
SAARLAND UNIVERSITY

Faculty of Mathematics and Computer Science
Department of Computer Science
Master Thesis



Identifying Skills in Constructed Responses with Explainable NLP Models

submitted by

Smita Bhattacharya (2581485)
M.Sc. in Data Science and Artificial Intelligence
Saarbrücken
July 2024

Advisors:

Prof. Dr. Hendrik Drachsler, Sebastian Gombert
Leibniz Institute for Educational Research and Educational Information (DIPF)
Frankfurt am Main, Germany

Reviewer 1: Prof. Dr. Vera Demberg

Computer Science and Computational Linguistics
Department of Computer Science
Saarland University

Reviewer 2: Prof. Dr. Hendrik Drachsler

Educational Technologies and Learning Analytics Lab
Department of Education Information Center
Leibniz Institute for Educational Research and Educational Information (DIPF)

Saarland University
Faculty MI – Mathematics and Computer Science
Department of Computer Science
Campus - Building E1.1
66123 Saarbrücken
Germany

Erklärung

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und keine anderen als die angegebenen Quellen und Hilfsmittel verwendet habe.

Statement

I hereby confirm that I have written this thesis on my own and that I have not used any other media or materials than the ones referred to in this thesis

Einverständniserklärung

Ich bin damit einverstanden, dass meine (bestandene) Arbeit in beiden Versionen in die Bibliothek der Informatik aufgenommen und damit veröffentlicht wird.

Declaration of Consent

I agree to make both versions of my thesis (with a passing grade) accessible to the public by having them added to the library of the Computer Science Department.

Saarbrücken, 19.07.2024
(Datum/Date)

Smita Bhattacharya
(Unterschrift/Signature)

Acknowledgements

To begin with, first, I would like to express my gratitude to Prof. Dr. Hendrik Drachsler for the opportunity to pursue my master's thesis in his group and for taking the time to review my work. I would like to thank Prof. Dr. Vera Demberg for having an interest in my thesis for providing me with invaluable feedback and for reviewing my work. I am deeply grateful to Sebastian Gombert for his useful advice, insightful discussions, and meticulous proofreading of my master thesis and lastly for supporting me in every situation during the entire time.

Further, I want to thank Prof. Dr. Walt Detmar Meurers for providing me with access to the CREG-TUE dataset for conducting the secondary evaluation for my thesis.

I also want to show my gratitude to all the researchers from Physics Education, IPN: Leibniz Institute for Science and Mathematics Education, Kiel and Educational Psychology and Technology, Ruhr University, Bochum, for providing support to design the Moodle units that are involved in data collection of AFLEK dataset.

Lastly, I especially thank my brother, parents and friends who have believed in me more than myself and supported, understood, and motivated me through thick and thin.

Abstract

Automated Short Answer Grading (ASAG) is a well-explored NLP task that evaluates students' active knowledge development. It was primarily approached using a range of feature-based machine learning algorithms to various deep learning architectures. However, along with active knowledge, assessing the development of scientific skills for school students, especially in STEM education, is crucial. The automatic identification of these skills in students' constructed responses enables students and teachers to monitor how each skill level develops throughout a unit. In this study, we aim to automate a specific case of short answer scoring, so-called analytic short answer scoring. This task can be considered as a sub-variant of ASAG. Moreover, assessment is inherently a high-stakes scenario where the system's explainability is essential for making its decision reliable to the stakeholders. In this study, we approach this problem of coding core scientific skills using short free-text responses from German K12 students employing an interpretable PDR framework. Each desideratum of this framework helps to interpret different aspects of model explanations and provides stakeholders with a comprehensive understanding of the model's behaviour. To instantiate the modules of this framework, three transformer-based models and their variants are finetuned along with a SHAP explanation model. We have coded these skills using both instance- and entailment-based scoring to evaluate whether supplying models with a sample solution can improve results. Furthermore, we assessed the quality of our explanation model by demonstrating the overlap of the important features identified by the model with human-annotated evidence span and loss in confidence in both the prediction and explainer model if the evidence spans are occluded in the data. Our evaluation results imply that our approaches are reasonable for identifying competence or skill development in students' responses, with the entailment-based scoring slightly outperforming the instance-based approach. Furthermore, the result shows that the explanation from the explainer model aligns well with the human-annotated evidence span with few exceptions. The occluding evidence spans decrease the transformer models' predictive performance and explainer model confidence, indicating the importance of the occluded evidence span in the model's learning.

Keywords: Automated Short Answer Grading, skill assessment, explainable AI

Contents

1	Introduction	1
2	Related Work	5
2.1	Automatic short answer assessment	5
2.1.1	Feature-based ASAG	6
2.1.2	Supervised and Unsupervised ASAG	6
2.2	Trusted learning analytics	8
2.3	Explainability	8
2.3.1	Methods	8
2.3.2	Framework	9
3	Background	10
3.1	PDR Framework	10
3.1.1	Predictive Accuracy	11
3.1.2	Descriptive Accuracy	12
3.1.3	Relevancy	12
3.1.3.1	Occlusion Test	13
3.2	Instance-based vs. entailment-based scoring	13
3.3	Architecture	14
3.4	Evaluation metrics	18
3.5	Interpretability through SHAP	20
3.6	Random effects of hierarchical data	21
4	Methodology	23
4.1	Datasets	23
4.1.1	AFLEK short answer dataset	23
4.1.2	CREG-TUE dataset	27
4.2	Experimental Setup	27
4.2.1	Predictive Accuracy	32
4.2.2	Descriptive Accuracy	34
4.2.3	Relevancy	35

5	Results	37
5.1	Predictive Accuracy:	38
5.1.1	Evaluation on AFLEK dataset	38
5.1.2	Secondary evaluation using CREG-TUE dataset	39
5.2	Descriptive Accuracy	41
5.2.1	Transformer’s Explainability	41
5.2.1.1	Skill: Constructing Explanations	41
5.2.1.2	Skill: Analyzing Data	48
5.2.1.3	Skill: Planning Investigations	53
5.3	Reliability	59
5.3.1	Constructing Explanations:	59
5.3.2	Analyzing Data	66
5.4	Relevancy	72
6	Discussion	76
7	Conclusion, Limitations and Future Directions, Ethical Considerations	79
7.1	Conclusion	79
7.2	Limitations and Future Directions	80
7.3	Ethical Considerations	80
	Bibliography	82
	Appendix	86

Chapter 1

Introduction

For optimizing the educational process, digital technologies such as the Learning Management System (LMS) (Dougiamas & Taylor, 2003) can play an essential role in creating personalized learning. They can enable students to acquire the competencies or skills required, especially in STEM education. The skills are primarily analytical and are driven to ensure that the student can construct a specific explanation, they can analyze experimental data, or plan an investigation if required. One such scenario to evaluate whether a student can construct an explanation while learning about an energy-related module of a physics class is as follows:

- **Question:** Warum wird ein Laptop manchmal heiß? Notiert euch beide eure Antworten.

Translation: Why does a laptop sometimes get hot? Both of you should write down your answers.

- **Sample solution:** In dem Prozessor eines Laptops laufen viele Stromkreise zusammen, welche verschiedene Dinge in dem Laptop steuern, z.B. ob der Bildschirm an oder aus ist. In diesen wird ein Teil der elektrischen Energie in thermische Energie umgewandelt. Dies macht sich durch eine Wärmeentwicklung bemerkbar. Je mehr Prozesse in einem Laptop ablaufen, desto wärmer wird ein Laptop. Auch wenn man auf dem Laptop Spiele spielt, wird in den Stromkreisen ein Teil der elektrischen Energie in thermische Energie umgewandelt und der Laptop wird heiß.

Translation: Many circuits run together in the processor of a laptop, which control various things in the laptop, e.g. whether the screen is on or off. In these circuits, part of the electrical energy is converted into thermal energy. This is noticeable

through the development of heat. The more processes are running in a laptop, the warmer a laptop becomes. Even when you play games on your laptop, part of the electrical energy is converted into thermal energy in the circuits and the laptop gets hot.

- **Response that could explain:** Ein Laptop wird manchmal heiß , da er überhitzt ist . Das Überhitzen kann z.B. von Apps kommen die alle gleichzeitig geöffnet sind .

Translation: A laptop sometimes gets hot because it is overheated. The overheating can be caused, for example, by apps that are all open at the same time.

- **Response that could not explain:** Vielleicht , weil der Laptop zu überlastet ist .

Translation: Maybe because the laptop is too overloaded.

To facilitate the development of such skills, teachers must be able to quickly identify and intervene to change unproductive learning trajectories - those that do not result in the development of competence - into productive ones. As it is challenging for teachers to do so for all the students in traditional sessions, the extensive data generated while progressing through digital learning units in a synchronous learning scenario helps to gain insights into areas learners are struggling with, which opens up new possibilities to intervene. This analytical approach is called Learning Analytics (Drachler & Greller, 2016). This should provide the basis for developing assistance systems for students' personalised learning, of which automatic scoring of student's answers is a major component.

The learning units can be structured with different questions, such as multiple-choice and open-ended short-answer questions. Among these categories, the open-ended items written in natural language are appropriate for assessing learners' knowledge (Livingston, 2009). Since automatically scoring these is not trivial, a range of natural language processing techniques, from rule-based to machine learning, have been developed for this purpose (Burrows et al., 2015). The previous works in this domain show that *BERT*-like Transformers-encoder-based architectures (Devlin et al., 2018) can achieve state-of-the-art performance for short answer scoring (Camus & Filighera, 2020; Bexte et al., 2023; Gombert et al., 2023)

With Transformers-based models, multiple possible setups exist to implement short-answer scoring systems. Bexte et al. (2023) distinguish instance-based scoring, one interprets short-answer scoring as a simple classification problem where a given response is classified according to a pre-defined rubric. For similarity-based scoring, one embeds a sample solution and a student's answer into a shared semantic vector space and measures their similarity.

Another way to implement similarity-based scoring is using entailment classification in the tradition of the shared task held by Dzikovska et al. (2013). For entailment-based short answer scoring, one trains models to classify whether a given sample solution is

semantically entailed in a student response. Camus & Filighera (2020) illustrated the Transformers-based approaches to the problem building upon this interpretation .

In past years, most of the ASAG works focused on improving the model’s predictive performance, which has led to the employment of complex and less interpretable architecture. These architectures’ black box or opaque nature is becoming less trustworthy for the stakeholders to deploy in real-life, high-stake scenarios (Belle & Papantonis, 2021) like education. Nevertheless, limited research has been done to determine the rationale for such a model’s predictions. Drachsler & Greller (2016) have coined the concept of "Trusted Learning Analytics", which stresses the importance of sound models that maintain ethical guidelines to provide transparency to the end-users (both teachers and students) for ASAG systems. Additionally, it is crucial to understand if the model’s explanation is reliable and matches the intuition of the stakeholders (Belle & Papantonis, 2021). To understand the model’s behaviour clearly and to make well-informed decision, the stakeholders must be provided with helpful interpretation from different perspectives (Murdoch et al., 2019). Considering the importance of scientific skill assessment, especially in STEM education of the students along with the active knowledge, in this study, we aim to automate a specific case of short answer scoring, so-called *analytic short answer scoring*. This means we aim to predict multiple content dimensions instead of predicting an overall grade (Mertler, 2001). In particular, we put into practice and assess methods for automatically identifying three scientific skills in the short answers of German K12 students, namely *Constructing Explanations*, *Analyzing Data*, and *Planning Investigations* as exemplified before. For this purpose, we evaluate fine-tuning established pre-trained transformer-encoder language models (Camus & Filighera, 2020) using both instance- and entailment-based scoring mechanisms. Furthermore, we conduct secondary evaluations using the CREG-TUE dataset (Ott, 2014), a control group version of the CREG dataset (Ott et al., 2012) (vividly used dataset in past works for short answer scoring in German.) From the limitation of the interpretability of high-performing models in ASAG systems, in this study we further employ explainable AI(XAI) in our analytical skill assessment system to make it more interpretable and reliable to the stakeholders.

To sum it up, one of the main contributions of the thesis is to develop an explainable analytical skill assessment system using existing state-of-the-art models and employing an interpretable PDR framework (Murdoch et al., 2019). We also want to assess whether adopting entailment-based scoring (Dzikovska et al., 2013) improves the predictive performance of the fine-tuned models than the models fine-tuned using the instance-based scoring (Bexte et al., 2023). The primary dataset we use in this study is AFLEK, which is expanded upon a dataset published by Gombert et al. (2023) and is collected in secondary school physics courses that deal with energy transformation. We further conducted a secondary evaluation to assess the effect of these techniques in another German short answer dataset named CREG-TUE (Ott, 2014). The details of the datasets are mentioned in the corresponding dataset section. Making the skill assessment system

interpretable is crucial in education to enhance trustworthiness. Hence, we adopted the current most human-intuition-aligned explainable technique SHAP (Lundberg & Lee, 2017) to explain our model's prediction globally and locally. Furthermore, we want to assess whether our model's logical conclusion matches human intuition. Additionally, to check the reliability of the model's explanation, we want to check whether the absence of such human-annotated evidence span in those responses affects the model's prediction. Together with the previous points, the final objective is to provide the stakeholders with relevant information on the model's decision-making behaviour. These leads to the following research questions:

- **RQ1:** To what extent can scientific skills be detected in students' free-text responses using different Transformers language models (predictive accuracy)?
- **RQ2:** Does adopting the entailment-based scoring improve the model's predictive performance over the instance-based scoring?
- **RQ3:** To what extent can the same models and the different scoring techniques be applicable to the domain-specific standard dataset(Secondary evaluation)?
- **RQ4:** To what extent do input words considered important by the models for their predictions match human-coded ones (descriptive accuracy)?
- **RQ5:** To what extent the interpretation of these models' decision-making behaviour is relevant to the stakeholders(Relevancy)?

