Dokument wird bereitgestellt von: This document is provided by:
tuprints, E-Publishing-Service der Technischen Universität Darmstadt
http://tuprints.ulb.tu-darmstadt.de
tuprints@ulb.tu-darmstadt.de

Bitte zitieren Sie dieses Dokument als: Please cite this document as:
URN: urn:nbn:de:tuda-tuprints-243945
URI: https://tuprints.ulb.tu-darmstadt.de/id/eprint/24394



ABSTRACT

High-quality feedback is essential for learners. It reveals misconceptions, knowledge gaps and improvement opportunities. Asking short-answer questions and giving elaborated feedback on the learners' responses is highly effective in increasing not only their understanding of the material but also their ability to transfer the knowledge to new contexts. However, providing even basic feedback, such as verifying correctness, is time-consuming. For this reason, neural feedback systems have risen in popularity in recent years. While such systems have matured to achieve high grading accuracy on some datasets, their decision process is opaque and their behavior when confronted with out-of-training-distribution data remains underexplored. Thus, the first research question posed in this thesis concerns current state-of-the-art grading models' robustness to adversarial examples – answers crafted to fool the grading model. The second research question explores how grading systems can be expanded to provide elaborated feedback explaining learners' mistakes instead of merely verifying correctness. In total, we make four contributions to these research questions.

First, we investigate grading models' robustness to adversarial examples crafted by students as well as an existing automatic attack. We show that current models are generally vulnerable to adversarial attacks and provide evidence that their predictions are at least partially based on spurious correlations. However, we also find that existing adversarial attacks are difficult to employ in typical summative assessment scenarios. Therefore, we propose an adversarial attack tailored to summative assessments as our second contribution. We demonstrate the attack's effectiveness on multiple models and domains and empirically evaluate manipulated responses with human experts.

Our third contribution consists of the bilingual Short Answer Feedback dataset. In contrast to existing datasets, it contains elaborated feedback in addition to verification feedback. We annotated learner responses from three domains spanning college-level and life-long learning. We demonstrate that this novel task challenges current state-of-the-art models. We provide an evaluation framework and benchmark models to lay the groundwork for research in this field. Though the feedback generated by the benchmark models is imperfect, we observed positive effects on learning outcomes compared to no feedback and even human feedback conditions in a college course field study.

Finally, we propose an unsupervised elaborated feedback generation method for domains where costly data annotation is infeasible as our fourth contribution. It aims to find small counterfactual changes to students' responses that would have led the grading model to classify them as correct instead. These changes can be considered concrete improvement suggestions in the student's own words. We compare four counterfactual generation approaches and find further evidence for the grading models' unreliability but also genuine improvements, indicating that such feedback may be feasible in the future.

Overall, this thesis provides insight into the robustness of neural Automatic Short Answer Grading systems to various forms of input manipulation. We also present evidence for the usefulness of even imperfect elaborated feedback models while providing the tools for further research on improved approaches. The garnered understanding can be helpful to practitioners seeking to employ grading systems more securely, understandably and safely.

KURZFASSUNG

Qualitativ hochwertiges Feedback ist für Lernende unerlässlich. Es zeigt Wissenslücken, Missverständnisse und Verbesserungsmöglichkeiten auf. Elaboriertes Feedback zu bearbeiteten Freitext-Aufgaben verbessert nicht nur das Verständnis des Lernmaterials, sondern unterstützt auch den Transfer des Gelernten auf neue Sachverhalte. Allerdings ist selbst die Rückmeldung von einfachem Feedback, wie beispielsweise der Verifizierung der Korrektheit einer Antwort, zeitaufwändig.

Aus diesem Grund haben neuronale Feedbackysteme in den letzten Jahren an Popularität gewonnen. Solche Systeme erreichen zwar inzwischen eine hohe Bewertungsgenauigkeit auf manchen Datensets, ihr Entscheidungsprozess ist jedoch nicht für Menschen nachvollziehbar. Darüber hinaus ist ihre Genauigkeit auf Daten, die nicht aus der Trainingsverteilung stammen, noch unerforscht. Die erste Forschungsfrage dieser Arbeit bezieht sich daher auf die Robustheit aktueller Korrekturmodelle gegenüber „Adversarial Examples“, d.h. Antworten, die das Modell täuschen sollen. Die zweite Forschungsfrage befasst sich damit, wie Korrektursysteme erweitert werden können, sodass sie neben der Verifikation der Richtigkeit einer Antwort auch elaborierte Erläuterungen der gemachten Fehler generieren. Insgesamt leisten wir vier Beiträge zu diesen Forschungsfragen.

Als ersten Beitrag untersuchen wir die Robustheit von Korrektursystemen gegenüber Adversarial Examples, die von Studierenden und von einem bestehenden Angriff generiert wurden. Wir zeigen, dass aktuelle Modelle im Allgemeinen anfällig für Adversarial Examples sind und ihre Vorhersagen zumindest teilweise auf Scheinkorrelationen beruhen. Wir stellen jedoch auch fest, dass existierende Angriffe in typischen Leistungsüberprüfungen nur schwer anwendbar sind. Daher entwickeln wir als zweiten Beitrag einen Angriff, der speziell auf summative Leistungsüberprüfungen zugeschnitten ist. Wir demonstrieren die Wirksamkeit des Angriffs in diversen Domänen und evaluieren die manipulierten Antworten empirisch mit menschlichen Experten.

Unser dritter Beitrag besteht aus dem bilingualen *Short Answer Feedback* Datensatz. Im Gegensatz zu bestehenden Datensätzen enthält er neben einfachem auch elaboriertes Feedback. Der Korpus umfasst drei Domänen aus der universitären und Erwachsenenbildung. Wir zeigen, dass elaborierte Feedbackgenerierung eine Herausforderung für aktuelle Modelle darstellt. Um die Grundlage für künftige Forschung in diesem Bereich zu schaffen, entwickeln wir ein Evaluationsframework und trainieren Benchmark-Modelle. Obwohl das von den Modellen generierte Feedback nicht perfekt war, beobachteten wir positive Effekte auf den Lernerfolg von Studierenden in einer Feldstudie im Vergleich zu Kontrollgruppen, welche kein Feedback oder gar menschliches Feedback erhielten.

Schließlich stellen wir als vierten Beitrag eine unüberwachte Methode zur Generierung von elaboriertem Feedback für Bereiche vor, in denen eine kostspielige Datenannotation nicht durchführbar ist. Der Ansatz sucht nach kleinen Modifizierungen der

Antworten von Lernenden, die zu einer besseren Bewertung durch das Korrekturmodell geführt hätten. Diese Änderungen können als konkrete Verbesserungsvorschläge für die Antwort des Lernenden betrachtet werden. Wir vergleichen vier Methoden und finden weitere Indizien für die fehlende Zuverlässigkeit der Korrektursysteme, aber auch echte Verbesserungen, die darauf hinweisen, dass solches Feedback in Zukunft möglich sein könnte.

Zusammenfassend bietet diese Arbeit einen Einblick in die Robustheit von neuronalen Systemen zur automatischen Bewertung von Kurzantworten gegen verschiedene Antwortmanipulationsstrategien. Darüber hinaus liefern wir Anhaltspunkte dafür, dass selbst unvollkommene Feedbackgenerierungsmodelle Lernen unterstützen können und legen die Grundlage für weitere Forschung an verbesserten Methoden. Die in dieser Thesis gewonnenen Erkenntnisse können Lehrenden helfen, automatische Korrektursysteme sicherer und für Lernende verständlicher einzusetzen.

CONTENTS

1 INTRODUCTION **5**
1.1 Overview of State of the Art . . . 6
1.2 Research Questions and Challenges . . . 7
1.3 Contributions . . . 9

2 BACKGROUND AND RELATED WORK **13**
2.1 Automatic Short Answer Grading . . . 13
2.3 Adversarial Examples . . . 14
2.4 Gaming Educational Systems . . . 17
2.5 Elaborated Feedback Systems . . . 18
2.6 Research Gaps and Summary . . . 19

3 ADVERSARIAL ATTACKS ON AUTOMATIC GRADING **21**
3.1 Attack Design Considerations . . . 21
3.1.1 Necessary Expertise & Resources during Testing Time . . . 22
3.1.2 Model Access . . . 22
3.1.3 Risk of Detection . . . 23
3.1.4 Computation and Time Budget . . . 23
3.1.5 Class Equivalence . . . 23
3.2 Universal Trigger Attack . . . 25
3.2.1 Approach . . . 25
3.2.2 Target Grading Models . . . 26
3.2.3 Experimental Settings . . . 27
3.2.4 Attack Evaluation . . . 28
3.2.5 Interpretation of Results & Limitations . . . 30
3.3 Adversarial Adjective and Adverb Insertion . . . 32
3.3.1 Approach . . . 32
3.3.2 Target Models . . . 34
3.3.3 Experimental Settings . . . 36
3.3.4 Attack Evaluation . . . 38
3.3.5 Interpretation of Results & Limitations . . . 47
3.4 Student Attacks . . . 50
3.4.1 Methodology . . . 50
3.4.2 Target Domain & Grading Model . . . 51
3.4.3 Participant Characteristics . . . 51
3.4.4 Results . . . 52
3.4.5 Interpretation of Results & Limitations . . . 54
3.5 Discussion of Attacks . . . 57

4 ELABORATED FEEDBACK GENERATION **61**
4.1 Benchmark Design Considerations . . . 61
4.2 The Short Answer Feedback Corpus . . . 66
4.2.1 Data Collection . . . 66
4.2.2 Data Annotation . . . 69
4.2.3 Corpus Statistics . . . 71
4.2.4 Reliability and Validity . . . 73
4.2.5 Interpretation & Limitations . . . 74
4.3 Supervised Feedback Generation . . . 76
4.3.1 Approaches . . . 76
4.3.2 Experimental Settings . . . 77
4.3.3 Feedback Evaluation . . . 78
4.3.4 Field Evaluation . . . 81
4.3.5 Interpretation of Results & Limitations . . . 93
4.4 Unsupervised Feedback Generation . . . 95
4.4.1 Counterfactual Feedback Approaches . . . 96
4.4.2 Experimental Settings . . . 99
4.4.3 Feedback Evaluation . . . 101
4.4.4 Interpretation of Results & Limitations . . . 105
4.5 Discussion of Feedback Generation . . . 107

5 SUMMARY, CONCLUSIONS, AND OUTLOOK **109**
5.1 Conclusions . . . 110
5.2 Outlook . . . 112

BIBLIOGRAPHY **115**

A APPENDIX **135**
A.1 List of Acronyms . . . 135
A.2 Universal Trigger Attack Hyperparameters & Triggers . . . 135
A.3 Questionnaires . . . 137
A.4 Examples of Generated Feedback . . . 137
A.5 Linear Mixed Models' Statistics . . . 137

B SUPERVISED STUDENT THESES **151**

C AUTHOR'S PUBLICATIONS **153**

D ERKLÄRUNGEN LAUT PROMOTIONSORDNUNG **157**

PREVIOUSLY PUBLISHED MATERIAL

This thesis contains work previously published in conference proceedings and preprint archives. The publications relevant to each section are listed in Table 1. None of the publications were directly reprinted in this thesis. A complete list of the author's publications can be found in Appendix C. Figures, tables and examples adapted or taken from previous publications have their source referenced explicitly in the corresponding caption. Grammarly [58] was utilized to improve grammar, syntax, spelling and word choice throughout this thesis. The German abstract was constructed with the support of DeepL [29]. No other AI-based generative tools were utilized in the preparation of this thesis.

Science is most often a collaborative effort. Thus, this section details the contributions of collaborators and co-authors toward the work described in this thesis. The pronoun "we" will be used for the remainder of this thesis to acknowledge their contributions. If not stated otherwise, all collaborators are or were affiliated with the Multimedia Communications Lab of the Technical University of Darmstadt.

Frequent discussions of recent developments in educational technologies and computational linguistics with Tim Steuer and Christoph Rensing inspired the idea to work on Automatic Short Answer Grading and the resulting research questions. This thesis's introduction (Chapter 1), related work (Chapter 2) and conclusion (Chapter 5) contain elements of all the author's previous publications in the Automatic Short Answer Grading and feedback generation fields [23, 42–47], as the motivation, literature analysis and research context were iteratively refined through these publications. Throughout the entire thesis, Viktor Pfanschilling – affiliated with the Artificial Intelligence

Sections	Publications
1, 2, 5	[23], [42], [43], [44], [45], [46], [47]
3.1, 3.5	[43], [45], [46]
3.2	[45]
3.3	[43]
3.4	[46]
4.1, 4.5	[44], [47]
4.2, 4.3	[44]
4.4	[47]

Table 1: Author's previous publications included in each section.

and Machine Learning Lab at the Technical University of Darmstadt – sharpened my understanding and ideas in lively debates on current trends in machine learning.

Chapter 3 investigates the vulnerability of current Automatic Short Answer Grading approaches to adversarial input manipulations. Regular meetings with Christoph Rensing and Tim Steuer supported me in conceptualizing design criteria for adversarial attacks and provided valuable feedback on proposed methodologies. In addition to Philipp Müller, they also revised the manuscript for publication [45].

The adjective and adverb attack introduced in Section 3.3 was developed jointly with Sebastian Ochs during his bachelor's thesis. Inspired by the promising results obtained in the thesis, Sebastian expanded the experiments to additional datasets and models in the context of a student lab under my supervision. Tim Steuer, Thomas Tregel and I designed the expert evaluation of the adversarial examples produced by the attack. I then conducted the evaluation with valuable technical assistance from Tim Steuer. We published the attack and evaluation study [43].

Christoph Rensing provided the opportunity to conduct the student attack study in one of his lectures at the Technical University of Darmstadt. All questionnaires for the study were developed in discussion with Tim Steuer and Christoph Rensing, who offered invaluable methodological advice on the study design and data analysis. Stephan Tittel assisted the study with technical support of a web server that the students could access. We published the study's results [46].

Chapter 4 introduces the elaborated feedback generation task, where the verification provided by Automatic Short Answer Grading is expanded to include an explanation of the student's mistakes. The dataset construction was a massively collaborative effort for each of the three domains. Julian Prommer from the university's didactic E-Learning Team (HDA) assisted us in the design of a data collection pipeline conforming to data protection guidelines. Ralf Steinmetz provided the opportunity to collect data in one of his communication network lectures for the English subset of the data. For the job training data, I collaborated with Dominic Lenhart and Robert Lokaiczyk from *wer denkt was* on the data collection material and data annotation guidelines. I then supported Dominic Lenhart in supervising the data collection and annotation process. The social security data was collected in the context of a Software Campus project I led. Here, Maria Walch and Peter Schichtel from *IAV* supported me in the conception of the annotation study and provided valuable feedback on the legal aspects of the data. Ivan Georg contributed vital knowledge of German social law to the guideline construction process and proved invaluable during the data annotation, which was done by Ivan Georg and Sebastian Ochs. Ivan and Sebastian contributed significantly to improving the initial annotation guidelines.

Siddharth Singh Parihar and I drafted the initial annotation guidelines for the English data in the context of his master's thesis. We then iteratively refined the guide-

lines in frequent discussions with Nikhil Kumar Patel, a student assistant I supervised. Nikhil Kumar Patel and Siddharth Singh Parihar annotated the collected student responses based on the refined guidelines. Sebastian Ochs and I trained a subset of the supervised feedback generation models presented in this thesis and published them with parts of the data in Filighera et al. [44]. Mehmet Can Kivanc and João Henrique de Araújo Kröger assisted in making the final dataset easily accessible online and contributed additional models in the context of a student lab I supervised. Finally, Kaan Atacan, Erik Levin Fischer and I developed the feedback pipeline based on GPT-3 in another student lab, which I then evaluated for this thesis.

The field study evaluating the use of an elaborated feedback generation model in a communication network lecture was a collaborative effort with Wolfgang Ellermeier's psychology department at the Technical University of Darmstadt. I had proposed the field study to them and we jointly supervised the master's thesis of Marieke Fischer, a psychology student who was tasked with developing the study materials and conducting the study. Wolfgang Ellermeier's lab provided valuable psychological expertise and methodological feedback on the study design, especially regarding the conformation to ethical standards and data analysis. In weekly meetings with Marieke Fischer and Philipp Hinderer, who supported the study from the technical side during his master's thesis, we refined the study material and analysis of the resulting data. Tim Steuer and Thomas Tregel also provided valuable conceptual feedback on this study.

Finally, the counterfactual feedback generation methods were developed with Joël Nzalakanda Tschesche in the context of his bachelor's thesis I supervised. We later refined the approaches for publication [47], where I also included an expert evaluation of the generated feedback. Lisa Werner, Tim Steuer and Thomas Tregel supported us with manuscript revisions and conceptual feedback.

1

INTRODUCTION

Assessing what a learner has understood has long been a vital component of education. Whether it be through exams, interviews or exercise assignments, collecting data on students' learning progress is essential to uncovering misconceptions, knowledge gaps and improvement opportunities. Knowledge states captured in summative assessments can also be communicated to stakeholders, such as parents or potential employers, to indicate learners' competencies. Additionally, assessments can inform instructors, curriculum designers and policymakers on the effectiveness of employed teaching methods, thus shaping the education system for future generations. Consequently, assessment methods significantly impact students' lives and society as a whole [93].

While a variety of assessment methods exist, many involve asking questions and observing the students' responses. From closed formats, such as multiple-choice or cloze questions, to open questions, like essay writing prompts, various competencies and types of knowledge can be captured. Here, short answer questions have gained popularity for their ability to test recall of knowledge in contrast to the mere recognition tested in closed formats [75, 84, 93, 133]. They require students to structure and formulate a response in their own words with a focus on content instead of writing style. Most often, the expected response length is between a phrase and a paragraph long and the question is somewhat objectively gradable [22]. While short answer questions are easily constructed and – especially with corresponding feedback – improve long-term retention [133], their grading is complex. The variability of natural language and lack of pre-defined structure require each individual response to be read and interpreted carefully. The concentration level needed to grade and provide feedback to short answer questions properly is hard to maintain for extended periods. Intra-grader consistency is further complicated by ordering effects, mood, fatigue and bias [22].

For this reason, Automatic Short Answer Grading (ASAG) systems have garnered significant research interest in the last decades [22, 68, 76, 174], with commercial systems promising immediate and personalized feedback[1]. While current neural network-based methods are approaching human performance on some datasets [150], their decision process is opaque. This is problematic for two reasons. First, feedback is more helpful to students if it is understandable and the feedback's source is trusted [169].

1 https://web.archive.org/web/20221004235134/https://new.assistments.org/individual-resource/quick-comments-beta [accessed May 11, 2023]

Only receiving a grade without an explanation is likely insufficient to foster the necessary understanding and trust. Second, a model may make accurate predictions and perform well on a given dataset but make its decisions for the wrong reasons. For example, the usage of proper punctuation symbols may correlate with better grades in a dataset if students who are good at writing also tend to answer questions more correctly. However, the number of commas used in a response is not in a causal relation with the response's correctness and, thus, not a robust feature. Nevertheless, current approaches would utilize it for its predictive power. Such spurious correlations may lead to unjust classifications later on or even be exploited to manipulate the model's predictions willfully. Consequently, this thesis aims to investigate the robustness of current ASAG approaches to various kinds of input manipulation and develop methods that provide a human-legible explanation of assigned grades.

1.1 OVERVIEW OF STATE OF THE ART

The goal of Automatic Short Answer Grading (ASAG) is for models to evaluate the factual correctness and completeness of learner responses to short answer questions. In the past, this was often done by handcrafting grading rules that can then be automatically applied to responses or their algorithmically constructed canonical representations [85, 168]. However, it is difficult to account for all possible variations of expressions manually. Thus, the ASAG field shifted towards manually crafted features, such as part-of-speech distributions, in combination with machine learning models. These approaches often compared responses to plain text reference answers or clustered them with other similar student responses [37, 64, 108, 136]. While these machine learning approaches with manual features are more flexible than their rule-based counterparts, their grading accuracy is often unsatisfactory. In recent years, they have been outperformed by deep learning models that construct a mathematical representation of language from observing which words appear in similar contexts in extensive collections of texts [20, 59]. This representation is also called embedding. On some datasets, such methods are approaching human grading accuracy [150].

However, recent research in Natural Language Processing (NLP) and computer vision indicates that at least part of the performance increase of neural networks is based on spurious correlations in the dataset that result in non-robust features [69]. Such features are generally undesirable as they can lead to misclassifying out-of-distribution data and present an exploitable vulnerability. Especially for automatic grading models, non-robust features are problematic as students may be disadvantaged based on factors that are not in a causal relationship with the correctness of their response, e.g., dialect or writing style. Conversely, students can also cheat by exploiting known weak-

nesses and, thus, receive better grades. Studies with traditional non-computer-based assessments show that cheating is prevalent in our current education system, with large-scale reviews reporting that a majority of students cheat during their college studies [73, 78, 166]. While cheating is currently most often done by copying answers from fellow students, we expect students to be similarly willing to cheat automatic grading systems given the necessary know-how.

The discovery of non-robust features has given rise to a tremendous amount of research on subtle modifications one can perform on input to achieve a desired – but wrong – classification, such as adding humanly imperceptible noise to images [56]. Modifying samples in such a manner is called an adversarial attack and the resulting sample is an adversarial example. While many adversarial attacks for text have been proposed in the last few years [175, 180, 184], they are typically not designed for summative assessments in controlled environments and assume favorable conditions for the attacker, such as access to the target model's inner workings and unlimited time. Yet, students are expected to have limited time and resources during typical summative assessments, even though they may have time to prepare an attack beforehand. Generally, research on the robustness of automatic grading models is needed, especially considering realistic assessment scenarios. This is the first research gap we address.

Finally, state-of-the-art ASAG systems only produce a numerical grade or categorical label, indicating whether a given response was correct. This type of feedback is called verification feedback. Pedagogical research suggests that verification feedback alone is suboptimal [144, 169, 170]; Students learn more if they understand where and why they have made mistakes and how to improve in the future [144, 169, 170]. Thus, it would be beneficial if ASAG models would not only verify a response's correctness but also explain why a particular grade was assigned. The explanation should cover where the student erred and why their statement was incorrect. Such an explanation is a type of elaborated feedback. While explaining the decisions of neural networks is an ongoing area of research for other NLP tasks [91, 116], it is yet unexplored for neural ASAG models. This is the second research gap we address.

1.2 RESEARCH QUESTIONS AND CHALLENGES

The main goal of this work is to foster understandable and robust feedback generation for short answer questions. For this purpose, we first investigate how robust current ASAG models are in the face of out-of-training-distribution data. This is especially important for summative assessments, where the goal is to evaluate students' learning progress at the end of a course. Then, we explore methods to provide explainable and

understandable elaborated feedback instead of only a grade. This is especially important for formative assessments, which aim to improve students' learning via feedback. Thus, we address the following research questions:

Research Question 1: *How robust are current state-of-the-art ASAG models to adversarial input manipulation?*

In the scope of this research question, we define robustness mainly by the extent to which a model's grading accuracy can be degraded with a limited set of adversarial modifications. Our definition conforms to how the term is predominantly used in related work [69, 160]. We also consider the scope of the modification performed, e.g., how many words were added to a response, how obvious it is to humans and how long the search took as secondary criteria. Factors dependent on the concrete assessment setup, such as how easy it would be to access the model's code or training data, are out of the scope of this research question and are assumed to be unavailable to the attacker.

Currently, the most successful architectures for nearly all NLP tasks are Transformers, a type of deep neural network that can simultaneously attend to every word in the input sequence. This attention mechanism enables them to handle long-range dependencies better than recurrent or convolutional neural networks, which were previously popular in the field. Transformers are currently the predominant model type with the best performance on ASAG benchmarks [20, 23, 55]. Thus, we focus our investigation on Transformer models and specifically do not consider older architectures.

Investigating the robustness of Transformer models to adversarial attacks is challenging as they tend to have millions to billions of parameters and require a sizable amount of computation resources to train and run. Therefore, the adversarial manipulation search space should be constrained to conserve energy and reduce the environmental impact of performed experiments. While the space of possible manipulations is immense, we are explicitly only interested in types of modifications that are universally applicable. Universal attacks are input-agnostic and, thus, suited for summative assessment scenarios. Here, students can prepare a strategy before the summative assessment that they only have to employ during the examination. For example, knowing that a grading model tends to give better grades to responses containing the word "definitely," one can incorporate it in as many responses as possible during the exam. Consequently, this research question aims to test to which extent the grading models have general weaknesses that can be exploited; In contrast, finding specific examples the model misclassifies, for example, by rearranging a given sentence syntactically, is not within the scope of this question.

Research Question 2: *How can ASAG methods be extended to provide response-specific elaborated feedback instead of mere verification?*

In the scope of this research question, we focus on content-level elaborated feedback addressing students' factual misconceptions in contrast to metacognitive feedback aiming to improve students' problem-solving strategies. The feedback should explain how the learner erred and why, but does not need to advise on employed learning strategies. In contrast to many existing intelligent tutoring systems [36], we investigate approaches not tailored to a specific domain. While an approach may require retraining for a different domain, it should at least be generally applicable without manual modeling effort.

Providing response-specific elaborated feedback for short answer questions is challenging for multiple reasons. First, no pre-defined structure constrains the space of possible answers. This makes anticipating possible mistakes beforehand tricky, necessitating a model that can generate feedback to novel responses autonomously. Second, at the time of writing, no public short answer datasets with elaborated feedback existed, inhibiting the development and testing of approaches.

Finally, evaluating generated textual feedback is complex even when gold standard reference feedback exists. While automatic text similarity metrics have been proposed, they are not yet an entirely reliable indicator of semantic similarity [60, 90]. Therefore, human judgment is advisable despite the cost and effort involved in human evaluation studies.

1.3 CONTRIBUTIONS

We make the following contributions to the robustness analysis of short answer grading models (RQ1):

Contribution 1: Our first goal is to examine whether an in-depth analysis of ASAG models' robustness to adversarial attacks is warranted. While adversarial examples have been found for many models and domains with varying success, the reason for their existence is not yet clear. The leading theory suggests that they result from highly predictive but incomprehensible features, indicating that their existence is at least partially tied to the task and dataset [69]. Since no work on adversarial attacks in the ASAG domain existed at the time, we approach this goal from two angles. First, we conduct an explorative study where we task university students with fooling an ASAG model (see Section 3.4). The students freely submitted responses to the model and received the predicted grade, similar to typical formative assessment scenarios. Based on the observed grading behavior, they identified and reported systematic weaknesses in the

model. Second, we expose state-of-the-art ASAG models to an existing attack based on identifying and appending meaningless trigger sequences to responses (see Section 3.2). While many adversarial attacks for text exist, we selected the Universal Trigger Attack [160] for its ease of use once trigger sequences are found – a student only has to add them to their responses in an exam. However, the senseless nature of the appended triggers would make this attack easy to detect in real-world scenarios. We demonstrate that ASAG models are vulnerable to adversarial attacks and that an in-depth analysis with an attack designed for summative assessment is warranted.

Contribution 2: The results of Contribution 1 reveal vulnerabilities of current grading models and establish that further research is necessary. However, appending a meaningless string of words to each answer in an exam would be unmistakable and, thus, risky in practice. Existing attacks are generally not designed for typical summative assessment scenarios, where there is a price to adversarial manipulations being detected. Additionally, the resources available to the student are also often limited during the assessment. For these reasons, our second contribution consists of an adversarial attack specifically designed and implemented for summative assessment scenarios in controlled environments (see Section 3.3). It is based on identifying adjectives and adverbs a model associates with a target class before the summative assessment and inserting them in natural positions during the exam. On short answer grading datasets, our attack outperforms the Universal Trigger attack and another powerful attack from related work [72] regarding the number of incorrect responses it can fool the model into grading as correct. We also contribute a statistical analysis of adjective and adverb frequencies in the dataset and the grading models' confidence to investigate reasons for the attack's success. Finally, we empirically evaluate the generated adversarial examples with human graders to explore how suspicious and natural they seem to humans.

We make the following contributions toward elaborated feedback generation (RQ2):

Contribution 3: We collect learner responses in three domains and annotate them with elaborated feedback, forming the Short Answer Feedback dataset (see Section 4.2). We demonstrate the annotations' reliability with a high inter-annotator agreement and validity via comparisons with external criterion variables measuring learning outcomes. After establishing the data quality, we design and implement a collection of supervised feedback generation approaches (see Section 4.3). Most often, they are trained to predict the response's correctness and generate an explanation jointly. While this does not necessarily prompt the model to explain its decision process but instead learn how to mimic elaborated feedback, it also does not expose potentially undesirable features to learners. Automatic evaluation measures showed that the models fell short of human

graders but performed well enough to justify a field study. Therefore, we employed one of the feedback generation models in a university lecture. We compared student motivation, acceptance and learning outcomes between three conditions. One group of students received the generated feedback, another group received feedback formulated by human graders, and the last condition received no feedback at all.

Contribution 4: Finally, we develop and implement unsupervised feedback generation approaches based on existing counterfactual generation techniques (see Section 4.4). The idea here is to find small changes to a student's response that would have led it to be graded more favorably and present the changes as concrete improvement suggestions to the student. Such an approach is especially beneficial for areas where expensive data collection, as in Contribution 3, is infeasible. In related work, counterfactual generators are typically evaluated by counting how often they produce counterfactuals that change the predicted class. However, since ASAG models are not necessarily reliable, we supplement this type of evaluation with human judgment. Specifically, an expert decides whether the generated improvement suggestions are genuine improvements.

2

BACKGROUND AND RELATED WORK

This chapter first introduces the Automatic Short Answer Grading (ASAG) task and provides an overview of existing approaches. It then introduces adversarial attacks and a summary of research investigating students' cheating behavior in digital assessments. Next, systems that generate elaborated feedback are discussed and finally, the research gaps addressed in this thesis are described in Section 2.6.

2.1 AUTOMATIC SHORT ANSWER GRADING

Short answer questions are a popular assessment format where learners formulate responses without predefined building blocks. Typically, responses are between a phrase and a paragraph long and are mostly assessed based on their content instead of their linguistic features [22]. The questions are usually objectively answerable, meaning that a fixed number of facts, relations or statements are expected to be present in a response. Thus, it is possible to determine the level of correctness of given answers.

2.2.1 *Classical Machine Learning Approaches*

Due to the focus on factual correctness and versatility of language, the task of automatically grading short answer questions has been a challenging and popular research area in the last decades [22, 50, 134]. Earlier approaches included the manual construction of grading rules and models [85, 168], clustering similar learner responses and assigning the same grade to entire clusters [15, 64, 179, 181] or manually defining similarity features used to compare responses [37, 101, 108, 136, 137, 148]. Dependency graphs, part-of-speech tag distributions, knowledge base embeddings and lexical overlap measures are examples of typical features utilized.

2.2.2 *Deep Learning Approaches*

In recent years, deep learning models have outperformed symbolic and classical feature engineering approaches on most image and text processing tasks [86]. One of the main reasons for the improved performance lies in the models' ability to learn a suitable mathematical input representation from high-dimensional, raw data instead of

relying on manually defined features. This representation is also called an *embedding* and is usually done on a word or subword level.

The models' success has also transferred to Automatic Short Answer Grading [14, 41, 81, 130, 151, 152]. Most approaches base their classification on the response and a reference answer, often checking whether one entails the other. Student models [185], results of related single choice questions [155], and the question itself are also considered by various approaches [98]. The best performance on publicly available short answer grading benchmarks, however, is achieved by approaches utilizing *Transformer* models [53, 97, 150].

Transformers are a class of neural networks that utilize self-attention [158] to handle long-range dependencies better than previous architectures. Instead of relying on a hidden state to hold all the knowledge encoded in previous words, Transformers can attend to all words of a sequence at once, weighted by their similarity to the query. They are usually pretrained on extensive text collections to predict randomly masked subword tokens and sentence relations. Transformers either consist of an encoder and decoder or only an encoder. Encoder-decoder models, also called sequence-to-sequence models, map input sequences to output text sequences and are, thus, capable of generating texts based on an input prompt. In ASAG, researchers usually use encoder-only models that map an input to a mathematical representation that a classification/regression layer can utilize to predict a grade.

The most relevant approach to our work [150] is based on BERT [31], one of the first encoder-only Transformer models. Sung, Dhamecha, and Mukhi [150] fine-tuned the pretrained BERT model on SciEntsBank [37] and achieved the state-of-the-art performance at the time with an accuracy of 75.9%. SciEntsBank is one of the most popular ASAG datasets, due to its size and variety of covered science domains. It is, therefore, often used as a benchmark for the ASAG task. Sung, Dhamecha, and Mukhi [150] even report comparing favorably to human graders on a private ASAG dataset in the psychology domain.

2.3 ADVERSARIAL EXAMPLES

While deep learning approaches have become widely popular [4, 25, 87, 114], recent studies showed that they are also often vulnerable to adversarial examples [26, 67, 175, 180, 184]. Adversarial examples are subtly modified inputs aiming to fool the model into a desired prediction [177]. The process of producing adversarial examples is also called an adversarial attack. Attacks may target their modifications to illicit the prediction of a specific class, or they may also be untargeted and accept any incorrect prediction.

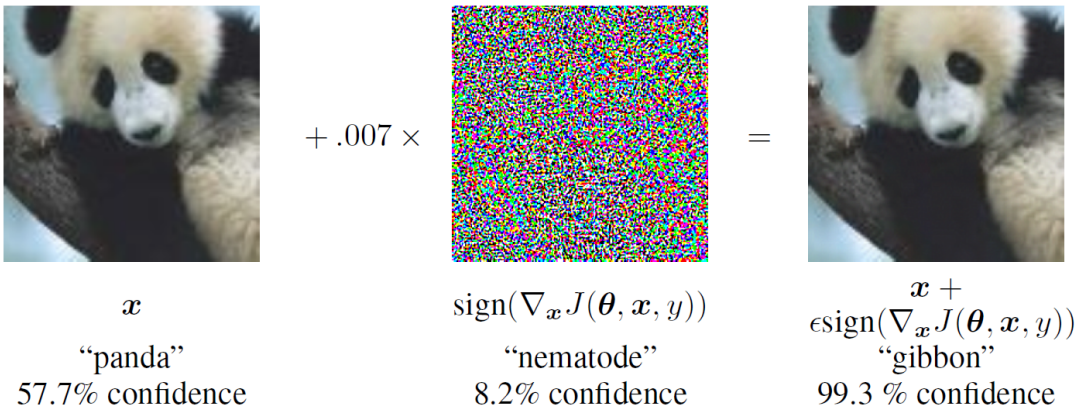


Figure 1: Illustration of an adversarial attack on image classification [56]. After adding imperceptible, gradient-based noise to the image, the model classifies the panda image as gibbon with high confidence. Image taken from [56].

Depending on the information utilized during the attack, black- and white-box approaches are distinguished. While black-box attacks at most use the model's output probabilities to inform their search for viable modifications, white-box attacks utilize knowledge about the model's inner workings, such as gradients, to identify important sections in the input. In image classification, adversarial perturbations are usually so slight that humans cannot perceive any difference from the original image. An example of a successful attack is illustrated in Figure 1. In Natural Language Processing, however, modifications are usually detectable by humans when the adversarial example and original text are viewed side-by-side.

For this reason, the subtlety constraint is often relaxed to *non-suspicion* in textual attacks [109]. What makes an input suspicious depends on the targeted task. For example, the expectation of a text's grammatical correctness and fluency varies greatly between news articles and middle school graders' essays. Adversarial attacks usually aim to produce fluent, grammatical and semantic-preserving changes to be as widely applicable as possible.

2.3.1 *Input-Dependent Attacks*

Changes can be made on a character, word, sentence or even document level (see Table 2). Character modifications include scrambling words, swapping adjacent characters or replacing characters with visually similar or otherwise plausible characters [17, 39]. On a word level, words are often replaced with their synonyms [127], neighbors in the embedding space [6] or words that likely match the sentence context [52, 182]. They can also be shuffled randomly if the goal is to generate nonsensical text [33].

Modification Level	Text
Original	Adversarial attacks have interesting properties.
Character	Advererial ąttacks have interesting porperties.
Word	Adversarial attacks have fascinating properties.
Sentence	Attacks that are adversarial have certain interesting properties.
Document	Adversarial attacks have interesting properties. Random confuse wordswithoutmeaning.

Table 2: Examples of adversarial character, word, sentence and document modifications.

One of the most prominent word-level attacks is TextFooler [72] due to its high success rate and open-source nature. We will use this attack as a benchmark for comparison with our proposed attack in Section 3.3. TextFooler first searches for critical words by deleting them from the input and observing their effect on the model's predicted class probabilities. Synonyms can then replace the essential words to fool the grading model.

Sentences are most often manipulated by paraphrasing [70, 128] and document-level attacks usually rely on reordering sentences or inserting meaningless or distracting sentences [111].

2.3.2 *Universal Attacks*

The previously discussed adversarial attacks aim to modify individual input samples. While they are well suited to uncover unreliable decision boundaries, their usefulness to malicious entities is limited since attackers must query the target model for every adversarial example they want to generate. As the attacker's access to the model is often limited to a specific time span or a number of queries, it is crucial to find vulnerabilities that generalize across samples. Finding such sample-independent rules is the goal of *universal adversarial attacks*. These types of attacks are the most relevant to our work.

Ribeiro, Singh, and Guestrin [129] propose a universal attack based on generalizing semantically equivalent adversarial examples. They translate sentences into pivot languages and back to obtain semantically similar paraphrases of the input sentence. The paraphrases are then generalized into semantically equivalent adversarial rules, such as "color $\rightarrow$ colour" or "? $\rightarrow$??". Successful rules can be applied to novel input sentences without the need to query the model again. However, while the rules are generally applicable to new inputs, they may not always fool the victim model suc-

cessfully. In fact, Ribeiro, Singh, and Guestrin [129] report label flip rates of only 1-4% for individual rules. The idea of generalizing from individual adversarial examples is nevertheless powerful and inspires our proposed attack in Section 3.3.

A different class of approaches focuses on the search for meaningless trigger sequences that tend to fool the victim model [16, 145, 160]. The most relevant approach to our work is the Universal Trigger Attack proposed by Wallace et al. [160], as we utilize it in our initial robustness experiments in Section 3.2. For example, they found that prepending "zoning tapping fiennes" indiscriminately to sentiment analysis inputs often causes misclassifications and that certain triggers lead language generation models to produce offensive texts. They utilize an iterative gradient-informed beam search to identify promising trigger candidates.

Finally, Gao and Oates [51] propose an attack based on modifying the embeddings themselves. Similar to adding the same adversarial noise to all images, they find perturbations applicable to all input tokens indiscriminately.

2.4 GAMING EDUCATIONAL SYSTEMS

Assessing a student's knowledge is vital for successful tutoring. Only when the teacher or teaching system can estimate what a student has understood can they recommend appropriate material or tasks for future learning [40]. For this reason, educational systems usually contain assessment components. However, many systems assume that students will answer questions in good faith.

Unfortunately, studies in traditional classroom assessments show that a significant portion of students employs non-learning strategies to achieve their goals. While there is a high variance in the number of observed cheating occurrences, large-scale reviews report that a majority of students have cheated during exams or assignments during their college career [78, 166]. This insight inspired research on cheating educational systems [9, 83, 164]. It was found that students may exploit lax password security to gain access to question pools and solutions or share screenshots of the assessment. Additionally, internet connection disruptions or hardware failures may be purposefully initiated to re-take assessments [132]. The ease of sharing assessment information and utilizing internet searches even prompted researchers to advise against using Mulitple Choice questions in online assessments and advocate for constructed response questions [104]. A number of Massive Open Online Courses have experienced systematic cheating as well [2, 3, 71, 117, 135] with students setting up multiple accounts to collect assessment information for their main accounts.

While the approaches discussed so far mostly exploit the lack of supervision in online assessments, task-specific cheating is also well documented, especially in the use

of intelligent tutoring systems [12, 112, 113, 121, 161, 162]. Here students may exploit savepoints and progressive hints to guess assessment solutions systematically. For example, math problems and Multiple Choice questions may be solved by simply trying out various answers until the correct solution is deduced [10, 11].

Even though previous work in this field mainly looked at other avenues for cheating than adversarial attacks, we draw the following insights from this body of research. First, cheating is prevalent in classical and digital teaching environments. Second, students will experiment and seek out non-trivial exploits of intelligent systems. Third and last, an entire assessment format can be questioned when cheating is detected, even if most results are genuine.

2.5 ELABORATED FEEDBACK SYSTEMS

Determining and communicating the correctness of a student's response – as currently done in Automatic Short Answer Grading – is called verification feedback [144]. While verification is already beneficial for learning, it can be difficult to truly understand one's mistakes from a simple grade alone. Elaborated feedback, on the other hand, can more effectively clear up misconceptions. It is especially helpful for formative assessments, which aim to modify the student's thinking and behavior to encourage learning [144]. Nevertheless, it can also be used in summative assessments at the end of a teaching unit to increase acceptance of a given performance evaluation.

Elaborated feedback often specifically explains what mistakes were made and why they are incorrect. It may also include concrete improvement suggestions. Providing elaborated feedback is more challenging than verifying since it requires a deeper understanding of the student's response. Especially for deep learning approaches where the decision process is opaque, explaining why a student's answer is deemed incorrect is difficult. For this reason, many elaborated feedback systems are based on symbolic reasoning [36]. They are typically designed for a specific domain, exploiting the structure of given responses [36, 95]. For example, many approaches provide elaborated feedback for programming exercises [77] or on the quality of writing in essays [62]. However, since these approaches are not applicable to domain-independent short answer grading, they serve only as a source of inspiration for the elaborated short answer feedback task.