The following chapters of this thesis are organized as follows:

- **Chapter 2:** Here, we will discuss the previous works in the related domains.
- **Chapter 3:** This chapter elaborates on different terminologies related to the task.
- **Chapter 4:** This chapter discusses the methodology used for our experiments.
- **Chapter 5:** This chapter illustrates and explains all the results we got from the various experiments.
- **Chapter 6:** Here we summarize all the main observations that we noticed in the result section and how these answer each of the research questions.
- **Chapter 7:** This chapter discusses the conclusion of this work, the limitations and the possible future direction for the same. Also, it discusses the ethical considerations.

Chapter 2

Related Work

In this chapter, we will discuss about the previous work done in the domains related to this thesis. The section is organized as follows:

- **2.1 Automated short answer assessment:** Here, we have discussed the previous work related to automated short answer assessment. Furthermore, we classified different types of algorithms that were used and how this domain evolved with time.
- **2.2 Trusted Learning Analytics** Here, we will discuss the importance of trusted learning analytics in the ASAG domain and the work related to it.
- **2.3 Explainability** Here we will discuss the works related to model explainability. Furthermore, we will discuss different methods and framework associated with it.

2.1 Automatic short answer assessment

Automatic short answer assessment (ASAA) or short answer grading (ASAG) is a well-established application of NLP in the educational context, dating back several years. Burrows et al. (2015) conducted a comprehensive literature review focusing on ASAG where they have summarized that most of the previous works on this problem were approached to predict holistic scores or grades for constructed responses by providing a discrete or continuous value that directly implies the response's quality. The methods adopted can be broadly categorized as feature-based, supervised and unsupervised machine learning algorithms.

2.1.1 Feature-based ASAG

One of the preliminary works for students' response assessment was based on keyword matching, where this predefined keyword lexicon represents different concepts. The presence of such lexicons in the response is regarded as correct (Callear et al., 2001). Significant work has been done applying pattern matching between students' responses and expert-annotated standard sample solutions using different features. Features like bag-of-words (Cutrone & Chang, 2010) of standard sample solution and the students' response are matched to assess the correctness of the student's response. Another feature-based approach, like a sub-segment of parse trees (Bachman et al., 2002), is used to solve the ASAG task. Bachman et al. (2002) illustrated how they replaced pen and paper assessment with the WebLAS (Web-based language assessment system) system to provide efficient administrative work and an authentic, interactive assessment system. The assessment system first extracts the important elements from the reference answer by tagging and parsing them. Similarly, it extracts important information from the students' responses. Finally, it matches the pattern of the tagged and parsed responses and reference answers and interactively proposes the assessment to the user. Furthermore, Hahn & Meurers (2012) presents the assessment of students' responses using formal semantics matching. They used lexical resource semantics (LRS) representation of the student's response, reference answer and question based on the parts-of-speech tags and the dependency parses. Then, the extracted LRS representations are further aligned with the student's response, reference answer, and questions. This LRS considered the local matching criteria using semantic similarity and the global matching criteria by measuring the degree of alignment between the dominance constraints and the structure on the variable level. This alignment is further used to determine the important elements in terms of information structure.

2.1.2 Supervised and Unsupervised ASAG

In later days, most solutions have shifted towards adapting different supervised and unsupervised machine learning algorithms. Various hand-crafted and semi-hand-crafted lexical and semantic similarity features are generated from sample solutions for the classification task of the ASAG system. Meurers et al. (2011) have demonstrated the necessity of replacing surface-based information with more linguistically informed content analysis. They extracted features like overlapping keywords, tokens, chunks, dependency triples, matching type, lemma, and synonyms between the student and target answers. Then, these features were used to train the memory-based learning algorithm Tilburg Memory-Based Learner, which is based on the k-nearest neighbour classification to evaluate the similarity between the response and the target answer. While most of the ASAG system focused on the similarity between the student answer and the reference answer, Horbach et al. (2013) illustrated the effect of including reading text in

addition to the reference answer and the student's response. Similarly to the previous work, they used features like matching among the tokens, chunks and dependency triples of reference answer, student response and the reading text to train the k-nearest classifier. Furthermore, Crossley et al. (2016) introduced a constructed response analysis tool as an ASAG system. This tool calculates different linguistic feature dimensions to find the similarity between the student response and the reference answer. They used hints provided to the students during the assessment as reference answers. This tool calculates lexical similarity like keywords, synonym overlaps, latent semantic analysis similarity, and phrasal similarity of bi-gram and trigram overlap. Further, psycholinguistic norms, such as the word information indices, and syntactic information, such as the frequency of nouns and adjectives, are used to train linear discriminant analysis models to score different tasks as a chemistry tutor. Marvaniya et al. (2018) demonstrates a unique, unsupervised way of generating the scoring rubrics for the ASAG system. They demonstrated that the correct answers provide additional information than the reference answer. Instead of using the reference answer as the similarity standard for assessment, they used a clustering-based representative selection of the student answer, which can be used as the reference answer for that particular grade category. Then, the lexical overlap and sentence-embedding-based similarity are calculated between the question, set of representative answers and the responses.

However, without hand-crafted features, deep learning algorithms have demonstrated impressive performance (Riordan et al., 2017). Trusting the evaluation of the ASAG system depends on its optimized performance. Maharjan et al. (2018) have used LSTM networks for response assessments. More recently, transfer learning techniques, for example, BERT (Devlin et al., 2018), have been further fine-tuned for particular tasks after being pre-trained in an unsupervised way. BERT is a Transformers encoder model trained with next-sentence prediction and masked language modelling. Transformers-based models (Sung et al., 2019; Camus & Filighera, 2020; Poulton & Eliens, 2021; Gombert et al., 2023) have been implemented in ASAG with encouraging outcomes. Sung et al. (2019) have applied it to short answer grading and illustrated its utility across domains using Multi-Genre Natural Language Inference(MNLI) (Williams et al., 2018) and two psychology domain datasets. Similarly, Camus & Filighera (2020) have also shown improved transfer learning performance of Transformers-based models based on knowledge distillation using MNLI and sciEntsBank of SemEval-2013 (Dzikovska et al., 2013).

A lot of development of ASAG systems using the Transformers-based models are focused exclusively on the English Datasets. Along with various feature-based models, Gombert et al. (2023) illustrated the performance of the German language-based Transformers models such as GBERT-base and GEBRT-large using the AFLEK dataset they collected. Pado & Kiefer (2015) have used a project-specific dataset, Computer Science Short Answers in German (CSSAG), along with the dataset Corpus of Reading Comprehension Exercises in

German (CREG) (Ott et al., 2012) to develop a flexible, cross-domain assessment system that pre-sorts responses based on filtered lemma similarity to a reference answer. Nath et al. (2023) have also used the previously mentioned datasets, CREG and CSSAG, to conduct a comparative study of the predictive performance of various automatic short answer grading models specifically focused on different German language-specific BERT models like GBERT-base, DBERT-base, GottBERT-base, xlm-RoBERTa-base, german-gpt-2. Also, a Logistic Regression model trained with a bag of words with cosine similarity and a frequency model is implemented as a baseline.

2.2 Trusted learning analytics

With the primary goal of supporting instructors and students in analytical assessment systems, learning analytics (LA) (Greller & Drachsler, 2012) uses computational and data-driven approaches for assessment, primarily incorporating predictive modelling using machine learning and natural language processing. A potential drawback of data-driven approaches in learning analytics is the presence of distributional biases in training sets, which can lead to models learning unintended shortcuts rather than precise regularizations. This phenomenon is called "Clever Hans modelling" (Anders et al., 2022). It is essential to keep this feedback system far from such effects, which can harm students.

Eventually, the ethical guidelines for LA (Slade & Tait, 2019) stress the importance of sound models that are free from algorithmic bias and maintain transparency and clarity for end-users.

Drachsler & Greller (2016) presented the concept of "Trusted Learning Analytics", which is connected to the previous ethical guideline idea. Although their primary focus is data privacy concerns, they also consider the issue of "Asymmetrical power dynamics" that arise in learning settings. Gombert et al. (2023) have considered "Trusted Learning Analytics" while implementing their assessment system.

2.3 Explainability

2.3.1 Methods

Machine learning models can be divided as glass-box or black-box (Murdoch et al., 2019). The glass-box models, like regression and tree-based models, often offer intrinsic explanations for predictions, which make them interpretable. The high-performing models like Transformers are often black-box, which makes it challenging to ensure they learn reliable patterns (Sun et al., 2021) from the input. Ongoing research aims to make the black-box models transparent and interpretable for making them accountable in high-

stakes real-life application scenarios (Murdoch et al., 2019). Bastings & Filippova (2020) have illustrated three types of explainable methods for black box models, i.e. gradient-, propagation- and occlusion-based. Layer-wise relevance propagation(LRP) pickups only the most important portion of the input relevant to the output (Binder et al., 2016). Gradient-based techniques get model gradients using backpropagation to determine the feature importance. Chefer et al. (2021) have introduced Transformers explainability based on LRP and gradient-based weighting, which has been incorporated in ASAG systems like Gombert et al. (2023) for illustrating the feature importance. Lundberg & Lee (2017) have described six additive feature attribution methods (LIME, DeepLIFT, Layer-wise Relevance Propagation(LRP), classic Shapley value estimations, Kernel SHAP, Deep SHAP) for model explanations. They also discussed that the unique solution to the whole class of additive feature attribution methods is game theory, which follows the three properties of local accuracy, missingness and consistency. Among all the methods mentioned previously of additive feature attribution, SHAP is the unified solution, and its estimation methods align the most with human intuition.

2.3.2 Framework

In light of the various approaches developed to explain models in different ways, it becomes very confusing because of the various viewpoints within the field of interpretability. Aiming to provide a general conceptual framework for approaching interpretable, explainable machine learning in an organized manner, the PDR framework was presented by Murdoch et al. (2019). In this context, PDR stands for predictive accuracy, descriptive accuracy, and relevance. The well-known machine learning evaluation techniques that rank the accuracy of predictions are called "predictive accuracy". Descriptive accuracy is whether a model's learning is credible and true to the relevant coding criteria. The word "relevance" signifies the qualities of descriptive and predictive accuracy regarding the models that are important to stakeholders and their demands.

Chapter 3

Background

Since automatically identifying various skills in students' constructed answers is executed following some non-trivial framework and methods, we will elaborate on the associated terminology and the concepts in this chapter.

The structure of this chapter is as follows:

- In section 3.1, we will elaborately discuss the three desiderata of the PDR framework and its importance in interpretable machine learning projects for proper explanations.
- In section 3.2, we will discuss how instance-based and entailment-based scoring works.
- In section 3.3, we will discuss the architecture and the models those are used in this work
- In section 3.4, the various evaluation metrics are discussed
- In section 3.5, we will discuss the interpretability through SHAP.

3.1 PDR Framework

The capacity of machine learning (ML) models to accurately comprehend a broad range of intricate tasks has drawn a lot of interest lately. However, there's a growing awareness that machine learning (ML) models may do more than just make predictions; they can also provide interpretations or knowledge about the domain relationships in the data.

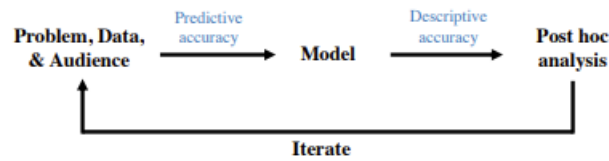


Figure 3.1: Overview of different data science life cycle stages(black text) where interpretability is important. Figure from the paper (Murdoch et al., 2019)

Without a clear definition of interpretability, a wide range of techniques with equally wide results (such as mathematical equations, natural language, and visualizations) have been categorized as interpretation. This has caused a great deal of misunderstanding on the concept of interpretability. Specifically, it is unclear how to choose an interpretation method for a given topic or audience, what it means to interpret something, and what connections exist between different ways.

Murdoch et al. (2019) have introduced the predictive, Descriptive, Relevant (PDR) framework, which can be used to fit most of the machine learning tasks and follow a standardised workflow to understand the interpretability of the models involved in such tasks.

They proposed that an interpretation method's output should accurately reflect the underlying process the practitioner is attempting to comprehend. Errors can occur in two different contexts related to machine learning (ML): (1) estimating the connections between underlying data and a model (predictive accuracy), and (2) approximating the model's learnt information using an interpretation approach (descriptive accuracy). (3) To ensure the reliability of an interpretation, it is advisable to optimize both accuracy metrics. Even in situations when the accuracy is low, the interpretations that are produced could still be helpful. But it's particularly crucial to verify their reliability by outside verification, like doing a second experiment. Finally, the higher reliability of the model's predictive and descriptive accuracy makes it relevant to the stakeholder's demand.

3.1.1 Predictive Accuracy

Figure 3.1 displays that the selection of an ML model is the initial source of mistakes during the model stage. Any information obtained from the model is unlikely to be valid if it learns an inaccurate approximation of the underlying connections in the data. Measures like test-set accuracy have been extensively researched in conventional supervised machine learning frameworks to assess a model's fit quality. This is referred to as Predictive accuracy. Interpretability using predictive accuracy is not limited to the model's average accuracy. For this reason, to understand the different perspective of the model's performance, evaluation measures like precision, recall and their harmonic mean F1 score are also considered necessary along with accuracy.

3.1.2 Descriptive Accuracy

Figure 3.1 illustrates that the second form of inaccuracy arises when a fitted model is analyzed using interpretable method in the post hoc analysis step. Sometimes the chosen interpretation techniques provide an inaccurate picture of the relationships that a model learns. This is particularly challenging for intricate black-box models which capture non-linear relationships between variables in complex ways, such as deep neural networks. In the context of interpretation, Murdoch et al. [2019] defines descriptive accuracy as the extent to which an interpretation technique accurately reflects the relationships that machine learning models have learnt. They also discussed two primary types of post hoc interpretation techniques that are most frequently employed: prediction-level and dataset-level interpretations, also known as local and global explainability.