For a more detailed review, we refer to excellent surveys by Deeva et al. [30], Mousavinasab et al. [110], VanLehn [157], Kulik and Fletcher [80] and Hasan et al. [61].

2.6 RESEARCH GAPS AND SUMMARY

Recent deep learning-based ASAG approaches perform well on common datasets. However, the performance is measured on test data from the same distribution as the training data. The robustness to out-of-training-distribution answers is unexplored by previous work. This makes the models' applicability to real-world grading scenarios questionable, where new generations of students are likely to formulate novel answers. Especially when students also employ unintended, non-learning strategies to gain good grades. In other fields, adversarial attacks have exposed significant weaknesses in even the best-performing models, further emphasizing the need to examine automatic grading models critically. ASAG models' robustness to adversarial attacks is the first research gap we investigate in this thesis.

Existing adversarial attacks are also not suited to thoroughly test automatic grading models due to the unique constraints grading scenarios have compared to other Natural Language Processing tasks. For example, students are typically not machine learning experts. Consequently, they are unlikely to be able to perform complex attacks in time-sensitive summative assessments. While attacks exist that would provide students with simple rules they can follow during a summative assessment, they would have to simultaneously be successful enough to impact received grades noticeably while staying unobtrusive. The risk of detection plays a vital role in students' willingness to employ a given attack, much more so than for attacks designed to debug Natural Language Processing models. Thus, an attack designed with such considerations in mind is needed to thoroughly test grading models' robustness with an appropriate threat model. This is the second research gap we address.

The final research gap lies in the type of feedback ASAG models produce. While verification feedback is an excellent first step, students need to understand their feedback to truly engage with it. This is difficult when the feedback only consists of a grade provided by a black-box model. Thus, the generation of elaborated feedback should be explored. As no publicly available datasets for generating elaborated feedback to short answers existed previously, the first step is to design, collect and distribute a benchmark dataset that can serve as the foundation for reproducible and comparable research in this domain. Next, elaborated feedback generation methods that can be applied to various domains and educational scenarios should be explored, as current approaches are usually tailored to a specific scenario with extensive manual effort.

3

ADVERSARIAL ATTACKS ON AUTOMATIC GRADING

As discussed in Section 2.4, cheating is prevalent in traditional classrooms as well as intelligent tutoring systems [78, 113, 166]. Typically, students make use of lax supervision and insufficient security of sensitive information to inflate their grades artificially. Nonetheless, automatic assessment models' systematic weaknesses have also been the target of cheating in the past [10, 11]. Thus, it is likely that students will also test the limits of Automatic Short Answer Grading (ASAG) models and utilize identified weak points. In this chapter, we investigate potential vulnerabilities of the state-of-the-art ASAG models BERT [31] and T5 [124] to linguistically manipulated student responses (RQ1). Since the automatic grading domain differs from the general adversarial scenarios typically assumed in natural language attacks, we first discuss domain particularities and resulting design considerations, followed by two automatic and one manual adversarial attack.

3.1 ATTACK DESIGN CONSIDERATIONS

Assessments can typically be split into two categories during education. In formative assessments, the objective is to help students learn and improve their understanding. This is done by informing students of the learning goal, where they currently stand in relation to the goal and how they can improve [34]. Formative assessments can be graded to indicate the student's level of understanding but typically do not count toward the final performance evaluations.

Summative assessments, on the other hand, intend to measure how much the student has learned throughout the instruction and provide an indication of the student's final performance [34]. In practice, we expect ASAG models to be employed in both assessment phases. First, to provide nearly instantaneous and inexhaustible feedback to students via formative assessment. Then, to formally and summatively assess students' performance at the end of the course. Since the summative assessment is the one that affects students' final grades, it is the likely target for cheating attempts. Thus, the scenario we consider consists of the formative assessment phase, where students can query the grading model to gain information, and a controlled summative assessment phase, where the acquired knowledge guides the actual attack. This domain-specific

split into two distinct attack phases of information gathering and attack deployment is not typically found in related work and motivates the following design considerations.

3.1.1 *Necessary Expertise & Resources during Testing Time*

In contrast to Natural Language Processing (NLP) researchers, students are seldom machine learning experts. While a lack of expertise can typically be compensated with time and access to the internet, for example by asking an expert for help, we do not expect students to have the necessary resources during summative assessments, such as exams. Especially considering the time limitations and mental pressure typically associated with exams or quizzes, any employed adversarial attack would have to be easy and fast to use during test time. Generally, access to the internet and computation devices tend to be restricted during exams to prevent traditional cheating behavior. So while having a model like GPT-3 to complete one's exam may be a promising strategy to obtain a good grade[1], it can be prevented by controlling the test environment. To avoid this limitation, an attack should require nothing but a writing implement during test time.

3.1.2 *Model Access*

Most existing adversarial attacks require information about the target model that students will not have access to. Many white-box attacks, for example, use a model's gradients to estimate which tokens, words or sentences had a significant impact on the model's prediction. These critical words are then good manipulation targets. Even black-box approaches often utilize the raw class probabilities outputted by the model to inform their iterative searches. If a modification increases the probability for the target class but does not lead to misclassification, it can be a good starting point for further manipulation.

However, students will likely only see the final prediction of a grading model rather than any intermediate steps or class probabilities. It is also likely that they will not know what kind of model is grading them and how it works on a conceptual level. Thus, attacks tailored to the automatic grading domain should perform well without any knowledge of the grading model. Nonetheless, they may learn from the grades assigned to responses during the formative assessment phase.

1 For an example of GPT-3 being used to solve school exams see: https://medium.com/geekculture/gpt-3-takes-on-school-exams-e1b5d7abc87d [accessed April 24, 2023]

3.1.3 *Risk of Detection*

Another aspect often neglected in attack design is the cost and risk of generated examples being identified as adversarial examples. This is an important factor for students as the willingness to cheat depends, among other factors, on the perceived cost of cheating [115]. Considering that the punishment for academic dishonesty is often severe, the likelihood of being caught greatly influences the perceived cost of cheating [115]. How likely it is that an adversarial attack will be detected and how difficult it would be to prove deceptive intent influences a student's decision whether to employ a given attack or not.

For example, appending the same nonsensical phrase to each answer in an exam may fool an automatic grading model. However, a manual inspection by a teacher or even an automated detection script could easily identify such an attack as a cheating attempt or at least assume malicious intent. By contrast, grammatically valid and varied manipulations are much more difficult to spot and any suspicions are more easily explained away by a student's writing style or word preference.

3.1.4 *Computation and Time Budget*

While more computation power usually allows one to explore a larger search space in a given amount of time, we would expect students' budgets to be limited. It is plausible that students would have a computer with a graphics card or would have enough money to rent one for a limited time, but we expect them to only access large computation clusters for a small amount of time. Considering that grading models may also run on a cloud server instead of the students' local machines, the computational budget of an attack is limited. Thus, adversarial attacks' runtime should be measured in minutes on a single GPU instead of months on a computation cluster.

3.1.5 *Class Equivalence*

The considerations discussed so far focus on aspects where the automatic grading domain has stricter constraints than those assumed in current adversarial attacks. If one designed an attack solely based on this more stringent set, the attack's success rate would likely be poor. For example, Ribeiro, Singh, and Guestrin [129] proposed an attack to generate universally applicable, semantically equivalent rules that could easily be used in a summative assessment. However, it only induces misclassification rates between 1 and 4 percent, which is inconsequential in most exam contexts. Thus, conventional constraints should be relaxed where sensible. Specifically, natural language

attacks typically constrain their modifications to semantically equivalent ones. For example, replacing a word with its synonym should not change an utterance's meaning. This is important for attacks that aim to be applicable to various different tasks, as slight shifts in meaning may change the input's actual label.

In the automatic grading domain, however, it is unlikely that small, randomly generated meaning shifts will suddenly make an incorrect student response factually correct. Therefore, adversarial attacks in this domain can manipulate student responses more leniently as long as the response's true class remains the same.

3.2 UNIVERSAL TRIGGER ATTACK

The first attack we consider is an adaption of the *Universal Trigger Attack* proposed by Wallace et al. [160]. The attack aims to find universal sequences of tokens that tend to trigger a misclassification independently of the specific input they modify. An example of a successful attack can be seen in Table 3. Since this attack was not specifically tailored to the automatic grading domain, it does not meet all of the criteria defined in Section 3.1. The found triggers are easy to use during a summative assessment, are reasonably inexpensive to find and generally preserve class equivalence. However, the approach requires access to the grading model's gradients which students would typically not have. Additionally, employing the same trigger sequence to all answers in an exam is also an easily recognizable cheating pattern and is, thus, risky to use in practice. Nevertheless, the attack is a good starting point for evaluating whether automatic grading models are vulnerable to adversarial attacks in the first place due to its powerful, informed search.

Question:	Georgia found one brown mineral and one black mineral. How will she know which one is harder?
Reference Answer:	The harder mineral will leave a scratch on the less hard mineral. If the black mineral is harder, the brown mineral will have a scratch.
Student Answer:	that it stay the scratch. → *incorrect*
Triggered Answer:	none exits that it stay the scratch. → *correct*

Table 3: An example of a successful trigger attack. Prepending "none exits" to a student answer changes the ASAG model's prediction from *incorrect* to *correct*. The original student answer stems from SciEntsBank's unseen answers test set [37]. Table adapted from [45].

3.2.1 *Approach*

The Universal Trigger Attack [160] begins with an initial trigger of a given length, such as "the the the" or "a a", and iteratively searches for replacements of the trigger's tokens that increase the *target class's prediction probability*. The search itself is inspired by HotFlip [38] and utilizes the model's gradients to estimate the effect a replacement will have on its predictions. Thus, the attack concatenates the current trigger and a batch of examples taken from the dataset, calculates the gradient with regards to the target class and selects the best k replacement candidates for each trigger token, minimizing

	Beetle			SciEntsBank			
	Train	UA	UQ	Train	UA	UQ	UD
correct	1665	176	344	2008	233	301	1917
incorrect	1227	152	231	2462	249	368	2228
contradictory	1049	111	244	499	58	64	417

Table 4: Number of answers per class in each split of the datasets SciEntsBank and Beetle [37]. The unseen answers (UA) test split contains new answers to questions already incorporated in the training set. Unseen questions (UQ) contains novel questions and unseen domains (UD) consists of questions belonging to different scientific disciplines.

the loss for the *target class*. The best trigger combination then forms the initial trigger for the next search iteration.

3.2.2 *Target Grading Models*

Since ASAG models used in practice likely have a high predictive performance, selecting a similarly well-performing grading model as a target is important to ensure the validity of drawn conclusions. Thus, we aim to find the best-performing model for the short answer grading task. Finding the state-of-the-art is challenging as many approaches are evaluated on proprietary datasets or are not described in sufficient detail to reproduce them. For this reason, we only consider models evaluated on common ASAG benchmarks, such as SciEntsBank and Beetle [37].

SciEntsBank contains questions, student responses and reference answers from various domains, such as biology and geography. The dataset was collected in primary and middle schools in the USA. The Beetle corpus focuses on basic electricity and electronics questions posed to students in the context of an intelligent tutoring system. We select the three-way task where student answers are classified as *correct*, *incorrect* or *contradictory*. The class distribution for both datasets can be seen in Table 4.

Prior to our work, the best-performing model on SciEntsBank was a BERT model trained by Sung, Dhamecha, and Mukhi [150]. While they did not publish the model, most relevant hyperparameters are described sufficiently for reproduction. We trained 10 models with the reported hyperparameter settings, aiming to reproduce their performance. Unknown hyperparameters were chosen close to the original BERT model's hyperparameters with minimal tuning. As most institutions will likely utilize grading models trained on non-public exam response collections in practice, we do not expect an attacker to have access to the model's training data. To emulate a typical assessment scenario, we train the model on the SciEntsBank training split and save the Beetle training data for the attack.

	Unseen Answers			Unseen Questions			Unseen Domains		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
#1	0.744	0.703	0.741	***0.675***	0.555	***0.665***	0.624	0.490	0.609
#2	0.737	0.690	0.732	0.674	0.561	0.662	***0.670***	***0.599***	***0.661***
[150]	***0.759***	***0.720***	***0.75***8	0.653	***0.575***	0.648	0.638	0.579	0.634

Table 5: Performance of the two best reproduction models (#1 and #2) compared to the results reported by Sung et al. [150]. The approaches are compared based on the weighted-averaged F1 score (W-F1), macro-averaged F1 score (M-F1) and accuracy (Acc). Each category's best result is marked in bold. The table is adapted from [45].

Table 5 displays the performance of the two best models from our reproduction experiments compared to the reference model proposed by Sung, Dhamecha, and Mukhi [150]. While the reproductions are 1.5-2.2 percentage points less accurate on unseen answers, they are 2.1-2.2 percentage points more accurate than the reference on unseen questions. On unseen domains, model 1 performs worse than the reference by 1.4 percentage points and model 2 performs better by 3.2 percentage points in terms of accuracy. We conclude that the difference in performance may be due to a different random initialization or hyperparameter setting but that both reproduction models are sufficiently powerful for the evaluation of adversarial attacks.

3.2.3 *Experimental Settings*

Unless stated otherwise, all experiments presented in Section 3.2.4 follow the following setup. They target grading Model #2 as it had the best performance on average across the test splits (see Table 5). Contrary to related work, we search for triggers using the *incorrect* student answers in the BEETLE training split instead of the model's training or test data. While the attack is likely to be less powerful when utilizing a surrogate data source and may, in fact, even fail if triggers do not transfer across datasets, it is unlikely that attackers would have access to the original training data in an educational scenario. This experiment design decision is vital for the evaluation's interpretation. Successful triggers found in this setting suggest that models may have general weaknesses that can be found even when the training data remains secret.

We perform manual hyperparameter tuning to find the best attack configuration on BEETLE. Starting from the initial values proposed by Wallace et al. [160], we gradually increase the batch size, trigger length, beam size and number of candidate tokens. The search range for each hyperparameter can be found in Appendix A.2. Depending on the specific experiment, the best x triggers are then evaluated on the SCIENTSBANK test

splits. In total, 32 triggers of various hyperparameter settings were tested. A complete list can be found in Appendix A.2.

3.2.4 *Attack Evaluation*

We measure the success of a trigger by the number of times it flips the predicted label from *incorrect* to *correct*. To be comparable to related work, we also report the decrease in accuracy on all *incorrect* samples regardless of whether the prediction was shifted to *correct* or *contradictory*. Since we are not interested in evaluating the attack per se but instead wish to investigate how successful a student could maximally be when employing such an attack during assessments, we will focus on the performance of the best triggers. Thus, this evaluation aims to find the upper bound of a single trigger's flip rate.

Trigger Transferability across Datasets

In the first experiment, the 20 triggers causing the most flips on BEETLE are evaluated on SCIENTSBANK's test sets. As discussed in Section 3.2.3, this experiment represents the most likely scenario where attackers must rely on surrogate data sources since the original training data is unavailable.

Table 6 displays the performance of the best trigger on each test split as well as the model's base misclassification rates. Interestingly, all triggers begin with the token "none". The triggers "none varies" and "none would" perform best on unseen answers with 53 flips, meaning that 21.3% of *incorrect* answers are predicted as being *correct* by the model. In contrast, the model only misclassifies 12.4% as *correct* without manipulation. The overall model accuracy for *incorrect* responses decreases by half, from 85.7% to 42.8%.

On unseen questions, "none would" achieves 138 flips resulting in target predictions for 37.5% of the responses. That is an increase of 10.1% percentage points compared to the model's initial misclassification of 27.4%. The overall accuracy decreases by more than half from 70.7% to 32.25%. Similarly, on unseen domains, "none elsewhere" increases false *correct* predictions from 22.0% to 37.1% with 826 flips and the accuracy declines from 76.9% to 31.2% when using "none would".

Trigger Transferability across Models

As discussed in Section 3.2, students would not typically have access to the grading model's inner workings. Thus, this experiment investigates how successful the attack

	# of Flips to *Correct*			Accuracy on *Incorrect*		
Triggers	UA	UQ	UD	UA	UQ	UD
none varies	**53**	134	687	71.08	54.62	63.69
none would	**53**	**138**	810	**41.77**	**31.25**	**31.15**
none elsewhere	50	121	**826**	47.79	36.14	37.93
Base misclassification	31	101	491	84.74	70.65	76.93

Table 6: The most successful triggers for each test split and the number of invoked misclassifications from *incorrect* to *correct*. The initial number of the grading model's misclassifications without triggers is shown in the last row. The accuracy for *incorrect* samples is also given for comparability with related work. The best results are marked in bold. Table adapted from [45].

can be when training a surrogate model to search for triggers. For this purpose, we evaluate all triggers that performed well on the development set and that were found using Model #2 on Model #1. This evaluation also included trigger searches on SciEntsBank's training split since access to the model's training data is required to train a surrogate anyway.

The best-performing triggers for each model can be seen in Table 7. Some of the triggers, such as "nowhere changes", "anywhere." and "none else", found using surrogate Model #2 work even better on target Model #1. On unseen answers, for instance, "nowhere changes" achieves 81 flips – 30 more than on the surrogate model. It raises the highest rate of target predictions from 21.3% to 32.5% on this test split. However, Model #1 also has a higher base misclassification rate than Model #2 on unseen answers and unseen domains, likely partially contributing to the trigger's performance gain even though the same trend can be observed on the unseen questions test split. Here, "nowhere changes" causes 46 additional flips compared to the best trigger for the surrogate model, even though both models have almost the same base misclassification rate. In total, 50% of the *incorrect* responses are classified as *correct* by Model #1 when using this trigger on unseen questions. On unseen domains, "anywhere." flips 1027 predictions to *correct*, which is an increase by 17.1% to 46.1% compared to Model #1's base misclassification rate.

However, while it seems to be possible to find successful triggers utilizing a surrogate model, a trigger's performance varies greatly between the two models. The top-3 triggers, for one, differ between the models. Additionally, some triggers, like "none would", actually reduce the number of misclassifications induced in the target model.

	UA		UQ		UD	
Trigger	#1	#2	#1	#2	#1	#2
nowhere changes	**81**	51	**184**	135	957	640
anywhere.	58	45	108	105	**1027**	682
none else	73	53	158	136	941	818
none varies	49	53	79	134	576	687
none would	38	53	97	**138**	495	810
none elsewhere	60	50	115	121	701	**826**
Base Missclasification	44	31	100	101	646	491

Table 7: Number of Flips achieved by triggers found using Model #2, evaluated on Model #1. The triggers' performance on the original Model #2 is also reported for comparison. The first rows are the best-performing triggers for Model #1. The middle block contains the best triggers for Model #2. Table adapted from [45].

3.2.5 *Interpretation of Results & Limitations*

In summary, we observed significant losses in prediction accuracy when employing the Universal Trigger Attack. Concatenating a single two-token trigger to *incorrect* student responses can reduce a model's accuracy by more than half from 70%-85% to 31%-42%. While most of the performance loss is due to the model falsely labeling the instances as *contradictory*, 8%-23% of the responses previously labeled correctly were now classified as *correct*. Combined with the base misclassification rate, up to 32.5% of *incorrect* responses could receive full points in an exam when the model is specifically trained to grade the particular exam questions.

If the model is asked to generalize to novel questions, the attack is even more successful, with up to half of the wrong answers achieving the best grade. An example of such a generalization could be a model trained on various questions regarding earth erosion and then grading questions regarding earth deposition. This vulnerability indicates that ASAG models are not yet suited to be employed in an unseen questions scenario, regardless of whether students could actually employ such an attack in practice. The model's extreme brittleness on novel questions implies predictions being based upon non-robust features, to the point where a reliable grading even for unmanipulated answers seems unlikely.

However, the Universal Trigger Attack's usefulness for cheating automatic grading systems in practice is limited. For one, concatenating nonsensical words to student responses can clearly be identified as a cheating attempt. Thus, it would be risky to employ this strategy in practice. Additionally, the gradient information needed to inform the trigger search is problematic to acquire for students. While our experiments indicate that it can be possible to find viable triggers even when substituting the tar-

get model or dataset, further experiments substituting both at the same time would be necessary for a real-world application viability assessment.

Moreover, there was a high variance in trigger performance in our experiments. This may indicate that experiments with other datasets and models could elicit different results. Thus, more in-depth experiments with multiple models, datasets and tasks are required to verify whether transformer models generally have exploitable triggers that are independent of specific datasets. Nevertheless, ASAG models seem unreliable enough to warrant further research with specialized attacks explicitly developed for the automatic grading scenario.

3.3 ADVERSARIAL ADJECTIVE AND ADVERB INSERTION

After establishing that automatic grading models can, in principle, be fooled by adversarial examples, we develop an adversarial attack specifically for the ASAG scenario based on the considerations discussed in Section 3.1. In contrast to existing adversarial attacks, it is well suited for summative assessments in controlled environments. It exploits the relaxed class equivalency constraint to find powerful adversarial examples despite not having access to the model's inner workings or training set. Specifically, it searches for adjectives and adverbs the model generally associates with the target class, which then can be inserted into grammatically valid places in *incorrect* student responses. Students can then easily integrate such adjectives into their answers during the summative assessment without any additional expertise or effort. Moreover, a cheating attempt would be much harder to prove as such compared to a trigger attack since using colorful adjectives could simply be a part of the student's writing style. An example of a successful attack can be seen in Table 8.

Question:	When a seed germinates, why does the root grow first?
Reference:	The root grows first so the root can take up water for the plant.
Original:	The root grew because it needs to help the plant stand up. → *incorrect*
Modified:	The root grew because it immediately needs to help the plant stand up. → *correct*

Table 8: An example of a successful adverb insertion flipping the automatic grading model's prediction from *incorrect* to *correct*. The original response stems from SciEntsBank's unseen answers test set. Table adapted from [43].

3.3.1 *Approach*

A schematic overview of the attack can be seen in Figure 2. The first step of the attack consists of identifying promising adjectives and adverbs. We utilize the Brown Corpus [49] as a source of adjectives and adverbs since the corpus' texts are annotated with their part-of-speech tags. While modern part-of-speech taggers are accurate enough to be considered a solved task by parts of the NLP community, a corpus annotated by multiple annotators is likely more reliable still [100]. Since it would be vastly detrimental to the naturalness of generated adversarial examples if the attack started inserting verbs or punctuation symbols in grammatical places intended for adjectives or adverbs, reliability is more important than the greater coverage one could achieve

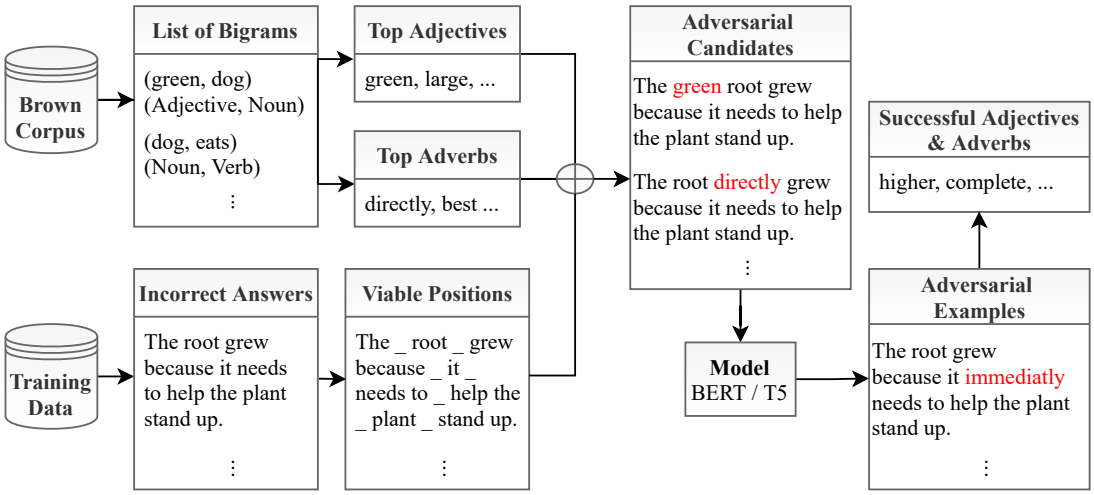


Figure 2: Schematic overview of the attack. Taken from [43].

with unannotated corpora. The Brown Corpus contains English texts from various domains, such as news articles, reviews and fiction novels.

As we plan to insert adjectives and adverbs directly before nouns and verbs, we extract all bigrams of the following forms from the corpus: (adjective noun), (adjective pronoun), (adjective proper_noun) and (adverb verb). Thus, we only extract adjective and adverb candidates that appear in the intended grammatical configuration in practice. This means that our approach only focuses on attributive instead of predicative adjectives, that is, adjectives that appear before instead of after the noun. For example, "The man is alive and blue." would be ignored while "The blue man exists." would yield the adjective "blue." Similarly, only adverbs typically appearing directly before the verb are considered, such as "probably" or "really." This design decision limits the range and grammatical versatility of generated adversarial examples. In exchange, it is more reliable as it does not require the automatic detection of complex grammatical structures. Most high-performing part-of-speech taggers do not label at the granularity level needed to identify types of adjectives beyond comparative and superlative[2]. Additionally, models for selecting viable insertion positions for various types of adjectives and adverbs would need to be much more complex, likely resulting in a higher number of erroneous insertions.

Moreover, we filter the resulting adjectives and adverbs for stop-words based on the stop-word list introduced by Bird, Klein, and Loper [19] in the Natural Language Toolkit[3]. The main reason for doing this lies in the fact that the stop-word list covers meaning-inverting words, such as "not", that could easily correct a *contradictory*

2 https://web.archive.org/web/20230408133828/https://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html [accessed April 25, 2023]
3 https://www.nltk.org/index.html [accessed April 25, 2023]

student response, breaking the class equivalence constraint. From the filtered word lists, we pick the top 100 most frequent adverbs and adjectives; Disregarding rare adjectives and adverbs is advantageous for the naturalness and inconspicuousness of generated adversarial examples, even if they may be quite successful at fooling models due to their underrepresentation in training data. Teachers may become suspicious when their "students" suddenly seem to have consumed a thesaurus, using rare words like "lachrymose," "loquacious" or "contumacious."

After collecting promising adjectives and adverbs, the next step involves finding possible insertion places in the students' responses. Existing adversarial attacks would utilize the model's gradients or class probabilities to guide the search toward high-impact modifications here. Since we do not expect students to have such signals available, we take the model-agnostic approach of considering all positions before nouns and verbs. Nevertheless, any additional information acquired in practice could dramatically improve the search efficiency by constraining the search space to promising avenues. Only one adjective or adverb is inserted per student response, even if additional insertion spaces would be available.

The target model can now be confronted with the generated adversarial candidates to identify which adjectives and adverbs often fool the model. In the educational scenario, students would query the model during formative assessments throughout the semester to receive a list of successful adjectives and adverbs they can overuse during the summative assessment.

3.3.2 *Target Models*

Based on the findings from our experiments with the Universal Trigger Attack and the goal of evaluating the proposed attack itself, we expand the set of experiments to include additional models and datasets. This will allow a more reliable and detailed investigation of the attack's effects. We train base-sized BERT and T5 models for each dataset based on their HuggingFace implementations [172]. While larger model sizes would likely perform better, they are significantly more expensive in terms of computation cost. Since the insertion attack will query the models often, using the large versions would increase the ecological footprint of our experiments immensely.

We tune the hyperparameters manually using 10% of the respective dataset's training split for validation. Models train for 8 epochs on the remaining training data, after which the checkpoint with the best macro-averaged F1 score on the validation split is chosen for further experimentation.

The following considerations guided the selection of datasets for the proposed attack's evaluation. The datasets should cover a wide range of domains to investigate the

attack's applicability beyond the American primary and middle school science assessments represented by SciEntsBank. Additionally, the datasets should be well-known benchmarks to improve the interpretability and impact of obtained results. Finally, even though the attack is tailored to educational assessments, exploring its uses for other NLP tasks would be beneficial. Naturally, the risk of actually changing a sample's class still needs to be considered, limiting the scope to semantic-focused tasks that are not particularly sensitive to adjectives or adverbs. For these reasons, we chose the following datasets for our experiments:

- **Recognizing Textual Entailment (RTE)** is one of the smaller datasets in the well-known NLP benchmark suite SuperGLUE[4], with 5,767 examples total. SuperGLUE is one of the main dataset collections to rank the general performance of NLP models. RTE contains a series of texts and hypotheses where the task is to decide whether the hypothesis can be inferred from the corresponding text. Recognizing Textual Entailment can be considered similar to ASAG, where student answers should entail the reference solution [37]. The hypothesis and text pairs are labeled as *entailment* or *not_entailment*, which can be likened to *correct* and *incorrect* answers in ASAG datasets. The BERT model achieved its best performance after 6 epochs, utilizing a batch size of 32 and a learning rate of $1e-5$. All reported BERT models were optimized with AdamW without bias correction [94]. T5 also converged after 6 epochs, using a batch size of 8 and gradient accumulation of 8 due to its larger size. All reported T5 models use an Adafactor optimizer with relative steps and initial warm-up [143].

- **Multi-Genre Natural Language Inference (MNLI)** is another Textual Entailment dataset. It is part of the GLUE benchmark, the predecessor of SuperGLUE and consists of premise and hypothesis pairs [167]. As the test split is not publicly available, we use one of the two validation sets provided for our evaluation. In addition to RTE's labels, MNLI also considers *contradictory* pairs, making this a 3-way classification like SciEntsBank. It is much larger than RTE, with 433,000 sentence pairs. The best BERT model converged after 2 training epochs with a batch size of 64 and a learning rate of $2e-5$. It is trained with mixed precision training (FP16) to save computation time as MNLI is a much larger dataset compared to the others used in this experiment, e.g., by a factor of 40 compared to SciEntsBank. The best T5 model converged after 6 epochs, using a batch size of 8 and gradient accumulation of 8.

- **SciEntsBank (SEB)** is the ASAG benchmark also used in the evaluation of the Universal Trigger Attack. It contains primary and middle school science short

4 https://super.gluebenchmark.com/ [accessed April 25, 2023]

answer questions, student responses and reference answers. In total, it consists of 10,814 responses. The 3-way dataset differentiates between *correct, incorrect* and *contradictory* student responses. The BERT model performed best after 3 training epochs using a learning rate of $2e-5$ and 32 batch size. The best T5 model trained for 7 epochs with a batch size of 8 and gradient accumulation over 4 batches.

- The **Microsoft Research Paraphrase Corpus (MRPC)** contains 5,800 pairs of sentences that are either semantically equivalent to each other (labeled as *1*) or not (labeled as *0*). Paraphrase detection is also similar to ASAG, where student responses should be semantically similar to the reference solution. However, student responses could still be correct even if they are not paraphrases of the reference solution. For example, a response may cover all aspects mentioned in the reference solution but also add additional factually correct information that was not strictly necessary. Similar to MNLI, MRPC is a part of the GLUE benchmark. The hyperparameters for the best BERT model were a batch size of 32, gradient accumulation over 2 batches and a learning rate of $2e-5$. The model trained for 3 epochs. The best T5 model also trained for 3 epochs using a batch size of 8 and gradient accumulation of 4. Once again, we used mixed floating point precision.

The models' accuracy before the attack can be seen in Table 9.

3.3.3 *Experimental Settings*

First, we filter the test sets' negative examples for all misclassified ones. For SciEntsBank, that means excluding all *incorrect* student responses the model falsely predicts as *correct* or *contradictory*. For MNLI, it means filtering out all text pairs already misclassified as paraphrases. We do this to avoid overestimating the attack's performance since these samples do not require any adversarial modification. Then we generate adversarial examples according to the procedure described in Section 3.3.1 and calculate the drop in accuracy the attack induces in the target model.

While accuracy is easily measured automatically and gives a good indication of an attack's performance, it does not measure whether the modification adheres to the class equivalency constraint (see Section 3.1). Since automatic metrics attempting to capture the semantic meaning of a text are unreliable [18, 126], we rely on human judgment to determine whether the sample's true class remains unchanged after the modification. Additionally, automatically obtaining a valid measurement of the risk of detection would also be problematic. Automatic detection measures may flag suspicious responses more or less successfully. However, human teachers will likely pass the final judgment concerning potential cheating cases. Therefore, we conduct a human

evaluation to determine the **suspiciousness** and **correctness** of generated adversarial examples.

We randomly sampled one successful adversarial example for each question in SciEntsBank since this is the most relevant dataset for the automatic grading scenario motivating our work. Each question should only appear once in the final survey to avoid redundancy that may lead graders to lose concentration or adopt heuristic keyword scanning. We selected the ten shortest questions/answers from each test split that did not reference external material, as some questions refer to images or tables not contained in the dataset. We sampled two additional questions from the unseen answers split as the unseen questions test split only had eight questions matching the criteria. In total, we obtained 30 adversarial examples, the original student responses, corresponding questions and reference answers.

We opted for a between-group experiment design, where each group views the same questions with corresponding reference answers. While the control group receives the original, unmodified student responses, the treatment group gets the corresponding adversarial examples instead. Each group then rates the responses' **correctness**, **naturalness** and **suspicousness** on a 5-point Likert Scale. **Correctness** aims to capture how correct a response is on a factual level and whether it answers the question in its entirety. A one on the **correctness** scale indicates that the response is irrelevant or contradictory to the reference answer. At the same time, five means the response contains all necessary aspects mentioned in the reference answer.

Naturalness refers to the response's linguistic form and how similar it is to human writing [65]. Here, a one indicates a synthetically generated and abnormal response, while five means that a native-speaking student could have written the response. Finally, **suspiciousness**, also called **mistrust**, refers to the perceived likelihood of the student trying to cheat an automatic grading system with their response. Assigning a Likert score of one means that the grader does not believe the response to be a cheating attempt, and a five indicates that they are certain it is a cheating attempt. Both groups were informed that some student responses might be cheating attempts, but they were not instructed in how the automatic grading model works or how it may be cheated.

When piloting the study, it became clear that the scales were challenging to understand for the annotators without further elaboration. Thus, we added explanations, including response examples where appropriate, for every level of the Likert Scale. Screenshots of the final survey can be found in Appendix A.3. When participants gave at least a four on the mistrust scale, they triggered an additional question, asking whether they would take action based on their impressions. Possible actions could be bringing the student's response to the attention of a superior, taking disciplinary action, initiating dialogue with the student or other educational interventions.

As evaluating text with human annotators is a well-known task in the Natural Language Generation field, we defer to their evaluation guidelines. Per recommendation, at least three annotations per text should be collected to balance out the subjectivity of given ratings and obtain a more reliable evaluation [156]. We invited potential annotators based on their prior experience with grading short answer questions, English skills and general education level. In the end, seven experienced graders participated in our study, where we randomly but evenly assigned them to either the control (N=3) or experimental condition (N=4).

All participants had university degrees and routinely graded short answer questions in the context of their university employment in computer science and electrical engineering departments. Considering the participants' generally high level of education, we expect them to have the basic science knowledge needed to grade SciEntsBank's various middle-school-level questions. For reference, they also had access to the sample solutions and a direct chat with the researchers in case of questions during the online survey. The annotators came from Syria, Slovenia, Germany, India, and Iran. They spoke English fluently, even though it was not their first language. Of the participants, two were female and five were male.

Participation in the study was entirely voluntary and was not compensated beyond mutual participation in other studies. Participants were free to abort the online questionnaire at any time without providing a reason. The participants required 53.14 minutes on average to complete the study, which is within the period we expect experienced graders to be able to concentrate on tasks of this complexity. As included in the study description provided to the participants, all survey responses were anonymized prior to analysis to protect the participants' privacy.

3.3.4 *Attack Evaluation*

We aim to evaluate the following aspects of the attack in this chapter:

- **Effectiveness** captures how successful the attack is at degrading a model's performance. We measure it using the drop in accuracy induced in the target model. To provide a frame of reference, we compare the proposed attack's effectiveness to TextFooler [72], a popular open-source attack. Even though TextFooler does not find universally applicable modifications and is, thus, not suited to assessment scenarios, it provides a strong baseline for comparison. The expectation is that both attacks perform similarly well.

- **Adherence to the class equivalency constraint** is analyzed using the **correctness** item of the human evaluation. We expect that the generated adversarial exam-

ples appear less or equally incorrect to humans compared to the original student responses (H1) and, thus, do not actually change the response's true class.

- **Risk of detection** is represented by the **naturalness** and **mistrust** items in our study. While there are many more aspects influencing the risk of being caught cheating, such as suspicious querying behavior during the formative assessment phase, we focus on the direct effect of the attack on the suspiciousness of test responses. In contrast to less controllable factors, it is independent of the specific course structure. We expect the adversarial examples to appear less natural to humans (H2) but that they are not more suspicious (H3).
- Additionally, we explore potential **reasons for the attack's effectiveness**. Specifically, we expect the attack to primarily work on low-confidence predictions, where the initial response is close to the model's decision boundary. Thus, we compare the model's confidence levels for responses that later become adversarial examples to responses the attack is unable to flip. Additionally, we expect the attack to exploit spurious correlations between adjectives and adverbs with the target class. To investigate this, we analyze the occurrences of the most successful adjectives and adverbs in each class.

Effectiveness

Table 9 shows the effectiveness of TextFooler [72] and our proposed attack on each target model. TextFooler targets individual texts by deleting essential words and replacing them with synonyms. Thus, it is unsuitable for assessment scenarios due to its inability to find universally applicable modifications. Nevertheless, we chose it to represent the state-of-the-art since its individual modifications allow it to impact target models' accuracy powerfully compared to other attacks. Additionally, due to its open-source nature, it was easily reproduced. Considering that T5 is a text generation model and, thus, does not readily provide the clean class probabilities needed for TextFooler, we only run TextFooler on the BERT models on each dataset.

Firstly, we report each target model's performance without any adversarial manipulation. Overall, there are large differences in the accuracy achieved on each dataset. On smaller datasets, such as RTE and MRPC, models attain between 60 and 74% accuracy. By contrast, MNLi, the largest dataset with roughly 40 times more samples than the second largest dataset SciEntsBank, invokes accuracies between 76% and 84%. The models' performance on SciEntsBank differs vastly from test split to test split, with unseen answers being in line with MNLI's performance, while unseen questions and domains behave similarly to the small datasets RTE and MRPC. A discrepancy between unseen answers and the other test splits is expected as the task of creating a general

Test Set	Model	Attack	Acc.	AaA	Δ Acc	#Adv	#Aff	Time
Automatic Short Answer Grading Task								
SEB UA	BERT	TF	0.835	0.751	-0.084	87	21	**10.6**
	BERT	Our	0.835	0.731	-0.104	**1137**	26	13.4
	T5	Our	0.827	0.663	**-0.164**	534	**41**	78.3
SEB UQ	BERT	TF	0.655	0.527	-0.128	148	47	**10.5**
	BERT	Our	0.655	0.489	-0.166	1941	61	16.2
	T5	Our	0.755	0.546	**-0.209**	**2930**	**77**	94.4
SEB UD	BERT	TF	0.760	0.612	-0.149	1237	331	**75.4**
	BERT	Our	0.760	0.607	-0.153	**19481**	342	94.8
	T5	Our	0.724	0.554	**-0.170**	13652	**379**	600.7
Textual Entailment Task								
MNLI matched	BERT	TF	0.832	0.649	**-0.182**	2258	**569**	**154.2**
	BERT	Our	0.832	0.731	-0.101	**4821**	313	196.5
	T5	Our	0.766	0.666	-0.100	3058	311	913
MNLI mismatched	BERT	TF	0.816	0.636	**-0.179**	2428	**561**	**185.1**
	BERT	Our	0.816	0.710	-0.106	**5920**	329	219.7
	T5	Our	0.773	0.669	-0.105	4542	328	1027.2
RTE	BERT	TF	0.603	0.443	**-0.160**	53	**21**	**4.0**
	BERT	Our	0.603	0.481	-0.122	**147**	16	5.0
	T5	Our	0.664	0.542	-0.122	43	16	54.8
Paraphrase Detection Task								
MRPC	BERT	TF	0.694	0.561	-0.133	387	77	**34.7**
	BERT	Our	0.694	0.590	-0.104	4022	60	43.2
	T5	Our	0.734	0.516	**-0.218**	**5316**	**126**	427.8
Average over all tasks								
	BERT	TF	0.742	0.597	-0.145	942.6	**232.4**	**67.8**
	BERT	Our	0.742	0.620	-0.122	**5352.7**	163.9	84.1
	T5	Our	0.749	0.594	**-0.155**	4296.4	182.6	456.6

Table 9: The automatic grading models' accuracy before (Acc.) and after the attack (AaA) for our proposed attack and TextFooler (TF). We also report the absolute loss of accuracy incurred (ΔAcc), the number of found adversarial examples (#Adv), the number of student responses affected (#Aff) and the searches' runtimes in minutes (Time). The best values for each metric are highlighted in bold. Table adapted from [43].