- **Local explainability/ Prediction-level Interpretability:** Interpretation techniques at the prediction level concentrate on elucidating specific predictions made by models, including the features and/or interactions that contributed to that prediction. A widely used method for explaining locally or at the prediction level is to provide importance values to specific attributes/ features. A variable with a high positive (negative) score implies higher positive (negative) contribution to a specific prediction.
- **Global explainability/ Dataset-level Interpretability:** It is the explainability that the model learns globally at an entire dataset level, such as which linguistic signal patterns are connected to a projected response.

3.1.3 Relevancy

Murdoch et al. (2019) described that only having high accuracy of an interpretation technique is insufficient; the extracted information must also be pertinent to the stakeholder's point of view. Enhancing the relevance of interpretations for audiences or stakeholders is another way to create better interpretation techniques. Typically, this is accomplished by presenting a highly reliable predictive and descriptive accuracy of the model output, like feature heat maps, feature hierarchies, rationales, or highlighting significant training set components and any further probable effect on the model's behaviour. For ASAG system relevancy, the model should produce an explanation that matches human intuition and highly reliable. For instance, the part of the response that the humans think is important is also marked as important by the ASAG system. It is also essential to check additional factors that might influence the model's decision.

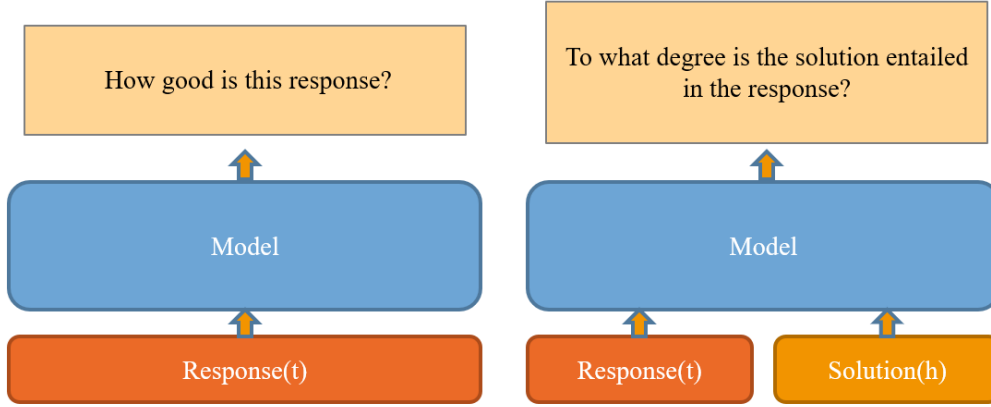


Figure 3.2: Instance-based vs. Entailment-based scoring
(Burrows et al., 2015; Dzikovska et al., 2013)

3.1.3.1 Occlusion Test

To understand the reliability of the previous predictive and descriptive accuracy methods, it is required to incorporate methods in the framework that elevate the trustworthiness of the stakeholders. Occlusion-based methods work by occluding or eliminating a part of the whole input features and measuring how that changes the model prediction (Bastings & Filippova, 2020). By intuitive understanding, it is evident that if we remove the features that do not have much impact, the model’s prediction should not change. In contrast, the opposite is true for important features. The occlusion-based method is a saliency method, which calculates the impact of a single input feature i as:

$$f_c(\mathbf{x}_{1:n}) - f_c(\mathbf{x}_{1:n}|_{x_i=0}) \quad (4)$$

computes saliency, where $\mathbf{x}_{1:n}|_{x_i=0}$ indicates that input embedding x_i was assigned with zero or dummy token, which does not have any impact on the model’s output, e.g. [MASK] token. In contrast, the rest of the inputs are unmodified.

3.2 Instance-based vs. entailment-based scoring

There are several ways to obtain models to create a short-answer scoring system that can learn an accurate approximation of the underlying data and have reliable predictive performance. According to Bexte et al. (2023), instance-based scoring is a form of short-answer scoring, which is best understood as a straightforward classification task in which a response is categorized using a predetermined rubric. One calculates similarity-based scoring by embedding the similarity between a student’s response and a sample solution into a common semantic vector space. In the context of the shared task

exhibited by Dzikovska et al. (2013), entailment-based classification is another method of implementing similarity-based scoring.

Korman et al. (2018) has formally defined the textual entailment as if " t textually entails h " typically, a human reading t would be justified in inferring the proposition expressed by h from the proposition expressed by t ".

$$t \text{ textually entails } h \equiv t \supset h \quad (1)$$

Similarly, the response (t) written by the students logically entails the standard reference answers (h). In automatic short-answer scoring, the entailment-based scoring is the textual entailment between the response and the standard reference answer. The responses are assessed according to their logical consistency and relevance to the reference answers. This strategy improves scoring accuracy by emphasizing the semantic alignment between student responses and the standard reference answers. We also evaluated both approaches to our problem. While, for instance-based scoring, models are only given a student response as input, for entailment-based scoring, the models are also given the corresponding sample reference solution in addition to a respective student response. Figure 3.2 illustrates how this works.

3.3 Architecture

As discussed in the related work, Transformers language models were already used by Camus & Filighera (2020); Gombert et al. (2023); Poulton & Eliens (2021), among others, to achieve state-of-the-art results for the well-known SemEval-2013 dataset. Accordingly, using this Transformers language model to identify scientific skills in student responses is a practical approach worth evaluating. We evaluate two German language models (Chan et al., 2020), namely *GermanBERT* (*GBERT*) and *GermanELECTRA* (*GELECTRA*), which are *BERT* (Devlin et al., 2018) and *ELECTRA* (Clark et al., 2020) versions, respectively pre-trained exclusively on German data. Also, we have fine-tuned the *XLM-RoBERTa* model. To the best of our knowledge, *GBERT* and *XLM-RoBERTa* have been used for automatic short-answer scoring in previous works. For this reason, we also wanted to evaluate the *GELECTRA* model for this task along with the *GBERT* and *XLM-RoBERTa* models. All these models are based on Transformers architecture. Here, we will briefly discuss the Transformers architecture and the pre-training of the various models based on it.

Transformers language models (Vaswani et al., 2017) are feed-forward neural networks that handle sequential input and overcome the parallelization bottleneck of architectures like *Recurrent Neural Networks* (*RNN*) and *Long Short-term Memory* (*LSTM*). The current state-of-the-art techniques for language tasks ranging from translation to language

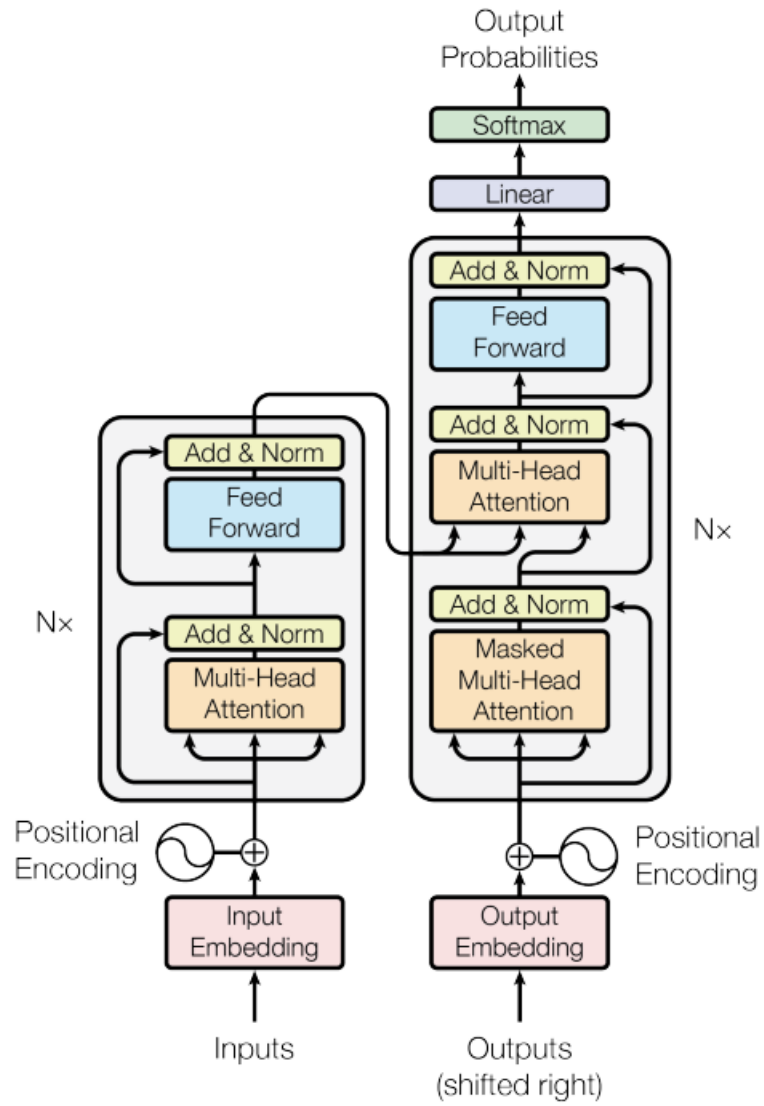


Figure 3.3: Transformer architecture (Vaswani et al., 2017)

generation are based on Transformers architecture. The Transformers consist of encoders and decoders. Figure 3.3 displays the architecture in detail. At first, the input is processed through the input embedding and positional encoding block to transform it into a vector representation. The positional encoder combines the sine and the cosine waves to retain the positional information associated with a given word. At the centre of the encoders, the self-attention mechanism uses a learnt weighted average of the vector representations of words in their context to learn how to represent words within a long sequence. These models can generate contextual word embeddings by focusing on various linguistic signals through numerous attention units, also known as attention heads. The output of the attention heads is then passed through the feed-forward neural network, which becomes the input for the multi-head attention block on the decoder side. This provides information regarding the current decoder state. The multi-head attention block of the decoding side does not allow future instances to be seen, and the output is generated in an auto-regressive manner.

- **GBERT(German Bidirectional Encoder Representation from Transformers):**

The *GermaBERT* (Chan et al., 2020) is based on the architecture of the *BERT* model (Devlin et al., 2018). The *BERT* is encoder stacks of Transformers architecture. Figure 3.4 displays the schematic representation of the training process of the *BERT* model. The model training consists of two phases. The first phase is called pre-training, which captures the underlying distribution of the language, and the second phase is fine-tuning the model to adapt to the specific downstream task. In the pre-training phase, a regular feed-forward layer is attached to the model and is trained with masked language modelling (Masked LM) and next-sentence prediction (NSP). In the masked language modelling, random words from the input sequence are masked, and the model learns to predict probable words for the masked position. In the next sentence prediction task, the model produces a 0 or 1 based on understanding whether the two input sentences follow each other. The Tok 1 to Tok N in figure 3.4 are the words in the input sequence to the model. E_1, E_2, \dots, E_N are the generated embeddings for the respective tokens. Afterwards, the task-specific linear neural network layer is attached to the model for fine-tuning to a downstream task, which takes input from the previous pretrained layers. *GBERT* is pre-trained on German Wikipedia data, OpenLegalData, and news using the same pre-training phase as *BERT*.

- **GELECTRA(German Efficiently Learning an Encoder that Classifies Token Replacement Accurately):**

The pre-training strategy of *BERT*, Masked language modelling, learns only from the masked token, which is fifteen per cent of the input tokens. It makes the model's learning very restricted and requires much computing to be effective. To overcome this problem, the *ELECTRA* model is introduced. *ELECTRA* is pretrained with a

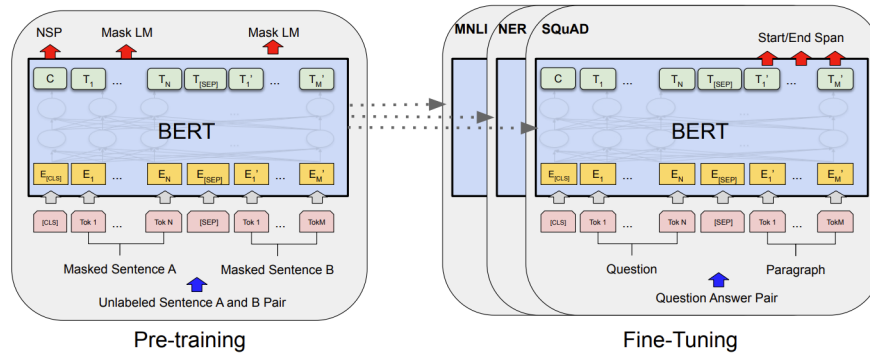


Figure 3.4: Overview of BERT pre-training process (Devlin et al., 2018)

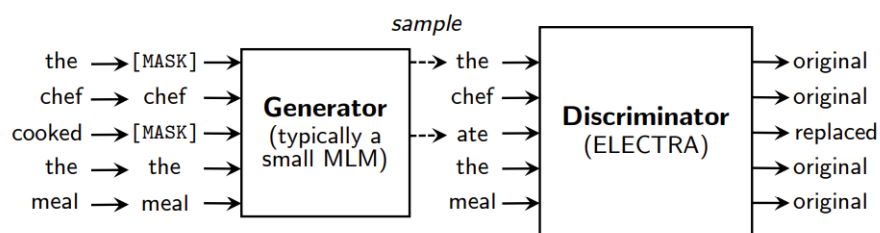


Figure 3.5: Overview of ELECTRA pre-training process: Replaced token detection (Clark et al., 2020)

more sample-efficient task called replaced token detection (RTD). Alternative to masking of tokens, RTD corrupts the input by replacing some tokens with possible alternative tokens sampled from a small generator network. Afterwards, instead of training the model to predict the original token in the masked position, a discriminative model is trained to predict whether a generator sample replaced each token in the corrupted input. Clark et al. (2020) illustrates that RTD is more effective than the MLM technique for pre-training as RTD is defined over all input tokens, whereas MLM only works with a subset of the masked tokens. 3.5 illustrates the Replaced token detection task overview. A small masked language model is trained jointly with the discriminator as a generator. The *ELECTRA* is intended for text input, so it is difficult to train the generator adversarially like other GAN models. For this reason, the generator is trained with maximum likelihood. After the pre-training, the generator is thrown away, and only the discriminator is finetuned for the downstream task. The *GermanELECTRA* (Chan et al., 2020) is based on the *ELECTRA* architecture. It is pre-trained with monolingual corpora, OSCAR, extracted from Common Crawl, Wikipedia dump for German, and OPUS project corpus, which contains texts from various domains like movie subtitles, parliament speeches and books. Open Legal data contains the German court decisions.