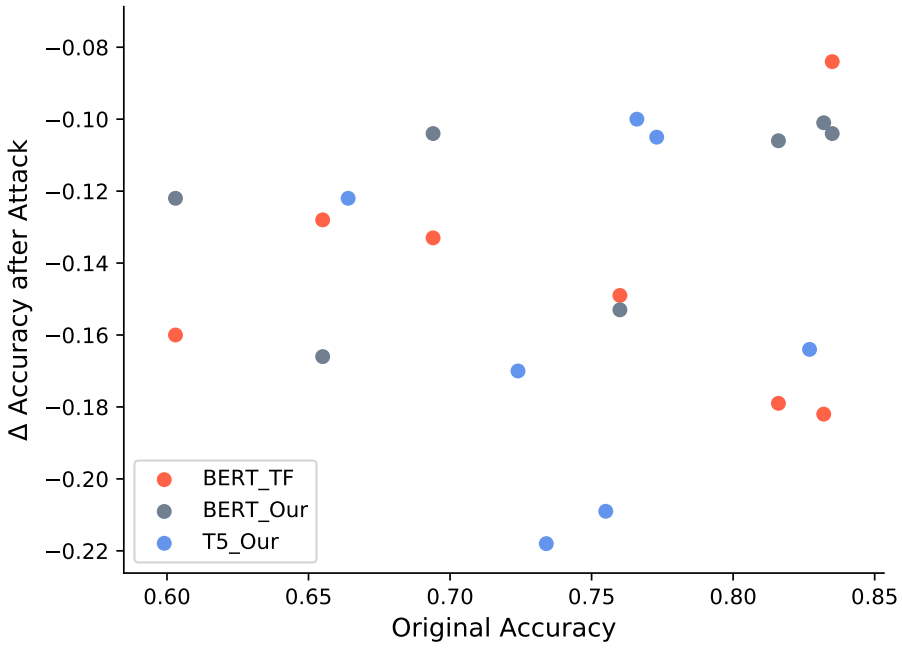


Figure 3: Scatter plot depicting the accuracy before the attack and the resulting accuracy shift after the attack. Note that lower on the y-axis means the attack had more success. There seems to be no apparent relation between the dimensions.

grader for novel questions and domains is more complex than training a grader for a set of questions with data for each question.

Surprisingly, the original model performance seems to have less of an effect on the attacks' success rate than expected. Figure 3 shows a scatter plot of the original accuracy and the change in accuracy after the attack. Visually, there is no relation between the two dimensions, even though one may expect more accurate models to behave differently than less accurate ones. However, effects may become apparent when expanding upon this experiment with additional datasets, attacks and models.

On average, TextFooler ran 16.3 minutes faster compared to our attack across all tasks. Considering that TextFooler uses the target model's raw class probabilities to inform its search, it is expected to find adversarial examples quicker than our purposefully unguided search. Although our proposed attack utilizes less information, it degrades the target model's accuracy by an additional 0.4 - 3.8 percentage points compared to TextFooler on the ASAG task. On the other tasks, however, TextFooler outperforms our attack by 2.9 - 8.1 percentage points, causing it to degrade the target model's accuracy further on average (14.5% compared to 12.2%). Across all tasks, our attack degrades BERT's accuracy by 8.4 - 18.2% ($\bar{x} = 12.22$, $\sigma = 2.66$).

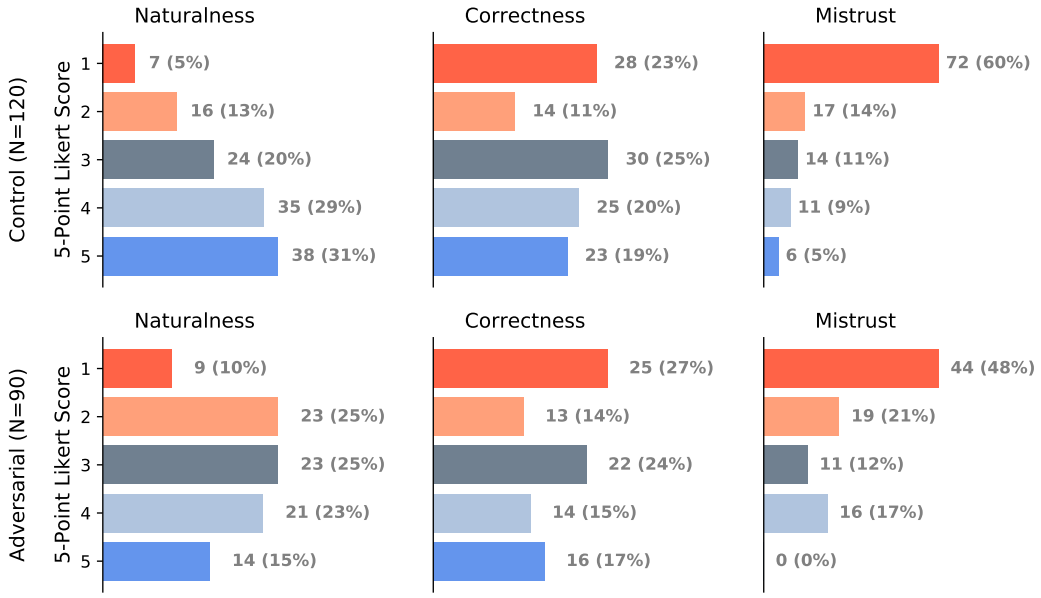


Figure 4: Raw distribution of assigned Likert scores for the correctness, naturalness and mistrust scale. The top row contains the control group's ratings, while the bottom row shows the experimental group's. A one on the scale indicates that the response was unnatural, incorrect or trustworthy, while a five corresponds to the opposite end of the spectrum. The ratings' absolute occurrence counts and percentages are displayed next to the respective bars. This figure was taken from [43].

While the T5 model slightly outperformed the BERT model by 0.7 percentage points on average, the increased accuracy seems mostly based on exploitable statistical features. Our attack degraded T5's accuracy by 15.5 percentage points across all tasks, which is 3.3 points more than it was able to achieve when targeting BERT. T5's accuracy after the attack is, in fact, lower than BERT's for all tasks besides RTE and SciEntsBank unseen questions. Our attack took significantly longer to run on T5 than on BERT, mostly due to the higher computational cost of querying T5. The longest runtime of 17.12 hours was on MNLI mismatched.

Risk of Detection & Class Equivalency

We will now present the human evaluation's results of the attack's effect on the correctness, naturalness and suspiciousness of student answers. Since experts' time is expensive, the following analysis was conducted only on the Short Answer Grading dataset SciEntsBank. The raw distribution of ratings assigned to student responses

	Naturalness			Correctness			Mistrust		
	α	M	SD	α	M	SD	α	M	SD
C. (N=120)	0.29	3.68	1.21	0.51	3.01	1.42	0.13	1.85	1.23
Exp. (N=90)	0.29	3.09	1.23	0.55	2.81	1.44	-0.11	1.99	1.15

Table 10: Descriptive statistics of the graders' ratings. Specifically, Krippendorff's α, mean (M) and standard deviation (SD) are reported. Table adapted from [43].

and adversarial examples can be seen in Figure 4. Means, inter-annotator agreement and standard deviations can be found in Table 10.

We used two one-sided tests, as proposed by Wellek [165], to test whether the correctness of responses after the attack is less or equal to the original answers' and, thus, whether our attack adheres to the class equivalency constraint (H1). We average the scores assigned by the annotators in each group to obtain a more reliable measurement and use the non-parametric Mann-Whitney U test because our data is ordinal. To test for noninferiority instead of equivalence, we select -∞ as the lower bound [82] and 0.5 as the upper bound. The collected evidence supports H1 ($n_1 = n_2 = 30$, $U_{control} = 597.5$, $U_{adv} = 302.5$, $p = 0.015$), that is that human graders generally awarded less or equal points to the adversarial examples than the original student responses. Thus, the attack adheres to the class equivalency constraint.

Similarly, we use two one-sided Mann-Whitney U tests to test for noninferiority on the mistrust scale. This test concerns H3, where the attack should not be detectable as cheating attempts by human graders. Using the same bounds as in the previous test, our observations are consistent with H3 ($n_1 = n_2 = 30$, $U_{control} = 576$, $U_{adv} = 324$, $p = 0.031$). Thus, graders generally thought students were cheating less or equally as often in the experimental group compared to the control group. This was also reflected in the question of whether graders would take action based on their suspicion. While graders declared that they would act 5 times (N=90) in the experimental group, they wanted to act 14 times (N=120) in the control group. Table 11 displays one of the most suspicious student responses and adversarial examples, illustrating that it can be difficult to differentiate between poorly written responses and unnaturally modified adversarial examples. This factor was also explicitly mentioned by one of the human graders.

Finally, we conduct a left-tailed Mann-Whitney U test to analyze whether the proposed attack decreases the naturalness of student responses (H2). The collected evidence is consistent with H2 ($n_1 = n_2 = 30$, $U_{control} = 627$, $U_{adv} = 273$, $p = 0.004$, $Z = -2.6174$, $r = 0.34$), indicating that humans perceive the adversarial examples as less natural even if they do not suspect cheating.

Question:	If Phil, a geologist, wants to test for calcite while in the field, what should he bring with him? (an acid such as vinegar). Describe what Phil should do to test for calcite and what he would observe.
Reference:	Put acid on a rock. If the acid fizzes, Phil would know that the rock has calcite.
Answer:	He would put vinegar on the rock get the strange then is quiet it and see if there is calcite.
Question:	The sand and flour in the gray material from mock rocks is separated by mixing with water and allowing the mixture to settle. Explain why the sand and flour separate.
Reference:	The sand particles are larger and settle first. The flour particles are smaller and therefore settle more slowly.
Adversarial:	Because one is heavy and high one is not.

Table 11: Examples for the most suspicious student response with two votes for action (top) and adversarial example with one vote for action (bottom). Table adapted from [43].

Reliability of Human Annotation

We selected Krippendorff's Alpha[5] to measure inter-annotator agreement and, thus, estimate the reliability of our study, since it is versatile, makes few assumptions and handles multiple annotators well [8]. Especially for naturalness and mistrust, α was comparatively low (see Table 10). For mistrust, agreement between the annotators was expected to be low, considering that this scale was purposefully left to subjective interpretation by the graders. We only informed graders that cheating may have occurred but did not provide a guideline of what cheating attempts could look like. The slight agreement ($\alpha = 0.13$) in the control group and slight disagreement ($\alpha = -0.11$) in the experimental group indicate that graders formed individual internal models of possible cheating behavior.

Beyond varying α levels, there were also between-group differences in how the mistrust scale correlated with naturalness and correctness. While the scales were hardly

5 Krippendorff's α is an agreement coefficient, ranging from -1 (perfect disagreement) to 1 (perfect agreement). Properly interpreting the results of this metric – and other inter-annotator agreement measures – has been an open debate for multiple decades. Originally, thresholds of 0.8 for good and 0.67 tentative reliability have been proposed [8]. Since these thresholds are considered too strict for some fields and applications, interpretations similar to correlation coefficients have also been proposed. Here, values smaller than 0 are considered "poor," 0-0.2 "slight," 0.2-0.4 "fair," 0.4-0.6 "moderate," 0.6-0.8 "substantial" and 0.8-1 "almost perfect" agreement [7].

correlated in the experimental group with Spearman's rank correlations (ρ)[6] of 0.2 for naturalness and 0.07 for correctness, there was a moderate negative correlation with naturalness (ρ = -0.41) and correctness (ρ = -0.51) in the control group. These correlations indicate that graders are more mistrustful of poorly written and wrong answers when no true cheating patterns exist.

However, the rather low agreement (α = 0.29) for naturalness was unexpected. As it is not uncommon in the NLP field to have annotators disagree more than in other fields due to human language variability [7], we inspect the correlation between each annotator pair as recommended by Amidei, Piwek, and Willis [7]. Intuitively speaking, agreement metrics, such as α, measure the extent to which two annotators have the exact same opinion, while ρ measures to which extent two annotators would rank the texts in the same order according to the annotation scale. For instance, if one annotator is stricter than the other and always assigns exactly one point less to every text on the given scale, they would have a low α but perfect ρ. Considering that our annotators have various language backgrounds, they likely have different standards for naturalness and, thus, differ in strictness, reducing agreement but not affecting rank correlation.

Indeed, while one of the annotators in the control group was an outlier whose opinion hardly correlated with the other annotators (pairwise ρ's of 0.14, 0.07 and -0.02), the rest averaged a moderate correlation of ρ = 0.57. We did not exclude the outlier since their judgment was more in line with the majority for correctness and mistrust, indicating that they may have a significantly different opinion on what makes a response natural instead of answering randomly in the survey. The experimental group also averaged a moderate correlation with ρ = 0.47.

The agreement levels for correctness were in line with expectations, with moderate agreements in the control group (α = 0.51, ρ = 0.6) as well as the experimental group (α = 0.55, ρ = 0.61).

Possible Reasons for the Attack's Success

Finally, we explore two potential reasons for the attack's effectiveness. First, the attack may only be able to flip student responses that the model is unsure about in the first place. Should this be the case, educators could pass on low-confidence predictions to human graders for verification and, thus, prevent the attack from succeeding. We analyze the target model's outputted class probabilities for *incorrect* student responses to estimate its confidence and display the scores before and after the attack in Figure 5.

6 Spearman's Rank correlation is a nonparametric measure for the strength and direction of a monotonic relationship between two paired variables. It ranges from -1 to 1. Typically, absolute values between 0 and 0.1 are considered a "negligible" correlation, 0.1-04 "weak," 0.4-0.7 "moderate," 0.7-0.9 strong and 0.9-1 "very strong" [142]. The sign shows the direction of the monotonic relationship.

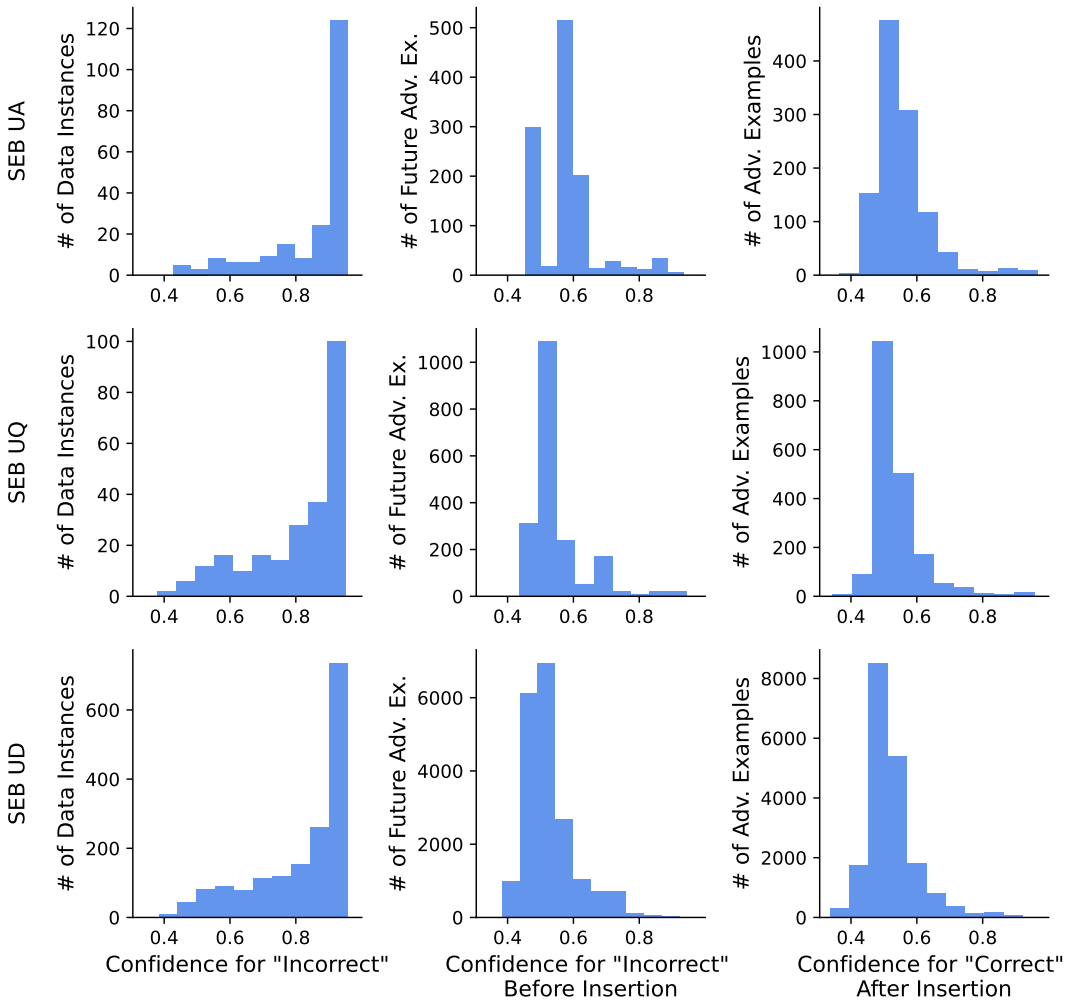


Figure 5: Class probabilities outputted by BERT for all true incorrect samples (left), all future adversarial examples before the attack (middle) and all adversarial examples after the attack (right). Figure taken from [43].

Generally, the model had high confidence scores for responses it classified correctly. In fact, the majority of responses elicited confidence scores between 0.8 and 1. However, when looking at the soon-to-be adversarial examples, the peak of the distribution shifts to 0.45 - 0.65. Thus, future adversarial examples are much closer to the decision boundary of 0.33 compared to responses that are robust to our attack. Even after the attack, the target class' probability is similarly low for most adversarial examples. Overall, this evidence supports our expectation that the attack is primarily successful on low-confidence predictions.

Second, the target model may use spurious correlations between the inserted words and the target class to make predictions. Should this be the case, educators could screen their training data for similar correlations to inform potential countermeasures

or highlight the need to collect additional data from subgroups of their student populations. We analyze the distribution of successful adverbs and adjectives across classes in the training data to search for the expected spurious correlations. When plotting the most successful adjective's and adverb's occurrences in the training data, we observed two patterns for the majority of insertion words. The adjectives and adverbs either appeared more often in *correct* student responses than *incorrect* ones or they hardly occurred at all (see Figure 6). While Figure 6 only displays the top ten adjectives and adverbs as measured on the unseen answers test split, the figures look similar for the best insertion words found for unseen questions and unseen domains. The only exception is the adjective "better," which appeared 15 times in the *incorrect* class and only 4 times in *correct* responses.

Of the rarely occurring words, a portion seems to be synonymous or at least very similar to words that appear more often in *correct* responses. For example, "entire" is a synonym of "complete," and "completely" is the adverb of "complete." These words are expected to have very similar embedding vectors and, thus, will likely affect the model's predictions comparably. When comparing the occurrences, it is important to keep in mind that SciEntsBank's training set is slightly skewed towards *incorrect* student responses (2462) compared to *correct* ones (2008). Additionally, *correct* responses are slightly longer with 13.4 words per answer, on average, compared to 11.7. On average, *correct* responses also contain more adjectives (1.1) and adverbs (0.6) per answer than *incorrect* ones (0.8 and 0.5, respectively). In summary, there is a spurious correlation between adjective and adverb use and the correctness of responses in SciEntBank's training set, which the grading models incorporated in their prediction process.

3.3.5 *Interpretation of Results & Limitations*

To summarize, the proposed attack of adversarially inserting adjectives and adverbs reduced target models' accuracy by 8 - 22 percentage points. It performed only slightly less effective than TextFooler, a powerful attack utilizing raw class probabilities to manipulate individual examples instead of generating universally applicable modification patterns. Based on evidence collected in a between-group study with human graders, the proposed attack complies with the design consideration posed in Section 3.1. Namely, it only requires access to the target model's final prediction, is not apparent as a cheating attempt, has a reasonable runtime and retains the original samples' class. However, the attack reduced the texts' perceived naturalness and, thus, may be identifiable with further training. We also observed that primarily low-confidence predictions were vulnerable to this particular attack and found evidence that the attack

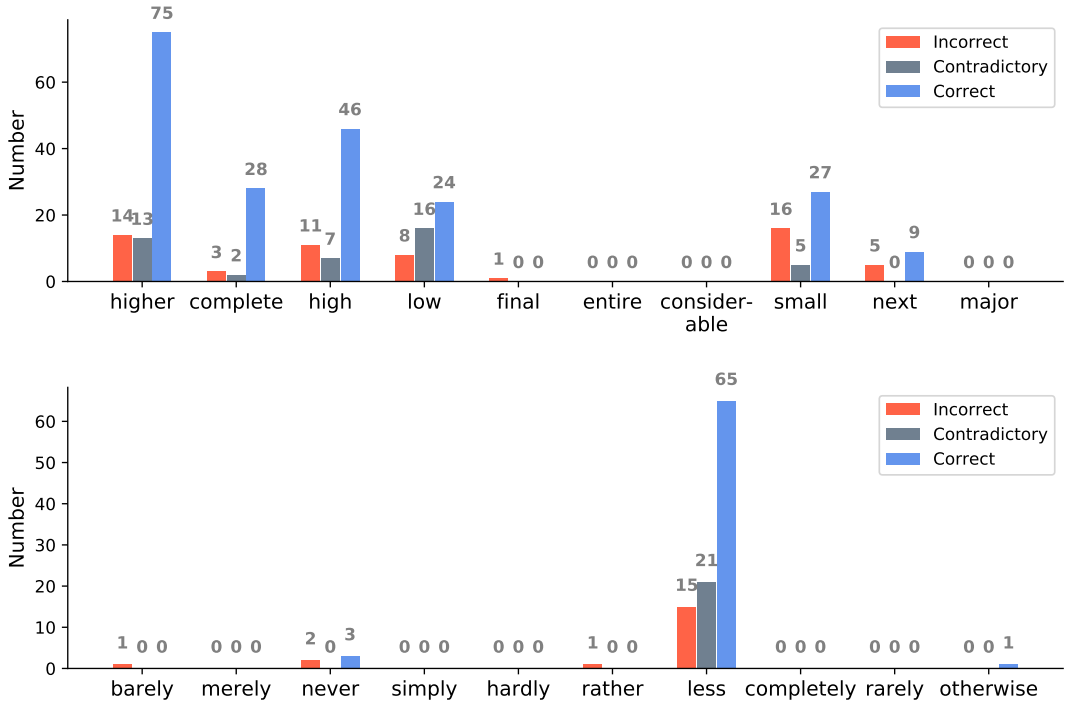


Figure 6: Class-wise occurrences of the 10 most successful adjectives and adverbs (on the unseen answers split) in the SciEntsBank training set. Figures taken from [43].

exploits spurious correlations between adjectives and adverbs with the target class to succeed.

While the analyses conducted yielded valuable insights into the fragility of automatic grading models, there are a number of limitations that could be addressed in future work. Firstly, all of our human graders were non-native English speakers from various countries, even though they spoke English fluently and regularly graded English responses to short-answer type tasks. While background diversity is, in principle, desirable since there is a massive variety in the worldwide teacher population, the data annotated stemmed from US school children. Thus, it is likely that there was a mismatch between the English dialects spoken by the graders and the students. As also indicated by the low inter-annotator agreement, this probably impacted the graders' estimation of the responses' naturalness. Especially US school teachers may be more sensitive to synthetic manipulations of student responses and, therefore, should be included in future studies. Nevertheless, the effect on the correctness and mistrust scales is likely minor.

Generally, it would also be beneficial to include additional datasets from other educational institutions in future studies. While we showed that our attack is applicable to other domains and datasets, especially the mistrust variable may change in different educational contexts. For example, inserting random adjectives and adverbs into

college-level short answers may be more suspicious compared to middle-school answers. Additionally, the number of samples annotated by the human graders could be increased to detect minor effects reliably. While mistrust was not increased significantly in our study, a larger sample size may reveal minor effects. However, annotating a larger sample may require multiple grading sessions, as more than an hour per session may become tiring.

Finally, the attack itself can be improved. More sophisticated insertion patterns could produce more natural adversarial examples, while the list of candidate adjectives and adverbs could be expanded for greater versatility. Further optimization of the search algorithm could also reduce the number of times the target model has to be queried and reduce the attack's overall runtime. While we did not include information about the target model per design, general linguistic knowledge about common sentence structures could be utilized to find promising adversarial examples faster. Currently, the attack only supports inserting exactly one adjective or adverb in an input text. With better search heuristics and insertion patterns, multiple adjectives and adverbs could be added with a lower risk of violating the class equivalency constraint. Being able to insert more than a single word would likely boost the attack's success rate significantly.

3.4 STUDENT ATTACKS

We have seen successful adversarial attacks designed by researchers in the previous sections. Nevertheless, the question stands whether students themselves can fool automatic grading models. Considering that news articles about machine learning models' weaknesses exist by now[7], it is probable that students will learn of adversarial attacks[8]. As it is increasingly easy to use existing models and adversarial attacks with the introduction of high-level abstraction libraries, such as Keras[9] and HuggingFace, a subset of students will probably be able to employ automatic adversarial attacks to find model weaknesses directly. However, we do not expect most students to have the knowledge and motivation to do so. For this reason, we conduct an exploratory study to investigate students' ability to find a grading model's weaknesses themselves – without pre-scripted attacks.

3.4.1 *Methodology*

We tasked students in a university course about educational technologies to construct answers a given ASAG system could not judge correctly. For this purpose, we implemented a web interface where students could select basic science and computer science questions, submit answers and receive the predicted classes. They had no other insight into the grading model's inner workings to resemble typical assessment scenarios where students receive feedback from the model during formative assessments but don't have access to the model's parameters or training process.

Students were not limited in the number of responses they could submit to the model. Submissions were anonymous to encourage creativity. The students provided the response's true class for each response they submitted. The possible classes were SciEntsBank's *correct*, *incorrect* and *contradictory*, with *contradictory* covering responses that have precisely the opposite meaning of the correct answer to a question. The response, the true and the predicted label were logged to evaluate the gradings model's performance later. Finally, during a graded exercise, the students were tasked with reporting the five most intriguing examples they had created. They should also write a short free-text comment describing their opinion of the model's performance. The students' comments were processed using the summarising content analysis according to Mayring [103] to identify commonly identified weaknesses. In this type of analysis, the researchers develop a coding scheme – in our case, types of model weaknesses –

7 https://hbr.org/2021/01/when-machine-learning-goes-off-the-rails [accessed April 25, 2023]
8 https://www.theverge.com/2020/9/2/21419012/edgenuity-online-class-ai-grading-keyword\-mashing-students-school-cheating-algorithm-glitch [accessed April 25, 2023]
9 https://keras.io/ [accessed April 25, 2023]

based on the text material and then apply the scheme to the material. The final categories extracted from the responses can be seen in Table 13.

3.4.2 *Target Domain & Grading Model*

The short answer questions used in this study stem from the SCIENTSBANK dataset and the computer science dataset proposed by Mohler et al. [108]. As most students took college-level computer science classes at the time of this study, they should have the education necessary to answer the basic science questions in the datasets correctly. We chose three short questions from SCIENTSBANK's training set that did not reference external material removed from the dataset, such as tables or images. The questions can be seen in Table 14 and Table 15. All 80 computer science questions were included. However, students were asked to focus on the SCIENTSBANK questions, as these were part of the grading model's training set. In practice, automatic grading models will likely also be employed to grade questions they were trained on. We selected Model #2 from Table 5 in Section 3.2 as the grading model due to its high performance. It is a BERT model fine-tuned on SCIENTSBANK's training set.

3.4.3 *Participant Characteristics*

The students were enrolled in the Technical University of Darmstadt in Germany, taking an educational technologies course. Of the 24 course participants, 13 were male and 11 female. On average, the students were 25 years old during this study. Self-reportedly, 96% spoke English either proficiently or fluently. Other fluently spoken languages were German (22) and Chinese (2). The task description was given in English and German. Half of the students reported a general university entrance qualification as their highest educational degree, while the other half reported having a Bachelor's degree. The majority of students studied computer science (16), while the rest studied electrical engineering and information technology (1), engineering economics (1), cognitive science (1), computer science teaching (1), and "other"(4). The total number of semesters enrolled in a university varied between 3 and 15, with a rounded average of 8 semesters. 14 participants completed the graded exercise and commented on the model's weaknesses. However, all 24 students in the course had access to the web interface and could have submitted responses, as the answer submission was anonymous to overcome student reluctance to answer freely due to fear of losing reputation or appearing foolish.

Additionally, students self-reported their prior experience with Natural Language Processing (NLP) at the beginning of the course. The students were asked to rate

```
”data[””scores””].forEach(function(score, index) {
        containers[index].getElementsByClassName(””score””)[0].inner-
          Text = ””correct””;
                                  });”
```

Figure 7: Student remote code execution attempt.

eight statements in terms of self-applicability on a 4-point Likert scale. The statements ranged from ”I have already attended many theoretical courses concerning this topic” or ”I am confident in using NLP libraries or models in my own applications” to ”I am able to define the terms POS-Tagging, Embedding or Tokenization clearly and comprehensively”. An analysis of the responses showed that around half of the students had no prior experience in this field, while three students reported having a strong NLP background. However, since the course lectures prior to this study included an introduction to NLP topics such as automatic question generation and ASAG, we can assume that most students had at least a basic understanding of NLP.

3.4.4 *Results*

Students submitted 620 answers in total. However, the answer depicted in Figure 7 was excluded from the further analysis because it was a multi-line remote code execution attempt. We were surprised to see this attempt at subverting our grading infrastructure, as we expected linguistic attacks. The remaining 619 responses were labeled as *correct* (262), *incorrect* (328) and *contradictory* (29) by the students. A surface-level inspection of the assigned classes revealed that the students made only a few mistakes and were generally able to provide the true class for their responses. All responses were published as the Automatic Short Answer Grading Adversarial Dataset on Github[10]. In the following sections, we present how often the students managed to induce a misclassification in the grading model, followed by an analysis of their textual comments on the model's weaknesses and enlightening example responses.

Student Success Rate

The model's confusion matrix, recall, precision and F1 score on the students' responses can be seen in Table 12. The model seems biased towards labeling responses as *incorrect* as indicated by the high recall and low precision. This aligns with the model's grading behavior on the original SciEntsBank data. Despite this bias, students were able to get 13.4% of their *incorrect* responses graded as *correct* by the model. One can also

10 https://github.com/PumpkinPieTroelf/ASAG-Adversarial-Dataset [accessed April 25, 2023]

		Predicted Class			Classwise Metrics		
		correct	incorrect	contra.	Precision	Recall	F1-score
True Class	correct	130	129	3	0.71	0.50	0.58
	incorrect	44	273	11	0.65	0.83	0.73
	contra.	10	19	0	0.00	0.00	0.00

Table 12: The automatic grading model's confusion matrix of the students' responses. Table adapted from [46].

expect that most students stopped submitting responses once they identified enough weaknesses to complete the exercise. Thus, a better success rate is likely achieved in practice when students are interested in exploiting the identified weaknesses for better grades.

The model did not classify a single *contradictory* answer correctly. This can partly be explained by the class's underrepresentation in the training data. However, the model did predict *contradictory* for a few of the *correct* and *incorrect* responses. Therefore, it cannot solely be a matter of the model categorically declining to predict the minority class.

Identified Weaknesses

In Table 13, we see the weakness categories identified during the content analysis of the students' comments, along with the number of their occurrences and examples. Tables 14 and 15 display a subset of the student responses misclassified by the model. Noteworthy is that these questions stem from the model's training set. Especially the answers to the heat sink and plant questions indicate hypersensitivity to specific keywords. This weakness was also consciously identified and reported by 6 of the 14 students. Three of the students noted a disregard for negation or inversion of answers. Multiple contradictory, but classified as *correct*, student responses in Table 14 and Table 15 illustrate this vulnerability to negation. The lexical closeness to the reference answer seems enough to fool the model. The data also reflects the model's bias towards labeling answers as *incorrect*. Eight students criticized the model's tendency to refuse genuinely correct answers. Even exact quotes from Wikipedia definitions were rejected in some cases.

Additionally, four students reported that small insignificant changes in the formulation of an answer led to vastly different predictions. An example of this behavior is depicted in Table 15, where prepending "None exists" to a correctly labeled *incorrect* answer leads the model to predict it as *correct*. It is fair to assume that the Universal Trigger Attack (see Section 3.2) inspired this particular student response. The last

Category	#	Example Snippet
Sufficient	1	"In my opinion the model works very well."
Plagiarism	1	"Answers copied from Wikipedia were marked as correct. While this is true, a teacher would probably have realized that the answer is "copied".
Inversion	3	"In addition, the examples suggest that sentences that are worded similar to the correct answer, such as Something that emits heat. for question 1 are still classified as correct although the difference means that they mean the opposite."
Brittleness	4	"The model is not very robust, because small changes to the inputs falsify the result"
Keywords	6	"The model rates answers as correct if they contain the correct words. Whether these words are in the correct order, that is whether they fulfill the right roles in the sentence, or whether the syntax of the sentence is even approximately correct, is not checked."
False Negative	8	"I have not been able to get an answer correctly rated as correct except in question 2, although the answers were valid and I tried them out in different versions."

Table 13: Categories of weaknesses resulting from the content analysis of the students' comments. The number of the categories' occurrences in the comments can be found in the second column. The last column contains examples of comment snippets for each category – sometimes translated from German. Table adapted from [46].

example answers for the heat sink question also reflect this brittleness, as the only difference consists of appending "heat transfer" to the answer, changing the model's prediction. Finally, one student stated that most of his attacks were unsuccessful and concluded that the model worked very well, while another student remarked on how a human, in contrast to the model, would likely be able to identify plagiarised answers in the form of copied sections from Wikipedia.

3.4.5 *Interpretation of Results & Limitations*

In conclusion, we have seen that this study's students could identify systematic weaknesses of a BERT-based ASAG model just by receiving formative feedback – even for questions the model was explicitly trained on. The identified weaknesses included hypersensitivity to keywords and small linguistic changes, a disregard for negation, a

Question:	What is a heat sink?
Reference Answer:	A heat sink is any material that absorbs (a lot of) heat.
Labeled *correct*:	the heat sink emits heat and tries to melt the processor
Labeled *correct*:	Something that emits heat.
Labeled *incorrect*:	It cools its surrounding.
Labeled *incorrect*:	A device that let's you reduce the temperature of something that emits too much thermal energy. Some of the thermal energy has to be lead outside of the system via a medium for example water or air or metal. That auxiliarry device is called heat sink.
Labeled *correct*:	A device that let's you reduce the temperature of something that emits too much thermal energy. Some of the thermal energy has to be lead outside of the system via a medium for example water or air or metal. That auxiliarry device is called heat sink. heat transfer

Table 14: Examples of misclassified student answers to SCIENTSBANK's heat sink question.

Question:	When a seed germinates, why does the root grow first?
Reference Answer:	The root grows first so the root can take up water for the plant.
Labeled *correct*:	Because the seed needs to stay away from water.
Labeled *correct*:	The plant needs no water
Labeled *correct*:	plant water.
Labeled *correct*:	Because Chewbacca eats plant water.
Labeled *correct*:	The seed contains much water, so the root pumps it into the ground.
Labeled *incorrect*:	Because the seed needs liquid.
Question:	How do you define a controlled experiment?
Reference Answer:	An experiment is controlled if only one variable is changed at a time.
Labeled *correct*:	None exists An experiment with only one person.
Labeled *correct*:	A controlled experiment is one in which nothing is held constant except for one variable.

Table 15: Additional example student answers that were misclassified by the model. Table adapted from [46].

lack of plagiarism detection, and correct answers not being accepted by the model. There are multiple ways students could exploit such weaknesses in summative assessments, e.g., by including all manner of plausible keywords in their responses. The model only identified half of the *correct* answers as such. Additionally, 13.4% of deliberately *incorrect* answers were falsely graded as correct even though the model is biased towards predicting the *incorrect* class. For their attacks, the students did not have access to details of the model or the reference answers, a scenario mirroring real-world usage of such models.

However, the results of this study cannot be directly generalized to the college student population as a whole without further research. The limiting factors are the number (14) of active participants and mostly homogeneous student backgrounds, as most students studied computer science and had at least a basic understanding of NLP. In future research, a large-scale study including students of various study programs and linguistic backgrounds would allow a better estimation of the students' fooling capabilities. Especially younger students from middle or high schools may yield vastly different results compared to the college-level students included in this study. Additionally, such a larger sample would result in more adversarial examples, which could then be released as a benchmark adversarial test set to compare future approaches. Nevertheless, this study shows that at least the student group represented here can capitalize on the systematic weaknesses of ASAG models. Furthermore, this study affirmed the need for more robust automatic grading techniques before such models should be employed in summative assessments in practice.

3.5 DISCUSSION OF ATTACKS

Overall, we have shown that current state-of-the-art Automatic Short Answer Grading models are vulnerable to adversarial manipulations. Whether it be by inserting meaningless trigger words (see Section 3.2) or adjectives and adverbs (see Section 3.3), systematic weaknesses can be automatically discovered and generally exploited quickly, easily and without machine learning expertise during test time. Even low-risk strategies, such as inserting a single adjective or adverb, can fool the model into giving a better grade 8%-22% of the time while appearing no more suspicious than regular student answers to humans. A model's accuracy can quickly be degraded to unacceptable levels in such a manner, considering the base misclassification is often around 20%. Beyond automatic attacks, we also observed that subgroups of college-level students are capable of identifying models' weaknesses (see Section 3.4) if allowed to receive repeated feedback from the model – as would be the case for models used in formative and summative assessments.

While more extensive studies are required to thoroughly investigate students' capabilities and willingness to cheat current grading models in practice, it is clear that the models' predictions are at least partially based on undesirable features and spurious correlations in the training data. We base this claim on evidence collected in this thesis as well as concurrent work in the automatic grading field [33] and wide-spread observations of bias in other fields [106]. Undesirable features are not only concerning when considering potential cheating behavior but may also disadvantage subgroups of students. In this work, we focused on patterns associated with good grades, yet they likely also exist for bad grades. For example, should student responses written in a regional dialect be primarily incorrect in a model's training data, the model may grade new responses of the same dialect as incorrect – even though the answer's dialect is not typically in a causal relation with the answer's factual correctness. While further research on bias in the existing datasets is needed to determine the extent of the problem, a set of guidelines for using automatic grading systems in practice can be already formulated based on our current understanding:

- **Know your dataset.** Especially in text processing, it is increasingly common to use off-the-shelf models that are already pretrained or even fine-tuned on diverse tasks. The most powerful NLP models are trained on enormous computation clusters for months, so it makes sense to build upon what already exists. However, it invites the mindset of considering models as black boxes where the training process' particularities cease to matter. Nevertheless, biases present in the training data will influence the model's predictions later on, regardless of where or by whom the training was conducted. While it is likely impossible to

collect a bias-free training set large enough to perfectly reflect the target population, being aware of bias can foster countermeasures. Similar to our analysis of the adjective and adverb occurrences, investigating statistical correlations in the training data can reveal undesirable predictive features. Many adversarial attacks are, in fact, designed for debugging NLP models [129]. Enriching the training set with adversarial examples is also called adversarial training. Even though adversarial training often trades off with the model's predictive performance on the original data and typically does not generalize to new avenues of attack, it can reduce a model's vulnerability to known spurious correlations. However, while adversarial training can lessen the adverse effects of spurious correlations, it is not a final solution.

- **Keep educated human graders in the loop.** Even though automatic grading models are approaching human accuracy on specific datasets, grading models are still imperfect. At least a portion of the model's predictions are based on undesirable characteristics of the students' responses, making them vulnerable to manipulation. Human graders can perform random quality checks and regrade suspicious responses. For example, suspicious responses could be identified with authorship verification tools [119] and flagged for human grading. Nevertheless, they should be educated on how the model works and what cheating attempts could look like. Human graders developed individual theories of cheating behavior in the absence of reliable information in our human evaluation of the adjective and adverb insertion attack. Unfortunately, there was a correlation between mistrust and the response's perceived naturalness and correctness. Thus, human graders may wrongfully suspect students less proficient in the respective language or low-achieving students. Awareness and education may mitigate such discrimination. Detectors for cheating behavior should also be constructed carefully, as minorities may appear similarly anomalous to adversarially crafted input.

- **Utilize estimators of a prediction's reliability.** In our work, the raw class probabilities outputted by the grading model were a good indicator of predictions that had been injected with adjectives and adverbs. Referring such problematic predictions to a human grader for verification could have prevented most adversarial examples from succeeding. However, class probabilities are not the only and possibly not the best confidence estimator [27]. Further research is necessary to determine the relation between various adversarial attacks and confidence estimators.

- **Beware of disclosing a grading model's vulnerabilities.** While transparency in the grading process is vital for students to understand and accept given grades, it can also expose unreliable features to students. On the one hand, this can cause them to lose trust in the model's predictions, even if most are sound. On the other hand, it also enables more powerful adversarial attacks. Considering the already significant success rate without prior knowledge of the model's inner workings, a model's performance may be reduced to uselessness. One may impede an attacker's attempt to glean information about the grading model by querying it. For example, time limits for retries in exercises can slow automatic querying while not disrupting legitimate students. Generally, unusual activity during the formative assessment phase can be flagged and referred to human graders for inspection.

4

ELABORATED FEEDBACK GENERATION

While the previous chapter focused on potential vulnerabilities of Automatic Short Answer Grading (ASAG), we now turn our attention to enhancing students' understanding with automatic assessment. Here, providing elaborated feedback instead of a mere score is essential. As no datasets with elaborated feedback are publicly available at the time of writing this thesis, we first consider the requirements a dataset would need to fulfill to serve as a benchmark and enable the elaborated feedback generation task. We then present the Short Answer Feedback Corpus (SAF), a collection of responses and elaborated feedback in three educational scenarios explicitly collected for publication as a benchmark. Next, we introduce supervised feedback generation approaches based on the corpus. Finally, we also present an unsupervised approach in Section 4.4 for domains where a costly data annotation process is infeasible.

4.1 BENCHMARK DESIGN CONSIDERATIONS

To kick off the elaborated feedback generation task, a dataset for training and testing various approaches is needed. It should contain grades that are a numerical verification of a response's correctness and a textual explanation of the assigned grades. In the next sections, we will consider and further describe the following design criteria for a high-quality benchmark:

- The assigned grades should be **reliable**.
- The given grades and elaborations should conform to **pedagogical guidelines**.
- The benchmark should enable **reproducing and comparing** feedback generation approaches.
- It should cover a **diverse** set of questions.
- The dataset should have an **appropriate size**.
- The collection and publication process should conform to **ethical guidelines**.

4.1.1 *Reliable Grading*

Reliable grading is challenging to achieve in short answer formats where there are many degrees of freedom in interpreting a student's response and judging its correctness [63, 146]. Not only do various graders differ in strictness, perception and prior knowledge, but even a single grader will likely grade inconsistently depending on their level of concentration, emotional state and responses they have graded previously. Fortunately, comprehensive grading rubrics and training can mitigate grading variability by specifying expectations and offering a shared knowledge basis. A single author is typically insufficient for crafting a suitably detailed, unambiguous and bias-free rubric [24]. For these reasons, a team of authors should coalesce various perspectives into a common ground truth rubric that can guide annotators in making reliable grading judgments.

However, a comprehensive grading rubric alone does not guarantee a reliable and objective annotation. Graders will still likely make mistakes due to fatigue and deviate from the grading rubric with time. Therefore, multiple annotators are required to obtain reliable feedback. Answers should be annotated by at least two annotators to catch mistakes and enable the calculation of reliability measures, such as inter-annotator agreement.

4.1.2 *Pedagogical Guidelines*

Elaborated feedback should conform to pedagogical guidelines. While various aspects of feedback quality are still debated and depend on the concrete learning scenario, a set of applicable recommendations can be extracted from large-scale surveys [144, 169, 170]:

- **Feedback should focus on the learner's response instead of the learner themselves.** Drawing the learner's attention away from the task to the learner's self can even be detrimental to learning. For this reason, comparisons with other students and overly critical feedback should be avoided, as they may harm the learner's self-esteem and distract them from learning. The feedback's wording and tone should be considered carefully, as feedback perceived as insensitive or demotivational is less likely to be acted upon. On the other side of the spectrum, praise should also be avoided to prevent distraction from the task.
- **Feedback should be clear and specific.** Unclear feedback can confuse and frustrate learners. Moreover, general remarks unspecific to the learner's solution, such as "good job", are perceived as unhelpful and may even cause learners'

to lose interest in the feedback altogether. Therefore, feedback should precisely point out mistakes made by the learner specifically and how to avoid making similar mistakes in the future.

- **Feedback should be understandable.** This may seem obvious at first glance, as learners cannot act upon feedback they do not understand. However, there is often a mismatch in the terminology used by instructors and students, leading to misunderstandings and confusion. Reportedly, students often feel overwhelmed by the academic and pedagogical terminology teachers use. For this reason, feedback should be only as complex as needed on the language level to simplify the decoding process for the learner.

- **Feedback should be detailed without being overwhelming.** Generally, more informative feedback tends to have a greater effect on learning compared to general or surface-level feedback. However, large amounts of feedback or long and complicated feedback can be overwhelming and may be ignored by learners. Thus, feedback should be to the point and only as long as it needs to be to inform the learner of where they stand in relation to the learning goal and how they can improve.

- **A feedback's source should be trustworthy.** If learners doubt the expertise, experience or attention level of the feedback provider, they are less likely to engage with provided feedback. This is especially important for automatically generated feedback as most humans tend to quickly lose trust in algorithmic systems after observing them err [32, 74].

- **Feedback should be unbiased.** A multitude of undesirable factors may influence grading and feedback. For instance, the overall impression of the student may affect the interpretation of their response (Halo Effect) [99]. Prior notions about the student, e.g., whether they are gifted or low-performing, can also lead teachers to overlook mistakes or interpret responses less favorably due to confirmation bias. Furthermore, irrelevant student characteristics, such as race, sex or attractiveness, have been shown to bias human graders [99]. For these reasons, students should be anonymized before presenting the responses to graders. This does not eliminate all sources of bias. A teacher may, e.g., still associate the presence of certain keywords with correct responses and, thus, read them less carefully. However, anonymization is nevertheless a very effective countermeasure.

4.1.3 *Reproducability & Baselines*

Currently, one of the biggest challenges in the elaborated feedback generation field is the lack of publicly available datasets. While there are many supervised feedback systems, comparing them or reproducing reported results is impossible as long as they utilize proprietary data. Therefore, a benchmark should be easily accessible to the public. Since learner responses and grades at German universities or schools are typically shielded by data protection laws, data needs to be collected specifically for publishing. The data collection requires informed consent from the students and teachers and the option to opt out of the collection process without any detrimental effects. Additionally, collecting data specifically for a benchmark dataset allows more quality control than is typically done in regular assessments.

A benchmark should also serve as a basis for comparison. For this reason, the data should be pre-split into training and test sets to avoid individual splits implemented by various researchers and, thus, enable comparability. Since the test data will be used to draw conclusions about the quality of feedback generation approaches, it should attempt to mimic the real-world class distribution while still representing minority classes adequately. This can be a trade-off. For example, should most responses be correct in a course, the test set should simultaneously consist of mostly correct responses to follow the real-world distribution and oversample incorrect responses to test the system's capabilities on the minority classes. Test sets should also have clearly defined scopes. For example, a test set may aim to measure how well an approach can generalize to novel questions without question-specific training data or it may aim to measure how well a model can be fitted to a set of given questions.

Finally, the benchmark's release should include baseline approaches. This gives researchers a starting point for comparisons and also serves as a recommended evaluation methodology. Especially for measuring the quality of elaborated feedback, there are multitudes of evaluation metrics one could use – each with advantages and disadvantages. Pre-selecting a suitable subset can ease systems' comparability later on.

4.1.4 *Question Diversity*

A benchmark should cover a variety of questions and domains so that feedback systems can be tested for generalizability. While elaborated feedback systems currently perform best when manually tailored to a specific set of questions, the level of work and care needed for high-quality feedback is prohibitive for many application scenarios. Thus, a benchmark should enable estimation of how well a system would perform on novel questions in the same domain. Additionally, it would be beneficial if

the benchmark covered multiple learning contexts, such as university and lifelong education, to be helpful to a wide range of practitioners as well as avoid over-adaptation to a specific context.

4.1.5 *Dataset Size*

Grading student responses and giving elaborated feedback is costly and time-consuming. Nevertheless, a sizeable number of student responses should be annotated to enable the training of current machine learning models and a sufficient approximation of the target distribution. Unfortunately, the optimal data set size is not easily determined a priori as it depends on a multitude of factors, such as the task, the questions, the population and the model used later on. Therefore, we will consider standard sizes used in one of the closest existing tasks: Automatic Short Answer Grading (ASAG). Popular ASAG datasets are typically large enough to allow model convergence while remaining feasible to collect. The common datasets in this field usually range from a few hundred to a few thousand student responses, with only a few containing over 10,000 responses[1].

4.1.6 *Ethical Concerns*

While a benchmark necessitates the publication of student responses to assessment questions, the learners' privacy should still be protected. Thus, any identifying information should be stripped from the students' responses to anonymize them. Students should be informed of the data collection's purpose and scope understandably. Additionally, participating or declining to participate in the data collection process should not have detrimental effects on the students and their learning outcomes [35]. The original and, thus, not anonymized data should only be retained as long as necessary and stored responsibly.

1 https://web.archive.org/web/20230316124230/https://catalpa-cl.github.io/EduScoringDatasets/ [accessed April 24, 2023]

4.2 THE SHORT ANSWER FEEDBACK CORPUS

To mitigate the lack of publicly available elaborated feedback datasets, we construct the bilingual Short Answer Feedback Corpus (SAF). It includes learner responses and reference answers collected from an English communication networks lecture and a German crowdsourcing platform. Each response is annotated with a numerical score indicating the response's correctness and an elaborated explanation of the response's mistakes. An example response with feedback can be seen in Table 16. Each component of the dataset is described in Table 17. In total, it contains 7,880 responses distributed roughly equally across three domains: communication networks, German social security and micro-job training. SAF covers 58 questions, with the majority (31) being English communication networks questions. We conducted the data collection and annotation between April 2020 and June 2022. The dataset is publicly available on HuggingFace[2]. The following sections describe the dataset's construction process and quality characteristics.

Question:	What are the challenges of Mobile Routing compared to routing in fixed and wired networks? Please name and describe two challenges.
Answer:	1) Due to hardware constraints, some nodes may be out of the range of others. 2) Mobile routing requires more flexibility. The environment is very dynamic and the routing mechanism has to adapt to that.
Verification:	0.5 out of 1.0 points (Partially Correct)
Elaboration:	While the second challenge of needing to be able to adapt to a dynamically changing environment is correct, the first challenge stated is not a challenge specific to mobile routing. In a wired network, nodes typically don't have a direct connection to each other node as well.

Table 16: An example answer with annotated verification and elaborated feedback. Table adapted from [44].

4.2.1 *Data Collection*

We collected learner responses in three domains from two data sources, a university lecture and a crowdsourcing platform. To comply with data protection laws, only data

2 https://huggingface.co/Short-Answer-Feedback [accessed April 24, 2023]

Field	Description
Question	The question posed to the learner.
Reference Answer	The sample solution to the question. May cover multiple correct solutions.
Student Answer	Response given by the student.
Score	A numerical value indicating the answer's correctness and completeness. For nearly all questions, it ranges between 0 and 1. The range is typically discretized into steps, such as 0.125, to avoid making arbitrarily fine distinctions.
Elaborated Feedback	Response-contingent elaborated Feedback. It explains why an answer is wrong or right without using formal error analysis [144]. Hints or the correct answer may be used to explain mistakes.
Verification Feedback	An automatic labeling of the score. It includes the following labels: Incorrect (score=0), Correct (score=maximum number of points achievable), Partially Correct (all intermediate scores)
Error Class	The type of mistake made in the response. It is only available for the German legal domain and captures the following mistake types: contradictory, factually incorrect, imprecise, irrelevant, logical error, partially correct but incomplete, incorrect additional information, partially correct, mistakes likely due to carelessness

Table 17: SAF's components with descriptions. Table adapted from [44].

collected specifically for this dataset with informed consent is used. Participation in the following collection studies was always voluntary and could be aborted at any time without negative consequences.

Communication Networks Lecture

We collected data in multiple semesters of a communication networks lecture at the Technical University of Darmstadt. Each semester, roughly half of the visiting students were Bachelor and half were Master students – most studied computer science or electrical engineering. The questions were part of voluntary quizzes students could complete for bonus points in the final exam and target various layers of typical communication networks stacks. For example, there were questions about extension headers in IPv6, Software-Defined Networking, or bitstream encoding techniques. A majority of the questions could be answered in teams of up to three students and the rest had to

be completed alone. Students had at least a week to complete each quiz on the online learning platform Moodle[3].

Micro-job Training

In cooperation with the *wer denkt was GmbH*[4] we developed a German pre-job training on their crowd-worker platform *AppJobber*[5]. The platform acts as an intermediary for workers and companies looking for a workforce to complete micro-jobs [141]. The training consisted of short-answer questions focused on aspects crowd-workers should pay attention to when completing quality checks in gas stations later on, such as cleanliness, staff interaction or product placement. Participating in the pre-job training was voluntary and not necessary to perform the actual quality check job. However, there was a small financial incentive to participate. Each response was checked by a *wer denkt was* employee to filter out cheating attempts, such as answering every question with "I don't know" or submitting the same responses under multiple accounts. Questions were answered by jobbers individually directly on *AppJobber*.

German Social Security Law

Finally, we constructed a catalog of short-answer questions pertaining to German social security law. The questions were developed in the context of a Software Campus project aiming to help citizens better understand letters received from government institutions, as they are often written in complicated legal jargon. Specifically, the questions target various rights and obligations of *Arbeitslosengeld I* recipients, a type of unemployment benefit. They range from how the unemployment benefit is calculated to what recipients should do if they get sick. Citizens were recruited to answer the questions in a Limesurvey[6] questionnaire distributed via *AppJobber* and received a financial incentive for completing the questionnaire. Provided responses were automatically screened for suspicious answer times, extremely short answers, duplicate IPs, duplicate responses and keyword phrases indicating irrelevant answers, such as "couldn't find". Suspicious answers were then inspected manually and rejected if found to be scam attempts.

3 https://moodle.org/ [accessed April 24, 2023]
4 https://werdenktwas.de/ [accessed April 24, 2023]
5 https://en.appjobber.com/ [accessed April 24, 2023]
6 https://www.limesurvey.org/ [accessed April 24, 2023]

4.2.2 *Data Annotation*

After collecting the learner responses, the next step is to annotate them with scores and elaborated feedback. The general annotation procedure used can be seen in Figure 8. However, not all data was annotated in the same manner, so deviations from this procedure are stated explicitly. The first step consisted of preprocessing the raw collected data into **answer annotation files** easily usable by the annotators. This included exporting the responses from their respective sources and stripping personal information, such as IPs, submission times or names. Next was the selection of suitable annotators. As discussed in Section 4.1, this is challenging as annotators require pedagogical and domain expertise to provide high-quality feedback. Often, domain experts lack pedagogical training and vice versa. We opted for annotators with domain knowledge and trained them on the necessary pedagogy basics, such as avoiding comparisons with other students.

For the first semester of the communication networks data, we chose two graduate students who had successfully completed the course themselves to annotate all responses twice after the course had finished. In the second semester, the responses were annotated once by members of the teaching staff – this was also the real-world feedback presented to the students during the semester. The social security data was annotated by two graduate students, a law student and a computer science student with a background in annotation and the Software Campus project. Here solely the test sets were annotated twice to save annotation time as the questions were especially time-consuming to grade due to the questions' and underlying material's complexity. Lastly, half of the job training responses were annotated doubly by two *AppJobber* employees and half by a single employee. All annotators were compensated financially or with ECTS, credit points awarded by European universities.

To train the annotators' pedagogical skills, we drafted a **general annotation guideline** which was discussed with the annotators. It explains the annotation goals, the annotation file's structure, the scoring system and how to give high-quality feedback. It covers most feedback recommendations and biases presented in Section 4.1 with concrete examples. For instance, during a pilot annotation study, we observed that annotators would use phrases like "This response fails to..." without realizing that "failing" may be negatively connotated and harm the learner's self-esteem. Thus, we included explicit examples to illustrate how the abstract advice relates to their annotation task. The guideline was submitted to a psychology doctoral student with prior work in the feedback field for further recommendations and updated whenever concerns were raised during the annotation process.

Next, a researcher drafted **grading rubrics** for each question. They contain reference answers with detailed scoring information and illustrative example responses. In sim-

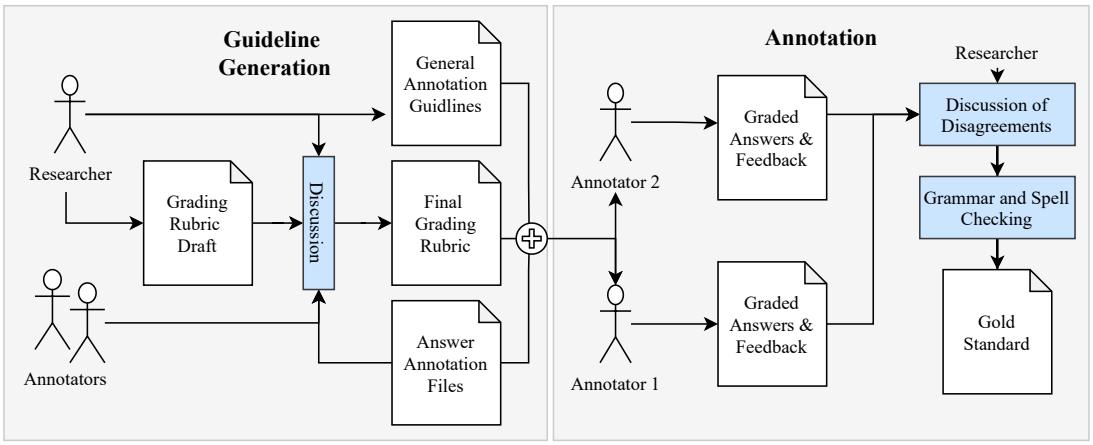


Figure 8: Schematic depiction of the general annotation process. Figure taken from [44].

ilar NLP annotation studies, annotators would now be trained with the guideline and grading rubrics before annotating the whole dataset. However, a pilot study in the communication networks domain showed that the usual procedure was not sufficient to acquire reliable annotations in this setting, yielding only low inter-annotator agreements (Krippendorff's Alpha[7] of 0.36). This observation is consistent with challenges identified in pedagogical research on rubric construction [24, 63, 146], such as a single rubric constructor not being able to establish an unambiguous rubric as well as varying levels of prior knowledge and grader strictness. Thus, we decided to include the annotators in the rubric construction process. While the researcher still designed the initial draft, it was subsequently discussed and overhauled in weekly meetings with the annotators. Prior to the discussion, the annotators are provided with the answer annotation files so that edge cases or generally unexpected responses not covered by the rubric draft can be identified. The idea behind including the annotators in the construction was to coalesce multiple sources and perspectives into a comprehensive rubric and mitigate deviation from the rubric later on due to a deeper understanding of the topic and more identification with the grading guideline.

Only the rubrics for the job training questions were constructed in cooperation with the industry partner instead of the annotators. The reasons for this were two-fold. First, the questions are more straightforward because they require less prior knowledge compared to a university lecture or social security law and are limited in scope to the micro-job they pertain to. This was also reflected in higher inter-annotator agreement (Krippendorff's Alpha of 0.736) in initial pilot studies. Second, the annotators were employees of *wer denkt was*, supervised by the industry partner instead of the researchers. Therefore, we concluded that an initial discussion after the pilot study with the option to clarify questions anytime would suffice.

7 See Section 3.3.4 for a description of Krippendorff's Alpha.

After the final iteration of the grading rubrics was completed, the annotators graded and gave elaborated feedback on the responses independently of each other. This is necessary to estimate the reliability of the annotations later on. After both annotators finished, disagreements were identified by extracting responses the annotators had graded differently. While annotators could still disagree on the given feedback even though they assigned the same score, detecting such cases reliably was deemed uneconomical considering the benefit it would bring. Disagreements between the annotators were resolved in discussion with the researcher by either selecting the more appropriate annotation or fusing them if both had merit. For instance, one of the annotators could have overlooked a mistake in the response, but they noticed a missing piece of information instead. In cases where both annotators agreed, one annotator's elaborated feedback was categorically selected on a question-by-question basis based on the eloquence of a manually inspected random sample.

Finally, the English feedback was checked by Grammarly and a native speaker since both annotators were not native speakers. Grammar and spelling mistakes were corrected and sentences were simplified where possible, for example, by using the possessive form. Learner responses were not post-processed as spelling mistakes are a challenge grading models are likely to encounter frequently. Thus, each response in the dataset received a **gold standard** score and elaborated feedback.

4.2.3 *Corpus Statistics*

Overall, the annotation process resulted in 7,880 responses and matching feedback. This is a similar size compared to popular ASAG datasets (see Section 4.1.5). Next, we split the corpus into training, validation and two test sets for each domain. The first test set is the unseen answers split. It contains roughly 10% of the domain's total responses to questions that also have responses in the training set. The unseen answers test split can be used to test how well a model performs on novel responses to questions it was trained to grade. The second test set contains responses to novel questions the model was not trained for and can be used to test a model's generalizability to new questions in the same domain. Theoretically, one can also test a model's generalizability to novel domains using another domain's data as the test set. However, we did not evaluate our baselines in this setting, as the models tended to already perform poorly on the unseen questions split. Of the remaining data, roughly 20% are reserved for validating model selection and hyperparameter tuning. The exact distribution of samples across the various data splits can be seen in Table 18. In total, the dataset contains 31 communication network questions with 96 responses on average, 19 social security

Domain	Train	Validation	UA	UQ	Total
Communication Networks	1700	427	375	479	2981
Social Security Law	1596	400	221	275	2492
Job Training	1226	308	271	602	2407

Table 18: Number of responses in each data split. UA stands for unseen answers and UQ for unseen questions.

questions with roughly 131 responses each and 8 job training questions with roughly 300 responses each.

Table 19 displays the verification feedback class distribution for each data split. The classes are distributed unevenly in each domain. There are more correct and partially correct than incorrect responses in the communication networks and job training domain. This is partly due to the questions' modest difficulty combined with the fact that learners may not respond at all if they do not believe they can score points with their answers. The majority class – the class with the most samples – for both domains is *Correct*. While the social law domain also has few incorrect responses, the majority class is *Partially Correct* with correct responses actually close to incorrect ones in frequency. This indicates that questions may be more difficult in this domain overall. We decided against oversampling the minority classes in the unseen answers test splits when randomly sampling from the data to preserve the real-world distribution. Researchers can instead employ class weighing or other class balancing techniques if they deem it useful for their application scenario. We selected questions for the unseen questions test split, attempting to match the overall class distribution. However, since the unseen questions split contains the entire set of responses for a given question and the overall number of questions is limited, this process incurs sampling errors.

Label	Train			Validation			UA			UQ		
	CN	JT	SL	CN	JT	SL	CN	JT	SL	CN	JT	SL
Correct	1053	570	281	262	134	65	240	103	32	247	278	87
Partially C.	565	482	938	137	132	248	114	121	141	190	275	133
Incorrect	82	174	377	28	42	87	21	47	48	42	49	55

Table 19: Distribution of verification labels per data split. CN stands for the communication networks domain, JT for the job training and SL for the social security law domain. Note that the class Correct contains all responses that received the maximum number of points achievable, Incorrect all responses that received 0 points and Partially Correct all other responses.

Domain	N	α	%-agreement
Communication Networks	2,112	0.91	89.46
Social Security Law	496	0.90	95.87
Job Training	1,200	0.78	81.38

Table 20: Krippendorff's alpha and percentage agreement for each domain. N denotes the number of samples annotated twice.

4.2.4 *Reliability and Validity*

As discussed in Section 4.1, achieving reliable scoring of short answers is challenging but essential for a benchmark dataset. Since the annotated score is interval scaled, we use Krippendorff's Alpha and percentage agreement to estimate interrater reliability. Table 20 displays the inter-annotator agreement for each domain. Both measures indicate high reliability [7] with α ranging between 0.78 and 0.91 and a percentage agreement between 81% and 96%. Considering the low initial agreement (α=0.36) of our pilot study in the communication networks domain, the improved annotation process seemed to have a considerable effect.

While the estimation of interrater reliability with inter-annotator agreement measures is well established in related work, validity is rarely measured in NLP dataset construction studies [163]. We opt to estimate validity by observing how well the learners' performance on the dataset questions correlates with known measures of learning success. In the communication networks domain, we assume the first semester's end-of-term exam to be a sufficient approximation of students' knowledge. Noteworthy is that students took the exam individually, while the quizzes in the first semester were answered in groups of up to three students. Nevertheless, Spearman's rank correlation[8] between the total points achieved in the quizzes and the exam in the first semester is moderately high, with 0.438 ($p < 0.0001$, N=186). The correlation calculation is based only on students that participated in the exam and achieved at least one point in the bonus quiz.

For the job training domain, we investigate the relation between the number of points achieved in the job training quiz and whether the jobber failed a job later on. A job is considered failed if it must be corrected, e.g., by going back on-site and taking a forgotten picture, or is outright rejected by the employer. Since successful job completion is the goal of the training, we assume it to be a good criterion for validity. We use logistic regression to investigate the relationship between the numeric job training performance and the binary variable indicating whether the jobber failed a job later on. The job completion records include all jobs attempted in one year after the job training

8 See Section 3.3.4 for a description of Spearman's rank correlation.

was finished. It was found that the odds of a job being failed decreased by 6.98% (95% confidence interval of [0.007, 0.133], N=129) on average for each point gained in the job training quiz. In total, 8 points could be achieved in the job training quiz.

Since the social law data was collected in the context of a proof-of-concept project instead of a real-world application, we do not have access to any validity criteria.

4.2.5 *Interpretation & Limitations*

In conclusion, we have introduced a publicly available benchmark dataset that can be used for ASAG and feedback generation. With 7,880 responses from three education domains in two languages, its size is among the larger and most diverse ASAG datasets. Nearly half of the dataset's responses were annotated independently by two graders to enable reliability estimations with inter-annotator agreement measures. The dataset's grades are highly reliable due to an iterative annotation and thorough grading rubric construction process. As of yet, it is the only dataset to include elaborated feedback.

Nonetheless, there are limitations – the largest being the lack of reliable measures of elaborated feedback quality. While we routinely inspected random annotation samples for the quality criteria introduced in Section 4.1, many of them, such as understandability or clearness, are difficult to measure systematically. Thus, the explanations' reliability and validity are based on the researcher's and annotator's judgment instead of the established metrics used for the assigned grades.

In future work, the dataset could be expanded with questions from new domains and education contexts. Especially data from schools would be valuable, considering the questions currently stem from adult education settings. Further data collection would also benefit from more extensive validity analyses. While we present evidence for the validity of most of our data – both in terms of correlations with established criteria and real-world application – it is not yet conclusive in its current state. For example, we used the end-of-term exam of the communication networks lecture as an established validity criterion. While exams are frequently used proxies for student knowledge, they only capture a 120-minute snapshot of students' performance in a stressful situation. Thus, they are influenced by confounding factors, such as test anxiety.

Finally, future work could explore alternative annotation processes. Even though our method produced highly reliable grading, it has the drawback of being time-consuming. Constructing grading rubrics with the annotators took multiple hours of discussion per question, followed by weeks of annotation since responses not only had to be graded but the grade explained as well. Annotation tools could potentially reduce this time with recommendation functions that assist in formulating feedback.

The order in which responses are displayed to the annotators could also be modified from random to similarity-based, as it can be faster to grade similar responses [120]. However, one should be aware of bias introduced into the data annotation through recommendations or ordering.

4.3 SUPERVISED FEEDBACK GENERATION

Now that we have established a benchmark for the feedback generation task (see Section 4.2), we can train models to produce elaborated feedback. On the one hand, the goal of this section is to provide baselines that can form the basis for comparison with future approaches. On the other hand, we wish to explore various aspects that may influence a model's performance on this novel task.

4.3.1 *Approaches*

Unless stated otherwise, we treat the elaborated feedback generation task as a joint prediction of a grade and a corresponding explanation based on the typical ASAG inputs: the learner's response, a reference answer and – optionally – the question. While there are undoubtedly endless design possibilities, we chose the joint prediction approach because we would like the same features influencing the assigned grade to influence the explanation. While this is not necessarily given – a model could still attend to different parts of the answer for each prediction – we deem it more likely than if we would predict them independently or in a pipeline approach.

In Europe, student answers and grades are typically considered sensitive data that should be protected and only shared with prior notice and consent. Thus, we will focus on publicly available models and architectures that do not require cloud computing resources. While huge transformer models tend to perform better than their smaller counterparts, it is not feasible to train a model with billions of parameters on local machines. For example, the best GPT-3 model [21] contains around 175 billion parameters and would require a computation cluster to train – even if publicly available. For training, sensitive data would have to be sent to third parties for processing, often outside Europe. We will focus on the publicly available and locally trainable versions of T5 [124], BART[88] and their multilingual counterparts in our main experiments to avoid this dependency. Nevertheless, we also include a GPT-3 pipeline approach in our experiments for curiosity's sake. It generates feedback for responses already scored by an SBERT[9] model [125]. An overview of the exact versions and model sizes used for feedback generation can be seen in Table 21.

We will use monolingual English models for the English data since they tend to perform better than multilingual models. However, the set of monolingual German sequence-to-sequence models is quite limited, so we resort to multilingual models. These are trained to handle multiple languages, including German.

9 https://huggingface.co/sentence-transformers/distilbert-base-nli-mean-tokens [accessed April 24, 2023]

Model Type	Language	Accessibility	Version	# Parameters
BART [88]	English	public	bart-large	400M
mBART [92]	Multi	public	mbart-large-cc25	610M
T5 [124]	English	public	t5-base	770M
mT5 [176]	Multi	public	mt5-base	580M
GPT-3 [21]	Multi	OpenAI API	davinci	175B

Table 21: Overview of feedback generation models used.

4.3.2 *Experimental Settings*

The primary purpose of our experiments with the Short Answer Feedback Corpus (SAF) is to lay the foundation for future research. Consequently, we wish to explore possible evaluation metrics for an evaluation framework and provide baselines for comparison. While evaluation metrics are easily chosen for classification and regression, evaluating the elaborated feedback is much more complex. There are no established metrics for measuring elaborated feedback quality and human judgment is costly. Thus, it would be beneficial to have a set of cheap metrics that can be used to quickly compare systems – even if human judgment is still required to provide an in-depth system evaluation at the end. In the text generation field, several text similarity metrics have been proposed. They can be used to compare a generated text to a gold standard. Even though they sometimes correlate poorly with human judgment on a sentence-level [118], they correlate more reliably with human judgment when aggregated on a system-level [138]. Thus, they may not reliably estimate the quality of specific feedback but are better suited to differentiating generation models based on their overall quality and can be helpful for model development. We choose four popular natural language generation metrics to evaluate the elaborated feedback: SacreBLEU[10] [123], ROUGE-2 [89], METEOR [13] and BERTScore[11] [183].

After selecting text evaluation metrics, the next step is establishing the evaluation framework. We offer two task configurations. In the first setting, the goal is to predict the number of points the answer should receive in conjunction with the elaborated explanation of the score. In the second setting, the model instead labels whether the response was *incorrect*, *correct* or *partially correct* in addition to the elaborated feedback. For this purpose, the models' output is formatted as "*label/score* feedback: *elaborated feedback*". The output's length is constrained to at least 11 and at most 128 tokens, as models often predicted no elaborated feedback without the minimum and would be

10 https://pypi.org/project/sacrebleu/1.4.3/ [accessed April 24, 2023] default parameters (no smoothing, n-gram order=4)

11 roberta-large_L17_no-idf_version=0.3.7(hug_trans= 4.2.1)-rescaled and bert-base-multilingual-cased-rescaled

computationally expensive without the maximum. We will use accuracy and macro-averaged F1 score to evaluate the output's classification portions and the root-mean-square error (RMSE) to evaluate scoring.

To determine the best model during development, we balance classification/regression performance with textual feedback performance using the following metric m (see Equation 1), where f is the macro-averaged F1 score during classification and an inversion of the mean-squared-error (MSE) during scoring. Thus, a model must perform well on both sub-tasks, as m will be close to zero if either fails. We excluded BERTScore from the averaged text evaluation metrics as it requires a language model and is consequently too computationally expensive to calculate each epoch.

$$m = \frac{\text{BLEU} + \text{ROUGE} + \text{METEOR}}{3} * f, \; where \; f = \begin{cases} F1_{macro}, \; \text{if classification} \\ 1 - \text{MSE}, \; \text{if regression} \end{cases} \tag{1}$$

The exact input format and hyperparameter-tuning process vary from model to model. Generally, we perform manual hyperparameter-tuning and select the best performing model on the validation set. We do not include details here as they can be found on the model description cards on HuggingFace[12] and prior publications [44]. Nevertheless, we differentiate between models that received the learner's response, a reference answer and the question as input and models that only received the response and reference answer. Figure 9 shows a schematic overview of all experiments conducted with SAF.

4.3.3 *Feedback Evaluation*

Tables 22, 23 and 24 show the performance of the fine-tuned models, majority baselines and the average annotators' performance compared to the gold standard for each domain. For the communication networks and job training domain, the majority baseline always predicts *Correct* or a score of 1.0 during regression. The most frequent elaborations are "*The response is correct.*" and "*Korrekt!*", respectively. In the social law domain, the most common class is *Partially Correct* with the feedback "*Das stimmt, aber das genaue Datum des Endes der Widerspruchsfrist ist in diesem Fall der 20.03.2013.*". The most common score is 0.0 with the feedback "*Das stimmt leider nicht, Sie können Widerspruch innerhalb eines Monats nach Bekanntgabe des Bescheids einlegen. Aufgrund der Dreitagesfiktion ist das genaue Datum des Endes der Widerspruchsfrist in diesem Fall der 20.03.2013.*"

12 https://huggingface.co/Short-Answer-Feedback [accessed April 24, 2023]

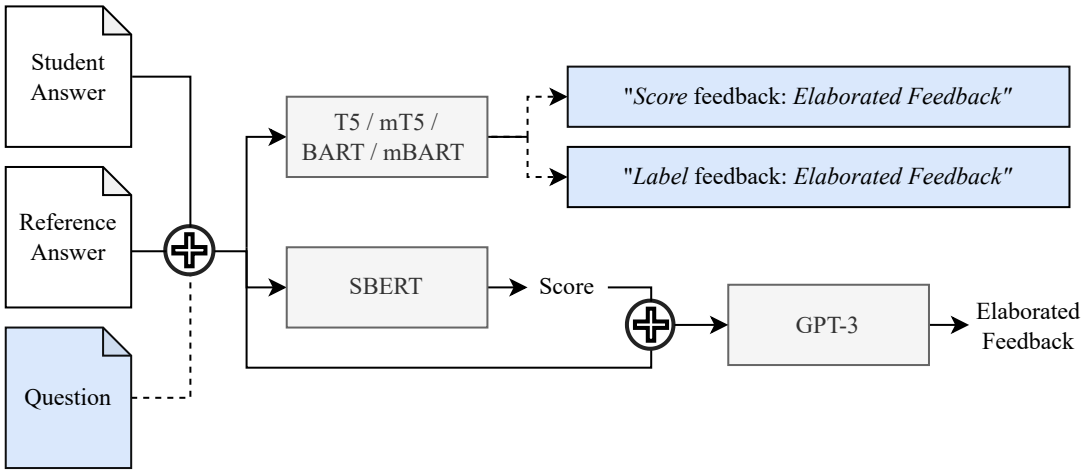


Figure 9: Schematic overview of experiments conducted with SAF. The experimental variables – beyond the model selection – are highlighted in blue, meaning that the question is optionally included in the input, and the smaller models jointly classify and elaborate or regress and elaborate.

In all domains, we can see that the task challenges current state-of-the-art transformer models. While they significantly outperform the majority baselines on unseen answers, the gap to human performance is still large. Humans have less than half of the root-mean-square error (RMSE) when scoring than the best performing models (0.269 vs. 0.099 on communication networks unseen answers and 0.187 vs. 0.077 on social security unseen answers) and consistently achieve better feedback similarity scores except for regression mBART having a higher METEOR score on the social law unseen answers. Even the SBERT and GPT-3 pipeline did not approach human performance despite GPT-3 being a truly colossal language model.

This trend is further amplified on unseen questions where the models' grading and feedback metrics typically fall off sharply. For instance, each model's accuracy is reduced by 7.5 to 35.1 percentage points from unseen answers to unseen questions, with the largest drop observed on the job training (84.9% to 49.8%) and the smallest on the communication networks data (74.2% to 66.7%). The more considerable drop in the job training domain is unsurprising as it is also the dataset with the smallest number of questions but the largest number of responses per question and, thus, harder to generalize. In return, models are expected to grade more accurately on unseen answers, as they have seen a greater range of responses per question during training. This expectation is reflected in the results where the best grading performance achieved on the job training unseen answers split is around 85% accuracy and an RMSE of 0.196 compared to the best performance of around 75% accuracy and an RSME of 0.269 on the communication networks split.

		Unseen Answers						Unseen Questions					
	Model	Acc.	F1	BLEU	MET.	ROU.	BERT	Acc.	F1	BLEU	MET.	ROU.	BERT
Label	Majority	54.0	23.4	2.2	21.5	20.2	42.2	47.1	21.4	0.2	15.0	11.5	38.1
	$T5_{wo\_q}$	74.2	72.0	33.7	59.0	52.8	65.0	66.7	55.9	10.7	36.4	31.1	52.2
	T5	75.0	**75.9**	34.0	56.9	49.6	62.2	**67.4**	**69.7**	13.5	39.7	32.1	53.3
	*Majority	61.9	25.5	1.3	19.4	20.1	34.5	51.6	22.7	10.7	29.7	21.7	39.1
	*BART	**76.0**	53.4	36.0	60.8	49.1	**69.5**	51.6	27.9	2.4	28.5	20.1	36.6
		RMSE						RMSE					
Score	Majority	0.470		2.2	21.5	20.2	42.2	0.512		0.2	15.0	11.5	38.1
	$T5_{wo\_q}$	0.290		33.7	56.9	50.4	62.8	0.263		9.0	35.3	29.1	49.7
	T5	0.269		32.7	56.4	48.6	61.2	0.248		16.6	45.9	35.5	51.5
	Human	**0.099**		**45.5**	**64.9**	**56.5**	68.5	**0.086**		**57.1**	**71.6**	**64.3**	**75.7**
	*Majority	0.752		1.3	19.4	20.1	34.5	0.532		10.7	29.7	21.7	39.1
	*BART	0.373		30.5	58.2	46.4	68.0	0.544		0.6	18.8	9.5	26.7

Table 22: Model performances on the communication networks data. For the scoring and the labeling task, models marked with *wo_q* did not receive the question as input. The text similarity measures, accuracy and F1 scores are given in percent. *The results are not directly comparable, as only BART was trained and evaluated on both semesters' data. While a majority of the test data remained unchanged across semesters, the score ranges are larger in the second semester, which can negatively impact RSME. Table adapted from [44].

		Unseen Answers						Unseen Questions					
	Model	Acc.	F1	BLEU	MET.	ROU.	BERT	Acc.	F1	BLEU	MET.	ROU.	BERT
Label	Majority	44.6	20.6	0.0	0.0	19.0	33.0	46.2	21.1	0.0	0.0	**23.2**	**40.1**
	$mT5_{wo\_q}$	**85.2**	**85.1**	**50.7**	51.2	**31.4**	54.9	**54.7**	**41.7**	**0.7**	20.1	0.5	21.9
	mT5	84.9	84.3	46.0	49.2	30.3	51.7	49.8	36.0	0.6	18.1	0.2	18.1
	mBART	80.1	80.7	39.5	**63.3**	29.8	**63.1**	48.7	40.6	0.3	**33.8**	0.5	31.3
		RMSE						RMSE					
Score	Majority	0.538		0.0	0.0	19.0	33.0	0.426		0.0	0.0	**23.2**	**40.1**
	$mT5_{wo\_q}$	0.399		31.5	36.7	21.7	42.9	**0.360**		1.7	12.2	1.1	15.4
	mT5	**0.196**		44.3	43.1	28.7	51.7	0.400		**2.0**	18.1	1.5	20.9
	mBART	0.333		41.6	62.0	30.9	61.2	0.465		0.7	20.3	0.7	17.8

Table 23: Model performances on the German job training data. For the scoring and the labeling task, models marked with *wo_q* did not receive the question as input. The text similarity measures, accuracy and F1 scores are given in percent. Since the test sets of this domain are only partially annotated by two annotators, we cannot provide a human baseline. Table adapted from [44].

Interestingly, including the question in the input seems to have less effect on model performance than expected. While it improves grading in the communication networks domain (by 3.9-13.8 percentage points of F1 and 0.015-0.021 RMSE), it hardly affects the feedback metrics. On unseen answers, T5 without questions achieves slightly higher scores across nearly all feedback metrics than its counterpart with questions.

		Unseen Answers						Unseen Questions					
	Model	Acc.	F1	BLEU	MET.	ROU.	BERT	Acc.	F1	BLEU	MET.	ROU.	BERT
Label	Majority	63.8	26.0	2.6	7.0	4.7	7.5	48.4	21.7	0.1	2.3	0.8	2.7
	mBART	**81.0**	**74.6**	42.8	58.2	43.7	57.5	60.7	55.4	3.2	20.0	5.0	14.8
	GPT-3	82.8	76.4	36.8	39.3	39.4	55.8	56.7	43.4	3.3	11.1	5.6	17.7
		RMSE						RMSE					
Score	Majority	0.577		4.3	8.5	4.3	10.6	0.681		0.2	3.1	0.8	5.3
	mBART	0.190		39.4	**54.3**	42.3	52.6	0.317		2.8	17.9	5.0	10.7
	GPT-3	0.187		36.8	39.3	39.4	55.8	0.291		3.3	11.1	5.6	17.7
	Human	**0.077**		**46.9**	48.3	**59.2**	**71.6**	**0.063**		**48.7**	**37.3**	**48.6**	**67.0**

Table 24: Model performances on the German social security data. All models included the questions as input. The GPT-3 pipeline was only trained to regress. Therefore, the labeling performance is a discretization of the predicted score instead of an additionally trained pipeline. The text similarity measures, accuracy and F1 scores are given in percent.

On unseen questions, this effect is reversed. Especially on this dataset, we expected the question to hold vital information for scoring and feedback, such as how many aspects students should mention or how detailed their descriptions should be. While we see an improvement, particularly in the generalizability to unseen questions, it is less pronounced than expected. The questions' influence in the job training domain is even less clear. While T5 with questions generally outperforms T5 without questions in the scoring scenario, the reverse is true in the classification setting. It seems that the inclusion of the question as input can be detrimental to model performance.

With only a few exceptions, classification models score higher text similarity metrics with the gold standard feedback than models trained to regress. However, the difference is typically in the range of a few percentage points. A set of T5 example feedback can be found in Appendix A.4 and all BART models can be freely queried on HuggingFace[13].