- **XLM-RoBERTa (Cross-Lingual Language Model with RoBERTa architecture):** *XLM-RoBERTa* is a multilingual model trained on 100 different languages. It is based on the *XLM multilingual model*, which has been architecturally adopted from the *RoBERTa* model. The full form of *RoBERTa* is *A Robustly Optimized BERT Pretraining Approach*. It is an extension of the *BERT* model with some modifications to make it more robust. Liu et al. (2019) illustrates that the *BERT* is significantly under-trained and can be improved. They incorporated a few changes in the model training and the data. *RoBERTa* was trained for longer with bigger batches and more data than the *BERT* model. The next sentence prediction objective was removed and trained with a longer sequence. It also includes dynamically changing the masking pattern during pre-training. *XLM-RoBERTa* is improved its performance over existing cross lingual models as it increases the number of languages which helps in better cross-lingual performances.

3.4 Evaluation metrics

To assess the models' predictive performance (predictive accuracy), we have used accuracy, precision, recall and their harmonic mean F1 score.

- **Accuracy:** It is defined as

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{False Positive} + \text{True Negative} + \text{False Negative}}$$

	Positive (actual)	Negative (actual)
Positive (predicted)	True positive	False positive
Negative (predicted)	False negative	True negative

Table 3.1: Confusion matrix illustrating every scenario that might occur in an experiment

To measure the predictive performance of machine learning models, we usually calculate the accuracy of that model, which is a very intuitive and straightforward method to get an idea of a model's performance. It is calculated as the percentage of the correct prediction of instances. However, high accuracy cannot always assure the model's performance. If the dataset used is biased for a particular class regarding the count of instances, using accuracy alone to assess that model's performance will not provide an optimal picture because the majority class will bias the model and will learn to classify most instances as majority class instances, ignoring the minor ones.

- **Precision:** It is defined as

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

Precision is the ratio of true positive predictions to predicted positive instances. In the educational context of automatic short answer scoring, precision indicates how many of the answers classified as correct by the ASAG system are actually correct and help to minimize the risk of assigning points for wrong answers. The system's high precision indicates that if it classifies an answer as correct, it is usually correct. It is important to have higher precision to maintain the credibility of the ASAG system. Also, using precision alone for the ASAG system is risky as it tends to ignore the false negatives by overlooking some potential correct answers.

- **Recall:** It is defined as

$$\text{Recall} = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

It is the ratio of true positive predictions to actual positive instances. In ASAG, using recall, we try to understand how well the model identifies all the correct answers. It is important to have a higher recall of the ASAG system to guarantee that all the students' correct answers are recognized and credited. Moreover, recall tends to overlook the false positive cases and overestimate the student's performance.

- **F1-score:** It is defined as

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

The F1 score is the harmonic mean of the previous two evaluation metrics, precision and recall. It provides a balance between these two. In an educational context, both precision and recall are very important. Ideally, the ASAG system must correctly identify as many right answers as possible (high recall) and avoid marking wrong answers as correct (high precision). The F1 score provides a fair and balanced judgement of model performance, especially when the label distribution of the data is not even (e.g., an imbalanced dataset). Although the F1 score tends to ignore true negatives, which might be less relevant for the ASAG system, it is still very important for comprehensive performance evaluation.

3.5 Interpretability through SHAP

Lundberg & Lee (2017) illustrated that if we want to get an explanation of a model's decision, then the best explanation can be provided by the model itself as it represents the best reason behind its decision. With increasing model complexity, the interpretability decreases and loses its explainability towards the decision. For instance, ensemble methods or deep networks. Due to the complexity of the prediction model, a simple model must be used as an interpretable approximation of the original prediction model. If we consider f is the original prediction model to be explained and g is the explanation model, the *local explainability* is designed to explain a prediction $f(x)$ based on a single input x . A *simplified inputs* x' is mostly applied to the explanation model that connects to the original exact inputs through a mapping function $x = h_x(x')$. In local explainability, the methods involved make sure that the $g(z') \approx f(h_x(z'))$ whenever $z' \approx x'$. Although x' may contain less information than x , $h_x(x') = x$ as h_x is specific to the current input x . The explanation methods such as LIME, DeepLIFT, Layer-Wise Relevance Propagation, Classic Shapley Value estimation, etc., follow the same additive feature attribution method under specific conditions. Lundberg & Lee (2017) defined the additive feature attribution model as follows:

Definition 1 *Additive feature attribution methods* have an explanation model that is a linear function of binary variables:

$$g(z') = \phi_0 + \sum_{i=1}^M \phi_i z'_i, \quad (2)$$

where $z' \in \{0, 1\}^M$, M is the number of input features which is simplified, and $\phi_i \in \mathbb{R}$.

The explanation models with a method that incorporates definition 2 to provide an effect ϕ_i to each feature; in that case, the summation of all the effects of all features approximates the output $f(x)$ of the original model.

The game theory assures a unique solution to the entire class of additive feature at-

tribution methods. To provide explainability for the complex black box model, the model-agnostic approach named "SHAP" (Shapley Additive exPlanations) (Lundberg & Lee, 2017) is used. The output of any machine learning model, including transformer models, can be explained using the game theoretic method. The explainable SHAP model retrains $S \subseteq F$ which is a subset of all possible features, where F is the set of all features. An importance value called *Shapley regression values* is provided to each feature that represents the effect on the model prediction of including that feature. This is done by perturbing the input features and observing the change in the model's prediction. To calculate the difference in the effect of each feature, a model $f_{S \cup \{i\}}$ is trained once with that feature which is present, and another model f_S is trained without the feature. Then, two models' predictions are compared on the input $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$, where x_S means the values of the input features in the set S . Because the effect of excluding a feature depends on other features in the model, all potential subsets have the preceding differences computed $S \subseteq F \setminus \{i\}$. The Shapley values are then calculated and applied as feature attribute values. They represent a weighted average of all potential differences:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (3)$$

As discussed, the mapping function h_x maps 1 or 0 to the original input space for Shapley regression values. If the input is included in the model, the value is 1 or 0 otherwise. If we consider $\phi_0 = f_\emptyset(\emptyset)$, then the Shapley regression values match the previous equation 2 and are an additive feature attribution method.

3.6 Random effects of hierarchical data

To assess any further effects which can affect the model's decision making behaviour, we employed further tests. For examining the effect of hierarchical data, we have implemented a generalized linear mixed effect model (GLMER) (Bates et al., 2015), also known as multilevel models in various fields. As our dataset contains multiple responses (multiple observations) from each student (each subject) and multiple responses (multiple observations) from each question (each item), this implies that individual responses are not completely independent from one another, but they are grouped under students and questions. So, using the GLMER model here, it can be observed whether there is any random effect of hierarchical questions and students on the performance of those models. One of such models is described below in detail.

$$\text{correct} \sim \text{model} + (1 \mid \text{question_id}) + (1 + \text{model} \mid \text{student_id}) \quad (3.1)$$

- correct: Difference between the actual and the predicted labels by the models. If the model predicts class 1 for an instance and the instance belongs to class 1, the

value of the correct variable is 1 or 0 otherwise.

- `model`: The name of the models whose prediction is being checked, for example, the best-performing model from the instance-based and the entailment-based scoring.
- `(1|question_id)`: Random intercept by the question id here means that the model might be more likely to be correct in predicting the label for different questions.
- `(1 + model|student_id)`: Random slope for the model under student means that some models might have an easier time predicting whether the responses of specific students are correct.

Chapter 4

Methodology

In this chapter, we will discuss the datasets we used for this study and will discuss the experimental setup for each module of the PDR framework. The section is organized as follows:

5.1 Datasets: Here we discuss in detail the AFLEK and the CREG-TUE dataset.

5.2 Experimental Setup: Here we will discuss how each stage of the PDR framework is instantiated.

4.1 Datasets

In this section, we will begin with the AFLEK dataset and next, we will discuss the CREG-TUE dataset.

4.1.1 AFLEK short answer dataset

The primary dataset we use in this study expands upon a dataset published by Gombert et al. (2023). Compared to the version used in this paper, the number of data points has nearly doubled. It is called the *AFLEK Short Answer Dataset*. The full form is *Analyse und Förderung von Lernverläufen zur Entwicklung von Kompetenzen* (*Analysis and promotion of learning processes for the development of skills*) It was collected using two Moodle courses dealing with energy transformation, spanning six periods. The courses were designed following the principles of project-based pedagogy (Sawyer, 2005). One sample question is described below:

- Original Question: Warum wird ein Laptop manchmal heiß? ist es dir schon einmal passiert, dass dein Laptop heiß wurde? Wann ist dir das passiert? Was hast du getan?
- Translated Question: Why does a laptop sometimes get hot ? has it ever happened to you that your laptop got hot? When did that happen to you? What did you do?

Every unit in those courses begins by posing a key question about energy and associated phenomena, which provides the basis or motivation for the subsequent sub-questions. The key question is further subdivided into three sub-questions. Each sub-question assesses each scientific skill separately, namely *Constructing Explanations*, *Analyzing Data* and *Planning Investigations*. Finally, the module ends with a final question that summarizes all the previous sub-questions.

1. **Constructing Explanations:** a response demonstrates that a student can combine multiple core ideas from energy physics to construct a coherent explanation. It is not important whether this explanation is scientifically perfect.
2. **Analyzing Data:** a response demonstrates that a student can analyse and interpret the data obtained from experiments. Whether they come to the right conclusions is not important
3. **Planning Investigations:** a response demonstrates that students can plan an experiment and justify their setup. It is not important whether the proposed setup is completely correct.

To develop the rubric of this dataset for coding the scientific skills in the students' answers, evidence-centered design(ECD) (Pellegrino et al., 2016) is used. Table 4.1 illustrates the rubrics used by the annotators to label or code the presence of scientific skills in the students' responses in detail.

The dataset is in German and comprises 31 different constructed response items. In total, 5159 responses were gathered from 620 students at various middle schools in the German state of Schleswig-Holstein(Gymnasium and Gemeinschaftsschule) (See also table 4.2). Each of the items comes with a sample solution. Each student's answer was labelled binarily, indicating whether a skill was identified — two annotators who reached a near-perfect agreement (Cohen's kappa = 0.96) did this. They are skilled student employees who independently iteratively coded the responses and verified their agreement. If there is a difference among the annotators in annotating specific responses, they discuss and resolve to find an agreement. Responses might have several codes since some constructed response items addressed numerous skills. Conversely, every student's answer does not have labels for all the skills. Instead, they were annotated with the particular skill which can be applied to them. Furthermore, the annotators also identified the specific portion or span of the response, which they considered to be the reason for

Skill	Goal	Reason	Evidence
Constructing Explanations <i>Figure 4.2 provides an example of the question measuring Constructing Explanations skill.</i>	Determine if students can construct coherent explanations. It is not important if they are completely correct from a normative perspective.	This label shall help us conclude whether the students understand the tasks and their prompts and can answer them.	Students use and combine ideas learned from the units to write a coherent explanation as an answer.
Analyzing Data <i>Figure 4.3 provides an example of the question measuring Analyzing Data skill.</i>	Trace whether students analyze and interpret the data obtained from the experiment	This label shall help us to conclude whether the students are trained to analyze and interpret data.	Students describe data and evaluate their ideas about the observed phenomenon based on that data.
Planning Investigations <i>Figure 4.4 provides an example of the question measuring Planning Investigation skill.</i>	Tracing whether students work on the tasks where they plan and carry out investigations.	This label shall help us to draw conclusions on whether the students performed or witnessed the experiments. We do not check whether they performed the experiments in a normative correct way obtaining normative correct results.	Students prepare the process for data acquisition and obtain as well as document data.