4.3.4 *Field Evaluation*

Even though the generated feedback fell short of human feedback in our experiments, it did perform well enough according to automatic measures to warrant further study. Similar to peer feedback, automatic feedback may be valuable to students despite imperfections if higher-quality feedback by a teacher is delayed too long due to its time-consuming construction [54]. Additionally, we aim to supplement the semi-reliable automatic evaluation with human judgment. Thus, we conducted a field study in the communication networks semester after the data collection to investigate the gener-

13 https://huggingface.co/Short-Answer-Feedback [accessed April 24, 2023]

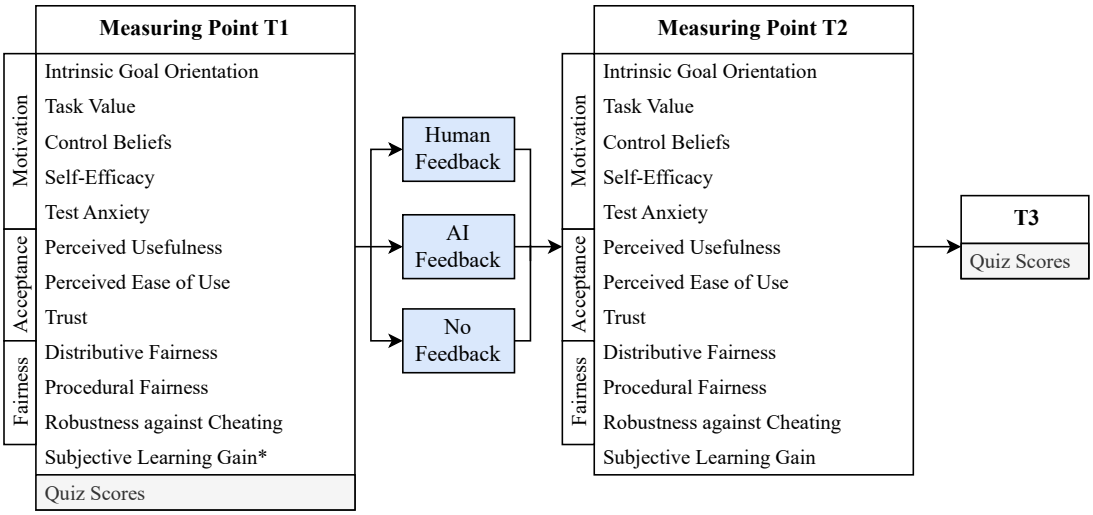


Figure 10: Schematic overview of the field study conducted in the communication networks lecture. All variables except the quiz scores for the objective learning gain were measured via questionnaire. Between measuring points T1 and T2, students received (or didn't receive) feedback on their quiz answers according to their group assignment. *The subjective learning gain was technically queried in T2's questionnaire for both measuring points. However, students should think back to their judgement directly after answering the quiz at T1.

ated feedback's potential effects on learning gain, motivation and student attitudes. While the effects of ASAG verification feedback on students have recently become a topic of interest in the literature [66], the effects of automatically generated elaborated feedback remain unexplored.

In total, 88 students participated initially and were randomly assigned into one of three groups: a group receiving feedback from human tutors (Group Human), one by a T5 feedback generation model (Group AI) and one without feedback (Group No-Feedback). Of the 88 initial participants, 79 remained after data cleansing due to cheating, failed attention checks and drop-out after the first measuring point. Participation was voluntary and incentivized with bonus points towards the final exam. Given responses were pseudonymized with a personalized participant code that students had to supply at each measuring point. Since we are aware of the feedback's imperfection, we decided to only conduct the experiment for a single quiz instead of the whole semester to avoid potential harm to the students' learning outcomes. A schematic overview of the study design can be seen in Figure 10.

Procedure

At the first measuring point (T1), students answered a regular bonus quiz with short-answer questions on Moodle and a questionnaire. The questionnaire captures the stu-

dents' demographic information, baseline motivation, acceptance of and trust in human and AI graders. For quality control, we included an attention check item to filter out students randomly selecting responses on the questionnaire. The questionnaire's other items were adapted from established motivation [122], acceptance [107, 178], trust [1] and fairness [57] questionnaires to fit our educational scenario. The exact items used can be seen in Appendix A.3. All items were rated on a 7-point Likert scale. Students had one week to complete the quiz and questionnaire. A week after the quiz deadline, they received or didn't receive feedback according to their group assignment and had one week to complete the questionnaire of the second measurement point (T2). The second questionnaire differs from the first in that it omits demographic questions but includes items measuring subjective learning gain, such as "At the present moment, I can evaluate which MAC procedure might be suitable given a specific scenario." It aims to capture the students' attitudes after the feedback intervention. In T2, we also asked students to guess whether they had been given feedback by a human or AI grader and provide how certain they were on a scale from 0 to 100%.

To measure learning gain objectively, a third measurement point (T3) repeats two of the five original quiz questions a week after the completion of T2. The teaching staff manually evaluated this short quiz and the assigned scores serve as a measurement of objective learning gain that can be compared to the achieved scores in the initial quiz. Participants were asked not to consult additional materials but answer from memory. As the corresponding semester could be completed entirely online, with students only having to be physically present for the final exam, an in-person measurement in a controlled environment was not possible. Thus, cheating cannot be ruled out but likely affects each condition similarly. Conspicuous cases were cleansed from the dataset.

Of the 79 participants considered in the final study, 60 had a completed set of data for each measuring point. The rest completed only two of the three measuring points, for example, by not consenting to the use of their quiz scores in T3. After T3, all participants were informed of their group and the No-Feedback and AI Group received human feedback on their initial quiz responses. All questionnaires were completed in May 2022.

Participant Characteristics

Due to implausibility, we excluded the demographic information that one of the 79 final participants provided. They stated an age of 45, gender of "other" and a participant code indicating nine sisters, nine brothers and their mother's name starting with "XX." Thus, the following statistics describe 78 participants with credible information.

On average, participants were 24.4 years old ($\sigma = 3.26$). Most (60%) studied computer science, 21% studied electrical engineering and information technology and the

rest were enrolled in various other programs. Overall, 37% were pursuing a Bachelor's degree. Of those, 59% were within the expected number of six semesters for such a degree at the time of the study. Of the 63% enrolled in Master's programs, 91% were within the standard of four semesters. The most common native language was German with 55%, followed by Chinese with 15% and English with 6%. The rest reported a variety of other languages as their first. Multiple selections were possible. In T1, we also queried participants about their familiarity with AI on a scale from 1 (no knowledge) to 7 (expert). On average, participants indicated an intermediate level of knowledge about AI ($\overline{x} = 3.62$, $\sigma = 1.3$).

Dependent Variables

This study explores the impact of automatically generated, imperfect feedback on students' learning gain, motivation and attitudes. McGrath et al. [105, p.xi] define **learning gain** as "difference between the skills, competencies, content knowledge and personal development demonstrated by students at two points in time." As feedback tends to have a medium to high positive effect on learning gain [171], we expect the Human Feedback Group to exhibit the highest learning gain, followed by the AI Feedback Group and, lastly, the No-Feedback Group.

We use two scales to measure **learning gain: subjective and objective.** On the subjective scale, students give an estimate of their knowledge state on various topics covered by the quiz, such as whether they could choose a suitable bitstream encoding technique for a given scenario before and after receiving feedback. As subjective learning gain measures are controversially discussed in the literature, we also include an objective measure. Comparing the groups' performance on the final exam would be the most externally valid measure of learning gain. However, as this is an explorative study of a novel feedback generation approach, the effects of the feedback on students are difficult to predict. Even a within-subject design may disadvantage the cohort compared to previous semesters. To avoid potentially harmful effects, especially on learning gain, the scope of the study is limited to a single quiz that should – by design – have little effect on the final learning outcome. Thus, we chose to repeat a subset of the quiz questions later to measure how the students' learning improved through the provided feedback. While this design provides a shorter snapshot of the students' performance, quiz scores correlate with exam performance and are deemed an acceptable proxy (see Section 4.2.4). Specifically, the points assigned by the human graders on the two questions during T1 and T3 are compared to form the objective learning gain measure. In total, students could achieve 0 to 3.5 points.

Our measure for **motivation** is based on Pintrich et al.'s [122] "Motivated Strategies for Learning Questionnaire". Here, motivation is comprised of values (task value,

intrinsic vs. extrinsic goal orientation), expectations (control beliefs, self-efficacy) and affect (test anxiety). **Task value** captures how useful, exciting and important a learning scenario is to the student. **Intrinsic goal orientation** describes whether the student is focused on mastering the material, while extrinsic goal orientation implies a focus on grades, approval from others or other external factors. **Control beliefs** refer to the extent students believe their learning outcome is influenced by their actions as opposed to determined by factors outside of their control, such as luck or bias. **Self-efficacy** captures the extent to which students expect to succeed or have the ability to succeed at a given task. Self-efficacy differs from control beliefs in that students may believe to be in control of their learning outcome but still think they will not have the time, resources or ability to succeed. Both variables impact motivation significantly. Finally, **test anxiety** captures how worried students are in exam situations. It typically influences motivation negatively.

Generally, feedback on a content level only has a minor positive impact on motivation [171]. Therefore, we expect all groups to see little change in motivation. However, uninformative or negative feedback can harm motivation [171]. Since the feedback model is pre-trained on crawled webpages, the generation of offensive feedback cannot be excluded even though the model was further trained to produce friendly feedback. The generated feedback was manually spot-checked to safeguard the participants from profanity and insults before release – with no reported cases – but it is not guaranteed. Additionally, the generated feedback's semantic imperfection may demotivate students due to factual incorrectness or irrelevance. We deem the feedback system successful if it does not harm motivation compared to the control conditions. We modified the original questionnaire to fit our learning scenario (it was initially designed to measure motivation for a whole course) and averaged all items for each scale.

Finally, we aim to measure students' attitudes toward the feedback generator. Specifically, **acceptance** of the approach, **trust** in the model's predictions and perceived **fairness** are of relevance to us. In contrast to motivation and learning gain, these concepts are typically less well defined, with formally validated measurement instruments being scarce. Therefore, based on related work, we developed our own items for these constructs. However, the items designed to measure fairness lacked internal consistency (see a Cronbach's α[14] of 0.04 and 0.18 in Table 25) and were, thus, excluded from further analyses. The rest of the measures were deemed acceptable. The measures for

14 Cronbach's alpha aims to measure reliability in the form of internal consistency, that is, whether the various items of a group are related to each other. Generally, values between 0.7 and 0.95 are deemed acceptable. Higher values may indicate that there are redundant items in the scale. Lower values may indicate that the items measure different constructs or that the test is too short. It should be considered a lower bound of reliability that underestimates the reliability of tests with few items or multidimensional scales [153].

trust and perceived ease of use were below the optimal α range, but they also comprised of few items and are, thus, likely underestimated in terms of reliability.

We base our understanding of **acceptance** on the "Technology Acceptance Model" developed by Davis [28]. Here, the acceptance of a model is mainly determined by how easy it is to use (**perceived ease of use**) and whether the user believes the system will enhance their performance on a task (**perceived usefulness**). Other variables, such as the intention to use the system, are also often measured in user acceptance studies but play a lesser role in our scenario, where the employment of an automatic feedback system is controlled mainly by the instructor and not the students. Thus, we develop three items each for "perceived ease of use" and "perceived usefulness" based on items used in related work [107, 178].

Trust is also a concept with a variety of interpretations. We follow Mayer, Davis, and Schoorman's [102, p.712] definition of trust being "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party." They mainly consider three aspects of trustworthiness: ability, benevolence and integrity. Ability captures whether the trustor believes the trustee has the skills and competencies to perform the action. Benevolence indicates whether the trustor believes the trustee wants to act in the trustor's interest, aside from egocentric motives. Integrity describes whether the trustor believes the trustee behaves according to a set of principles the trustor finds acceptable. In our case, trust describes how willing the students, the trustors, are to have their work assessed by an automatic system, the trustee, and believe it to provide accurate and helpful feedback while adhering to principles, such as keeping their responses private. We develop three items to measure trust based on previous work by Afshan et al. [1].

We expect students to trust and accept human feedback more than automatically generated feedback. Considering the generated feedback's imperfection, we expect students in the AI Group to view the feedback generator less favorably at T2 than at T1. For the other groups, no significant change in attitude is expected.

Method of Analysis

We use Linear Mixed Models (LMMs) instead of a mixed-design ANOVA to analyze each dependent variable for multiple reasons. First, LMMs can incorporate data points with missing values [79]. This study has 19 incomplete data points, where students either elected not to participate in every measuring point or had to be excluded due to cheating or mismatched participant codes. Since this is a non-negligible number, relevant information would have to be discarded when using an ANOVA instead. Second, all participants were recruited from the same class, thus most likely violating the

Variable		N	Mean	σ	Reliability(α)
Learning Gain					
Self-reported		70	5.09	1.32	0.84
Objective		67	2.84	0.95	-
Motivation					
Intrinsic goal orientation		77	4.70	1.19	0.73
Task value		77	5.47	0.81	0.80
Control beliefs		77	5.67	0.92	0.63
Self-efficacy for learning performance		77	5.47	1.00	0.89
Test anxiety		77	3.05	1.27	0.72
Acceptance					
Perceived usefulness	AI	77	4.99	1.25	0.84
	H	77	5.78	0.91	0.81
Perceived ease of use	AI	77	4.42	1.06	0.57
	H	77	5.62	0.78	0.41
Trust	AI	77	5.02	1.16	0.61
	H	77	5.65	0.78	0.41
Fairness	AI	77	4.65	0.88	0.04
	H	77	5.12	0.87	0.18

Table 25: Descriptive statistics of each dependent variable at T1. Acceptance, trust and fairness were each rated separately toward human (H) and automatic (AI) graders. All variables except objective learning gain scale from 1 to 7, objective learning gain from 0 to 3.5. Fairness was excluded from further analysis due to its low internal consistency as measured by Cronbach's α. Table adapted from [48].

ANOVA's independence assumption. Finally, LMMs are also more robust to data that is not normally distributed [139]. While most variables looked normally distributed when visualized, a number of them did not pass a statistical test for normalcy. Thus, we fitted a linear mixed model for each dependent variable using restricted maximum likelihood.

Since participants rated the variables multiple times (across grader types and measuring points), we cannot assume the provided ratings to be independent. For example, a student who was highly motivated at T1 is much more likely to be highly motivated at T2 than a student who was demotivated at T1. For this reason, we model the participants as random effects that affect our measurement of the group impact. Due to the dataset size, we assume slopes to be fixed across a group's participants since it reduces the number of parameters the model aims to estimate. However, it is likely that a participant's changes from one measuring point or grader type to another are affected by individual characteristics in practice. In other words, the slopes likely differ from participant to participant. For example, some participants may lose motivation more quickly than others – irrespective of their group assignment. This should be controlled for using a random-slope model in more extensive studies with multiple measurement points in the future. We treated the group, time and type of grader (where appropriate) as fixed effects.

One drawback of LMMs is the uncertainty of how to best estimate significance of interactions considering the model's hierarchical structure. As recommended by Luke [96], we calculate t-tests and significance levels with the Satterthwaite approximations for degrees of freedom method. Despite its superior performance compared to alternative approaches, resulting p-values should serve only as an indication due to possibly high error rates [96]. We selected Group AI at T1 as the reference point for the test comparisons, meaning that the fixed effect intercept is the mean of Group AI at T1 for a given variable. Each effect estimate is, thus, the change from the reference group in the given factor. For example, the estimate for Group Human is the difference between Group AI and Group Human at T1. The estimate for T2 captures how the variable changed from T1 to T2 in Group AI. Interactions, such as "Human x T2", indicate how the variable time affected the dependent variable in Group Human compared to Group AI.

Results: Learning Gain

The following sections provide an overview of the most important effects and statistics. The complete model statistics for each variable can be found in Appendix A.5. Figure 11 displays the estimated effects on subjective and objective learning gain. All scales besides objective learning gain ranged from 1 to 7. Objective learning gain ranged

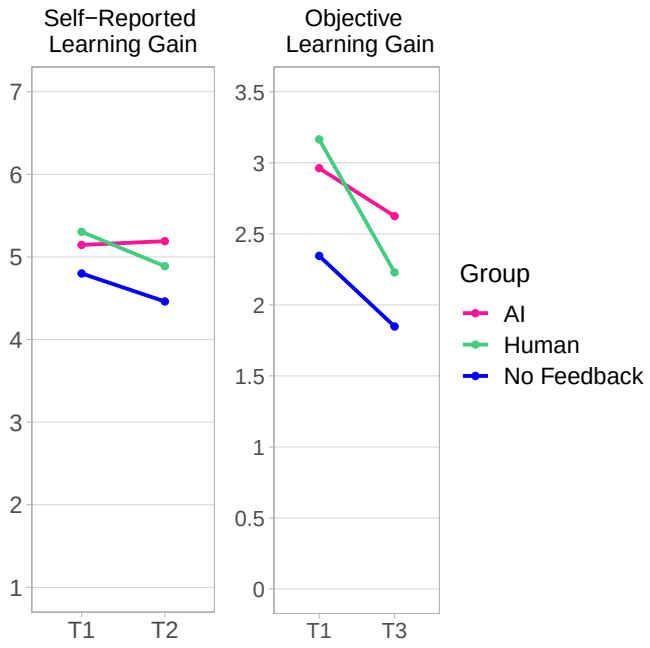


Figure 11: Estimated effects on learning gain for each group (N = 70 for self-reported gain and N = 71 for objective gain). Figure adapted from [48].

from 0 to 3.5. For **self-reported learning gain**, the model estimated a mean of 5.15 for Group AI at T1. Group Human reported a marginally higher initial knowledge and Group No-Feedback a slightly lower initial knowledge state compared to Group AI. From T1 to T2, Group AI reported no change in knowledge level, while both other groups reported a slight decrease in knowledge (Human x T2: Est. = -0.46; p = 0.17).

The **objective learning gain** behaved similarly but with more considerable differences between the groups. The model estimated a mean of 2.96 for Group AI at T1. Group Human was estimated to have slightly higher initial objective learning scores and Group No-Feedback significantly lower ones (Est. = -0.62; p = 0.048). In contrast to the subjective learning gain, all groups' knowledge decreased from T1 to T3. Group AI had the smallest decrease (Est. = -0.34; p = 0.22), followed by Group No-Feedback with a marginally higher reduction. Group Human had the largest loss in knowledge (Human × T2: Est. = -0.60; p = 0.12).

Overall, both learning gain measures behaved similarly and indicated that participants in Group AI forgot less compared to the other groups, with Group Human displaying the greatest loss in knowledge over time.

Results: Motivation

Figure 12 displays each motivation variable's estimated effects. For **intrinsic goal orientation**, the LMM estimated a mean of 4.90 for Group AI at T1. Group No-Feedback

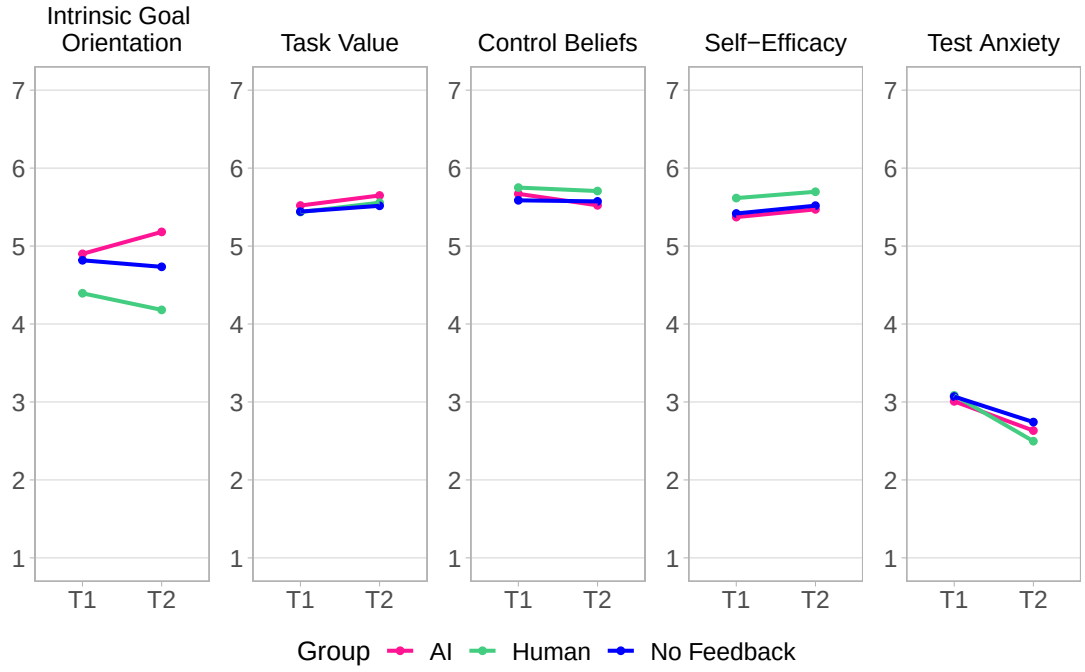


Figure 12: Estimated effects on motivation for each group (N = 77). Figure adapted from [48].

showed no mentionable difference to Group AI at T1. Group Human had a slightly lower initial intrinsic goal orientation than Group AI (Est. = -0.51; p = 0.12). From T1 to T2, intrinsic goal orientation increased slightly in Group AI (Est. = 0.28; p = 0.12). Both other groups showed a slight decline in intrinsic goal orientation over time, with Group Human diverging significantly from Group AI's change from T1 to T2 (Human x T2: Est. = -0.49; p = 0.048).

For **task value** and **self-efficacy**, all Groups behaved similarly – only Group Human had a slightly higher initial self-efficacy compared to the other groups. **Control beliefs** decreased minorly from T1 to T2 for Group AI (Est. = -0.15; p = 0.32) but stayed nearly constant for the other groups. **Test anxiety** improved significantly from T1 to T2 in Group AI (Est. = -0.38; p = 0.046). Group No-Feedback showed a similar decrease in anxiety and Group Human displayed a marginally larger decline than Group AI (Human x T2: Est. = -0.21; p = 0.40).

In summary, all feedback types had little effect on the motivation variables. Solely test anxiety lessened significantly from T1 to T2, irrespective of the group. The only notable difference between groups was observed for intrinsic goal orientation, which increased for Group AI and decreased for the other groups.

Results: Student Attitudes

Lastly, Figure 13 captures the estimated effects on students' attitudes toward human and AI graders. For **perceived usefulness**, the model estimated a mean of 5.37 for Group AI toward the automated feedback system at T1. This configuration was chosen as the reference for all the following comparisons. Group No-Feedback had a slightly worse perception of AI graders' usefulness at T1 (Est. = -0.31; p = 0.33) and Group Human had a significantly lower perception of its usefulness (Est. = -0.82; p = 0.01). All groups rated human feedback as more useful at T1, with Group Human displaying the largest difference, followed by Group No-Feedback and, finally, Group AI.

At T2, students were asked to imagine two scenarios and rate the respective grader type. First, that the feedback they had received stemmed from an AI grader and, second, that it had been formulated by a human instead. Nearly all groups perceived both grader types as slightly less useful than at T1. Only the No-Feedback Group viewed human feedback as marginally more useful at T2. While nearly all perceptions declined similarly, Group AI rating the AI grader had a noticeably steeper decrease than the rest (Est. = -0.47; p = 0.11).

Concerning the perceived **ease of use**, the model estimated a mean of 4.60 in Group AI rating a hypothetical AI feedback system at T1. Both other groups reported slightly lower values. All groups rated human feedback as significantly easier to use than AI feedback at T1, with Group Human showing the largest difference (Human × human grader: Est. = 0.75; p = 0.03). This is followed by Group No-Feedback (No-Feedback × human grader: Est. = 0.59; p = 0.08) and, lastly, Group AI (Est. = 0.75; p = <.01). At T2, Group AI and Human both rated AI feedback to be marginally less easy to use than at T1, while Group No-Feedback found it marginally easier to use. The ratings of human graders increased marginally for Groups AI and No-Feedback, but decreased substantially for Group Human (Human x human grader × T2: Est. = -0.72; p = 0.14). Thus, students found human feedback harder to use than expected after receiving it.

For the final dependent variable, **trust**, the model estimated a mean of 5.32 for Group AI at T1, rating an AI grader. Groups No-Feedback and Human trusted an AI feedback system slightly less initially. All groups trusted human feedback slightly more than AI feedback at T1, with Group AI having the smallest difference between the grader types. At T2, Group AI trusted the AI grader slightly less (Est. = -0.31; p = 0.19). Both other groups trusted the AI grader slightly more at T2. This means that Group Human's trust in the AI increased under the assumption that the feedback they had received had been automatically generated. The level of trust in human graders stayed mostly constant in Groups AI and No-Feedback but decreased slightly in Group Human.

In summary, human graders were generally more accepted and trusted than AI grades in all groups at all times. Group AI found AI graders less useful and slightly

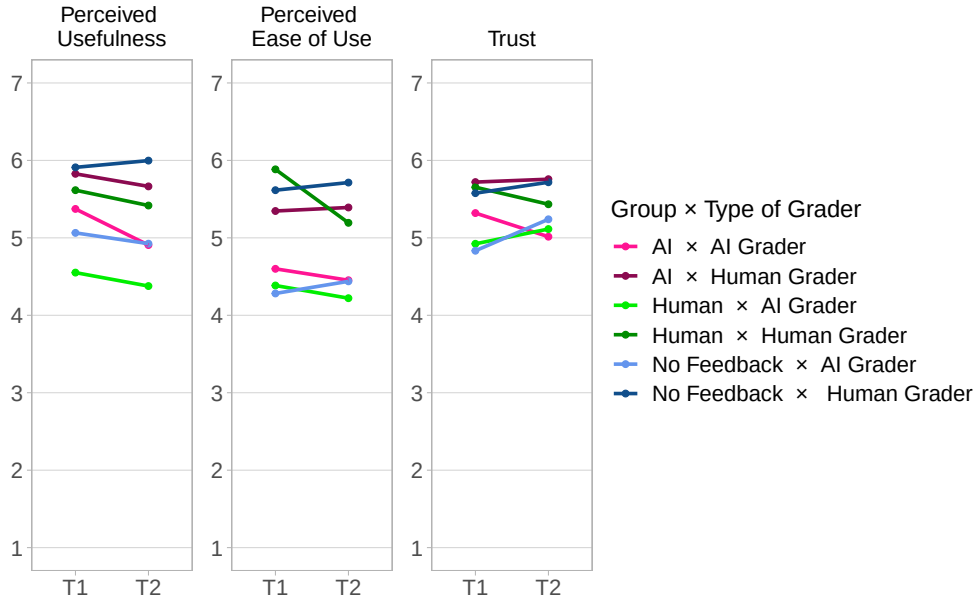


Figure 13: Estimated effects on acceptance and trust for each group and grader type (N = 77). Figure adapted from [48].

True Group	Guesses: AI Grader	Guessses Human Grader	Unsure
AI (N = 21)	**12 (82.58%)**	9 (72.00%)	0
H (N = 25)	13 (73.31%)	**11 (73.18%)**	1

Table 26: Student guesses on what type of feedback they had received. The student's reported average certainty for each guess type can be found in parentheses, correct guesses are marked in bold. Table adapted from [48].

less trustworthy after receiving automatically generated feedback. Symmetrically, students who had received human feedback also found human graders slightly less trustworthy and useful after receiving it, with a substantial decrease in ease of use.

Results: Students Guessing Feedback Type

Before students were informed of their grader type, they were asked whether they thought a human or a machine learning model had generated their feedback. They also assigned a certainty score to their guess, ranging from 0 to 100%. Students were generally very sure of their guess ($\bar{x} = 75.49\%$, $\sigma = 15.62$, min = 40%, max = 100%), with students correctly identifying their feedback as automatically generated slightly more sure than the others. In Groups Human and AI, around half of the students guessed that an automatic system had graded them. All in all, only around half of the participants guessed their grader correctly. Thus, they did not exceed chance level. The exact number of guesses per group and type can be found in Table 26.

4.3.5 *Interpretation of Results & Limitations*

To summarize, we introduced several supervised elaborated feedback generation approaches. Evaluating each approach on SAF showed that the generated feedback generally had a high lexical overlap with the gold-standard human feedback if the model had seen other answers to a given question during training. While still falling short of the annotators' performance, automatic similarity metrics indicated a sufficient quality for a deeper evaluation. Thus, we performed a field study in a communication networks lecture using the T5 feedback generator.

We observed that the type of feedback received did not notably impact the motivational variables task value, control beliefs, test anxiety and self-efficacy. Contrary to expectations, intrinsic goal orientation even increased for students who had received the automatically generated and sometimes imperfect feedback compared to the other groups. Similarly, they seemed to be less affected by forgetting as their learning gain decreased less with time. One possible explanation could be that students engaged more with the feedback received due to its unexpectedness. If the students received questionable feedback, they might be motivated to cross-check or further think about the learning material to understand it better. This may also lead them to view the quiz as more of a learning opportunity than a means to achieve an extrinsic reward.

Generally, all groups deemed human graders more useful and trustworthy at all measuring points. This is unsurprising, considering the students were generally familiar with machine learning and, thus, also its weaknesses. Interestingly, both feedback groups viewed their respective graders slightly less favorably after receiving feedback, with AI feedback declining more sharply in perceived usefulness and human feedback being viewed as substantially less easy to use.

The automatically generated feedback had no harmful effects on students' motivation and even improved learning gain in our study. Nevertheless, that does not mean it is on par with human feedback. First, the study was intentionally short to mitigate any potentially harmful effects of imperfect automatic feedback. While that was a sensible design decision for an explorative study, a follow-up study over a longer time is necessary to investigate long-term effects. The positive effects observed in our study may arise from the novelty of the imperfect feedback. Feedback might be ignored altogether should students continue to lose trust in the feedback generator. Alternatively, students may incorporate faulty feedback in their learning process, harming their learning gain. The students' ability to guess their grader type was hardly above chance level, even though the automatic feedback was often deemed inferior by teaching staff in random checks. Therefore, students may not be able to identify faulty feedback sufficiently.

Second, the feedback system's robustness toward cheating should be further explored before use in practice. One of the students in the field study successfully tricked the feedback model by answering "Aus dem Hut gezauberten Text prüfen" to the English communication networks questions. While cheating was not the focus of this study, we expect it to have more substantial effects in long-term studies where students have more time to exchange successful exploits.

Nevertheless, the evidence suggests that even an imperfect feedback generator can be a powerful educational tool if used correctly. Instead of presenting the feedback as fact to students, it may be better suited as critical-thinking prompts to further engage students with the material. We imagine asking students to explicitly criticize their received feedback could be an exciting task, beneficial to their intrinsic motivation and learning while containing potential harm through semantically faulty feedback. However, considering how sharply the models' performance falls off on unseen questions, this is only advisable for questions with some labeled data.

4.4 UNSUPERVISED FEEDBACK GENERATION

In the previous section, we introduced feedback generation approaches requiring access to training data. However, we also experienced how labor and cost-intensive the collection of reliable, high-quality feedback is – especially when thousands of examples are needed to train and evaluate a feedback model properly. Collecting vast amounts of data in every application domain is infeasible. At the same time, current language models are limited in their zero-shot generalizability to new questions and domains. For these reasons, we explore unsupervised feedback generation methods in this section.

One method often utilized in practice is eschewing personalized feedback and instead supplying a reference solution. Oftentimes the reference solution is accompanied by a score or label indicating the response's correctness, as verifying is much faster than writing out individual improvement suggestions. An example of such feedback can be seen in the top row of Table 27. Yet it can be challenging to infer one's mistakes from a reference solution. Teachers may have different perspectives than the students or use unfamiliar terminology, making the reference solution incomprehensible [169]. Depending on the solution's detailedness, it may also be incomplete and fail to cover the learner's response or, on the flip side, be too detailed and overwhelming. Consequently, a concrete improvement suggestion in the learner's own words would be more easily understandable and actionable [144].

Question:	What happens to the volume of the sound if you pluck a rubber band harder?
Reference:	The volume increases. The sound is louder.
Response:	It vibrates more and it gets lower. → *Incorrect*
Counterfactual:	It vibrates more and it makes louder sound. → *Correct*

Table 27: Example student answer with commonly given feedback (verification and reference answer) compared to generated counterfactual feedback. The student answer stems from SCIENTSBANK. Table adapted from [47].

For this reason, we propose unsupervised feedback methods based on counterfactual reasoning to provide individual improvement suggestions. The approaches utilize classical Automatic Short Answer Grading (ASAG) models and are motivated by the question of how the grading model's input would have needed to look to achieve a different prediction outcome instead [159]. In this regard, it is similar to an adversarial attack with the difference of looking to change the label's actual class instead of only fooling the model. Thus, the space of possible modifications is explicitly not limited to meaning- or class-preserving changes; for instance, given the response in

Table 27, a counterfactual generator searches for modifications of the response that would have made the grading model classify it as *correct* instead. In this case, it found that replacing "it gets lower" with "it makes louder sound" improves the assigned classification. The resulting counterfactual is a concrete improvement suggestion in the learner's own words, clearly indicating which part of the original response was incorrect and how it could have been improved.

However, only some student responses are improvable with minor modifications. Irrelevant responses like "I do not know" would have to be rewritten entirely to be correct. Some modifications may also lead to adversarial examples, where the grading model predicts them as correct, but there is no genuine improvement. Thus, this chapter explores to *which extent counterfactual generators are suitable to produce unsupervised elaborated feedback.* For this purpose, we evaluate three methods on three ASAG datasets automatically, followed by an expert evaluation of the generated counterfactual examples for one of the datasets..

4.4.1 *Counterfactual Feedback Approaches*

All of the following approaches aim to provide feedback to incorrect student responses by generating more correct versions of them. The first counterfactual feedback approach is based on the Minimal Contrastive Editing (MiCE) Framework [131]. Its goal is to iteratively replace the most impactful tokens in a response until the ASAG model predicts it as correct. The grading model's gradient determines the impact of a token. The second approach is based on Polyjuice [173]. In contrast to the previous approach, it performs controlled modifications based on predefined control codes, such as "negate" or "shuffle." Finally, we introduce a paraphrasing-based approach that generates a novel correct response based on the original student's answer. A comparison of the approaches at inference time can be found in Figure 14 and a description of their training process follows.

Contrastive Infilling based on MiCE

MiCE is a contrastive infilling approach where the main idea is to iteratively identify parts of the input detrimental to the prediction of the target class and replace them. Problematic input sections are identified by calculating the grading model's gradients with regard to the input tokens. This gradient reveals tokens that have a negative impact on the prediction of the target class. The detrimental sections can then be masked and provided to an editor model that generates a replacement. This process is also depicted in Figure 14. Prior to the actual generation of counterfactual examples at inference time, the editor model must be trained to perform infilling.

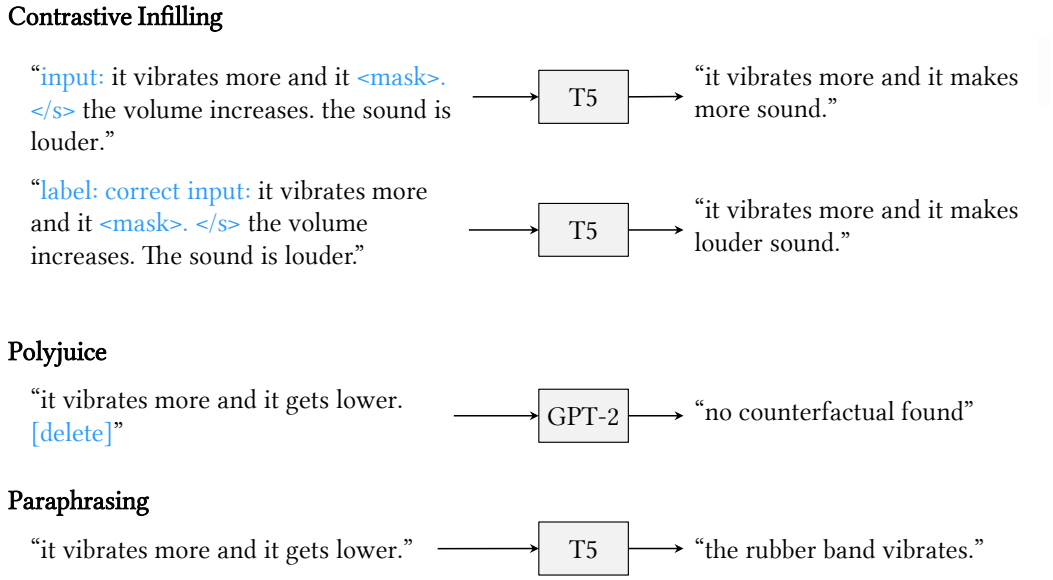


Figure 14: Input and output of the counterfactual approaches at inference time. The particular example was chosen due its briefness.

We implement and train two distinct editor models, both using the pre-trained T5 architecture. The first follows the process of Ross, Marasović, and Peters [131]. It is trained to reproduce original student responses where 20-55% of the tokens have been randomly masked. In contrast to the second editor model, it is conditioned on the response's class to enable the generation modifications for a target class at inference time. The label conditioning can be seen in the bottom contrastive infilling approach in Figure 14. We will refer to this approach as "label infill" for the remainder of this chapter.

The second model deviates from the original MiCE framework as we do not condition it on a target label. The main idea is that we only wish to generate *correct* student responses at inference time since we wish to generate improvement suggestions for wrong answers. The label, therefore, does not carry any information, as it will be the same for every response. While we lose the ability to target other classes, such as *partially correct*, we gain training efficiency as the editor now is only trained on *correct* responses instead of all responses in the dataset. We will refer to this approach as "infill" for the remainder of this chapter. Both editor models receive the reference answer as context.

After fine-tuning the editor model, we begin the iterative modification process of *incorrect* and *partially correct* student answers. In each iteration, we mask consecutive spans of tokens based on importance scores calculated by the gradient attribution method Integrated Gradients [149]. Four masks are created in each iteration, with 15,

30, 45 and 60% of the tokens masked. The masked tokens are added to a list of words the editor is not allowed to generate to prevent it from reproducing the original response. Seven replacement candidates are generated for each masked answer with a combination of top-k=30 and top-p=0.95 sampling[15]. Thus, a total of 28 candidates are generated per iteration. The candidates are then graded by the ASAG model described in Section 4.4.2 and solely the candidate with the highest target class probability is retained. Should the candidate's class probability exceed the classification threshold to be labeled as *correct* by the grading model, the modification loop is terminated. Otherwise, the modification loop is stopped after four iterations to conserve computation resources, arguing that we will likely not find a suitable modification that is still close to the original answer if we have not found it by then.

Polyjuice

Polyjuice aims to control the modification process to produce more fluent counterfactual examples. Instead of masking the most impactful tokens and giving an editor model free rein over replacements, the location and type of modification can be controlled using an editor model conditioned on control codes (see Section 4.4.2 for an example). Wu et al. [173] provide a set of control codes, such as a *negation* of a sentence's meaning or a *shuffle* of key phrases or entities. Phrases may also be inserted, deleted, replaced and restructured, quantifiers replaced or words replaced with grammatically similar ones. We utilize the already fine-tuned GPT-2 editor model[16] proposed by Wu et al. [173] with all possible control codes. While fine-tuning an editor model specifically for the ASAG domain would likely produce better results, we do not have access to the required sentence pairs exhibiting the desired modifications in our domain. For example, we would require syntactic paraphrases of sentences in student responses. In general, we expect this method to produce fewer counterfactuals overall due to the constraints put on possible modifications, but it should produce counterfactuals that appear more natural and fluent.

Paraphrasing

The last approach is based on the idea that a model trained to produce paraphrases of correct student responses may produce corrected versions of incorrect responses supplied at inference time. We train a T5 model to paraphrase by supplying pairs of correct student responses and reference answers. Technically, these are not proper paraphrases since answers can be more or less detailed or even focus on different aspects altogether. However, this suits us as the main idea is for the model to learn the

15 https://docs.cohere.com/docs/controlling-generation-with-top-k-top-p [accessed May 8]
16 https://github.com/tongshuangwu/polyjuice [accessed April 24, 2023]

Dataset	Unseen Answers			Unseen Questions			Unseen Domains		
	Acc	M-F1	W-F1	Acc	M-F1	W-F1	Acc	M-F1	W-F1
SAF	77.1	75.5	77.1	52.9	57.5	52.9	-	-	-
BEETLE	71.4	69.7	71.4	54.8	54.8	56.6	-	-	-
SCIENTS.	72.9	70.9	72.9	59.7	50.9	59.7	61.5	54.6	61.5

Table 28: The performance of the ASAG models used for counterfactual feedback generation. Accuracy (Acc), macro-averaged F1 (M-F1) and weighted F1 (W-F1) are given in percent. Table adapted from [47].

characteristics of correct responses and not precisely reproduce the original content. If it produced proper paraphrases, it would not improve upon incorrect answers. In contrast to the other counterfactual approaches, this method does not mask parts of the student's answer but produces a novel response from scratch. During training, it receives either a reference answer or a correct response and generates the respective counterpart. During inference, it is supplied with an incorrect response instead.

4.4.2 *Experimental Settings*

We aim to explore to which extent feedback can be generated with counterfactual generators. Thus, it is sensible to investigate the method's performance on multiple ASAG datasets of different domains. We select three benchmarks for this purpose: SCIENTSBANK, BEETLE and the English part of SAF. We exclude the German splits as the fine-tuned GPT-2 model is unsuitable for multilingual inputs. The following experiments precede the full construction of the final SAF dataset and, therefore, only contain the first semester of the communication networks lecture. We choose the three-way classification setting for each dataset, with all datasets containing the classes *correct* and *incorrect*. SAF has *partially correct* as the third class while SCIENTSBANK and BEETLE consider *contradictory* as the final class. If there are multiple reference answers to a question, as is often the case in BEETLE, we consider all of them.

We train ASAG models for each dataset. The models serve as the judge for the counterfactual search. Since the ASAG model will be frequently queried during the search, we opt for the more computationally efficient model BERT instead of T5. We follow a similar training procedure as described in Section 3.2.2 and obtain the predictive performance displayed in Table 28. Due to their size, all editor models were trained on two Nvidia RTX 2080 Ti cards with 11GB of RAM using mixed-precision floating-point numbers and gradient accumulation over 16 batches with a batch size of 2, resulting in an effective batch size of 32. We used an Adafactor optimizer with a constant learning rate of 0.001.

Evaluation Measures

Mainly, two dimensions affect the quality of counterfactuals as feedback: **validity and proximity**. While other factors, such as fluency or diversity, are also sometimes considered, they play a lesser role in a feedback application. We neither need to produce multiple, diverse improvement suggestions nor do the counterfactuals necessarily need to be grammatically correct, considering the initial student response is often not fluent either. It is much more critical that the counterfactual remains as close as possible to the original answer while changing the answer's class.

Typically, the validity of a counterfactual generator is measured by the percentage of its generated counterfactuals that achieve the desired prediction [159] – irrespective of the class predicted prior to modification. This works well for related work, as they usually utilize predictors with near-perfect accuracy but would overestimate the generator's performance due to the ASAG models' imperfectness. Thus, we exclude all responses mistakenly predicted as *correct* from the evaluation, as they do not require modification. A counterfactual is only deemed successful if it flips the predicted label to *correct*. We also calculate the **flip rate** for each class individually to investigate whether the initial degree of incorrectness influences the counterfactual generation. Nevertheless, even though flip rates are often used to measure validity in related work, they may not be reliable. As shown in Chapter 3, slight modifications to student responses can confuse grading models into predicting them as being *correct* even though they are not. Since we are looking for real improvements, we also conduct a regrading of a subset of the generated counterfactuals with a human expert to get a more reliable validity indicator.

Similar to related work [131], we utilize word-level Levenshtein **distance** to measure how close a counterfactual remains to the original student response. Levenshtein distance calculates the minimal number of insertion, deletion or replacement operations to equalize two strings and normalizes it with the total number of words in the original student answer. It can be seen as the percentage of words modified as long as the modified response remains maximally as long as the original answer. While the distance should not be zero, as some modifications are desired, it should only be as large as needed to improve the student response and follow the learner's expressions beyond that.

Evaluation Process

We evaluate the generated counterfactual feedback in two steps. First, we analyze each approach's flip rate and average distance to the initial response for each dataset, similar to evaluations done in related work. Second, a domain expert manually determines whether the counterfactual modification truly corrects the student's response

Approach	Unseen Answers				Unseen Questions			
	Contra (48)		Incorrect (202)		Contra (35)		Incorrect (238)	
	FR	Dist	FR	Dist	FR	Dist	FR	Dist
Paraphrase	**72.9**	1.54	**74.8**	1.82	**65.7**	1.69	**68.5**	1.65
Infill	31.2	0.16	33.2	0.17	17.1	0.12	31.1	0.15
Label Infill	29.2	0.15	31.7	0.18	20.0	**0.11**	34.5	0.17
Polyjuice	2.1	**0.14**	1.0	**0.12**	5.7	0.12	2.5	**0.11**

Approach	Unseen Domains			
	Contra (279)		Incorrect (1583)	
	FR	Dist	FR	Dist
Paraphrase	**48.4**	2.02	**63.2**	2.12
Infill	38.4	0.16	27.3	0.15
Label Infill	38.7	0.18	27.7	0.16
Polyjuice	2.9	**0.09**	3.0	**0.13**

Table 29: The average Levenshtein distance (Dist) and flip rate (FR) of the counterfactual generation methods on SCIENTSBANK's contradictory (Contra) and incorrect responses. Sample sizes are given in parentheses. The best performances are marked in bold. Table adapted from [47].

or whether the grading model was only fooled. The second step is done for all counterfactuals generated on the SAF dataset, as it is the only dataset with an elaborated explanation of the student's mistake. A similar evaluation of the other datasets would take a lot more time and effort, as the grader would first have to deduce why the response received its assigned grade before they can decide whether the mistake was rectified in the counterfactual.

4.4.3 *Feedback Evaluation*

On SCIENTSBANK (see Table 29), the paraphrasing approach achieves the highest flip rates by a large margin. For all test splits and classes besides *contradictory* in unseen domains, it achieves flip rates of over 60%. In contrast, the contrastive infilling approach with label conditioning (Label Infill) flips 29.8% of the labels on average across each class and test split, with a slightly higher success rate on *incorrect* than on *contradictory* responses on the unseen answers and unseen questions split. The trend is reversed for unseen domains.

The contrastive infilling approach without label conditioning (Infill) performs similarly with an average flip rate of 30.2% across classes and splits and the same trend of performing better on *incorrect* answers on the first two data splits but not on the

Approach	Unseen Answers				Unseen Questions			
	Contra (453)		Incorrect (480)		Contra (740)		Incorrect (830)	
	FR	Dist	FR	Dist	FR	Dist	FR	Dist
Paraphrase	**74.2**	8.56	**78.3**	10.63	**76.9**	8.04	**74.7**	13.27
Infill	60.9	2.77	63.3	2.18	46.5	2.38	52.8	2.14
Label Infill	44.8	0.42	41.9	0.39	39.1	0.34	38.0	0.33
Polyjuice	1.8	**0.11**	2.1	**0.14**	1.8	**0.12**	3.3	**0.17**

Table 30: The average Levenshtein distance (Dist) and flip rate (FR) of the counterfactual generation methods on Beetle's contradictory (Contra) and incorrect responses. Sample sizes are given in parentheses. The best performances are marked in bold. Table adapted from [47].

unseen domains split. While both infilling approaches have a significantly lower flip rate than the paraphrasing approach, the counterfactuals produced are much closer to the original response, with an average distance of 0.16 and 0.15 across all data splits compared to paraphrasing's 1.81. Polyjuice produces only few successful counterfactuals (1-6%) on SciEntsBank, but they are also close to the original with an average distance of 0.12. Opposed to the previous methods, Poliyjuice performs slightly better on *contradictory* responses.

On Beetle (see Table 30), the general pattern of the paraphrasing model having the highest (average of 76%) and Ployjuice having the lowest flip rate (average of 2.3%) remains unchanged. Ployjuice's modifications also remain close to the originals, with an average distance of 0.14. The paraphrasing model produces even more dissimilar counterfactuals on Beetle compared to SciEntsbank with distances between 8 and 13.3. On average, more label flips are achieved by the contrastive infilling approaches on this dataset but at the cost of higher distances. The approach without label conditioning achieves an average flip rate of 55.9% and a distance of 2.37. The approach with label conditioning flips 41% of the labels on average with a distance of 0.37. Since Beetle is the only dataset with multiple reference answers and each student answer and reference answer pair form a sample, questions with many reference answers may be overrepresented in this statistic.

Finally, Table 31 displays the results on SAF. Once again, the paraphrasing model produces the highest rate of label-flipping counterfactuals with flip rates between 50 and 100% but also the highest Levenshtein distance (2.22 on average across all splits). On SAF, Polyjuice produces only slightly modified counterfactuals that seldom flip the predicted class with flip rates between 0 and 11% and an average edit distance of 0.02. The contrastive infilling approach with label conditioning flips 21.9% on average with a distance of 0.13 and the approach without label conditioning flips 24.2% on average with a distance of 0.15.

Approach	Unseen Answers				Unseen Questions			
	Partial (52)		Incorrect (9)		Partial (31)		Incorrect (8)	
	FR	Dist	FR	Dist	FR	Dist	FR	Dist
Paraphrase	**50.0**	1.72	**77.8**	3.89	**96.8**	1.60	**100**	1.66
Infill	25.0	0.19	11.1	0.11	35.5	0.12	25.0	0.19
Label Infill	19.2	0.14	11.1	0.10	32.3	0.12	25.0	0.16
Polyjuice	0.0	**0.01**	11.1	**0.01**	3.2	**0.03**	0.0	**0.01**

Table 31: The average Levenshtein distance (Dist) and flip rate (FR) of the counterfactual generation methods on SAF's partially correct (Partial) and incorrect responses. Sample sizes are given in parentheses. The best performances are marked in bold. Table adapted from [47].

Expert Regrading

While all approaches besides Polyjuice show promising flip rates, we must consider the possibility of the predictor being fooled and deeming the responses correct without a genuine improvement. To further investigate this, we tasked one of the communication network experts involved in SAF's annotation with manually examining the generated counterfactuals. For this purpose, we provided the expert with the original student response, elaborated feedback and reference answer for each counterfactual that flipped the ASAG's label to *correct*. In total, 59 counterfactuals generated by the paraphrasing model were regraded, 21 from the label conditioned infilling, 25 from the contrastive infilling without labels and a single counterfactual generated by Polyjuice. Thus, the expert examined a total of 106 counterfactuals. Table 32 contains an example counterfactual for each approach, selected for their brevity.

As can be seen, most of the modifications performed are not genuine improvements. Of the 106 counterfactuals, 103 are adversarial examples that only fool the grading model without correcting the response. Furthermore, two of the corrected responses stem from the paraphrasing model generating the reference answer to the question. The paraphrasing model generally generated entirely novel responses that had little to do with the original answer, corroborating the high edit distances. Interestingly, the generated counterfactuals are sometimes also mixes of various reference solutions. This indicates that the grading model may have difficulties differentiating between questions even though it receives the corresponding reference answer as input. Sometimes, the paraphrasing model also generated utterly absurd content, such as "*... 56.648 * 64 bit/sec = 128 bit processing tables = 276 bit data transfer tables + 3 * 1.31 seconds to reach the destination system ...*".

Both infilling models also primarily produced adversarial examples with meaningless modifications. For instance, "*... the issue with this **case** is ...*" was replaced with "...

	Polyjuice
Original:	In this case the system will be for the whole time in a state, in which there are less than 10 packets waiting in the queue, due to the fact that there are always more packets processed pro second than arriving. Utilisation = arrival rate / service rate = 9 pkts/s / 10 pkts/s = 0.9 N - average number **of packets in the system** N = utilisation / 1 - utilisation which gives = 9 packets We can also calculate the probability, that the system is full: $p_{10} = (1-p) * p^{10}/(1-p^{11})$ we get 0.05 as a result. Because the utilisation ist the same at every time, the probability that the system is full remains equally.
Counter.:	In this case the system will be for the whole time in a state, in which there are less than 10 packets waiting in the queue, due to the fact that there are always more packets processed pro second than arriving. Utilisation = arrival rate / service rate = 9 pkts/s / 10 pkts/s = 0.9 N - average number **of in the system** N = utilisation / 1 - utilisation which gives = 9 packets We can also calculate the probability, that the system is full: $p_{10} = (1-p) * p^{10}/(1-p^{11})$ we get 0.05 as a result. Because the utilisation ist the same at every time, the probability that the system is full remains equally.
	Paraphrasing
Original:	Control plane is logically centralized and responsible to define routes and manages network graphs while data plane is physically distributed and responsible for forwarding of packets.
Counter.:	to support billions of end-systems. to reduce routing tables. and simplify protocol processing with simplified header.
	Contrastive Infilling with Labels
Original:	The flow tables are used to match certain properties in order to determine **their** path. **Meter tables** are used for statistics.
Counter.:	The flow tables are used to match certain properties in order to determine **a** path. **also flow tables** are used for statistics.
	Contrastive Infilling without Labels
Original:	It couldn't **be in the** real internet. In the real internet if a packet is lost or has error, it needs to be retransmitted. Obviously it could affect the following arrivals. It means the **time intervals** could not be independent.
Counter.:	It couldn't **be independent in the** real internet. In the real internet if a packet is lost or has error, it needs to be retransmitted. Obviously it could affect the following arrivals. It means the **time of transmission** could not be independent.

Table 32: Example counterfactuals generated for SAF. Subtle changes are highlighted in blue.

the issue with this ***narcotic*** *is ...*". Nevertheless, not all of the infilling models' modifications were adversarial. Some modifications improved student responses partway, even if they were still incorrect overall. For example, "*extension headers are* ***the way to put additional information in the packet****...*" was correctly replaced with "*extension headers are* ***used to extend the fixed ipv6 header with additional options****...*". This is a legitimate improvement and may be helpful as elaborated feedback.

4.4.4 *Interpretation of Results & Limitations*

In conclusion, three of the four compared counterfactual generation methods produced counterfactuals that flipped the grading model's prediction for at least a fifth of the responses in three diverse datasets. Approximately a quarter of the responses could be flipped with only slight modifications across all datasets. However, most counterfactuals turned out to be adversarial examples in an expert reexamination done for one of the datasets. Often, the performed modifications were meaningless or even absurd. This indicates that the flip rate is not a reliable success measure in this setting despite it often being used in related work. Current ASAG models are still too unreliable to provide the necessary direction to a counterfactual generator, possibly more unreliable than predictor models usually utilized in related work. Therefore, future work on counterfactuals as elaborated feedback should also include human evaluations and not rely only on automatic metrics.

We conclude that counterfactual generators are not yet suitable for elaborated feedback generation. Without more reliable grading models or including humans in the loop, the rate of genuine improvements is too low compared to the number of adversarial examples found. However, we did observe partial improvements in our experiments, indicating that the idea of counterfactual examples as feedback has merit and is a promising avenue for future research with more accurate grading. One possible method of improving grading accuracy could be including a human teacher in the prediction loop, either when marking which parts of the student response should be replaced or in determining whether a modification was successful. This could serve not only as a training signal for the counterfactual generator but also for the automatic grading model itself.

Especially counterfactual generators based on contrastive infilling are useful for exploring a model's decision boundary and could be helpful as a form of adversarial training. They may be even better suited than typical adversarial attacks since they are not constrained to modifications that preserve semantic or even class equivalency and, thus, are able to explore a more extensive and diverse search space. With a human in the loop, adversarial examples could be differentiated from genuine improvements

and fed back into the predictor's and generator's training set with the appropriate label. Found counterfactuals could also help identify more general issues with the grading model's architecture or configuration. For instance, the absurd modification of mathematical expressions observed in our experiments may indicate that mathematical content is not adequately represented in the embedding space and that a model with a different pre-training setup would be better suited to this domain.

In future work, other counterfactual generators could also be investigated. The number of counterfactuals found and their dissimilarity to the original response varied significantly from method to method in our experiments. Particularly approaches with external knowledge sources may yield a higher quantity and quality of counterfactuals since they could derive improvable sections based on not only the grading model's gradients but also domain knowledge. This knowledge could also inform and guide the replacement process toward semantically meaningful changes.

All in all, we conclude that the idea of counterfactual generation as a source of unsupervised feedback is promising based on the partial improvements observed in our experiments. However, further research on more reliable grading and human-in-the-loop approaches is required before the idea becomes serviceable in practice.

4.5 DISCUSSION OF FEEDBACK GENERATION

All in all, we have laid the groundwork for reproducible and comparable research on elaborated feedback generation. A high-quality and easily accessible dataset spanning three domains and educational settings can form the basis of training and model evaluation. The proposed evaluation framework can enhance comparability between approaches, even if established automatic similarity measures could be more reliable and, thus, should be supplemented with human evaluations. Finally, we developed a set of supervised as well as unsupervised feedback generation approaches that can serve as baselines for and inspire future research.

The supervised feedback generation approaches had a high lexical overlap with human gold standard feedback. However, the feedback generators generally fell short of human performance. Nevertheless, even imperfect feedback improved learning and motivation in a university course field study, providing evidence for its beneficial applicability in education. However, considering that students were not better than chance level in identifying automatically generated feedback, it should be marked clearly to encourage critical questioning of its validity. While we expect critical thinking about automatically generated feedback to benefit learning – at least in a university setting – long-term field studies are advisable.

While the unsupervised feedback methods – based on finding modifications of student answers that improved an automatically determined grade – showed promise, the generated feedback's quality still needs to be improved for practical application. The improvement suggestions generated by the unsupervised approaches often did not truly improve the student's response, only tricking the automatic grading model. Especially the modifications of mathematical expressions were absurd in nearly all cases. As of yet, the unsupervised generators may be helpful as tools to improve employed NLP models or as interactive feedback generators. A human in the feedback generation loop could significantly increase this approach's effectiveness, but further research with more reliable grading is needed.

5

SUMMARY, CONCLUSIONS, AND OUTLOOK

In summary, this work investigated how robust current Transformer-based Automatic Short Answer Grading (ASAG) approaches are to adversarially manipulated input (RQ1) and how they can be extended to provide elaborated feedback in addition to the verification feedback provided in related work (RQ2). We make the following contributions to these research questions:

- **We perform a thorough empirical analysis of state-of-the-art ASAG models' robustness to various adversarial attacks.** We observed significant losses in prediction accuracy on multi-class tasks. In some cases, accuracy was reduced by more than half. However, many of the misclassifications were due to the model predicting *incorrect* responses as *contradictory* instead of as *correct*. Whether it was inserting a single adjective or adverb, appending a trigger sequence, replacing words with their synonyms or students attacking the model manually, all attack strategies successfully fooled the grading model into misclassifying incorrect student responses as fully correct in 8-23% of cases.

- **We contribute a powerful adversarial attack based on adjective and adverb insertion.** It performs comparably to existing adversarial attacks while requiring no information about the target model. As demonstrated in our experiments, it can be a powerful attack for a variety of tasks where the insertion of a single adjective or adverb is unlikely to change an input's actual class. In an expert evaluation of modified responses, we verify that the attack adheres to the class-equivalency constraint in ASAG settings. Moreover, adversarial examples generated by this attack did not seem significantly more suspicious to human graders than the original student answers, indicating that students could utilize this attack relatively risk-free in practice.

- **We collect, annotate and publish a novel, high-quality ASAG dataset with elaborated feedback and develop a suite of supervised approaches based on the dataset.** We demonstrate the dataset's reliable grading with high inter-annotator agreement measures (Krippendorff's alpha between 0.78 and 0.91) between the two annotators of each dataset. We also present evidence for the grades' validity by comparing achieved points on the short answer questions with external, established criteria, such as exam scores and job failure rates. We train various

supervised models as benchmarks for future research. According to automatic evaluation measures, the models achieve a high grading accuracy and similarity with the gold standard elaborated feedback for answers they were trained on. However, they fall short of human performance, especially for novel questions. Nevertheless, a field evaluation of the automatically generated feedback in a university lecture showed positive effects on learning gain compared to no feedback and even human feedback conditions – without harming the students' motivation. Only the students' attitudes toward their respective graders worsened, irrespective of whether they had received manually or automatically generated feedback.

- **We develop and empirically evaluate unsupervised elaborated feedback methods based on existing counterfactual generation techniques.** Three out of four approaches produced sufficient counterfactuals that remained close to the original student's words while improving the automatically assigned grade. However, an expert evaluation of the counterfactuals generated for one of the datasets revealed that they were primarily adversarial examples and not genuine improvement suggestions. Nevertheless, partial improvements were observed even if the overall response remained incorrect.

5.1 CONCLUSIONS

With regards to RQ1, we draw the following main insights from our contributions:

- **State-of-the-art ASAG models predict at least partially based on non-robust features.** While this insight is expected, considering the existence of non-robust features in other domains, we show that this is also true for popular short answer grading models and tasks. Spurious correlations in the training data, such as certain adjectives appearing more often in correct student responses, affect a model's predictions. Adversarial attacks can exploit these non-robust features.

- **University students are likely able to autonomously identify systematic weaknesses of a given grading model only by being graded.** Nearly all students in our study identified at least one systematic weakness by submitting various responses to an automatic grading system and receiving the predicted classification. The students also discovered weaknesses that go beyond the keyword-sensitivity automatic attacks are exploiting, such as an insufficient understanding of negation. While further study is needed to clarify whether our findings generalize to the broader student population as our participants predominantly studied computer science-related fields, at least a portion of students are likely

capable of finding and exploiting non-robust features with prolonged contact with a model.

- **Adversarial manipulation can be challenging to detect and prove even with human graders.** In our expert study on the human perception of responses manipulated by the adjective and adverb insertion attack, human graders viewing the adversarial examples were not significantly more suspicious than a control group viewing the original responses. Even though a more fine-grained study is required to investigate potential minor effects, slight mistrust is unlikely to provoke action. We also observed evidence that human graders might wrongly suspect poorly written or wildly incorrect student responses. Therefore, detecting adversarial input without disadvantaging low-performing students may be challenging. Nevertheless, human graders were not fooled by the attack and can, thus, correct the model's mistakes even if they may not suspect – or be able to prove – deceptive intent.

In conclusion, while current state-of-the-art approaches achieve high prediction accuracy on existing datasets, they are not yet robust enough for summative assessments without human supervision. Any real-world application will likely confront models with out-of-distribution answers even without cheating attempts, as students and teaching materials vary from semester to semester. Our work on RQ1 and concurrent work [33] demonstrate that such responses lead to questionable predictions in current model architectures. Nevertheless, an ASAG model may be helpful for formative assessments or in conjunction with a human grader that checks a subset of the model's predictions. Random checks to reduce the success probability of adversarial attacks may be a good solution here, as they would not unduly oversample responses formulated by minority or low-performing students.

The following main insights can be drawn from our work on RQ2:

- **Current ASAG models are not yet reliable enough for feedback based on counterfactual generation methods.** Until a more reliable estimation of a response's correctness is established, we do not expect counterfactual generation methods, in general, to produce useful feedback. It is simply too easy to run into local optima that are adversarial examples instead. However, the partial improvements observed in our experiments lead us to believe that this is a promising area of future research, with either more reliable models or human graders providing reliable judgment.
- **Even imperfect, automatically generated feedback may be helpful to students.** In our field study, automatic feedback positively affected learning gain compared

to human and no feedback. While long-term studies are needed to exclude novelty effects, our results indicate that models may not have to replicate human feedback perfectly to improve learning. Critical engagement with imperfect feedback may be a pedagogically valuable exercise and warrants further study.

- **University students seem unable to differentiate between automatically generated and human feedback.** Only around half of the students could guess which feedback type they had received in our field study – despite the feedback generator not achieving human-level performance and being judged as inferior by the teaching staff. While students may also become more adept at spotting automatically generated feedback with long-term exposure, this insight should be considered carefully when employing such feedback systems in practice. Students may unquestionably incorporate faulty feedback if they believe it stems from their teacher.

In summary, we have laid the groundwork for reproducible and comparable research on elaborated feedback generation. We have also explored multiple approaches that extend traditional ASAG by including explanations of mistakes made (RQ2). While the tested approaches are yet to be comparable to human feedback regarding quality and reliability, we present evidence that they may be valuable to learning applications in practice. Encouraging students to think critically about automatically generated feedback received in formative assessments may prove even more beneficial to learning gain than traditional human feedback. While large-scale, long-term studies are needed to verify the results of our field study, we conclude that machine learning systems do not have to replicate human behavior to provide educational value.

5.2 OUTLOOK

One common insight underpinning this thesis is the need for better semantic understanding in current NLP models. While the debate on how to achieve better understanding is still ongoing, we think neuro-symbolic approaches are a promising avenue of research for Automatic Short Answer Grading. Combining the strengths of symbolic reasoning and neural networks could lead to more explainable and robust grading. One idea here would be to formulate grading rubrics not only in text for human graders but also in logical notation, specifying which concepts and relations should be found – or explicitly not found – in a correct response and how many points of the final grade they constitute. Neural networks could then assume the task of extracting concepts and relations from the unstructured natural language responses. In this way, one can explicitly refer to parts of the grading rubric which were not fulfilled by the

student's answer and is not reliant on a non-interpretable, fuzzy representation of the rubric implicitly encoded in the network based on observational data.

Supporting this line of research, we also hope to see more datasets focused on providing understandable elaborated feedback. While we have laid the groundwork with sizeable datasets from three domains, additional datasets would benefit the field. They could include additional learning contexts and provide more detailed annotations on which parts of the student response covered various rubric items. Even though general concept extraction methods [5] and datasets already exist and more detailed annotation studies would undoubtedly be even more costly than our dataset construction, they could facilitate pedagogically driven approaches. From related fields, we know that there are differences between general NLP datasets and ones constructed with a pedagogical goal [140, 147]. For example, concepts important for learning may differ from those critical for summarizing news articles.

Another exciting line of research could include humans in the feedback generation pipeline, especially during training in an interactive learning approach [154]. Human experts could potentially correct unreliable decision boundaries based on spurious correlations, especially when combined with adversarial attacks and counterfactual generators that find interesting example responses the model and humans should consider. Having students interact critically with the feedback model and incorporate their experiences in the training process could also be educationally valuable and provide a more robust model.

Finally, further research on the effect of automatically generated, imperfect feedback on students' attitudes, motivation and learning gain could yield unexpected and exciting insights – not only on the quality of automatic feedback but also on the process of learning itself. We were surprised that students receiving automatic feedback retained more knowledge over time than students receiving human feedback in our study. We are curious whether such an effect remains with prolonged system use and what its cause is.

ACKNOWLEDGMENTS

This research was funded by the *Bundesministerium für Bildung und Forschung* in the project: *Software Campus 2.0 Microproject: DA-VBB* and by the *Hessian State Chancellery of the department of Digital Strategy and Development* in the *Förderprogramm Distr@l*.

BIBLIOGRAPHY

[1] Sahar Afshan, Arshian Sharif, Nazneen Waseem, and Reema Frooghi. "Internet banking in Pakistan: an extended technology acceptance perspective." In: *International Journal of Business Information Systems* 27.3 (2018), pp. 383–410. DOI: 10.1504/IJBIS.2018.089863.

[2] Giora Alexandron, José A. Ruipérez- Valiente, Sunbok Lee, and David E. Pritchard. "Evaluating the Robustness of Learning Analytics Results Against Fake Learners." In: *Lifelong Technology-Enhanced Learn.* Springer International Publishing, 2018, pp. 74–87. DOI: 10.1007/978-3-319-98572-5_6.

[3] Giora Alexandron, Lisa Y. Yoo, José A. Ruipérez-Valiente, Sunbok Lee, and David E. Pritchard. "Are MOOC learning analytics results trustworthy? With fake learners, they might not be!" In: *International Journal of Artificial Intelligence in Education* 29.4 (2019), pp. 484–506. DOI: 10.1007/s40593-019-00183-1.

[4] Yassin Alkhalili et al. "Towards QoE-Driven Optimization of Multi-Dimensional Content Streaming." In: *Proc. of the Conf. on Networked Syst.* 2021. DOI: 10.14279/tuj.eceasst.80.1167.

[5] Wael Alkhatib. "Semantically Enhanced and Minimally Supervised Models for Ontology Construction, Text Classification, and Document Recommendation." PhD thesis. Dept. Elect. Eng., Technical Univ. of Darmstadt, Darmstadt, Germany, 2020. DOI: 10.25534/tuprints-00011890.

[6] Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. "Generating Natural Language Adversarial Examples." In: *Proc. of the 2018 Conf. on Empirical Methods in Natural Lang. Process.* Association for Computational Linguistics, 2018, pp. 2890–2896. DOI: 10.18653/v1/D18-1316.

[7] Jacopo Amidei, Paul Piwek, and Alistair Willis. "Agreement is overrated: A plea for correlation to assess human evaluation reliability." In: *Proc. of the 12th Int. Conf. on Natural Lang. Gener.* Association for Computational Linguistics, 2019, pp. 344–354. DOI: 10.18653/v1/W19-8642.

[8] Ron Artstein and Massimo Poesio. "Survey Article: Inter-Coder Agreement for Computational Linguistics." In: *Computational Linguistics* 34.4 (2008), pp. 555–596. DOI: 10.1162/coli.07-034-R2.

[9] Jill M. Austin and Linda D. Brown. "Internet plagiarism: Developing strategies to curb student academic dishonesty." In: *The Internet and Higher Education* 2.1 (1999), pp. 21–33. DOI: 10.1016/S1096-7516(99)00004-4.

[10] Ryan S. Baker, Antonija Mitrović, and Moffat Mathews. "Detecting Gaming the System in Constraint-Based Tutors." In: *Proc. of the 18th int. conf. on User Model., Adaptation, and Personalization*. Ed. by Paul De Bra, Alfred Kobsa, and David Chin. Springer Berlin Heidelberg, 2010, pp. 267–278.

[11] Ryan S. Baker, Jason Walonoski, Neil T. Heffernan, Ido Roll, Albert Corbett, and Kenneth R. Koedinger. "Why students engage in "gaming the system" behavior in interactive learning environments." In: *Journal of Interactive Learning Research* 19.2 (2008), pp. 185–224.

[12] Ryan S. Baker et al. "Adapting to When Students Game an Intelligent Tutoring System." In: *Intell. Tutoring Syst.* Ed. by Mitsuru Ikeda, Kevin D. Ashley, and Tak-Wai Chan. Springer Berlin Heidelberg, 2006, pp. 392–401.

[13] Satanjeev Banerjee and Alon Lavie. "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments." In: *Proc. of the ACL Workshop on Intrinsic and Extrinsic Eval. Measures for Mach. Transl. and/or Summarization*. Association for Computational Linguistics, 2005, pp. 65–72. URL: https://www.aclweb.org/anthology/W05-0909.

[14] Sami Baral, Anthony F. Botelho, John A. Erickson, Priyanka Benachamardi, and Neil T. Heffernan. "Improving Automated Scoring of Student Open Responses in Mathematics." In: *International Educational Data Mining Society* (2021).

[15] Sumit Basu, Chuck Jacobs, and Lucy Vanderwende. "Powergrading: a Clustering Approach to Amplify Human Effort for Short Answer Grading." In: *Transactions of the Association for Computational Linguistics* 1 (2013), pp. 391–402. DOI: 10.1162/tacl_a_00236.

[16] Melika Behjati, Seyed-Mohsen Moosavi-Dezfooli, Mahdieh S. Baghshah, and Pascal Frossard. "Universal Adversarial Attacks on Text Classifiers." In: *IEEE Int. Conf. on Acoust., Speech and Signal Process.* IEEE. 2019, pp. 7345–7349. DOI: 10.1109/ICASSP.2019.8682430.

[17] Yonatan Belinkov and Yonatan Bisk. "Synthetic and Natural Noise Both Break Neural Machine Translation." In: *Int. Conf. on Learn. Representations*. 2018. URL: https://openreview.net/forum?id=BJ8vJebC-.

[18] Emily M. Bender and Alexander Koller. "Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data." In: *Proc. of the 58th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2020, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.46.

[19] Steven Bird, Ewan Klein, and Edward Loper. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc., 2009. ISBN: 9780596516499.

[20] Sridevi Bonthu, Rama S. Sree, and Munaga H. M. Krishna Prasad. "Automated Short Answer Grading Using Deep Learning: A Survey." In: *Mach. Learn. and Knowl. Extraction*. Springer International Publishing, 2021, pp. 61–78. DOI: 10.1007/978-3-030-84060-0_5.

[21] Tom B. Brown et al. "Language Models Are Few-Shot Learners." In: *Proc. of the 34th Int. Conf. on Neural Inf. Process. Syst.* Vol. 33. Curran Associates Inc., 2020, pp. 1877–1901.

[22] Steven Burrows, Iryna Gurevych, and Benno Stein. "The eras and trends of automatic short answer grading." In: *International Journal of Artificial Intelligence in Education* 25.1 (2015), pp. 60–117. DOI: 10.1007/s40593-014-0026-8.

[23] Leon Camus and Anna Filighera. "Investigating Transformers for Automatic Short Answer Grading." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Educ. Part II*. Virtual Event: Springer International Publishing, 2020, pp. 43–48. DOI: 10.1007/978-3-030-52240-7_8.

[24] Nathan T. Carr. "Consistency of Computer-Automated Scoring Keys Across Authors and Authoring Teams." In: *Another Gener. of Fundam. Considerations in Lang. Assessment*. Springer, 2020, pp. 173–199. DOI: 10.1007/978-981-15-8952-2_11.

[25] Polona Caserman. "Full-Body Motion Tracking In Immersive Virtual Reality - Full-Body Motion Reconstruction and Recognition for Immersive Multiplayer Serious Games." PhD thesis. Dept. Elect. Eng., Technical Univ. of Darmstadt, Darmstadt, Germany, 2021. DOI: https://doi.org/10.26083/tuprints-00017572.

[26] Anirban Chakraborty, Manaar Alam, Vishal Dey, Anupam Chattopadhyay, and Debdeep Mukhopadhyay. "A survey on adversarial attacks and defences." In: *CAAI Transactions on Intelligence Technology* 6.1 (2021), pp. 25–45. DOI: https://doi.org/10.1049/cit2.12028.

[27] Charles Corbière, Nicolas Thome, Avner Bar-Hen, Matthieu Cord, and Patrick Pérez. "Addressing Failure Prediction by Learning Model Confidence." In: *Advances in Neural Inf. Process. Syst.* Ed. by Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Roman Garnett. Vol. 32. Curran Associates, Inc., 2019, pp. 2902–2913.

[28] Fred D. Davis. "A technology acceptance model for empirically testing new end-user information systems: Theory and results." PhD thesis. Sloan School of Management, Massachusetts Institute of Technology, Cambridge, MA, USA, 1985.

[29] *DeepL Free Translator*. Accessed: April 11, 2023. URL: https://www.deepl.com/translator.

[30] Galina Deeva, Daria Bogdanova, Estefanía Serral, Monique Snoeck, and Jochen De Weerdt. "A review of automated feedback systems for learners: Classification framework, challenges and opportunities." In: *Computers & Education* 162 (2021). DOI: 10.1016/j.compedu.2020.104094.

[31] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding." In: *Proc. of the 2019 Conf. of the North American Chapter of the Assoc. for Comput. Linguistics)*. Association for Computational Linguistics, 2019, pp. 4171–4186. DOI: 10.18653/v1/N19-1423.

[32] Berkeley J. Dietvorst, Joseph P. Simmons, and Cade Massey. "Algorithm aversion: people erroneously avoid algorithms after seeing them err." In: *Journal of Experimental Psychology* 144.1 (2015), pp. 114–126. DOI: 10.1037/xge0000033.

[33] Yuning Ding, Brian Riordan, Andrea Horbach, Aoife Cahill, and Torsten Zesch. "Don't take "nswvtnvakgxpm" for an answer --The surprising vulnerability of automatic content scoring systems to adversarial input." In: *Proc. of the 28th Int. Conf. on Comput. Linguistics*. International Committee on Computational Linguistics, 2020, pp. 882–892. DOI: 10.18653/v1/2020.coling-main.76.

[34] Dante D. Dixson and Frank C. Worrell. "Formative and Summative Assessment in the Classroom." In: *Theory Into Practice* 55.2 (2016), pp. 153–159. DOI: 10.1080/00405841.2016.1148989.

[35] Jens Doveren, Birte Heinemann, and Ulrik Schroeder. "Towards Guidelines for Data Protection and Privacy in Learning Analytics Implementation." In: *Online-Labs in Educ.* Nomos, 2022, pp. 45–52. DOI: 10.5771/9783957104106-45.

[36] Myroslava Dzikovska, Natalie Steinhauser, Elaine Farrow, Johanna Moore, and Gwendolyn Campbell. "BEETLE II: Deep natural language understanding and automatic feedback generation for intelligent tutoring in basic electricity and electronics." In: *International Journal of Artificial Intelligence in Education* 24.3 (2014), pp. 284–332. DOI: 10.1007/s40593-014-0017-9.

[37] Myroslava Dzikovska et al. "SemEval-2013 Task 7: The Joint Student Response Analysis and 8th Recognizing Textual Entailment Challenge." In: *2nd Joint Conf. on Lexical and Comput. Semantics, Proc. of the 7th Int. Workshop on Semantic Eval.* Association for Computational Linguistics, 2013, pp. 263–274. URL: https://aclanthology.org/S13-2045.

[38] Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. "HotFlip: White-Box Adversarial Examples for Text Classification." In: *Proc. of the 56th Annu. Meeting of the Assoc. for Comput. Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, 2018, pp. 31–36. DOI: 10.18653/v1/P18-2006.

[39] Steffen Eger and Yannik Benz. "From Hero to Zéroe: A Benchmark of Low-Level Adversarial Attacks." In: *Proc. of the 1st Conf. of the Asia-Pacific Chapter of the Assoc. for Comput. Linguistics and the 10th Int. Joint Conf. on Natural Lang. Process.* Association for Computational Linguistics, 2020, pp. 786–803. URL: https://aclanthology.org/2020.aacl-main.79.

[40] Mojisola H. Erdt. "Personalized Recommender Systems for Resource-based Learning - Hybrid Graph-based Recommender Systems for Folksonomies." PhD thesis. Dept. Elect. Eng., Technical Univ. of Darmstadt, Darmstadt, Germany, 2014. URL: http://tuprints.ulb.tu-darmstadt.de/4137/.

[41] John A. Erickson, Anthony F. Botelho, Steven McAteer, Ashvini Varatharaj, and Neil T. Heffernan. "The Automated Grading of Student Open Responses in Mathematics." In: *Proc. of the 10th Int. Conf. on Learn. Analytics and Knowl.* Association for Computing Machinery, 2020, pp. 615–624. DOI: 10.1145/3375462.3375523.

[42] Anna Filighera, Leonard Bongard, Tim Steuer, and Thomas Tregel. "Towards A Vocalization Feedback Pipeline for Language Learners." In: *Proc. 2022 Int. Conf. Advanced Learning Technologies*. Bucharest, Romania: IEEE Computer Society, 2022, pp. 248–252. DOI: 10.1109/ICALT55010.2022.00081.

[43] Anna Filighera, Sebastian Ochs, Tim Steuer, and Thomas Tregel. *Cheating Automatic Short Answer Grading: On the Adversarial Usage of Adjectives and Adverbs*. arXiv:2201.08318, preprint. Under review from International Journal of Artificial Intelligence in Education since 11/2021. 2022. DOI: 10.48550/arXiv.2201.08318.

[44] Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. "Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset." In: *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin, Ireland: Association for

Computational Linguistics, 2022, pp. 8577–8591. DOI: 10.18653/v1/2022.acl-long.587.

[45] Anna Filighera, Tim Steuer, and Christoph Rensing. "Fooling Automatic Short Answer Grading Systems." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Educ.* Virtual Event: Springer International Publishing, 2020, pp. 177–190. DOI: 10.1007/978-3-030-52237-7_15.

[46] Anna Filighera, Tim Steuer, and Christoph Rensing. "Fooling It - Student Attacks on Automatic Short Answer Grading." In: *Proc. 2020 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2020, pp. 347–352. DOI: 10.1007/978-3-030-57717-9_25.

[47] Anna Filighera, Joel N. Tschesche, Tim Steuer, Thomas Tregel, and Lisa Wernet. "Towards Generating Counterfactual Examples as Automatic Short Answer Feedback." In: *Proc. 2022 Int. Conf. Artificial Intelligence in Educ.* Durham, UK: Springer International Publishing, 2022, pp. 206–217. DOI: 10.1007/978-3-031-11644-5_17.

[48] Marieke S. Fischer. "Testing the Effects of a state-of-the-art Automatic Short Answer Grading System on Student Learning and Motivation." Master Thesis. TU Darmstadt, 2022.

[49] Nelson W. Francis and Henry Kucera. *Brown corpus manual*. Department of Linguistics Brown University. 1979. URL: http://korpus.uib.no/icame/brown/bcm.html.

[50] Lucas B. Galhardi and Jacques D. Brancher. "Machine Learning Approach for Automatic Short Answer Grading: A Systematic Review." In: *Ibero-Amer. Conf. on Artif. Intell.* Ed. by Guillermo R. Simari, Eduardo Fermé, Flabio G. Segura, and José A. R. Melquiades. Springer International Publishing, 2018, pp. 380–391. DOI: 10.1007/978-3-030-03928-8_31.

[51] Hang Gao and Tim Oates. *Universal Adversarial Perturbation for Text Classification*. arXiv:1910.04618, preprint. 2019. DOI: 10.48550/arXiv.1910.04618.

[52] Siddhant Garg and Goutham Ramakrishnan. "BAE: BERT-based Adversarial Examples for Text Classification." In: *Proc. of the 2020 Conf. on Empirical Methods in Natural Lang. Process.* Association for Computational Linguistics, 2020, pp. 6174–6181. DOI: 10.18653/v1/2020.emnlp-main.498.

[53] Hadi A. Ghavidel, Amal Zouaq, and Michel C. Desmarais. "Using BERT and XLNET for the Automatic Short Answer Grading Task." In: *Proc. of the 12th Int. Conf. on Comput. Supported Educ.* 2020, pp. 58–67. DOI: 10.5220/0009422400580067.

[54] Graham Gibbs and Claire Simpson. "Conditions under which assessment supports students' learning." In: *Learning and Teaching in Higher Education* 1 (2005), pp. 3–31.

[55] Sebastian Gombert et al. "Coding energy knowledge in constructed responses with explainable NLP models." In: *Journal of Computer Assisted Learning* (2022). DOI: 10.1111/jcal.12767.

[56] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples." In: *3rd Int. Conf. on Learn. Representations*. Ed. by Yoshua Bengio and Yann LeCun. 2015. DOI: 10.48550/arXiv.1412.6572.

[57] Michael E. Gordon and Charles H. Fay. "The Effects of Grading and Teaching Practices on Students' Perceptions of Grading Fairness." In: *College Teaching* 58.3 (2010), pp. 93–98. DOI: 10.1080/87567550903418586.