Table 4.1: This rubric lists various codes used for coding the AFLEK data for scientific skills

Skill	Number of students	Number of responses	Avg. number of responses per Student	Avg. Number of words per Response
All	620	5159	8.1	25.11
Constructing Explanations	615	4032	6.55	26.33
Analyzing Data	431	972	2.25	21.01
Planning Investigations	155	155	1	33.45

Table 4.2: AFLEK length Summary

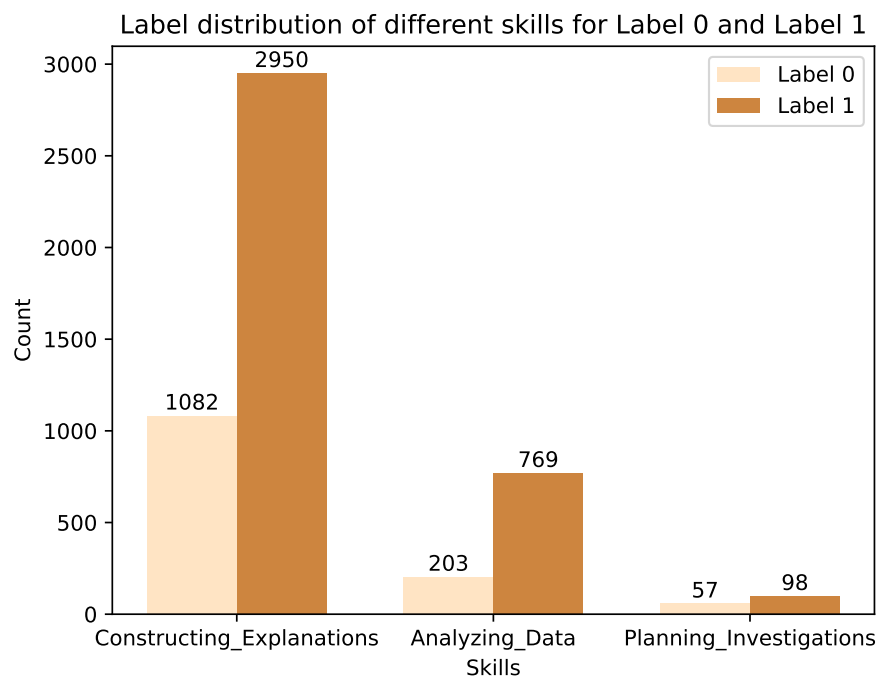


Figure 4.1: Label distributions for the three skills. *Label 0* refers to cases where a skill could not be detected in a response, while *Label 1* refers to cases where a skill could be detected.

classifying it as a positive response. This portion or span of the response is called the evidence span. For example, a student is asked to write about his/her observation after watching a video. As an answer, he/ she writes, "First a video game is played and then **a notification appears that overheating has been detected**. The laptop then switches off automatically". The annotators marked the bold portion as evidence span to demonstrate why they think the student could construct a response properly. Figure 4.1 portrays the label distribution of each skill. A complete set of questions, reference answers, and corresponding positive responses with marked evidence span and negative responses for each skill is discussed in the result section's table 5.3.

4.1.2 CREG-TUE dataset

The second dataset that we have used for our study is the Corpus of Reading Comprehension Exercises in German-TUE (Ott, 2014)(CREG-TUE), which is an addition to the Corpus of Reading Comprehension Exercises in German (CREG) (Ott et al., 2012). It was designed as a control group obtained by letting German native speakers answer a subset of the same CREG dataset that the German learners at Ohio State University and Kanas University had to respond to by doing the same exercises and being rated by the same annotators as the learners at the US universities. We have used this dataset for our secondary evaluation to validate our methods on an established benchmark for the task of short answer scoring.

During the summer of 2010, information was gathered at the computer lab of Tübingen University's Seminar für Sprachwissenschaft. One hundred participants, most of them being German native speakers, settled in to work on the Moodle activities designed to improve their reading comprehension. Upon completion, each participant received 5€ in payment. Before the exercise, an attempt was made to increase participant motivation by rewarding the greatest exercise submission. The winner could receive thirty euros as a voucher for a restaurant or a bookstore, both popular among students in town.

It consists of 6516 student responses with 180 reference answers. 143 questions were answered by 100 students. A sample from the dataset is illustrated in table 4.3. Additionally, table 4.4 illustrates further details. For each answer except for a binary true and false label, the annotators have further classified each true as "Extra Concept", "Correct", and each false as "Missing Concept", or "Blend". Figure 4.5 displays the label distribution for both binary labels and the 4-class labels.

4.2 Experimental Setup

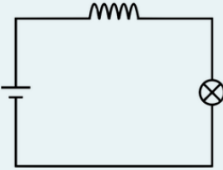
Figure 4.6 illustrates the thesis workflow schematically. Each stage shows the working of each desideratum of the PDR framework. In the following section, we will discuss how

Figure 4.2: Example question for assessing Constructing Explanations Skill.
Question Translation: Explain why we don't always align solar cell with the position of the sun

Figure 4.3: Example question for assessing Analyzing Data Skill.
Question Translation: Compare the series of measurements for brightness with the series of measurements for electrical voltage. What do you notice? Also name the places where you noticed something.

Baut einen Stromkreis nach der unten stehenden Schaltskizze auf. Folgende Anleitung soll euch weiterhelfen:

1. Schließt zunächst die Lampe mithilfe von zwei Kabeln und zwei Krokodilklemmen an die Batterie an.
2. Unterbrecht den Stromkreis.
3. Bringt in den Stromkreis den aufgedrehten Draht mit den zwei Stativen ein.
4. Ihr sollt nun die Temperatur im Inneren des aufgedrehten Drahtes messen. Schiebt dafür das Thermometer in den aufgedrehten Draht und messt die Temperatur 1min lang alle 10s für einen geschlossenen und einen nicht geschlossenen Stromkreis. Notiert euch die Werte in der unten stehenden Tabelle (Tipp: Beginnt beide Messungen mit der gleichen Starttemperatur).



Hinweis: Wenn du dir beim Aufbau unsicher bist, findest du [hier](#) das Bild des aufzubauenden Stromkreises.

Rich text editor toolbar with icons for undo, redo, bold, italic, bulleted list, numbered list, link, unlink, image, and H-P. Below the toolbar is a large empty text area for writing answers.

Figure 4.4: Example question for assessing Planning Investigation Skill.

Translated question: "Build a circuit according to the circuit diagram below. The following instructions should help you:

1. First connect the lamp to the battery using two cables and two alligator clips.
2. Break the circuit.
3. Insert the twisted wire with the two tripods into the circuit.
4. You should now measure the temperature inside the untwisted wire. To do this, push the thermometer into the untwisted wire and measure the temperature every 10 seconds for 1 minute for a closed and non-closed circuit.

Write down the values in the table below (tip: start both measurements with the same starting temperature). Note: If you are unsure about how to set it up, you can find the picture of the circuit to be set up here.

	Original Text	Translated Text
Text Title	Fatima Yützel erzählt	Fatima Yützel tells
Text	<p>Ich heie Fatima Yützel, bin fünfzehn Jahre alt und bin in Berlin geboren. Meine Eltern kommen aus einem kleinen Dorf in der Türkei und leben seit zwanzig Jahren in Berlin. Wir wohnen in einer Wohnung und haben dort viele Nachbarn, Türken und Deutsche. Am Abend und am Wochenende besuchen wir oft unsere türkischen Nachbarn, denn unsere Nachbarn sind auch unsere Freunde. Meine Eltern sprechen nicht gut Deutsch und sie haben auch deshalb nur wenig Kontakt mit Deutschen. Aber wie so viele junge Türken in Deutschland spreche ich besser Deutsch als Türkisch. Meine Freundin Melanie ist Deutsche. Ich bin auch sehr gern und sehr oft bei Melanie und ich finde sie und auch ihren Bruder David sehr nett. Das sage ich meine Eltern nicht, denn laut meinen Eltern hat ein ordentliches Mädchen keinen Freund. Die Eltern finden den Mann für ihre Tochter. Aber ich bin hier in Deutschland geboren und vielleicht heirate ich sogar mal einen Deutschen.</p>	<p>My name is Fatima Yützel, I am fifteen years old and was born in Berlin. My parents come from a small village in Turkey and have lived in Berlin for twenty years. We live in an apartment and have lots of neighbours there, Turks and Germans. In the evenings and at weekends we often visit our Turkish neighbours, because our neighbours are also our friends. My parents don't speak German well and therefore have little contact with Germans. But like so many young Turks in Germany, I speak better German than Turkish. My friend Melanie is German. I also like to visit Melanie very often and I find her and her brother David very nice. I don't tell my parents that, because according to my parents, a decent girl doesn't have a boyfriend. The parents find the man for their daughter. But I was born here in Germany and maybe I'll even marry a German one day.</p>
Question	Woher kommt Fatima Yützel?	Where does Fatima Yützel come from?
Reference Answer	Fatima kommt aus Berlin.	Fatima comes from Berlin.
Positive Response and extra concept	Sie ist in Berlin geboren. Ihre Eltern kommen aber aus der Türkei.	She was born in Berlin, but her parents come from Turkey.
Positive Response and correct	aus Berlin.	from Berlin.
Negative Response and blend	aus einem kleinen Dorf in der Türkei	from a small village in Turkey
Negative Response and missing concept	Not answered	Not answered

Table 4.3: A sample example of CREG-TUE dataset with sample text, question, reference answer, and various students' responses

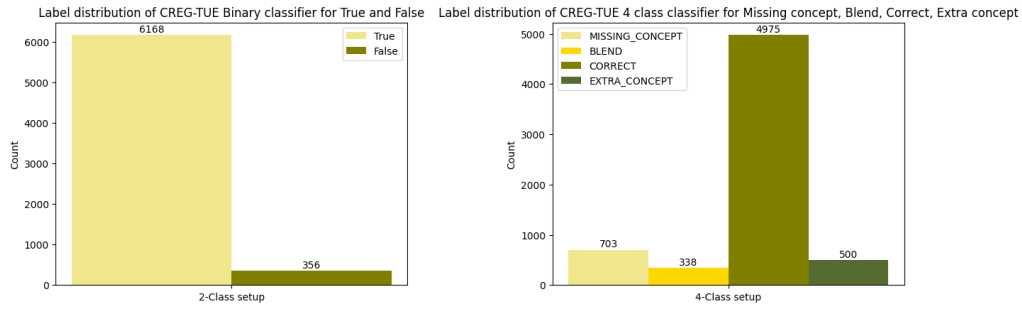


Figure 4.5: Label distributions for binary and four class classification of CREG-TUE dataset. For binary classifier, *True* refers to cases where a student could provide information in response properly, while *False* refers to cases where a student could not. For the four class classifiers, the labels explain the amount of information provided in the responses.

	No. of Student	No. of Response	Avg. No. of Response per student	Avg. No. of Words per Response
All	100	6516	65.16	7.01
2-Class setup				
True	100	6168	61.68	7.02
False	80	356	4.45	6.70
4-Class setup				
Extra Concept	92	500	5.43	13.76
Correct	100	4975	49.75	6.3
Blend	89	338	3.80	9.58
MissingConcept	95	703	7.40	5.63

Table 4.4: CREG dataset details

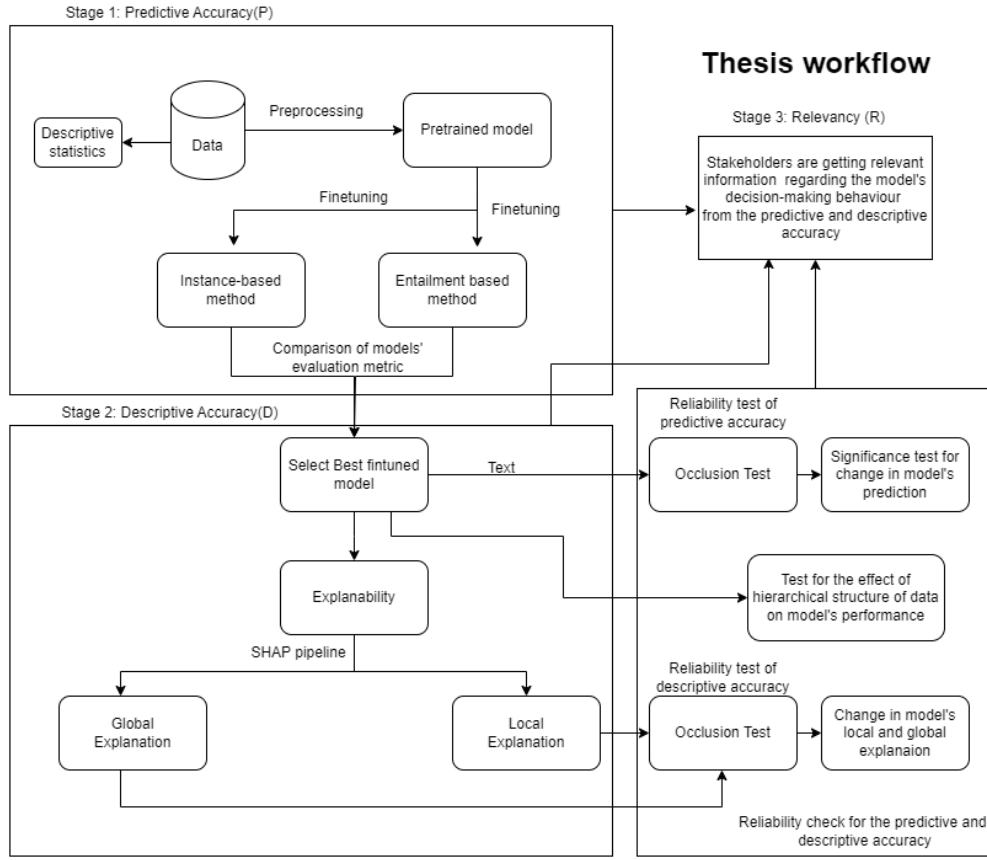


Figure 4.6: Thesis workflow

we have instantiated each one of the stages of the PDR framework.

4.2.1 Predictive Accuracy

As discussed in the previous chapter, we fine-tuned two German language models (Chan et al., 2020), namely *GBERT* and *GELECTRA*, which are *BERT* (Devlin et al., 2018) and *ELECTRA* (Clark et al., 2020) versions respectively, pre-trained exclusively on German data. Additionally the multilingual model, *XLM-RoBERTa* is fine-tuned. Both the base and the large variants are used for all the models. Although the base model contains 12 transformer layers, the large model has 24 transformer layers and doubles memory consumption compared to the base model. We have used PyTorch¹ for the implementation. We incorporated the deepsets pretrained *GermanBERT base*, *GermanBERT large*, *GermanELECTRA base*, *GermanELECTRA large* models and the Facebook AI's *XLM-RoBERTa base* and the *XLM-RoBERTa large* model from the Hugging Face². The input

¹<https://pytorch.org/docs/stable/index.html>

²<https://huggingface.co/>

data is pre-processed before using for fine-tuning. We have removed the responses which do not have proper labels or improper entries. Inspired by previous work in short answer scoring, we explored both instance-based and entailment-based scoring. While, for instance-based scoring, models are only given a student response as input, for entailment-based scoring, the models are also given the corresponding sample reference solution appended with the respective student response. The pre-processed reference solution and the responses are appended along with the [SEP] token. The appended text is sent as input to the models. We train distinct models for each of the skills in the AFLEK dataset. For the AFLEK dataset, this results in thirty-six experiments for each combination of the three skills, two different setups, two different models and their different model sizes. They are trained and evaluated in a 5-fold cross-validation setup using scikit-learn library³. The optimizer used is AdamW (Loshchilov & Hutter, 2019). Each model is trained for three epochs with a learning rate 1e-5, a weight decay of 1e-2, warm up steps of 400 and a batch size of 3. Also, mixed precision training and gradient accumulation are used for memory optimization during the training. We have used Google Colab's Tesla T4 GPU⁴ to fine-tune our models. We have incorporated accuracy, precision, recall, and F1-score from the scikit-learn library as evaluation metrics to compare the model performances.