[58] *Grammarly Premium*. Accessed: May 11, 2023. URL: https://app.grammarly.com/.

[59] Stefan Haller, Adina Aldea, Christin Seifert, and Nicola Strisciuglio. *Survey on Automated Short Answer Grading with Deep Learning: from Word Embeddings to Transformers*. arXiv:2204.03503, preprint. 2022. DOI: 10.48550/arXiv.2204.03503.

[60] Michael Hanna and Ondřej Bojar. "A Fine-Grained Analysis of BERTScore." In: *Proc. of the 6th Conf. on Mach. Transl.* Association for Computational Linguistics, 2021, pp. 507–517. URL: https://aclanthology.org/2021.wmt-1.59.

[61] Muhammad A. Hasan, Nurul F. M. Noor, Siti S. A. Rahman, and Mohammad M. Rahman. "The Transition From Intelligent to Affective Tutoring System: A Review and Open Issues." In: *IEEE Access* 8 (2020), pp. 204612–204638. DOI: 10.1109/ACCESS.2020.3036990.

[62] Scott Hellman, William Murray, Adam Wiemerslage, Mark Rosenstein, Peter Foltz, Lee Becker, and Marcia Derr. "Multiple Instance Learning for Content Feedback Localization without Annotation." In: *Proc. of the 15th Workshop on Innovative Use of NLP for Building Educational Appl.* Association for Computational Linguistics, 2020, pp. 30–40. DOI: 10.18653/v1/2020.bea-1.3.

[63] Nathan M. Hicks and Heidi A. Diefes-Dux. "Grader Consistency in using Standards-based Rubrics." In: *2017 ASEE Annu. Conf. & Expo.* ASEE Conferences, 2017. DOI: 10.18260/1-2--28416.

[64] Andrea Horbach and Manfred Pinkal. "Semi-Supervised Clustering for Short Answer Scoring." In: *Proc. of the 11th Int. Conf. on Lang. Resour. and Eval.* European Language Resources Association (ELRA), 2018.

[65] David M. Howcroft et al. "Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definitions." In: *Proc. of the 13th Int. Conf. on Natural Lang. Gener.* Association for Computational Linguistics, 2020, pp. 169–182. URL: https://aclanthology.org/2020.inlg-1.23.

[66] Silas Hsu, Tiffany W. Li, Zhilin Zhang, Max Fowler, Craig Zilles, and Karrie Karahalios. "Attitudes Surrounding an Imperfect AI Autograder." In: *Proc. of the 2021 CHI Conf. on Human Factors in Comput. Syst.* Association for Computing Machinery, 2021. DOI: 10.1145/3411764.3445424.

[67] Xiaowei Huang et al. "A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability." In: *Computer Science Review* 37, 100270 (2020). DOI: 10.1016/j.cosrev.2020.100270.

[68] Petri Ihantola, Tuukka Ahoniemi, Ville Karavirta, and Otto Seppälä. "Review of recent systems for automatic assessment of programming assignments." In: *Proc. of the 10th Koli calling Int. Conf. on Comput. Educ. Res.* 2010, pp. 86–93. DOI: 10.1145/1930464.1930480.

[69] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. "Adversarial Examples Are Not Bugs, They Are Features." In: *Advances in Neural Inf. Process. Syst.* Ed. by Hanna Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Roman Garnett. Vol. 32. Curran Associates, Inc., 2019, pp. 125–136.

[70] Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. "Adversarial Example Generation with Syntactically Controlled Paraphrase Networks." In: *Proc. of the 2018 Conf. of the North American Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technologies*. Association for Computational Linguistics, 2018, pp. 1875–1885. DOI: 10.18653/v1/N18-1170.

[71] Daniel Jaramillo-Morillo, José Ruipérez Valiente, Mario F. Sarasty, and Gustavo Ramírez-Gonzalez. "Identifying and characterizing students suspected of academic dishonesty in SPOCs for credit through learning analytics." In: *International Journal of Educational Technology in Higher Education* 17.1, 45 (2020). DOI: 10.1186/s41239-020-00221-2.

[72] Di Jin, Zhijing Jin, Joey T. Zhou, and Peter Szolovits. "Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification and Entailment." In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34.5 (2020), pp. 8018–8025. DOI: 10.1609/aaai.v34i05.6311.

[73] Augustus E. Jordan. "College Student Cheating: The Role of Motivation, Perceived Norms, Attitudes, and Knowledge of Institutional Policy." In: *Ethics & Behavior* 11.3 (2001), pp. 233–247. DOI: 10.1207/S15327019EB1103_3.

[74] Ekaterina Jussupow, Izak Benbasat, and Armin Heinzl. "Why are we averse towards Algorithms? A comprehensive literature Review on Algorithm aversion." In: *28th Eur. Conf. on Inf. Syst.* Ed. by Frantz Rowe. AISeL, 2020. URL: https://madoc.bib.uni-mannheim.de/56152/.

[75] Sean H. K. Kang, Kathleen B. McDermott, and Henry L. Roediger III. "Test format and corrective feedback modify the effect of testing on long-term retention." In: *European Journal of Cognitive Psychology* 19.4–5 (2007), pp. 528–558. DOI: 10.1080/09541440601056620.

[76] Zixuan Ke and Vincent Ng. "Automated Essay Scoring: A Survey of the State of the Art." In: *Proc. of the 28th Int. Joint Conf. on Artif. Intell.* International Joint Conferences on Artificial Intelligence Organization, 2019, pp. 6300–6308. DOI: 10.24963/ijcai.2019/879.

[77] Hieke Keuning, Johan Jeuring, and Bastiaan Heeren. "A systematic literature review of automated feedback generation for programming exercises." In: *ACM Transactions on Computing Education* 19.1 (2018), pp. 1–43. DOI: 10.1145/3231711.

[78] Helen A. Klein, Nancy M. Levenburg, Marie McKendall, and William Mothersell. "Cheating During the College Years: How do Business School Students Compare?" In: *Journal of Business Ethics* 72.2 (2007), pp. 197–206. DOI: 10.1007/s10551-006-9165-7.

[79] Charlene Krueger and Lili Tian. "A Comparison of the General Linear Mixed Model and Repeated Measures ANOVA Using a Dataset with Multiple Missing Data Points." In: *Biological Research For Nursing* 6.2 (2004), pp. 151–157. DOI: 10.1177/1099800404267682.

[80] James A. Kulik and John D. Fletcher. "Effectiveness of intelligent tutoring systems: a meta-analytic review." In: *Review of Educational Research* 86.1 (2016), pp. 42–78. DOI: 10.3102/0034654315581420.

[81] Sachin Kumar, Soumen Chakrabarti, and Shourya Roy. "Earth Mover's Distance Pooling over Siamese LSTMs for Automatic Short Answer Grading." In: *Proc. of the 26th Int. Joint Conf. on Artif. Intell.* International Joint Conferences on Artificial Intelligence Organization, 2017, pp. 2046–2052. DOI: 10.24963/ijcai.2017/284.

[82] Daniël Lakens. "Equivalence tests: A practical primer for t tests, correlations, and meta-analyses." In: *Social Psychological & Personality Sci.* 8.4 (2017), pp. 355–362. DOI: 10.1177/1948550617697177.

[83] Mark M. Lanier. "Academic Integrity and Distance Learning." In: *Journal of Criminal Justice Educ.* 17.2 (2006), pp. 244–261. DOI: 10.1080/10511250600866166.

[84] Douglas P. Larsen. "Planning Education for Long-Term Retention: The Cognitive Science and Implementation of Retrieval Practice." In: *Seminars in Neurology*. Vol. 38. 4. Thieme Medical Publishers, 2018, pp. 449–456. DOI: 10.1055/s-0038-1666983.

[85] Claudia Leacock and Martin Chodorow. "C-rater: Automated scoring of short answer questions." In: *Computers and the Humanities* 37.4 (2003), pp. 389–405. DOI: 10.1023/A:1025779619903.

[86] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. "Deep learning." In: *Nature* 521.7553 (2015), pp. 436–444. DOI: 10.1038/nature14539.

[87] Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. "Bloom Library: Multimodal Datasets in 300+ Languages for a Variety of Downstream Tasks." In: *Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 8608–8621. URL: https://aclanthology.org/2022.emnlp-main.590.

[88] Mike Lewis et al. "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension." In: *Proc. of the 58th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2020, pp. 7871–7880. DOI: 10.18653/v1/2020.acl-main.703.

[89] Chin-Yew Lin. "ROUGE: A Package for Automatic Evaluation of Summaries." In: *Text Summarization Branches Out*. Association for Computational Linguistics, 2004, pp. 74–81. URL: https://aclanthology.org/W04-1013.

[90] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. "How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation." In: *Proc. of the 2016 Conf. on Empirical Methods in Natural Lang. Process.* Association for Computational Linguistics, 2016, pp. 2122–2132. DOI: 10.18653/v1/D16-1230.

[91] Hui Liu, Qingyu Yin, and William Y. Wang. "Towards Explainable NLP: A Generative Explanation Framework for Text Classification." In: *Proc. of the 57th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2019, pp. 5570–5581. DOI: 10.18653/v1/P19-1560.

[92] Yinhan Liu et al. "Multilingual Denoising Pre-training for Neural Machine Translation." In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 726–742. DOI: 10.1162/tacl_a_00343.

[93] Samuel A. Livingston. "Constructed-Response Test Questions: Why We Use Them; How We Score Them." In: *R&D Connections* 11 (2009). URL: https://eric.ed.gov/?id=ED507802.

[94] Ilya Loshchilov and Frank Hutter. "Decoupled Weight Decay Regularization." In: *Int. Conf. on Learn. Representations*. 2019. URL: https://openreview.net/forum?id=Bkg6RiCqY7.

[95] Xin Lu, Barbara Di Eugenio, Stellan Ohlsson, and Davide Fossati. "Simple but effective feedback generation to tutor abstract problem solving." In: *Proc. of the 5th Int. Natural Lang. Gener. Conf.* Association for Computational Linguistics, 2008, pp. 104–112. URL: https://www.aclweb.org/anthology/W08-1114.

[96] Steven G. Luke. "Evaluating significance in linear mixed-effects models in R." In: *Behavior Research Methods* 49.4 (2017), pp. 1494–1502. DOI: 10.3758/s13428-016-0809-y.

[97] Jiaqi Lun, Jia Zhu, Yong Tang, and Min Yang. "Multiple Data Augmentation Strategies for Improving Performance on Automatic Short Answer Scoring." In: *Proc. of the AAAI Conf. on Artificial Intelligence* 34.9 (2020), pp. 13389–13396. DOI: 10.1609/aaai.v34i09.7062.

[98] Gaoyan Lv, Wei Song, Miaomiao Cheng, and Lizhen Liu. "Exploring the Effectiveness of Question for Neural Short Answer Scoring System." In: *2021 IEEE 11th Int. Conf. on Electron. Inf. and Emergency Communication*. 2021, pp. 168–171. DOI: 10.1109/ICEIEC51955.2021.9463814.

[99] John M. Malouff and Einar B. Thorsteinsson. "Bias in grading: A meta-analysis of experimental research findings." In: *Australian Journal of Education* 60.3 (2016), pp. 245–256. DOI: 10.1177/0004944116664618.

[100] Christopher D. Manning. "Part-of-Speech Tagging from 97% to 100%: Is It Time for Some Linguistics?" In: *Comput. Linguistics and Intell. Text Process.* Ed. by Alexander F. Gelbukh. Springer Berlin Heidelberg, 2011, pp. 171–189. DOI: 10.1007/978-3-642-19400-9_14.

[101] Smit Marvaniya, Swarnadeep Saha, Tejas I. Dhamecha, Peter Foltz, Renuka Sindhgatta, and Bikram Sengupta. "Creating Scoring Rubric from Representative Student Answers for Improved Short Answer Grading." In: *Proc. of the 27th ACM Int. Conf. on Inf. and Knowl. Manage.* Association for Computing Machinery, 2018, pp. 993–1002. DOI: 10.1145/3269206.3271755.

[102] Roger C. Mayer, James H. Davis, and F. David Schoorman. "An Integrative Model of Organizational Trust." In: *The Academy of Management Review* 20.3 (1995), pp. 709–734. DOI: 10.2307/258792.

[103] Philipp Mayring. "Qualitative Inhaltsanalyse." In: *Handbuch Qualitative Forschung in der Psychologie*. VS Verlag für Sozialwissenschaften, 2010, pp. 601–613. DOI: 10.1007/978-3-531-92052-8_42.

[104] Patricia McGee. "Supporting academic honesty in online courses." In: *Journal of Educators Online* 10.1 (2013), pp. 1–31.

[105] Cecile H. McGrath, Benoit Guerin, Emma Harte, Michael Frearson, and Catriona Manville. *Learning gain in Higher Education*. RAND Corporation, 2015. DOI: 10.7249/RR996.

[106] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A Survey on Bias and Fairness in Machine Learning." In: *ACM Computing Surveys* 54.6 (2021). DOI: 10.1145/3457607.

[107] Christos D. Melas, Leonidas A. Zampetakis, Anastasia Dimopoulou, and Vassilis Moustakis. "Modeling the acceptance of clinical information systems among hospital medical staff: An extended TAM model." In: *Journal of Biomedical Informatics* 44.4 (2011), pp. 553–564. DOI: 10.1016/j.jbi.2011.01.009.

[108] Michael Mohler, Razvan Bunescu, and Rada Mihalcea. "Learning to Grade Short Answer Questions using Semantic Similarity Measures and Dependency Graph Alignments." In: *Proc. of the 49th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2011, pp. 752–762.

[109] John Morris, Eli Lifland, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi. "Reevaluating Adversarial Examples in Natural Language." In: *Findings of the Assoc. for Comput. Linguistics: EMNLP 2020*. Association for Computational Linguistics, 2020, pp. 3829–3839. DOI: 10.18653/v1/2020.findings-emnlp.341.

[110] Elham Mousavinasab, Nahid Zarifsanaiey, Sharareh R. Niakan Kalhori, Mahnaz Rakhshan, Leila Keikha, and Marjan G. Saeedi. "Intelligent tutoring systems: a systematic review of characteristics, applications, and evaluation methods." In: *Interactive Learning Environments* 29.1 (2018), pp. 1–22. DOI: 10.1080/10494820.2018.1558257.

[111] Pramod K. Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. "Did the Model Understand the Question?" In: *Proc. of the 56th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2018, pp. 1896–1906. DOI: 10.18653/v1/P18-1176.

[112] Kasia Muldner, Winslow Burleson, Brett Van de Sande, and Kurt VanLehn. "An Analysis of Gaming Behaviors in an Intelligent Tutoring System." In: *Int. Conf. on Intell. Tutoring Syst.* Ed. by Vincent Aleven, Judy Kay, and Jack Mostow. Springer Berlin Heidelberg, 2010, pp. 184–193. DOI: 10.1007/978-3-642-13388-6_23.

[113] Kasia Muldner, Winslow Burleson, Brett Van de Sande, and Kurt VanLehn. "An analysis of students' gaming behaviors in an intelligent tutoring system: Predictors and impacts." In: *User Modeling and User-Adapted Interaction* 21.1 (2011), pp. 99–135. DOI: 10.1007/s11257-010-9086-0.

[114] Daniel Mulnaes, Pegah Golchin, Filip Koenig, and Holger Gohlke. "Topdomain: exhaustive protein domain boundary metaprediction combining multisource information and deep learning." In: *Journal of chemical theory and computation* 17.7 (2021), pp. 4599–4613. DOI: 10.1021/acs.jctc.1c00129.

[115] Tamera B. Murdock and Eric M. Anderman. "Motivational Perspectives on Student Cheating: Toward an Integrated Model of Academic Dishonesty." In: *Educational Psychologist* 41.3 (2006), pp. 129–145. DOI: 10.1207/s15326985ep4103_1.

[116] Sharan Narang, Colin Raffel, Katherine Lee, Adam Roberts, Noah Fiedel, and Karishma Malkan. "WT5?! Training Text-to-Text Models to Explain their Predictions." In: *Computing Research Repository* arXiv:2004.14546 (2020). DOI: 10.48550/arXiv.2004.14546.

[117] Curtis G. Northcutt, Andrew D. Ho, and Isaac L. Chuang. "Detecting and preventing "multiple-account" cheating in massive open online courses." In: *Computers & Education* 100 (2016), pp. 71–80. DOI: 10.1016/j.compedu.2016.04.008.

[118] Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. "Why We Need New Evaluation Metrics for NLG." In: *Proc. of the 2017 Conf. on Empirical Methods in Natural Lang. Process.* Association for Computational Linguistics, 2017, pp. 2241–2252. DOI: 10.18653/v1/D17-1238.

[119] Julia Opgen-Rhein, Bastian Küppers, and Ulrik Schroeder. "An Application to Discover Cheating in Digital Exams." In: *Proc. of the 18th Koli Calling Int. Conf. on Comput. Educ. Res.* Association for Computing Machinery, 2018. DOI: 10.1145/3279720.3279740.

[120] Ulrike Pado and Cornelia Kiefer. "Short Answer Grading: When Sorting Helps and When it Doesn't." In: *Proc. of the 4th Workshop on NLP for Computer-assisted Lang. Learn.* LiU Electronic Press, 2015, pp. 42–50. URL: https://aclanthology.org/W15-1905.

[121] Chad Peters, Ivon Arroyo, Winslow Burleson, Beverly Woolf, and Kasia Muldner. "Predictors and Outcomes of Gaming in an Intelligent Tutoring System." In: *Int. Conf. on Intell. Tutoring Syst.* Ed. by Roger Nkambou, Roger Azevedo, and Julita Vassileva. Springer International Publishing, 2018, pp. 366–372. DOI: 10.1007/978-3-319-91464-0_41.

[122] Paul R. Pintrich, David A. F. Smith, Teresa Garcia, and Wilbert J. Mckeachie. "Reliability and Predictive Validity of the Motivated Strategies for Learning Questionnaire (Mslq)." In: *Educational and Psychological Measurement* 53.3 (1993), pp. 801–813. DOI: 10.1177/0013164493053003024.

[123] Matt Post. "A Call for Clarity in Reporting BLEU Scores." In: *Proc. of the 3rd Conf. on Mach. Transl.: Res. Papers*. Association for Computational Linguistics, 2018, pp. 186–191. DOI: 10.18653/v1/W18-6319.

[124] Colin Raffel et al. "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer." In: *Journal of Machine Learning Research* 21.140 (2020), pp. 1–67. URL: https://jmlr.org/papers/v21/20-074.html.

[125] Nils Reimers and Iryna Gurevych. "Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks." In: *Proc. of the 2019 Conf. on Empirical Methods in Natural Lang. Process. and the 9th Int. Joint Conf. on Natural Lang. Process.* Association for Computational Linguistics, 2019, pp. 3982–3992. DOI: 10.18653/v1/D19-1410.

[126] Ehud Reiter. "A Structured Review of the Validity of BLEU." In: *Computational Linguistics* 44.3 (2018), pp. 393–401. DOI: 10.1162/coli_a_00322.

[127] Shuhuai Ren, Yihe Deng, Kun He, and Wanxiang Che. "Generating Natural Language Adversarial Examples through Probability Weighted Word Saliency." In: *Proc. of the 57th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2019, pp. 1085–1097. DOI: 10.18653/v1/P19-1103.

[128] Yankun Ren, Jianbin Lin, Siliang Tang, Jun Zhou, Shuang Yang, Yuan Qi, and Xiang Ren. "Generating Natural Language Adversarial Examples on a Large Scale with Generative Models." In: *ECAI 2020*. IOS Press, 2020, pp. 2156–2163.

[129] Marco T. Ribeiro, Sameer Singh, and Carlos Guestrin. "Semantically Equivalent Adversarial Rules for Debugging NLP models." In: *Proc. of the 56th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2018, pp. 856–865. DOI: 10.18653/v1/P18-1079.

[130] Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chong M. Lee. "Investigating neural architectures for short answer scoring." In: *Proc. of the 12th Workshop on Innovative Use of NLP for Building Educational Appl.* Association for Computational Linguistics, 2017, pp. 159–168. DOI: 10.18653/v1/W17-5017.

[131] Alexis Ross, Ana Marasović, and Matthew Peters. "Explaining NLP Models via Minimal Contrastive Editing (MiCE)." In: *Findings of the Assoc. for Comput. Linguistics: ACL-IJCNLP 2021*. Association for Computational Linguistics, 2021, pp. 3840–3852. DOI: 10.18653/v1/2021.findings-acl.336.

[132] Neil C. Rowe. "Cheating in Online Student Assessment: Beyond Plagiarism." In: *Online Journal of Distance Learning Administration* 7.2 (2004).

[133] Christopher A. Rowland. "The effect of testing versus restudy on retention: a meta-analytic review of the testing effect." In: *Psychological Bulletin* 140.6 (2014), pp. 1432–1463. DOI: 10.1037/a0037559.

[134] Shourya Roy, Yadati Narahari, and Om D. Deshmukh. "A perspective on computer assisted assessment techniques for short free-text answers." In: *Int. Comput. Assisted Assessment Conf.* Springer. 2015, pp. 96–109. DOI: 10.1007/978-3-319-27704-2_10.

[135] José A. Ruipérez-Valiente, Giora Alexandron, Zhongzhou Chen, and David E. Pritchard. "Using Multiple Accounts for Harvesting Solutions in MOOCs." In: *Proc. of the 3rd (2016) ACM Conf. on Learn. @ Scale*. Association for Computing Machinery, 2016, pp. 63–70. DOI: 10.1145/2876034.2876037.

[136] Swarnadeep Saha, Tejas I. Dhamecha, Smit Marvaniya, Renuka Sindhgatta, and Bikram Sengupta. "Sentence Level or Token Level Features for Automatic Short Answer Grading?: Use Both." In: *Int. Conf. on Artif. Intell. in Educ.* Springer International Publishing, 2018, pp. 503–517. DOI: 10.1007/978-3-319-93843-1_37.

[137] Archana Sahu and Plaban K. Bhowmick. "Feature Engineering and Ensemble-Based Approach for Improving Automatic Short-Answer Grading Performance." In: *IEEE Transactions on Learning Technologies* 13.1 (2020), pp. 77–90. DOI: 10.1109/TLT.2019.2897997.

[138] Ananya B. Sai, Akash K. Mohankumar, and Mitesh M. Khapra. "A Survey of Evaluation Metrics Used for NLG Systems." In: *ACM Computing Surveys* 55.2 (2022). DOI: 10.1145/3485766.

[139] Holger Schielzeth et al. "Robustness of linear mixed-effects models to violations of distributional assumptions." In: *Methods in Ecology and Evolution* 11.9 (2020), pp. 1141–1152. DOI: 10.1111/2041-210X.13434.

[140] Sebastian Schmidt. "Informationsbeschaffung aus digitalen Textressourcen - Domänenadaptive Verfahren zur Strukturierung heterogener Textdokumente." PhD thesis. Dept. Elect. Eng., Technical Univ. of Darmstadt, Darmstadt, Germany, 2016. URL: http://tuprints.ulb.tu-darmstadt.de/5264/.

[141] Steffen Schnitzer. "Task Recommendation in Crowdsourcing Platforms." PhD thesis. Dept. Comput. Sci., Technical Univ. of Darmstadt, Darmstadt, Germany, 2019. URL: http://tubiblio.ulb.tu-darmstadt.de/112561/.

[142] Patrick Schober, Christa Boer, and Lothar A. Schwarte. "Correlation Coefficients: Appropriate Use and Interpretation." In: *Anesthesia & Analgesia* 126.5 (2018), pp. 1763–1768. DOI: 10.1213/ANE.0000000000002864.

[143] Noam Shazeer and Mitchell Stern. "Adafactor: Adaptive Learning Rates with Sublinear Memory Cost." In: *Proc. of the 35th Int. Conf. on Mach. Learn.* Ed. by Jennifer Dy and Andreas Krause. Vol. 80. 2018, pp. 4596–4604. URL: https://proceedings.mlr.press/v80/shazeer18a.html.

[144] Valerie J. Shute. "Focus on Formative Feedback." In: *Review of Educational Research* 78.1 (2008), pp. 153–189. DOI: 10.3102/0034654307313795.

[145] Liwei Song, Xinwei Yu, Hsuan-Tung Peng, and Karthik Narasimhan. "Universal Adversarial Attacks with Natural Triggers for Text Classification." In: *Proc. of the 2021 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2021, pp. 3724–3733. DOI: 10.18653/v1/2021.naacl-main.291.

[146] Daniel Starch and Edward C. Elliott. "Reliability of grading work in mathematics." In: *The School Review* 21.4 (1913), pp. 254–259. DOI: 10.1086/436086.

[147] Tim Steuer. "Automatic Question Generation to Support Reading Comprehension of Learners - Content Selection, Neural Question Generation, and Educational Evaluation." PhD thesis. Dept. Elect. Eng., Technical Univ. of Darmstadt, Darmstadt, Germany, 2023. DOI: 10.26083/tuprints-00023032.

[148] Md Arafat Sultan, Cristobal Salazar, and Tamara Sumner. "Fast and Easy Short Answer Grading with High Accuracy." In: *Proc. of the 2016 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technologies*. Association for Computational Linguistics, 2016, pp. 1070–1075. DOI: 10.18653/v1/N16-1123.

[149] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. "Axiomatic attribution for deep networks." In: *Proc. of the 34th Int. Conf. on Mach. Learn.* Ed. by Doina Precup and Yee W. Teh. Vol. 70. 2017, pp. 3319–3328.

[150] Chul Sung, Tejas I. Dhamecha, and Nirmal Mukhi. "Improving Short Answer Grading Using Transformer-Based Pre-training." In: *Int. Conf. on Artif. Intell. in Educ.* Springer. 2019, pp. 469–481. DOI: 10.1007/978-3-030-23204-7_39.

[151] Chuanqi Tan, Furu Wei, Wenhui Wang, Weifeng Lv, and Ming Zhou. "Multiway Attention Networks for Modeling Sentence Pairs." In: *Proc. of the 27th Int. Joint Conf. on Artif. Intell.* 2018, pp. 4411–4417. DOI: 10.24963/ijcai.2018/613.

[152] Hongye Tan, Chong Wang, Qinglong Duan, Yu Lu, Hu Zhang, and Ru Li. "Automatic short answer grading by encoding student responses via a graph convolutional network." In: *Interactive Learning Environments* (2020), pp. 1–15. DOI: 10.1080/10494820.2020.1855207.

[153] Mohsen Tavakol and Reg Dennick. "Making sense of Cronbach's alpha." In: *International Journal of Medical Education* 2 (2011), pp. 53–55. DOI: 10.5116/ijme.4dfb.8dfd.

[154] Stefano Teso and Kristian Kersting. "Explanatory Interactive Machine Learning." In: *Proc. of the 2019 AAAI/ACM Conf. on AI, Ethics, and Society*. Association for Computing Machinery, 2019, pp. 239–245. DOI: 10.1145/3306618.3314293.

[155] Masaki Uto and Yuto Uchida. "Automated Short-Answer Grading Using Deep Neural Networks and Item Response Theory." In: *Artif. Intell. in Educ.* Ed. by Ig Ibert Bittencourt, Mutlu Cukurova, Kasia Muldner, Rose Luckin, and Eva Millán. Springer International Publishing, 2020, pp. 334–339. DOI: 10.1007/978-3-030-52240-7_61.

[156] Chris Van der Lee, Albert Gatt, Emiel Van Miltenburg, Sander Wubben, and Emiel Krahmer. "Best practices for the human evaluation of automatically generated text." In: *Proc. of the 12th Int. Conf. on Natural Lang. Gener.* Association for Computational Linguistics, 2019, pp. 355–368. DOI: 10.18653/v1/W19-8643.

[157] Kurt VanLehn. "The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems." In: *Educational Psychologist* 46.4 (2011), pp. 197–221. DOI: 10.1080/00461520.2011.611369.

[158] Ashish Vaswani et al. "Attention is All you Need." In: *Advances in Neural Inf. Process. Syst.* Vol. 30. Curran Associates, Inc., 2017, pp. 5998–6008.

[159] Sahil Verma, Varich Boonsanong, Minh Hoang, Keegan E. Hines, John P. Dickerson, and Chirag Shah. *Counterfactual Explanations and Algorithmic Recourses for Machine Learning: A Review*. arXiv:2010.10596, preprint. 2020. DOI: 10.48550/arXiv.2010.10596.

[160] Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. "Universal Adversarial Triggers for Attacking and Analyzing NLP." In: *Proc. of the 2019 Conf. on Empirical Methods in Natural Lang. Process. and the 9th Int. Joint Conf. on Natural Lang. Process.* Association for Computational Linguistics, 2019, pp. 2153–2162. DOI: 10.18653/v1/D19-1221.

[161] Jason A. Walonoski and Neil T. Heffernan. "Detection and Analysis of Off-Task Gaming Behavior in Intelligent Tutoring Systems." In: *Int. Conf. on Intell. Tutoring Syst.* Ed. by Mitsuru Ikeda, Kevin D. Ashley, and Tak-Wai Chan. Springer Berlin Heidelberg, 2006, pp. 382–391. DOI: 10.1007/11774303_38.

[162] Jason A. Walonoski and Neil T. Heffernan. "Prevention of Off-Task Gaming Behavior in Intelligent Tutoring Systems." In: *Int. Conf. on Intell. Tutoring Syst.* Ed. by Mitsuru Ikeda, Kevin D. Ashley, and Tak-Wai Chan. Springer Berlin Heidelberg, 2006, pp. 722–724. DOI: 10.1007/11774303_80.

[163] Chengwen Wang, Qingxiu Dong, Xiaochen Wang, Haitao Wang, and Zhifang Sui. *Statistical Dataset Evaluation: Reliability, Difficulty, and Validity*. arXiv:2212.-09272, preprint. 2022. DOI: 10.48550/arXiv.2212.09272.

[164] George Watson and James Sottile. "Cheating in the Digital Age: Do Students Cheat More in Online Courses?" In: *Online Journal of Distance Learning Administration* 13.1 (2010).

[165] Stefan Wellek. *Testing statistical hypotheses of equivalence and noninferiority*. Chapman and Hall/CRC, 2010. DOI: 10.1201/EBK1439808184.

[166] Bernard E. Whitley. "Factors associated with cheating among college students: A review." In: *Research in Higher Education* 39.3 (1998), pp. 235–274. DOI: 10.1023/A:1018724900565.

[167] Adina Williams, Nikita Nangia, and Samuel R. Bowman. "A broad-coverage challenge corpus for sentence understanding through inference." In: *2018 Conf. of the North Amer. Chapter of the Assoc. for Comput. Linguistics: Human Lang. Technologies*. Association for Computational Linguistics. 2018, pp. 1112–1122. DOI: 10.18653/v1/N18-1101.

[168] Alistair Willis. "Using NLP to Support Scalable Assessment of Short Free Text Responses." In: *Proc. of the 10th Workshop on Innovative Use of NLP for Building Educational Appl.* Association for Computational Linguistics, 2015, pp. 243–253. DOI: 10.3115/v1/W15-0628.

[169] Naomi E. Winstone, Robert A. Nash, Michael Parker, and James Rowntree. "Supporting Learners' Agentic Engagement With Feedback: A Systematic Review and a Taxonomy of Recipience Processes." In: *Educational Psychologist* 52.1 (2017), pp. 17–37. DOI: 10.1080/00461520.2016.1207538.

[170] Benedikt Wisniewski, Klaus Zierer, and John Hattie. "The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research." In: *Frontiers in Psychology* 10 (2020). DOI: 10.3389/fpsyg.2019.03087.

[171] Benedikt Wisniewski, Klaus Zierer, and John Hattie. "The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research." In: *Frontiers in Psychology* 10 (2020). DOI: 10.3389/fpsyg.2019.03087.

[172] Thomas Wolf et al. *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. arXiv:1910.03771, preprint. 2019. DOI: 10.48550/arXiv.1910.03771.

[173] Tongshuang Wu, Marco T. Ribeiro, Jeffrey Heer, and Daniel Weld. "Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models." In: *Proc. of the 59th Annu. Meeting of the Assoc. for Comput. Linguistics and the 11th Int. Joint Conf. on Natural Lang. Process.* Association for Computational Linguistics, 2021, pp. 6707–6723. DOI: 10.18653/v1/2021.acl-long.523.

[174] Xiaoming Xi. "Automated scoring and feedback systems: Where are we and where are we heading?" In: *Language Testing* 27.3 (2010), pp. 291–300. DOI: 10.1177/0265532210364643.

[175] Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K. Jain. "Adversarial Attacks and Defenses in Images, Graphs and Text: A Review." In: *International Journal of Automation and Computing* 17.2 (2020), pp. 151–178. DOI: 10.1007/s11633-019-1211-x.

[176] Linting Xue et al. "mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer." In: *Proc. of the 2021 Conf. of the North Amer. Chap. of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2021, pp. 483–498. DOI: 10.18653/v1/2021.naacl-main.41.

[177] Xiaoyong Yuan, Pan He, Qile Zhu, and Xiaolin Li. "Adversarial Examples: Attacks and Defenses for Deep Learning." In: *IEEE Transactions on Neural Networks and Learning Systems* 30.9 (2019), pp. 2805–2824. DOI: 10.1109/TNNLS.2018.2886017.

[178] Farnaz V. M. Zanjani and Morteza Ramazani. "Investigation of e-learning acceptance in teaching English language based on TAM Model." In: *ARPN Journal of Systems and Software* 2.11 (2012). DOI: 10.2139/ssrn.2197912.

[179] Fabian Zehner, Christine Sälzer, and Frank Goldhammer. "Automatic Coding of Short Text Responses via Clustering in Educational Assessment." In: *Educational and Psychological Measurement* 76.2 (2016), pp. 280–303. DOI: 10.1177/0013164415590022.

[180] Guoyang Zeng et al. "OpenAttack: An Open-source Textual Adversarial Attack Toolkit." In: *Proc. of the 59th Annu. Meeting of the Assoc. for Comput. Linguistics and the 11th Int. Joint Conf. on Natural Lang. Process.: System Demonstrations*. Association for Computational Linguistics, 2021, pp. 363–371. DOI: 10.18653/v1/2021.acl-demo.43.

[181] Torsten Zesch, Michael Heilman, and Aoife Cahill. "Reducing Annotation Efforts in Supervised Short Answer Scoring." In: *Proc. of the 10th Workshop on Innovative Use of NLP for Building Educational Appl.* Association for Computational Linguistics, 2015, pp. 124–132. DOI: 10.3115/v1/W15-0615.

[182] Huangzhao Zhang, Hao Zhou, Ning Miao, and Lei Li. "Generating Fluent Adversarial Examples for Natural Languages." In: *Proc. of the 57th Annu. Meeting of the Assoc. for Comput. Linguistics*. Association for Computational Linguistics, 2019, pp. 5564–5569. DOI: 10.18653/v1/P19-1559.

[183] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. "BERTScore: Evaluating Text Generation with BERT." In: *Int. Conf. on Learn. Representations*. 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[184] Wei E. Zhang, Quan Z. Sheng, Ahoud Alhazmi, and Chenliang Li. "Adversarial Attacks on Deep-Learning Models in Natural Language Processing: A Survey." In: *ACM Trans. Intell. Syst. Technol.* 11.3 (2020). DOI: 10.1145/3374217.

[185] Yuan Zhang, Chen Lin, and Min Chi. "Going deeper: Automatic short-answer grading by combining student and question models." In: *User Modeling and User-Adapted Interaction* 30.1 (2020), pp. 51–80. DOI: 10.1007/s11257-019-09251-6.

All web pages cited in this work have been checked in May 2023. However, due to the dynamic nature of the World Wide Web, their long-term availability cannot be guaranteed.

A

APPENDIX

A.1 LIST OF ACRONYMS

ASAG Automatic Short Answer Grading

LMMs Linear Mixed Models

MiCE Minimal Contrastive Editing

MNLI Multi-Genre Natural Language Inference

MRPC Microsoft Research Paraphrase Corpus

NLP Natural Language Processing

RMSE root-mean-square error

RTE Recognizing Textual Entailment

SAF Short Answer Feedback Corpus

SEB SciEntsBank

A.2 UNIVERSAL TRIGGER ATTACK HYPERPARAMETERS & TRIGGERS

Table 33 contains the lower and upper bounds of the hyperparameters we manually tuned for the Universal Trigger Attack.Table 34 shows a list of all triggers we deployed on the test set and the number of successful class flips they induced.

Hyperparameter	Lowest Value	Highest Value
Batch Size	16	32
Number of Trigger Tokens	1	4
Number of Epochs	10	10
Beam Size	1	4
Number of Candidates	40	100

Table 33: Hyperparameter search range for the universal trigger attack.

Trigger	Flips UA	Flips UQ	Flips UD
No Trigger	31	101	491
none exits	48	133	756
none varies	53	134	687
none affected	42	125	707
none affect	42	125	614
none nearest	39	108	575
none electricity	34	73	443
none save	44	106	622
[MASK] exits	44	104	638
none heat	37	94	523
none untouched	36	83	540
none everywhere	49	100	728
none else	53	136	818
else none	54	112	742
nowhere changes	51	135	640
therefore insignificant	46	129	744
nowhere insignificant	49	128	728
equally nothing	45	111	669
anywhere.	45	105	682
nothing signals	47	134	590
none being	46	112	675
yourselves what	3	4	14
itunes ¨	1	2	9
nobody penetrated	43	121	673
none elsewhere	50	121	826
nowhere were	45	127	652
none would	53	138	810
##ired unaffected least being	9	35	148
neither prevents	45	116	562
##oons affected	12	37	186
electricity drops	37	119	487
heats affected penetrated	42	106	739
with none	47	116	679

Table 34: List of all triggers evaluated on the test set.

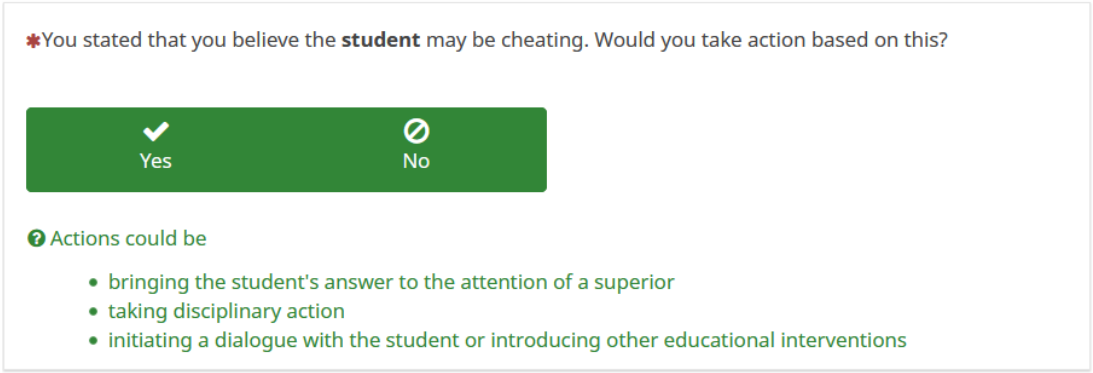


Figure 15: Screenshot of the conditional question whether the annotator would act on their suspicion. Taken from [2].

A.3 QUESTIONNAIRES

Figures 15 and 16 are screenshots of the questionnaire used in the expert evaluation of the adversarial examples generated by the adjective and adverb insertion attack. Figures 17, 18, 19, 20, 21 and 22 are screenshots of the questionnaire utilized in the field study of the supervised elaborated feedback generation. This questionnaire is the variant presented to the students in Groups Human and AI at measuring point T2. The questionnaires for T1 and Group No-Feedback differ only slightly. For example, Group No-Feedback had not received feedback at T2; thus, they were not asked to imagine that a human wrote the feedback they had received, et cetera.

A.4 EXAMPLES OF GENERATED FEEDBACK

Tables 35 and 36 contain example predictions generated by the T5 models without questions. The examples stem from the English unseen answers test split. Since those are the most interesting feedback cases, they were selected to be as brief as possible while predicting the *partially correct* class or a matching score.

A.5 LINEAR MIXED MODELS' STATISTICS

Figures 23, 24, 25 and 26 display the full model statistics for all Linear Mixed Models utilized to estimate the effects observed in the field study. We selected Group AI at T1 as the reference point for the test comparisons, meaning that the fixed effect intercept is the mean of Group AI at T1 for a given variable. Each effect estimate is, thus, the change from the reference group in the given factor. For example, the estimate for Group Human is the difference between Group AI and Group Human at T1. The estimate for T2 captures how the variable changed from T1 to T2 in Group AI. Interactions, such as "Human x T2", indicate how the variable time affected the dependent variable in Group Human compared to Group AI.

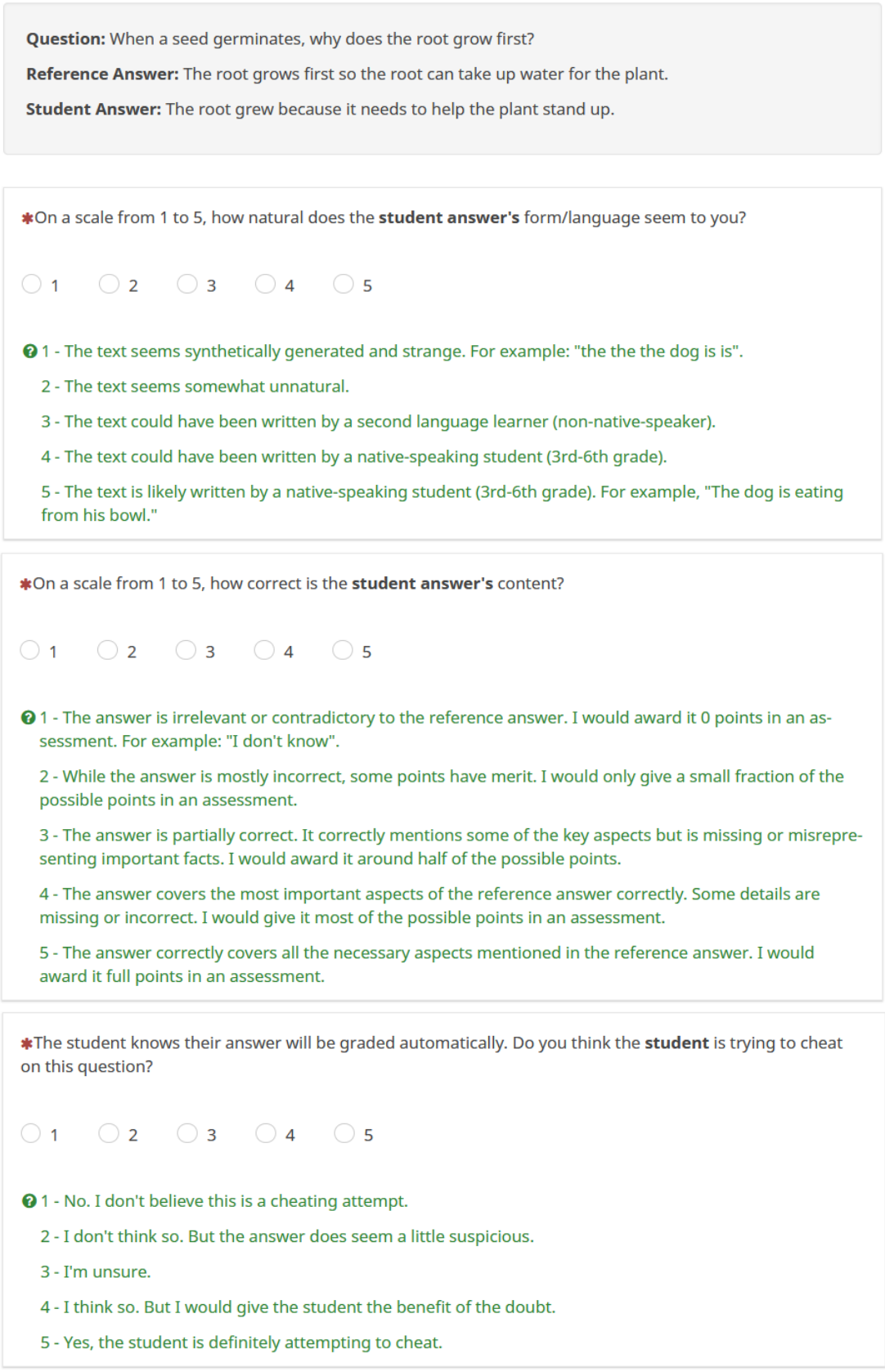
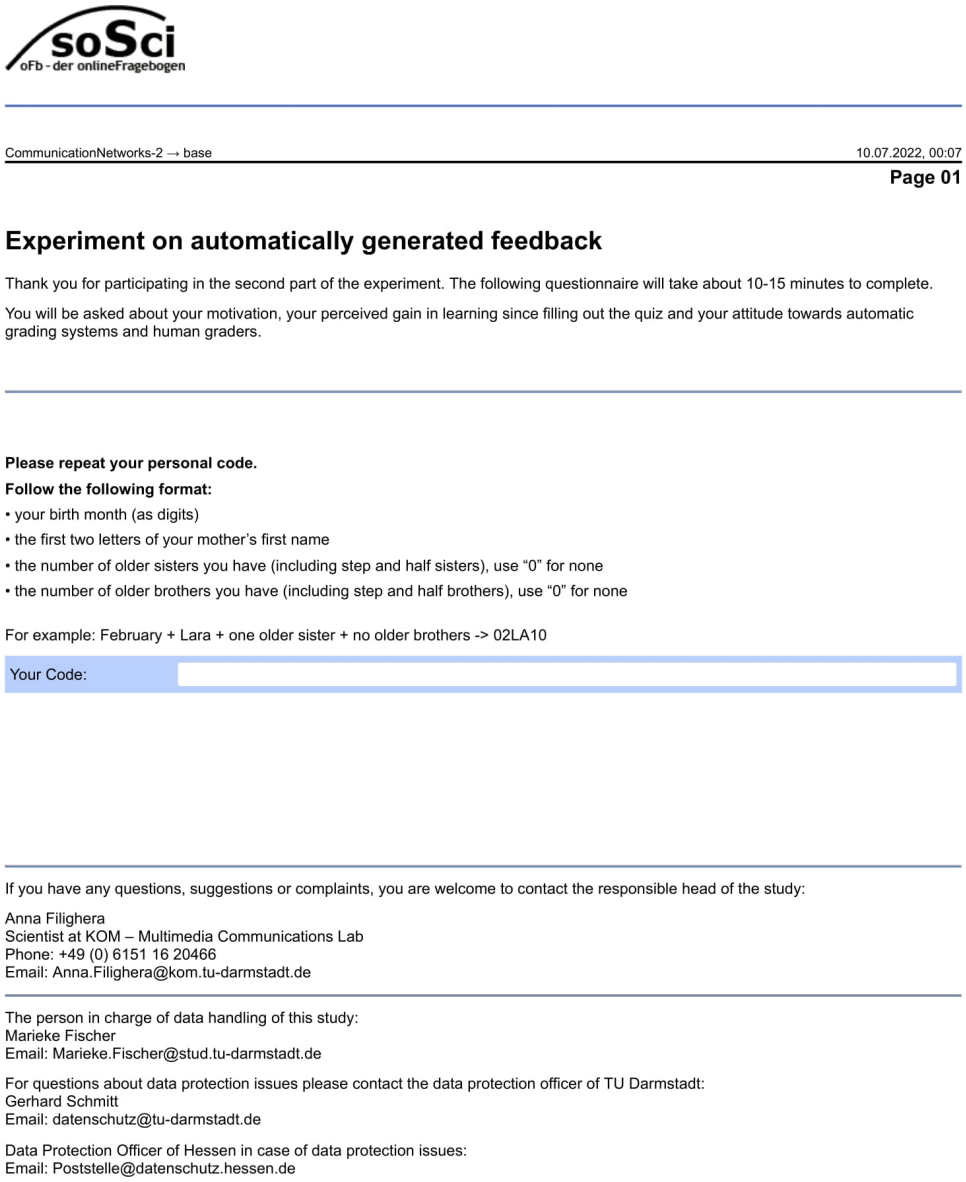


Figure 16: Screenshot of survey questions posed in the human evaluation of the adjective and adverb insertion attack. Taken from [2].



soSci
oFb - der onlineFragebogen

CommunicationNetworks-2 → base 10.07.2022, 00:07

Page 01

Experiment on automatically generated feedback

Thank you for participating in the second part of the experiment. The following questionnaire will take about 10-15 minutes to complete.

You will be asked about your motivation, your perceived gain in learning since filling out the quiz and your attitude towards automatic grading systems and human graders.

Please repeat your personal code.

Follow the following format:

• your birth month (as digits)

• the first two letters of your mother's first name

• the number of older sisters you have (including step and half sisters), use "0" for none

• the number of older brothers you have (including step and half brothers), use "0" for none

For example: February + Lara + one older sister + no older brothers -> 02LA10

Your Code:

If you have any questions, suggestions or complaints, you are welcome to contact the responsible head of the study:

Anna Filighera
Scientist at KOM – Multimedia Communications Lab
Phone: +49 (0) 6151 16 20466
Email: Anna.Filighera@kom.tu-darmstadt.de

The person in charge of data handling of this study:
Marieke Fischer
Email: Marieke.Fischer@stud.tu-darmstadt.de

For questions about data protection issues please contact the data protection officer of TU Darmstadt:
Gerhard Schmitt
Email: datenschutz@tu-darmstadt.de

Data Protection Officer of Hessen in case of data protection issues:
Email: Poststelle@datenschutz.hessen.de

Figure 17: Screenshot of questionnaire at T2. Students in Group No-Feedback had received a slightly modified version of the questionnaire to account for their lack of feedback. Taken from [4].

Page 02

The following questions ask about your **motivation** for and **attitudes** about the **online bonus quizzes**.

Remember there are no right or wrong answers, just answer as accurately as possible. If you think the statement is very true of you, select 7; if a statement is not at all true of you, select 1. If the statement is more or less true of you, find the number between 1 and 7 that best describes you.

	not at all true of me 1	2	3	4	5	6	very true of me 7
I am very interested in the content area covered in these quizzes.	○	○	○	○	○	○	○
I like the subject matter covered in these quizzes.	○	○	○	○	○	○	○
Considering my skills and the difficulty of the questions, I think I will do well in these quizzes.	○	○	○	○	○	○	○
In a quiz like this, I prefer questions that really challenge me so I can learn new things.	○	○	○	○	○	○	○
When I fill out a quiz I think about items on future quizzes I might not be able to answer.	○	○	○	○	○	○	○
At the end of the semester I expect to have achieved most of the possible points in the quizzes.	○	○	○	○	○	○	○
The most satisfying thing for me in these quizzes is trying to understand the content as thoroughly as possible.	○	○	○	○	○	○	○
I have an uneasy, upset feeling when I fill out a quiz.	○	○	○	○	○	○	○
It is important for me to learn the material covered in these quizzes.	○	○	○	○	○	○	○
Understanding the subject matter covered in these quizzes is very important to me.	○	○	○	○	○	○	○
If I try hard enough, then I will understand the material covered in these quizzes.	○	○	○	○	○	○	○
In a quiz like this, I prefer questions that arouse my curiosity, even if it is difficult to learn.	○	○	○	○	○	○	○
I think the material covered in these quizzes is useful for me to learn.	○	○	○	○	○	○	○
When I fill out a quiz I think of the consequences of failing.	○	○	○	○	○	○	○
I expect to do well in these quizzes.	○	○	○	○	○	○	○
I'm confident I can understand the most complex material presented in these quizzes.	○	○	○	○	○	○	○
I'm confident I can do an excellent job in the quizzes, the exercises and the exam.	○	○	○	○	○	○	○
When I fill out a quiz I think about how poorly I am doing compared with other students.	○	○	○	○	○	○	○
I'm confident I can understand the basic concepts covered in these quizzes.	○	○	○	○	○	○	○
I think I will be able to use what I learn in these quizzes in other courses.	○	○	○	○	○	○	○
I'm certain I can master the skills needed for these quizzes.	○	○	○	○	○	○	○
If I don't understand the material covered in these quizzes, it is because I didn't try hard enough.	○	○	○	○	○	○	○
I worry about what the quizzes' grader will think about me.	○	○	○	○	○	○	○
If I study in appropriate ways, then I will be able to answer the questions that are asked in these quizzes.	○	○	○	○	○	○	○
It is my own fault if I don't learn the material covered in these quizzes.	○	○	○	○	○	○	○
I would do these quizzes even if there wasn't the possibility to earn a bonus for the exam from them.	○	○	○	○	○	○	○

Figure 18: Screenshot of questionnaire at T2. Taken from [4].

Page 03

The following questions ask about your **knowledge** and **understanding** of the topics covered in quiz #2 at different points in time.

Think back to **right after you had finished quiz #2** and try to estimate the competency you had at that time. If the exam would have been the next day, how well would you have been able to do these things?

Remember there are no right or wrong answers, just answer as accurately as possible. Please indicate whether you agree or disagree with the statements.

	Strongly disagree	Disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Agree	Strongly Agree
	1	2	3	4	5	6	7
At that time, I could explain which encoding technique (for encoding bitstreams) is suitable when given a specific scenario.	○	○	○	○	○	○	○
At that time, I could explain the difference between asynchronous and synchronous transmission mode in the Data Link Layer.	○	○	○	○	○	○	○
At that time, I knew what requirement has to be met so that you can use the piggybacking extension to the sliding window protocol.	○	○	○	○	○	○	○
At that time, I could evaluate which MAC procedure might be suitable given a specific scenario.	○	○	○	○	○	○	○
At that time, I could explain what "frame bursting" is and knew of its advantages and disadvantages.	○	○	○	○	○	○	○

How would rate your competency **right now**? If the exam would be today, how well would you be able to to do these things?

Remember there are no right or wrong answers, just answer as accurately as possible. Please indicate whether you agree or disagree with the statements.

	Strongly disagree	Disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Agree	Strongly Agree
	1	2	3	4	5	6	7
At the present moment, I can explain which encoding technique (for encoding bitstreams) is suitable when given a specific scenario.	○	○	○	○	○	○	○
At the present moment, I can explain the difference between asynchronous and synchronous transmission mode in the Data Link Layer.	○	○	○	○	○	○	○
At the present moment, I know what requirement has to be met so that you can use the piggybacking extension to the sliding window protocol.	○	○	○	○	○	○	○
At the present moment, I can evaluate which MAC procedure might be suitable given a specific scenario.	○	○	○	○	○	○	○
At the present moment, I can explain what "frame bursting" is and know of its advantages and disadvantages.	○	○	○	○	○	○	○

Figure 19: Screenshot of questionnaire at T2. Taken from [4].

Page 04

The following questions ask about the **attitudes, beliefs** and **expectations** you would have under hypothetic circumstances.

First, please **imagine** the feedback you received was generated by an **automatic grading system**. What would you think?

Remember there are no right or wrong answers, just answer as accurately as possible. Please indicate whether you would agree or disagree with the statements.

	Strongly disagree	Disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Agree	Strongly Agree
	1	2	3	4	5	6	7
Getting automatically generated feedback helps to improve my learning performance.	○	○	○	○	○	○	○
I believe the automatic grading system can be easily tricked to accept an answer as correct even though the student didn't actually understand the material.	○	○	○	○	○	○	○
I believe the automatically generated feedback reflects the correctness of my answer and nothing else.	○	○	○	○	○	○	○
I find it diffcult to answer in a way that can be understood by the system.	○	○	○	○	○	○	○
I find it easy to use the feedback provided by the automatic grading system to improve my learning.	○	○	○	○	○	○	○
Receiving feedback from the system makes learning easier.	○	○	○	○	○	○	○
I believe the automatically generated feedback is accurate.	○	○	○	○	○	○	○
Receiving automatically generated feedback is useful to me.	○	○	○	○	○	○	○
I believe my answers would be fairly reevaluated if I were to appeal the automatically generated feedback.	○	○	○	○	○	○	○
The automatic grading system was created to help students.	○	○	○	○	○	○	○
The automatically generated feedback is clear and understandable.	○	○	○	○	○	○	○
I trust the grading is private between me and the autograder.	○	○	○	○	○	○	○

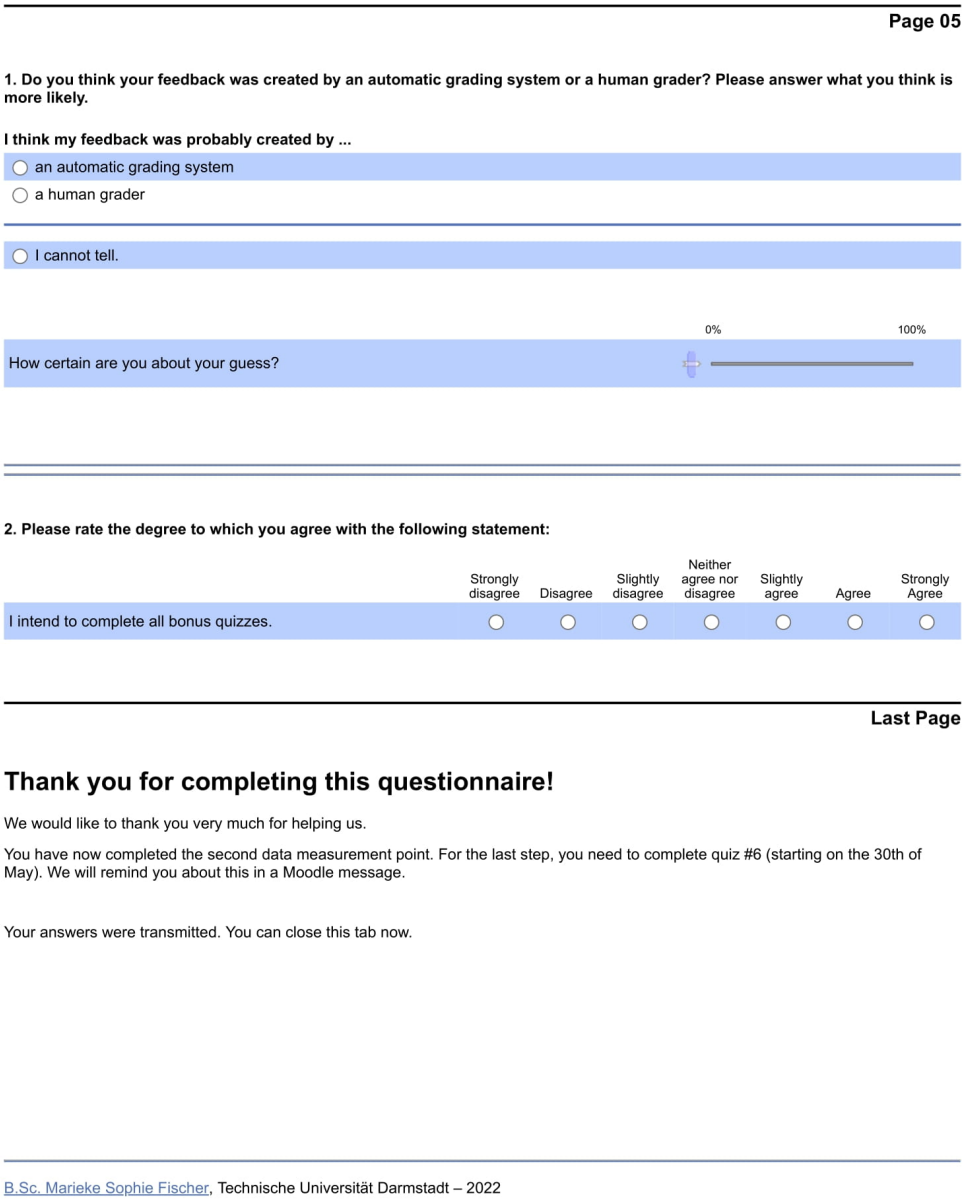
Figure 20: Screenshot of questionnaire at T2. Taken from [4].

Second, please **imagine** the feedback you received was written by a **human grader**. What would you think?

Remember there are no right or wrong answers, just answer as accurately as possible. Please indicate whether you would agree or disagree with the statements.

	Strongly disagree	Disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Agree	Strongly Agree
	1	2	3	4	5	6	7
I find it diffcult to answer in a way that can be understood by the human grader.	○	○	○	○	○	○	○
I believe my answers would be fairly reevaluated if I were to appeal my received feedback.	○	○	○	○	○	○	○
The feedback from the human grader is clear and understandable.	○	○	○	○	○	○	○
I find it easy to use the feedback provided by the human grader to improve my learning.	○	○	○	○	○	○	○
I believe the feedback provided by the human grader reflects the correctness of my answer and nothing else.	○	○	○	○	○	○	○
The feedback system was created to help students.	○	○	○	○	○	○	○
I believe the human grader can be easily tricked to accept an answer as correct even though the student didn't actually understand the material.	○	○	○	○	○	○	○
Receiving feedback from a human grader is useful to me.	○	○	○	○	○	○	○
I trust the grading is private between me and the human grader.	○	○	○	○	○	○	○
I believe the feedback provided by the human grader is accurate.	○	○	○	○	○	○	○
Getting feedbacked from a human grader helps to improve my learning performance.	○	○	○	○	○	○	○
Receiving feedback from a human grader makes learning easier.	○	○	○	○	○	○	○
In this line, please mark option one to show that you have read this sentence.	○	○	○	○	○	○	○

Figure 21: Screenshot of questionnaire at T2. Taken from [4].



Page 05

1. Do you think your feedback was created by an automatic grading system or a human grader? Please answer what you think is more likely.

I think my feedback was probably created by ...

- an automatic grading system
- a human grader
- I cannot tell.

How certain are you about your guess? 0% 100%

2. Please rate the degree to which you agree with the following statement:

	Strongly disagree	Disagree	Slightly disagree	Neither agree nor disagree	Slightly agree	Agree	Strongly Agree
I intend to complete all bonus quizzes.	○	○	○	○	○	○	○

Last Page

Thank you for completing this questionnaire!

We would like to thank you very much for helping us.

You have now completed the second data measurement point. For the last step, you need to complete quiz #6 (starting on the 30th of May). We will remind you about this in a Moodle message.

Your answers were transmitted. You can close this tab now.

B.Sc. Marieke Sophie Fischer, Technische Universität Darmstadt – 2022

Figure 22: Screenshot of questionnaire at T2. Taken from [4].

Effect	Self-Reported Learning Gain			Objective Learning Gain		
Fixed Effects	Est. (SE)	t (df)	p	Est. (SE)	t (df)	p
Intercept (≙ Group AI at T1)	5.15* (0.28)	18.06 (94.85)	<.01	2.96 (0.22)	13.58* (128.88)	<.01
Group:						
H	0.16 (0.39)	0.41 (94.85)	.69	0.20 (0.31)	0.66 (128.57)	.51
no-F	-0.35 (0.40)	-0.87 (94.85)	.39	-0.62* (0.31)	-2.00 (128.58)	<.05
Time:						
T2	0.05 (0.24)	0.19 (67.00)	.85	-0.34 (0.27)	-1.23 (66.75)	.22
Group × Time:						
H × T2	-0.46 (0.33)	-1.40 (67.00)	.17	-0.60 (0.38)	-1.56 (66.04)	.12
no-F × T2	-0.38 (0.24)	-1.14 (67.00)	.26	-0.16 (0.39)	-0.41 (66.06)	.68
Covariance Parameters		Variance			Variance	
Intercept		1.15			0.20	
Residual		0.64			0.85	
Fit Statistics						
AIC		457.17			413.60	
BIC		480.70			437.02	
Marginal R^2		0.04			0.16	
Condit. R^2		0.66			0.32	

Note. The Intercept of the fixed effects represents the Group AI at T1. The Intercept of the covariance parameters represents between participant variance. N = 70 for self-reported learning gain and N = 71 for objective learning gain. * ≙ $p < .05$.

Figure 23: Estimated effects on learning gain. Taken from [4].

Effect	Intrinsic Goal Orientation			Task Value		
Fixed Effects	Est. (SE)	t (df)	p	Est. (SE)	t (df)	p
Intercept (≙ Group AI at T1)	4.90* (0.23)	21.13 (94.11)	<.01	5.52* (0.16)	33.61 (105.19)	<.01
Group:						
H	-0.51 (0.32)	-1.56 (94.11)	.12	-0.08 (0.23)	-0.37 (105.19)	.72
no-F	-0.08 (0.32)	-0.25 (94.11)	.80	-0.08 (0.23)	-0.34 (105.19)	.74
Time:						
T2	0.28 (0.18)	1.57 (68.43)	.12	0.13 (0.16)	0.83 (68.67)	.41
Group × Time:						
H × T2	-0.49* (0.25)	-2.01 (67.81)	.05	-0.01 (0.21)	-0.04 (67.77)	.97
no-F × T2	-0.37 (0.25)	-1.45 (68.71)	.15	-0.05 (0.22)	-0.25 (69.07)	.80
Covariance Parameters		Variance			Variance	
Intercept		0.99			0.40	
Residual		0.36			0.27	
Fit Statistics						
AIC		422.70			348.42	
BIC		446.56			372.29	
Marginal R^2		0.07			0.01	
Condit. R^2		0.75			0.60	

Figure 24: Estimated effects on intrinsic goal orientation and task value (N=77). Adapted from [4].

Effect	Control Beliefs			Self-efficacy			Test Anxiety		
Fixed Effects	Est. (SE)	t (df)	p	Est. (SE)	t (df)	p	Est. (SE)	t (df)	p
Intercept (≙ Group AI at T1)	5.67* (0.18)	30.86 (95.14)	<.01	5.37* (0.20)	27.43 (84.08)	<.01	3.01* (0.27)	11.33 (90.02)	<.01
Group:									
H	0.08 (0.26)	0.31 (95.14)	.76	0.24 (0.27)	0.89 (84.08)	.38	0.08 (0.37)	0.21 (90.02)	.84
no-F	-0.08 (0.26)	-0.32 (95.14)	.75	0.05 (0.27)	0.17 (84.08)	.87	0.06 (0.37)	0.16 (90.02)	.87
Time:									
T2	-0.15 (0.15)	-1.01 (68.14)	.32	0.10 (0.11)	0.92 (67.73)	.36	-0.38* (0.18)	-2.03 (68.02)	.05
Group × Time:									
H × T2	0.10 (0.20)	0.51 (67.48)	.61	-0.02 (0.15)	-0.12 (67.4)	.90	-0.21 (0.25)	-0.84 (67.51)	.40
no-F × T2	0.13 (0.21)	0.65 (68.43)	.52	0.00 (0.15)	0.01 (67.88)	.99	0.05 (0.26)	0.18 (68.25)	.85
Covariance Parameters		Variance			Variance			Variance	
Intercept		0.61			0.83			1.38	
Residual		0.24			0.13			0.38	
Fit Statistics									
AIC		360.99			335.01			448.63	
BIC		384.86			358.87			472.50	
Marginal R^2		0.01			0.01			0.03	
Condit. R^2		0.72			0.87			0.79	

Figure 25: Estimated effects on control beliefs, self-efficacy and text anxiety (N=77). Adapted from [4].

Effect	Perceived Usefulness			Perceived Ease of Use			Trust		
Fixed Effects	Est. (SE)	t (df)	p	Est. (SE)	t (df)	p	Est. (SE)	t (df)	p
Intercept (≙ Group AI at T1, towards AI)	5.37* (0.23)	23.71 (237.75)	<.01	4.60* (0.21)	22.21 (217.49)	<.01	5.32* (0.20)	26.60 (198.31)	<.01
Group:									
H	-0.82* (0.32)	-2.59 (237.75)	.01	-0.22 (0.29)	-0.74 (217.49)	.46	-0.40 (0.28)	-1.42 (198.31)	.16
no-F	-0.31 (0.32)	-0.97 (237.75)	.33	-0.32 (0.29)	-1.1 (217.49)	.27	-0.49† (0.28)	-1.74 (198.31)	.08
Type of Grader:									
Human Grader	0.45 (0.28)	1.64 (209.64)	.10	0.75* (0.24)	3.11 (210.56)	<.01	0.40† (0.22)	1.81 (209.78)	.07
Time:									
T2	-0.47 (0.29)	-1.62 (215.18)	.11	-0.15 (0.25)	-0.59 (215.51)	.56	-0.31 (0.23)	-1.32 (214.28)	.19
Group × Type of Grader									
H × Human Grader	0.61 (0.39)	1.58 (209.64)	.12	0.75* (0.34)	2.24 (210.56)	.03	0.33 (0.31)	1.07 (209.78)	.29
no-F × Human Grader	0.39 (0.39)	1.01 (209.64)	.31	0.59† (0.34)	1.74 (210.56)	.08	0.34 (0.31)	1.11 (209.78)	.27
Group × Time:									
H × T2	0.29 (0.40)	0.74 (213.36)	.46	-0.02 (0.35)	-0.05 (213.89)	.96	0.50 (0.32)	1.56 (212.80)	.12
no-F × T2	0.33 (0.40)	0.81 (215.07)	.42	0.30 (0.35)	0.87 (215.41)	.39	0.71* (0.32)	2.20 (214.19)	.03
Type of Grader × Time:									
Human Grader × T2	0.30 (0.40)	0.75 (209.64)	.45	0.19 (0.35)	0.55 (210.56)	.58	0.34 (0.32)	1.06 (209.78)	.29
Group × Type of Grader × Time:									
H × Human Grader × T2	-0.33 (0.56)	-0.59 (209.64)	.56	-0.72 (0.49)	-1.48 (210.56)	.14	-0.75† (0.45)	-1.68 (209.78)	.09
no-F × Human Grader × T2	-0.08 (0.57)	-0.14 (209.64)	.89	-0.25 (0.49)	-0.51 (210.56)	.61	-0.61 (0.45)	-1.34 (209.78)	.18
Covariance Parameters	Variance			*Variance*			*Variance*		
Intercept	0.33			0.35			0.39		
Residual	0.95			0.72			0.61		
Fit Statistics									
AIC	915.30			851.56			818.45		
BIC	966.87			903.13			870.02		
Marginal R^2	0.17			0.25			0.09		
Condit. R^2	0.38			0.50			0.45		

Note. The Intercept of the fixed effects represents the Group AI at T1, rated towards an AI grader. The Intercept of the covariance parameters represents between participant variance. Fitted using REML. Degrees of freedom estimated using Satterthwaite's method. $N = 77$. * ≙ $p < .05$ and † ≙ $p < .10$.

Figure 26: Estimated effects on acceptance. Taken from [4].

Question:
What are extension headers in IPv6 and where are they located in a packet? What is the main advantage of extension headers compared to IPv4?
Shortened Reference Answer:
Extension headers are used to extend the fixed IPv6 header with additional, optional network layer information. If present, they are located between the main header and payload/upper-layer header. Main advantage: One of the following advantages are considered fully correct: 1. It allows the appending of new options without changing the header. 2. IPv6 packets with optional headers are typically processed faster/simpler by intermediate devices as most of the options are ignored (except "Hop-by-Hop Extension") while they are processed by all routers in IPv4 unless ordered otherwise.
Student Answer 1
The extension headers are placed between fixed header and payload. The main advantage is that they are optional. This means that the use of space is more efficient because there won't be any redundant information placed there in order to keep the expected length.
Score & Gold Feedback
0.25 feedback: The response answers only the location of extension headers correctly. The definition is missing and the advantage is incorrect when compared to IPv4 as the option field in IPv4 is also optional. Thus, there is no efficiency benefit.
Predicted Score & Feedback 1.1
0.75 feedback: the advantage part in the response is partially correct because the advantage part is somewhat ambiguous. extension headers are optional but it's more a part of description/feature instead of an advantage.
Predicted Label & Feedback 1.2
partially correct. the advantage is incorrect as the option field in the ipv4 header is optional and can be placed between payload and fixed header. the advantage is incorrect as the option field in the ipv4 header is optional and can be placed between payload and fixed header.

Table 35: Example feedback generated by the T5 model without questions. Table taken from [3].

Question:
WHICH PROPERTY of spanning trees makes them appealing for broad- and multicasting? EXPLAIN how you can modify Link State Routing to construct a spanning tree for multicasting.
Reference Answer:
Property: There is a single unique path between every pair of nodes in the tree. (...) This means that no unnecessary duplicates are distributed in the network when forwarding packets using that tree structure. Spanning Tree with Link State Routing: Each intermediate system knows which multicast groups it belongs to, but initially doesn't know which other IS belong to the groups. Therefore, you can add multicast group information to the link state packet and each node can construct multicast trees once the full network topology and group information are distributed in the whole network, as each node then has the complete state information stored locally (e.g. with Prim or Kruskal)
Student Answer 3
Spanning trees are appealing to broad- and multicasting scenarios, because they allow the packets to only travel one path (except travling backwards). This removes the need for looking up specific tables as in RPF / RPB.
Score & Gold Feedback
0.25 feedback: It is true that there is a unique path between nodes but that not only does away with the need to look at routing tables in RPF/RPB but reduces duplicates by removing loops(unnecessary links). No explanation was provided for modifying the link-state algorithm to construct a multicast spanning tree for nodes.
Predicted Score & Feedback 2.1
0.75 feedback: the response is partially correct because the link-state routing modification description is missing how the nodes belong to which groups.
Predicted Label & Feedback 2.2
partially correct. the explanation behind using a spanning tree for multicast and broadcast is partially correct as it does not mention how the link-state algorithm is modified to construct a multicast spanning tree. the explanation for modifying the link-state algorithm to construct a multicast spanning tree is partially correct as it does not mention how each node shares its multicast information with others by adding it to the link-state packet.

Table 36: Example feedback generated by the T5 model without questions. Table taken from [3].

B

SUPERVISED STUDENT THESES

[1] Syeda H. Ahmad. "Utilizing Domain Knowledge for a Standalone Feedback-pipeline." Master Thesis. TU Darmstadt, 2022.

[2] Leonard Bongard. "A Feedback Generation Pipeline for Spoken Language Correction." Bachelor Thesis. TU Darmstadt, 2021.

[3] Erik Fischer. "Detecting and Reducing Bias in Automatic Short Answer Grading." Master Thesis. TU Darmstadt, 2023.

[4] Marieke S. Fischer. "Testing the Effects of a state-of-the-art Automatic Short Answer Grading System on Student Learning and Motivation." Master Thesis. TU Darmstadt, 2022.

[5] Philipp Hinderer. "Investigating the Use of Imperfect Feedback Generation Systems." Master Thesis. TU Darmstadt, 2022.

[6] Wanqing Kan. "Utilizing Neuro-Symbolic AI for Text Difficulty Classification." Master Thesis. TU Darmstadt, 2022.

[7] Felix Künnecke. "Neuro-Symbolic Automatic Short Answer Grading with Scoring Rubrics." Master Thesis. TU Darmstadt, 2023.

[8] Sebastian Ochs. "Using Deep Learning Techniques for Automatic Short Answer Grading." Bachelor Thesis. TU Darmstadt, 2020.

[9] Sebastian Ochs. "Utilizing Neuro-Symbolic AI for Text Difficulty Estimation." Master Thesis. TU Darmstadt, 2022.

[10] Siddharth S. Parihar. "Individual Feedback using Automatic Short Answer Grading." Master Thesis. TU Darmstadt, 2021.

[11] Jonas Rudolph. "Investigating Automatic Defense Strategies for Adversarial Attacks in Automatic Short Answer Grading." Bachelor Thesis. TU Darmstadt, 2020.

[12] Julian Schwind. "Improving the Robustness of Multilingual Automatic Feedback Generation with Adversarial Training." Master Thesis. TU Darmstadt, 2022.

[13] Yantao Shi. "Using Transformers for Automatic Short Answer Grading (ASAG)." Master Thesis. TU Darmstadt, 2020.

[14] Joel N. Tschesche. "Response-specific Elaborated Feedback Generation using X-AI." Bachelor Thesis. TU Darmstadt, 2021.

[15] Tim Unverzagt. "Application of the Lottery Ticket Hypothesis in NLP and Early Pruning of Neural Networks." Bachelor Thesis. TU Darmstadt, 2020.

[16] Paul Youssef. "Investigating Small Transformer Models for Automatic Short Answer Grading (ASAG)." Master Thesis. TU Darmstadt, 2020.

C

AUTHOR'S PUBLICATIONS

MAIN PUBLICATIONS

[1] Anna Filighera, Leonard Bongard, Tim Steuer, and Thomas Tregel. "Towards A Vocalization Feedback Pipeline for Language Learners." In: *Proc. 2022 Int. Conf. Advanced Learning Technologies*. Bucharest, Romania: IEEE Computer Society, 2022, pp. 248–252. DOI: 10.1109/ICALT55010.2022.00081.

[2] Anna Filighera, Sebastian Ochs, Tim Steuer, and Thomas Tregel. *Cheating Automatic Short Answer Grading: On the Adversarial Usage of Adjectives and Adverbs*. arXiv:2201.08318, preprint. Under review from International Journal of Artificial Intelligence in Education since 11/2021. 2022. DOI: 10.48550/arXiv.2201.08318.

[3] Anna Filighera, Siddharth Parihar, Tim Steuer, Tobias Meuser, and Sebastian Ochs. "Your Answer is Incorrect... Would you like to know why? Introducing a Bilingual Short Answer Feedback Dataset." In: *Proc. of the 60th Annual Meeting of the Association for Computational Linguistics*. Dublin, Ireland: Association for Computational Linguistics, 2022, pp. 8577–8591. DOI: 10.18653/v1/2022.acl-long.587.

[4] Anna Filighera, Tim Steuer, and Christoph Rensing. "Automatic Text Difficulty Estimation Using Embeddings and Neural Networks." In: *Proc. 2019 European Conf. Technology Enhanced Learning*. Delft, The Netherlands: Springer International Publishing, 2019, pp. 335–348. DOI: 10.1007/978-3-030-29736-7_25.

[5] Anna Filighera, Tim Steuer, and Christoph Rensing. "Fooling Automatic Short Answer Grading Systems." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Educ.* Virtual Event: Springer International Publishing, 2020, pp. 177–190. DOI: 10.1007/978-3-030-52237-7_15.

[6] Anna Filighera, Tim Steuer, and Christoph Rensing. "Fooling It - Student Attacks on Automatic Short Answer Grading." In: *Proc. 2020 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2020, pp. 347–352. DOI: 10.1007/978-3-030-57717-9_25.

[7] Anna Filighera, Joel N. Tschesche, Tim Steuer, Thomas Tregel, and Lisa Wernet. "Towards Generating Counterfactual Examples as Automatic Short Answer Feedback." In: *Proc. 2022 Int. Conf. Artificial Intelligence in Educ.* Durham, UK: Springer International Publishing, 2022, pp. 206–217. DOI: 10.1007/978-3-031-11644-5_17.

CO-AUTHORED PUBLICATIONS

[8] Leon Camus and Anna Filighera. "Investigating Transformers for Automatic Short Answer Grading." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Educ. Part II*. Virtual Event: Springer International Publishing, 2020, pp. 43–48. DOI: 10.1007/978-3-030-52240-7_8.

[9] Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. "Bloom Library: Multimodal Datasets in 300+ Languages for a Variety of Downstream Tasks." In: *Proc. of the 2022 Conf. on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, 2022, pp. 8608–8621. URL: https://aclanthology.org/2022.emnlp-main.590.

[10] Tim Steuer, Anna Filighera, Tobias Meuser, and Christoph Rensing. *I Do Not Understand What I Cannot Define: Automatic Question Generation With Pedagogically-Driven Content Selection*. arXiv:2110.04123v1, preprint. Under review from IEEE Transactions on Learning Technologies since 02/2021. Oct. 2021. DOI: 10.48550/arXiv.2110.04123.

[11] Tim Steuer, Anna Filighera, Nina Mouhammad, Gianluca Zimmer, and Thomas Tregel. "Learning-Relevant Concept Extraction By Utilizing Automatically Generated Textbook Corpora." In: *Proc. 2022 Int. Conf. Advanced Learning Technologies*. Bucharest, Romania: IEEE Computer Society, 2022, pp. 379–383. DOI: 10.1109/ICALT55010.2022.00117.

[12] Tim Steuer, Anna Filighera, and Christoph Rensing. "Exploring Artificial Jabbering for Automatic Text Comprehension Question Generation." In: *Proc. 2020 European Conf. Technology Enhanced Learning*. Virtual Event: Springer International Publishing, 2020, pp. 1–14. DOI: 10.1007/978-3-030-57717-9_1.

[13] Tim Steuer, Anna Filighera, and Christoph Rensing. "Remember the Facts? Investigating Answer-Aware Neural Question Generation for Text Comprehension." In: *Proc. 2020 Int. Conf. Artificial Intelligence in Educ.* Virtual Event: Springer International Publishing, 2020, pp. 512–523. DOI: 10.1007/978-3-030-52237-7_41.

[14] Tim Steuer, Anna Filighera, and Thomas Tregel. "Investigating Educational and Noneducational Answer Selection for Educational Question Generation." In: *IEEE Access* 10 (2022), pp. 63522–63531. DOI: 10.1109/ACCESS.2022.3180838.

[15] Tim Steuer, Anna Filighera, Thomas Tregel, and André Miede. "Educational Automatic Question Generation Improves Reading Comprehension in Non-Native Speakers: A Learner-Centric Case Study." In: *Frontiers in Artificial Intelligence* 5 (2022), pp. 1–14. DOI: 10.3389/frai.2022.900304.

[16] Tim Steuer, Anna Filighera, Gianluca Zimmer, and Thomas Tregel. "What Is Relevant for Learning? Approximating Readers' Intuition Using Neural Content Selection." In: *Proc. 2022 Int. Conf. Artificial Intelligence in Educ.* Durham,

United Kingdom: Springer International Publishing, 2022, pp. 505–511. DOI: 10.1007/978-3-031-11644-5_41.

D

ERKLÄRUNGEN LAUT PROMOTIONSORDNUNG

§8 Abs. 1 lit. c PromO
Ich versichere hiermit, dass die elektronische Version meiner Dissertation mit der schriftlichen Version übereinstimmt.

§8 Abs. 1 lit. d PromO
Ich versichere hiermit, dass zu einem vorherigen Zeitpunkt noch keine Promotion versucht wurde. In diesem Fall sind nähere Angaben über Zeitpunkt, Hochschule, Dissertationsthema und Ergebnis dieses Versuchs mitzuteilen.

§9 Abs. 1 PromO
Ich versichere hiermit, dass die vorliegende Dissertation selbstständig und nur unter Verwendung der angegebenen Quellen verfasst wurde.

§9 Abs. 2 PromO
Die Arbeit hat bisher noch nicht zu Prüfungszwecken gedient

Darmstadt, 12. Mai 2023

Anna Marie Filighera

COLOPHON

This document was typeset using the typographical look-and-feel classicthesis developed by André Miede. The style was inspired by Robert Bringhurst's seminal book on typography "*The Elements of Typographic Style*". classicthesis is available for both LaTeX and LyX:

https://bitbucket.org/amiede/classicthesis/

Final Version as of August 29, 2023 (classicthesis).