For secondary evaluations on the CREG-TUE dataset, we carry out twenty-four additional experiments to assess each combination of model types, their size, two scoring type and two different class setups.

To ensure the reliability of our predictive accuracy, we have conducted an occlusion test for the skill *Constructing Explanation* and *Analyzing data*. This study requires a marked evidence span in the dataset for each instance. We have this annotated evidence span only for the two skills mentioned above. Unfortunately, due to the absence of the evidence span, we could not perform the occlusion study on *Planning investigation* skill. Table 5.5 illustrates sample occluded example for both the skills.

For this study, we have generated a new version of the dataset. We have occluded the evidence span by "MASK" ing it. We have received the token-wise evidence span as an XML file for the positively classified responses. The responses are tokenized, and if the corresponding skill label is 1 (positively classified responses), then the tokens, if considered as evidence, are denoted as 1 or 0 otherwise. All of the tokens for the negatively classified responses are marked as 0. We checked if the token is denoted as 1, and then we replaced that token with the [MASK] token. We kept all the other tokens as they were. Finally, we saved all the newly generated occluded positive and non-occluded negative responses for use in the study. We used the same fine-tuned model and passed the occluded data to check for model performance change. Moreover,

³<https://scikit-learn.org/stable/>

⁴https://colab.research.google.com/github/d2l-ai/d2l-tvm-colab/blob/master/chapter_gpu_schedules/arch.ipynb

we have performed a statistical significance test using the GLMER model to check the difference in model performance.

4.2.2 Descriptive Accuracy

As discussed in the background section, to incorporate the transformer's explainability, the game theoretic method named "SHAP" (SHapley Additive exPlanations) Lundberg & Lee (2017) is used. The way it operates is by assigning a SHAP value, which is a measure of the average impact of a feature on the output of the model. It also explains how every feature in the input data is attributed. When utilizing transformer models for automatic skill identification tasks, we have utilized SHAP values to explain how specific words or phrases influence the model's prediction. We have used the SHAP⁵ package to implement it which employs "partitioning" to calculate SHAP values for a transformer model. Partitioning requires gradually breaking the input text into smaller chunks and calculating the impact of each piece on the output of the model. These steps are repeated until every word in the text has been found. There are several ways to display the SHAP values for a specific text instance, including force plots, bar charts, and text plots. Using these visualizations, one may determine which words or phrases are most crucial to the text and how the model predicts that word.

We have picked up the best-performing model based on predictive accuracy for each skill and fitted it in the SHAP explainability pipeline to get global and local explanations. Here, we trained the SHAP explainer using the non-occluded original dataset.

- **Local explainability/ Interpretability:** To demonstrate the explanation at the prediction level to elucidate the model understanding locally for a specific prediction, we have used instance-wise text, force and waterfall plots. Also, we illustrated the additive attribution of the SHAP method using the waterfall plot for each skill.
- **Global explainability/ Interpretability** For a global explanation or a dataset-level explanation, we have plotted the features with the highest positive and the lowest negative SHAP values for each skill and provided a detailed understanding. As each student's responses have varying lengths, the SHAP explainer generates SHAP vectors length-wise. This SHAP vector contains three arrays. The first array contains the base value of each instance, the second array contains arrays of the SHAP value of each feature in an input test, and the third array also contains arrays of feature names. To calculate the global positive and the negative feature contribution, we padded the SHAP array of each feature to the highest length. Then, we collected all the features with SHAP values and compared their positive and negative contribution. Then, the positive and negative SHAP features are sorted according to values in increasing order for plotting.

⁵<https://shap.readthedocs.io/en/latest/>

- **Reliability test for Descriptive accuracy:** Similarly to check the reliability of the descriptive performance of the model, we have used the global and local explainability. The same fine-tuned model is used to fit in the SHAP explainer pipeline. However, the SHAP explainer is trained with the occluded dataset to check the change in the model prediction. We further checked the global and local explanations using the occluded dataset.

4.2.3 Relevancy

Here we further discuss, whether the interpretability of the predictive and descriptive accuracy is relevant to the stakeholders. The model is relevant to the stakeholders if the reliability of the predictive and descriptive accuracy of the model is higher and if it provides a clear interpretation of the model's decision-making behaviour. In both the predictive and descriptive sections, we performed the reliability test. Furthermore, to get different plausible effects on the model's behaviour, we conducted additional analysis.

As described in the background section, we have implemented a generalized linear mixed effect model (GLMER) (Bates et al., 2015) for investigating further to understand whether the hierarchical questions and students affect the models' performance. We have implemented three models with random intercept and random slope.

```
question_model <-- glmer(correct ~ model + (1 | question_id), data = df, family
  = binomial)

student_model <-- glmer(correct ~ model + (1 | question_id) + (1 | student_id),
  data = df, family = binomial)

full_model <-- glmer(correct ~ model + (1 | question_id) + (1 + model |
  student_id), data = df, family = binomial)
```

- The `question_model` includes the random intercept for the question to check if, for different questions, the models might be more likely to be correct in predicting the label.
- The `student_model`, along with various questions, accounts for the variability of different students on the models' performance by including additional random intercepts for students.
- Lastly, the `full_model` is the most complex one. It includes a random slope for models under students to check whether some models might have an easier time predicting the responses of specific students being correct.

We have used here the binomial family in the GLMER model because our response variable is binary, and the error distribution is not Gaussian, it is binomial. In the

GLMER model, we have used logit as a link function which allows prediction of response variables with non-Gaussian distribution.

The model was implemented in R studio using the `glmer` function from the `lme4` package in R (R Core Team, 2024) with the `lmer` 4_1.1-35.5 version and used R version 4.4.1.

Chapter 5

Results

In this chapter, we will discuss the results for each module of the PDR framework. The section is organized as follows:

- **5.1 Predictive accuracy:** Here we discuss the results of different models using the AFLEK and the CREG-TUE datasets.
- **5.2 Descriptive accuracy:** Here we will discuss the transformer’s explainability for each skill, *Constructing Explanations*, *Analyzing Data* and *Planning Investigations*. For each skill, the local and global explainability is discussed separately.
- **5.3 Reliability:** Here we will discuss the reliability of our predictive and descriptive accuracy of the models. We will demonstrate the different results we get by performing the occlusion test for the skills *Constructing Explanations* and *Analyzing Data*. Furthermore, for each skill, we have demonstrated the occlusion test on predictive and descriptive accuracy.
- **5.4 Relevancy:** Here we have discussed the other possible interpretation of the models to provide relevant information regarding models’ behaviour. For this reason, here we assessed if there is any effect of hierarchical data on models’ performance. We performed skill-wise analysis.

	Skills	Instance						Entailment					
		GBb	GEb	XRb	GBl	GEI	XRI	GBb	GEb	XRb	GBI	GEI	XRI
Accuracy	C	84.72	81.99	83.38	84.74	82.30	79.41	84.10	81.22	79.98	85.36	82.73	79.63
	A	83.54	79.21	80.86	82.41	81.37	80.04	81.59	79.11	79.11	80.36	81.43	79.12
	P	84.51	78.06	73.54	85.16	86.45	81.93	79.35	75.48	72.90	84.51	82.58	70.96
Precision	C	85.88	84.45	84.35	86.04	84.44	80.19	86.29	84.01	82.07	86.48	83.60	80.08
	A	84.99	79.19	80.97	84.46	84.42	80.20	83.35	79.11	79.11	84.09	82.93	79.12
	P	83.85	77.52	70.58	84.88	86.87	74.08	82.22	73.98	82.07	86.39	82.50	80.08
Recall	C	94.71	92.59	94.88	94.48	95.28	95.55	93.07	92.02	93.18	95.00	96.10	97.25
	A	96.35	1.0	99.34	95.31	94.15	99.18	95.79	1.0	1.0	94.93	96.61	1.0
	P	92.82	93.34	1.0	93.14	93.83	74.95	85.47	95.34	96.25	90.82	93.23	63.10
F1-Score	C	90.05	88.25	89.30	90.05	89.51	87.41	89.53	87.73	87.20	90.48	89.18	87.80
	A	90.24	88.38	89.15	89.51	88.87	88.67	89.11	88.31	88.33	90.63	89.16	88.30
	P	87.89	84.12	82.32	88.43	89.78	74.34	83.40	82.73	81.16	87.16	87.12	65.78

Table 5.1: Results for the AFLEK dataset. The abbreviated model names such as **GBb**, **GEb**, **XRb**, **GBI**, **GEI**, **XRI**, correspond to *GBERTbase*, *GELECTRAbase*, *XLm-RoBERTabase*, *GBERTlarge*, *GELECTRALarge* *XLm-RoBERTAlarge* respectively. In terms of skills, **C** corresponds to *Constructing Explanations*, **A** corresponds to *Analyzing Data*, and **P** corresponds to *Planning Investigations*. **Bold indicates the best model within a particular method and the bold italic indicates the best model among the methods.** The green colour indicates the larger version of the model has performed better than its base counterpart and for the red colour, it is vice versa.

5.1 Predictive Accuracy:

5.1.1 Evaluation on AFLEK dataset

Table 5.1 presents the results of all the finetuned models using the AFLEK dataset, where the best performance within one scoring mechanism is highlighted in bold and the best between both scoring mechanism is highlighted in bold italic. The *GBERT large* model is the most accurate in identifying the *Constructing Explanations* skill for both the instance and entailment-based scoring whereas the *GBERT base* model (marked in bold italic) is the most accurate among the scoring for the skill *Analyzing Data*. For the *Planning Investigations* skill, the larger models(*GBERT-large*, *GELECTRA-large*) have better accuracy for both scoring, with one exceptional case with *XLm-RoBERTa large* in the entailment-based scoring.

Furthermore, we can see in the figure 4.1 that for all the skills, the number of positive responses is larger than the number of negative responses. To compare the performance among the models, accuracy alone cannot be relied on as an evaluation metric because of the majority class influence. For this reason, other evaluation metrics such as precision, recall, and their harmonic mean F1-score have been considered and reported here.

From the table 5.1, we can observe three pairs of models (*GBERT*, *GELECTRA*, *XLm-RoBERTa*) where each pair contains its base and the large counterpart. Those are finetuned following two scoring (Instance and entailment-based) for the three skills (*Constructing Explanations*, *Analyzing Data* and *Planning Investigations*). These lead to 18 pairs of base model vs. large model comparison. The larger models (*GBERT-large*, *GELECTRA-large* and *XLm-RoBERTa large*) consistently achieved at least equal or better F1-score than their

base counterparts in 12 (marked in green) out of 18 pairs. This indicates that the increased model capacity probably allows for a more comprehensive understanding of task complexities and leads to more accurate predictions. On the other hand, the *XLM-RoBERTa base* model outperformed the *XLM-RoBERTa large* model in 5 out of 6 (marked in red) remaining pair cases. Mostly, these 5 pairs involve identifying *Analyzing Data and Planning Investigations* skills. From table 4.2, it is observed that the length of the responses in *Analyzing Data* is relatively shorter than those written by the students in the *Constructing Explanations* skill. These relatively shorter response lengths for the skill could result in less contextually dense input for the models to process. The *XLM-RoBERTa base* model with fewer parameters can plausibly generalize from these shorter texts better than the *large* model and is less prone to overfitting. It is observed from the dataset that the *Planning Investigations* skill presents unique challenges for the models. The responses in this skill have a particular template-like format, and the students must fill in numerical values observed in an experiment. Table 5.3 illustrates a sample response of this skill. The repetitive text and inconsistent numerical data likely pose a substantial challenge for all models, particularly for *XLM-RoBERTa large*. Being pretrained on multilingual data, *XLM-RoBERTa* finds it more difficult to generalize these linguistic patterns of the German data than the other models, which are particularly pretrained with the German data, resulting in overfitting and decreased performance. These problems are exacerbated by the bigger model’s greater capacity, which makes it more likely to overfit when exposed to numerically focused input with a repetitious single language dataset.

The choice of model and scoring mechanism is a crucial factor in achieving high performance. *GBERT-large*, in an entailment-based scoring, consistently performed best in identifying two out of three skills and therefore we will use this model for all further experiments of descriptive accuracy for these two skills. This underscores the importance of the model’s approach in achieving high performance. Figure 4.1 reveals a positive class bias across all the skills in the AFLEK dataset. This class imbalance might lead models to prioritize achieving high recall to capture as many positive instances as possible. This could explain the generally higher recall scores across all models and skills. Notably, the entailment-based scoring results in more precise models, while the instance-based scoring leads to higher recall.

5.1.2 Secondary evaluation using CREG-TUE dataset

Table 5.2 presents the results of all the finetuned models using the CREG-TUE dataset, where the best performance within the scoring mechanism is highlighted in bold, and the best between both scoring mechanisms is highlighted in bold italic. The table shows that the two-class setup models are more accurate in predicting the correct responses than the models in the four-class setup. Mostly, in the two-class setup, the entailment-based models have slightly outperformed the instance-based models, whereas, in the four-class

2 Class Setup												
	Instance						Entailment					
	GBb	GEb	XRb	GBI	GEI	XRI	GBb	GEb	XRb	GBI	GEI	XRI
Accuracy	95.63	95.30	94.71	95.12	95.57	94.54	96.91	96.45	95.60	96.79	96.62	96.07
Precision	96.51	95.46	95.13	95.80	96.39	95.09	97.24	96.38	94.42	98.18	96.72	94.82
Recall	98.98	99.77	99.49	99.18	99.02	1.0	99.54	96.45	95.60	98.42	96.62	96.07
F1-score	97.72	97.57	97.26	97.46	97.68	97.19	98.37	96.40	94.92	98.30	96.62	95.38

4 Class Setup												
	Instance						Entailment					
	GBb	GEb	XRb	GBI	GEI	XRI	GBb	GEb	XRb	GBI	GEI	XRI
Accuracy	78.72	76.35	76.54	77.24	79.66	76.51	83.39	82.96	78.19	84.17	83.85	82.58
Precision	74.63	76.35	60.98	62.09	79.66	62.03	83.45	82.96	71.37	83.80	83.85	82.57
Recall	78.72	58.31	76.54	77.24	72.89	76.51	83.39	83.24	78.19	84.17	83.44	82.58
F1-score	73.76	66.12	67.43	68.38	75.67	67.56	83.31	81.65	73.84	83.70	83.41	82.10

Table 5.2: Results for the two- and four-class evaluation setups of the CREG-TUE dataset. The abbreviated model names such as **GBb**, **GEb**, **XRb**, **GBI**, **GEI**, **XRI**, correspond to *GBERTbase*, *GELECTRAbase*, *XML-RoBERTabase*, *GBERTlarge*, *GELECTRALarge*, *XML-RoBERTalarge* respectively. In terms of Class Setup, **2 Class Setup** corresponds to *Binary Classification*, and **4 Class Setup** corresponds to *4-Class Classification*.

setup, the entailment-based models have notably performed better than the models in the instance-based method. In both two-class and four-class setups, the entailment-based models, i.e., the *GBERT base* and the *GBERT large* model are the most accurate respectively. Furthermore, the two-class classifier setup achieved considerably higher F1 scores, mainly than the four-class setup for all models (bold entries), which is an expected result. While the instance-and entailment-based scoring achieves similar results for the two-class setup, using the entailment-based scoring leads to notably higher results in the four-class setup. F1 scores dropped considerably in the four-class setup, with all models falling below 84.00. *GBERT-large* emerged as the best-performing model. This highlights the increased difficulty for models to accurately classify the responses in four-class categories.

With a few exceptions for both datasets, larger models outperformed base models across most skills (AFLEK) or classifier setups (CREG-TUE). Both datasets exhibit a positive class bias, which might explain the generally higher recall scores observed across models and skills/setup in most cases. The impact of data size was more evident in AFLEK. Skills with less data (e.g., *Planning Investigations*) seemed to benefit more from large models. The CREG-TUE experiment demonstrates the prominent effect of classifier complexity. The four-class setup proved more challenging for all models than the two-class setup in AFLEK, where skills have varying difficulty levels. While a single best model or scoring can not be determined, the overall trend suggests that prioritizing large models (e.g. *GBERT-large*, *GELECTRA-large*) and Leveraging the entailment-based scoring whenever possible can enhance the possibility of predicting precise class.

5.2 Descriptive Accuracy

5.2.1 Transformer's Explainability

As an initial analysis, positively and negatively classified students' responses only for the *Constructing Explanations* skill were passed through the SHAP pipeline separately. The pretrained GBERT base model with randomly initialised weight was used as the classifier.

Subsequently, we advanced our analysis by replacing the pretrained model with the finetuned models in the SHAP explainability pipeline. This is done for each skill individually to provide a more accurate skill-specific explanation. The features with the highest positive and the lowest negative SHAP values are plotted to obtain the global explanations for each skill. We have used the text, force and waterfall plot to demonstrate the local explainability instance-wise. Table 5.3 shows the details of an instance belonging to a positive class and an instance belonging to a negative class for all the skills. The table contains the original German and translated English versions of the sample question, solution, a response classified as positive for the sample question mentioned, and a negatively classified response.

5.2.1.1 Skill: Constructing Explanations

The table 5.1 from the previous section shows that the *GBERT large* model in entailment-based scoring with greater contextual input, achieved higher F1 scores. We have selected this finetuned model to fit SHAP explainer pipeline for the SHAP partition training.

- **Local Explainability:**

The table 5.3 illustrates one response belonging to a positive class and another to a negative class for the same question for the *Constructing Explanations* skill. It shows the students are asked to construct an explanation regarding their observations about a video shown to them. For the positive response of the student properly demonstrates his/her explanation regarding the observation of the video. The important part of the response that led it to be classified as a positive instance by the human annotator, i.e. the evidence spans, are marked in bold in the table. The annotators have denoted those parts following certain guidelines during the annotation phase. The text plot 5.1 reveals how each feature in the text instance contributes to the model's prediction. This figure contains two text plots, i.e. a text plot for a positive response and a text plot for a negative response for the same question. The plot, on top of the text, explains the positive response. From the top plot, we can see that the model's base value ϕ_0 is -5.35358 (equation reference 2), which denotes the average logit output of the model when there is

Constructing Explanations		
	Original Text	Translated Text
Sample Question	Notiere die anschließend deine Beobachtungen unter dem Video.	Then note your observations below the video.
Solution	In dem Video ist ein Laptop zu sehen , auf dem ein Spiel läuft . Nach einiger Zeit erscheint die Meldung , dass eine Überhitzung erkannt wurde und der Laptop gleich in einen Standby Modus wechselt . Dabei ertönt ein lautes wiederkehrendes Signal . Nach Ablauf weniger Sekunden geht der Bildschirm aus und das Signal hört auch auf .	The video shows a laptop running a game. After a while, a message appears saying that overheating has been detected and the laptop will go into standby mode. A loud, recurring signal is heard. After a few seconds, the screen goes off and the signal stops.
Response classified as positive	Zuerst wird ein Videospiel gespielt und dann kommt die Benachrichtigung , dass einer Überhitzung erkannt wurde . Anschließend schaltet sich der Laptop automatisch aus .	First a video game is played and then a notification appears that overheating has been detected . The laptop then switches off automatically.
Response classified as negative	Der Laptop verhielt sich anfangs recht gut leider sagt der er uns nicht , ob der heiß gelaufen ist und ob das Programm überhaupt lief. Aber trotzdem kann man sagen , dass der Laptop ein Signal geschickt das er es nicht mehr halten kann und schließt das Programm automatisch.	The laptop behaved quite well at first, but unfortunately it doesn't tell us whether it was overheating or whether the program was even running. But you can still say that the laptop sent a signal that it couldn't handle it any longer and closed the program automatically.
Analyzing Data		
	Original Text	Translated Text
Sample Question	1. Was beobachtest du wenn der Leiter stromdurchflossen ist? 2. Erkläre deine Beobachtung mithilfe von Energiebegriffen. 3. Was bedeutet deine Beobachtung für den heiß werdenden Laptop?	1. What do you observe when current flows through the conductor? 2. Explain your observation using energy terms. 3. What does your observation mean for the laptop that is heating up?
Solution	Wenn der Leiter stromdurchflossen ist, leuchtet die Lampe und die Temperatur T (in $^{\circ}\text{C}$) des Drahtes steigt mit der Zeit t (in s) an. Die Höchsttemperatur liegt bei $x^{\circ}\text{C}$. Wenn ein Leiter stromdurchflossen ist, wird elektrische Energie transportiert. Ein Teil dieser elektrischen Energie wird in thermische Energie umgewandelt. Diese thermische Energie wird durch die erhöhte Temperatur deutlich .	When current flows through the conductor, the lamp lights up and the temperature T (in $^{\circ}\text{C}$) of the wire increases over time t (in s). The maximum temperature is $x^{\circ}\text{C}$. When current flows through a conductor, electrical energy is transported. A part of this electrical energy is converted into thermal energy. This thermal energy is evident from the increased temperature.
Response classified as positive	Bei dem offenen Stromkreis hat man gesehen das die Temperatur die ganze Zeit bei dem selben wert geblieben ist und bei dem geschlossenen Stromkreis ist die Temperatur gestiegen was daran lag dass der Stromkreis geschlossen war	With the open circuit, it was seen that the temperature remained at the same value the whole time and with the closed circuit, the temperature rose , which was due to the fact that the circuit was closed.
Response classified as negative	Wenn ein Leiter	If a conductor
Planning Investigations		
	Original Text	Translated Text
Sample Question	Notiert euch die Werte in der unten stehenden Tabelle.	Write down the values in the table below.
Solution	Werte für Zeit und Temperatur .	Values for time and temperature .
Response classified as positive	VERSUCHSERGEBNISSE:OFFENER STROMKREIS ZEIT T IN S TEMPERATUR T IN $^{\circ}\text{C}$ 10 20,6 20 20,4 30 20,2 40 20,1 50 20,0 60 19,8 70 19,7geschlossenener Stromkreis ZEIT T IN S TEMPERATUR T IN $^{\circ}\text{C}$ 1028,62030,63032,64031,65031,46035,27036,8	TEST RESULTS: OPEN CIRCUIT TIME T IN S TEMPERATURE T IN $^{\circ}\text{C}$ 10 20.6 20 20.4 30 20.2 40 20.1 50 20.0 60 19.8 70 19.7 CLOSED CIRCUIT TIME T IN S TEMPERATURE T IN $^{\circ}\text{C}$ 1028.6 2030.6 3032.6 4031.6 5031.4 6035.2 7036.8
Response classified as negative	VERSUCHSERGEBNISSE:OFFENER STROMKREIS ZEIT T IN S TEMPERATUR T IN $^{\circ}\text{C}$ geschlossenener Stromkreis ZEIT T IN S TEMPERATUR T IN $^{\circ}\text{C}$	TEST RESULTS: OPEN CIRCUIT TIME T IN S TEMPERATURE T IN $^{\circ}\text{C}$ CLOSED CIRCUIT TIME T IN S TEMPERATURE T IN $^{\circ}\text{C}$

Table 5.3: Student's response instance for the skill *Constructing Explanations*, *Analyzing Data* and *Planning Investigations*

no input feature. This serves as a reference point from which the contributions of individual tokens are assessed. The model's output logit value $g(z')$ is 5.25248, the sum of the base value ϕ_0 and the SHAP value contribution of each token ϕ_i in the text. Negative SHAP values cause the logit value to decline, while positive SHAP values push the prediction towards the positive class by increasing the logit value. The colour gradient of the red and blue text chunks represents the magnitude of the impact of the corresponding words. Darker red and blue impact the model's output, while the lighter portions contribute less. From the table 5.3 we can see the evidence span of this instance in bold. The evidence span is "Benachrichtigung, dass einer Überhitzung erkannt wurde". If the model's learning for classifying this response as positive matches the human intuition, then the SHAP explainer should assign positive SHAP values to the features in the evidence span. By comparing the evidence span and the text plot 5.1, it can be observed that there is a noteworthy overlap between the features that are explained as important by the model using the positive SHAP values and the evidence span. Some highly contributing features are "hi"(9.893), "Über"(0.039), "tz"(0.051) of the word Überhitzung, "Benach"(0.03), "richtung"(0.07), "dass"(0.097). Unfortunately, the model also learned some shortcuts and could not learn some part of the evidence as important, for example, "einer"(-0.17) "ung"(-0.099) of the word Überhitzung, erkannt(-0.19), wurde(-0.154). Also, other important features outside of the evidence span, which is generally used for constructing explanations, are "Zuerst", "dann", etc., which also provide positive SHAP values of 0.078 and 0.065 towards the model's decision. This detailed explanation provides a transparent interpretation of why the fine-tuned model decided that the student could construct the explanations required for this question.

The plot at the bottom of 5.1 shows the text explanation of a response classified as negative. This instance's base value is $\phi_0 = -5.3544$. The plot shows that the SHAP explainer has learned to identify the features influencing the negative class prediction and has assigned a SHAP value to each feature accordingly. Despite the presence of a highly positive feature like heiß(1.995), the SHAP explainer produces the logit output $g(z')$ of -3.78299 and classifies this instance as a negative instance.

Figure 5.16 contains the force plots of the same positive and negative instances, which summarize the overall features' contribution linearly. These plots also contain the direction to visualize how the positive and negative features push the model's prediction. The length of each red and blue chunk shows the magnitude of that feature and the color decides the sign and direction. Red being the positive whereas the blue being the negative. The plot at the top explains the positive instance, and the force plot at the bottom explains the negative instance. The model output ($g(z') = 5.25, -3.78$) in the force plot is slightly different than the text plot ($g(z') = 5.25248, -3.78299$) due to the rounding of the value.



Figure 5.1: Constructing Explanation: Text plot
Top: Response belongs to positive class, Bottom: Response belongs to negative class

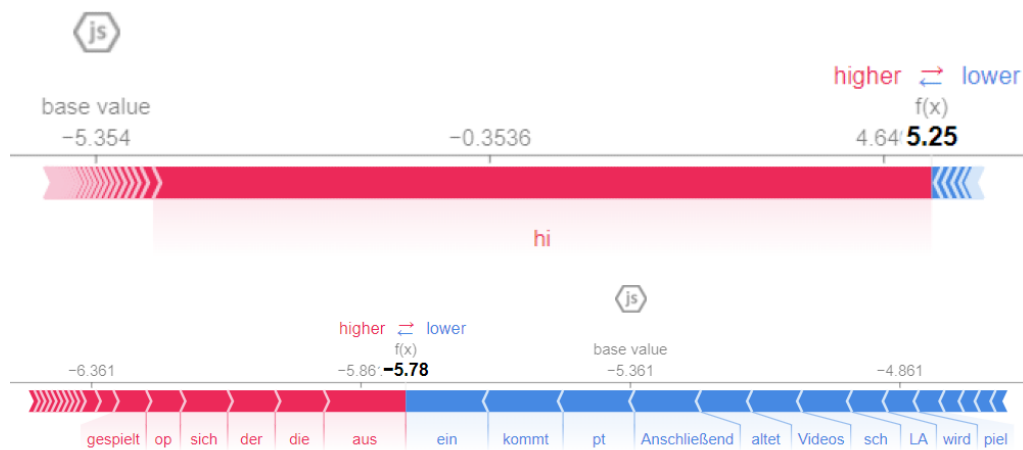


Figure 5.2: Constructing Explanation: Force plot
Top: Response belongs to positive class, Bottom: Response belongs to negative class

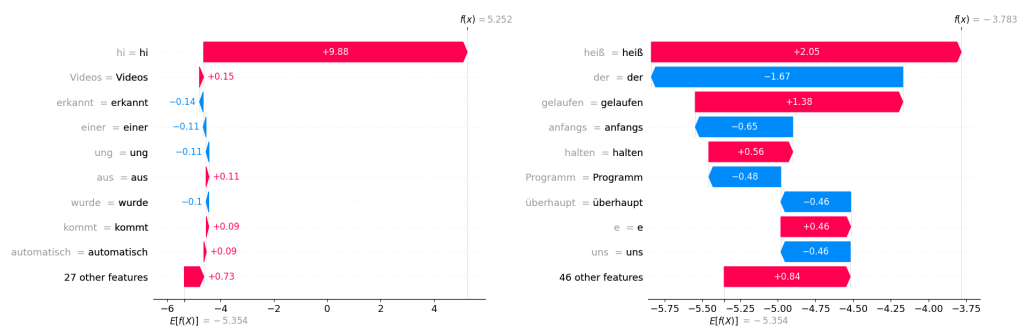


Figure 5.3: Constructing Explanation: Waterfall plot
Left: Response belongs to positive class, Right: Response belongs to negative class

The waterfall plot 5.3 shows detailed additive contributions of the features step by step. This displays the proof of SHAP's Additivity axiom of the additive feature attribution method, which asserts that the difference between the model's prediction for a given instance and the mean predicted over the training dataset equals the sum of the SHAP values for all features. This figure contains the plot of the previous positive response and the negative respectively. The plot on the left explains the instance classified as positive, and the plot on the right explains the negative instance.

Positive Class Instance :

The model output $g(z') = 5.252$

The base value ϕ_0 = represents the average model output over the training dataset: -5.354

Feature Contributions: Each feature's SHAP value (ϕ_i) represents its contribution to the deviation from the base value: ϕ_0

Additive property of the additive feature attribution method SHAP:

$$g(z') = -5.354 + \phi_{hi} + \phi_{Videos} - \phi_{erkannt} - \phi_{einer} - \phi_{ung} + \phi_{aus} - \phi_{wurde} + \phi_{kommt} \\ + \phi_{automatisch} + \phi_{27 \text{ other features}}$$

$$g(z') = -5.354 + 9.88 + 0.15 - 0.14 - 0.11 - 0.11 + 0.11 - 0.1 + 0.09 \\ + 0.09 + 0.73$$

$$= 5.236 \simeq 5.252 \text{ (Model output)}$$

Negative Class Instance :

The model output $g(z') = -3.783$

The base value ϕ_0 = represents the average model output over the training dataset: -5.354

Feature Contributions: Each feature's SHAP value (ϕ_i) represents its contribution to the deviation from the base value: ϕ_0

Additive property of the additive feature attribution method SHAP:

$$g(z') = -5.354 + \phi_{heiß} - \phi_{der} + \phi_{gelaufen} - \phi_{anfangs} + \phi_{halten} - \phi_{Program} - \phi_{überhaupt} + \phi_e \\ - \phi_{uns} + \phi_{46 \text{ other features}}$$

$$\begin{aligned}
g(z') &= -5.354 + 2.05 - 1.67 + 1.38 - 0.65 + 0.56 - 0.48 - 0.46 + 0.46 \\
&\quad - 0.46 + 0.84 \\
&= -3.784 \simeq -3.783 \text{ (Model output)}
\end{aligned}$$

The plots provide a transparent and additive explanation of the model's output for this instance complying the SHAP's axiom.

- **Global Explainability:**

To determine the global explainability of the *Constructing Explanations* skill, the top 60 positive SHAP features and the bottom 60 negative SHAP features in increasing order are calculated from the SHAP explainer. Figure 5.4 plots the feature distribution of the positive and the negative SHAP negative features. From the positive feature distribution, it can be observed that many features related to the energy domain of physics received a higher positive SHAP value. For instance, "strahlung" (> 4.0), "heiß" ($\simeq 12.0$), "hi" ($\simeq 8.0$), "hei" (> 6.0), "heiße" ($\simeq 8.0$), "Heiß" (> 4.0), "warme" ($\simeq 12.0$), "warm" ($\simeq 12.0$), "Sonnenaufgang" (> 3.0), "licht" (3.0) etc. as the responses are collected from the open-ended tasks based on modules of the transformation of the different forms of energy. Also, there are certain features related to direction which received a considerably high positive SHAP value. For instance, "schräg" ($\simeq 12.0$), "Richtungen" (> 8.0), "Isol" ($\simeq 9.0$), "Norden" ($\simeq 8.0$) "positionen" (6.0), "ausgerichtet" (> 5.0), "Richtung" (> 3.0) etc. This illustrates that students explained different energy-related observations using these features.

The lower plot of the figure 5.4 shows the features from the smallest negative SHAP value to the larger negative SHAP value. This gives an overall idea of what features are pushing away the model's decision to classify an instance as positive. From the plot, it can be observed that these negative features are very generalised and provide very little domain-specific information. However, the plot shows, these features mainly describe a situation or can be used to construct certain explanations. For instance, "für" (< -0.001), "anschließend" (> -0.001), "viele" (> -0.001), "dafür" ($\simeq -0.006$), "erstmal" (> -0.003), "hoch" (-0.007) etc. Moreover, the model has also learnt a few words as negative features like "warmen", where the different forms of this same stem are also considered as positive features as described before.

As a global explanation, both of the plot shows how the model has learnt a domain-specific feature importance at a dataset level and explains quite reasonably. This global analysis by the SHAP offers the stakeholders a crucial understanding of the model's decision-making process and trust in this automatic skill identification system.

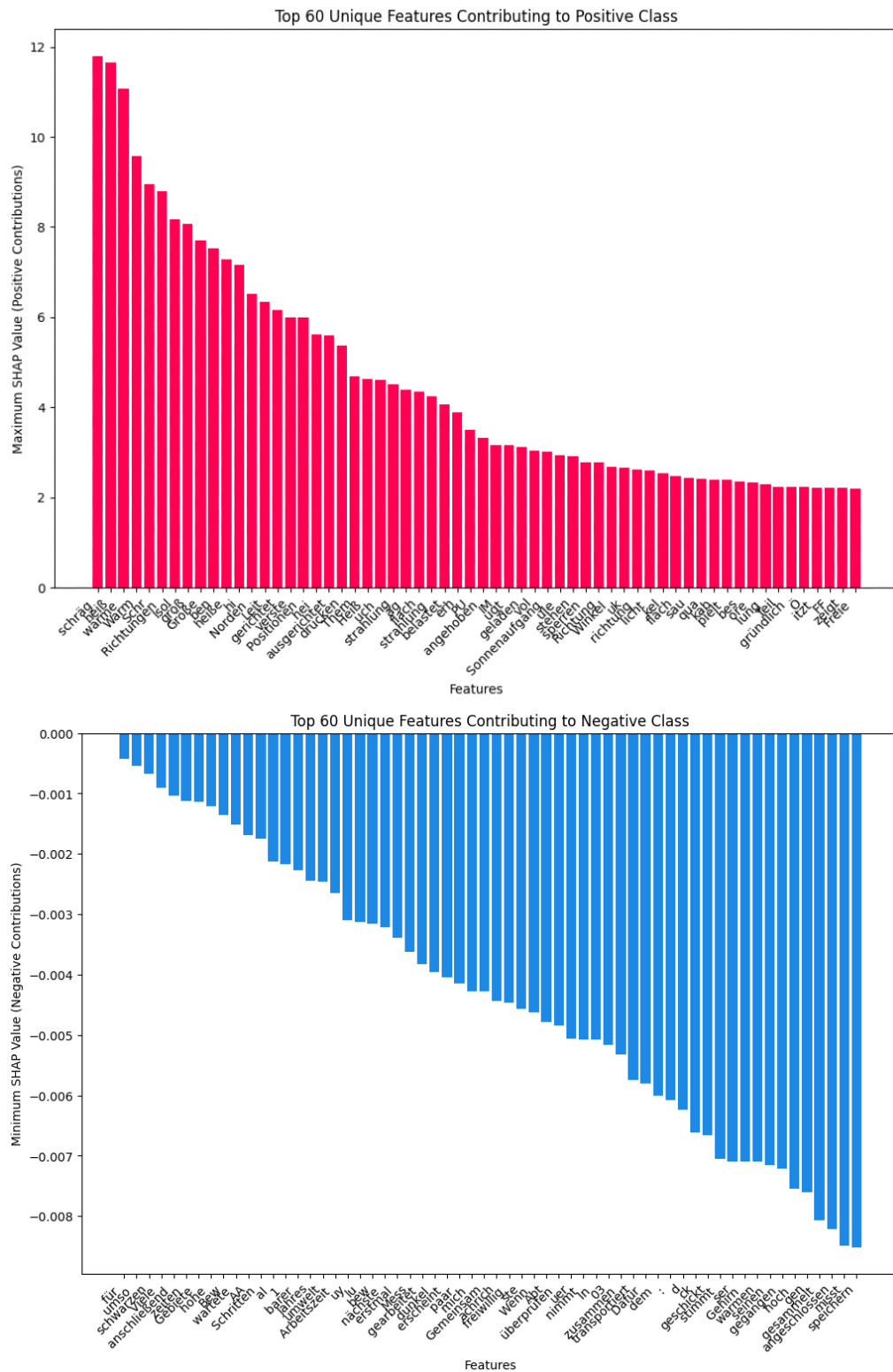


Figure 5.4: Constructing Explanation: Global explanation
Top: Positive features, Bottom: Negative features

5.2.1.2 Skill: Analyzing Data

Table 5.1 from the previous section shows that the *GBERT large* model, which is entailment-based scoring with greater contextual input, achieved higher F1 scores. We have selected this finetuned model to fit the SHAP explainer pipeline for the SHAP partition training.

- **Local Explanability:**

The table 5.3 illustrates one response belonging to a positive class and another to a negative class for the same question for the *Analyzing Data* skill. Here the students are asked to analyse their observations when current flows through the conductor. For the positive response, the student's response properly explains his/her analysis and reports it and for the negative one the student's response is incomplete and does not report the analysis. As explained for the previous skill, the evidence spans are marked in bold for the positive instance. The text plot 5.5 reveals how each feature in the text instance contributes to the model's prediction. This figure contains two text plots, i.e. a text plot for a positive response and a text plot for a negative response for the same question. As explained for the previous skill, the model's base value ϕ_0 for the positive instance is -0.312558 and the model's output logit value $g(z')$ is 5.03812, the sum of the base value ϕ_0 and the SHAP value contribution of each token ϕ_i in the text. Negative SHAP values cause the logit value to decline, while positive SHAP values push the prediction towards the positive class by increasing the logit value. The colour gradient of the red and blue text chunks represents the magnitude of the impact of the corresponding words. Darker red and blue impact the model's output, while the lighter portions contribute less. The evidence span for the instance is **"Bei dem", "das die Temperatur die ganze Zeit bei dem selben wert geblieben ist", "bei dem", "ist die Temperatur gestiegen"** (Taken from table 5.3). If the model's learning for classifying this response as positive matches with the human intuition, then the SHAP explainer should assign positive SHAP values to the features in the evidence span. By comparing the evidence span and the text plot 5.20, it can be observed that there is a noteworthy overlap between the features that are explained as important by the model using the positive SHAP values and the evidence span. Some highly contributing tokens are "Temperatur"(0.193), "die"(0.22), "ganze"(0.202), "Zeit"(0.2020), "gestiegen"(0.15). This detailed explanation provides a transparent interpretation of why our fine-tuned model decided that the student could analyse the data.

The plot at the bottom of 5.5 shows the text explanation of a response classified as negative. Probably the student did not finish the answer as required. This instance's base value is $\phi_0 = -0.472423$. We can observe the presence of a highly positive feature like "Liter"(0.156), "ein"(0.057) and a negative feature "Wenn"(-0.132). The SHAP explainer produces the logit output $g(z')$ of -0.440861 and classifies this instance as negative.

Figure 5.6 contains the force plots of the same positive and negative instances, which summarize the overall features' contribution linearly. These plots also contain the direction to visualize how the positive and negative features push the model's prediction. The length of each red and blue chunk shows the magnitude of that feature and the color decides the sign and direction. Red being the positive whereas blue is the negative. The plot at the top explains the positive instance, and the force plot at the bottom explains the negative instance. The model output ($g(z') = 5.04, -0.44$) in the force plot is slightly different than the text plot ($g(z') = 5.03812, -0.440861$) due to the rounding of the value.

The waterfall plot 5.7 shows detailed additive contributions of the features step by step. This displays the proof of SHAP's Additivity axiom, which asserts that the difference between the model's prediction for a given instance and the mean predicted over the training dataset equals the sum of the SHAP values for all features. This figure contains the plot of the previous positive response and the negative respectively. The plot on the left explains the instance classified as positive, and the plot on the right explains the negative instance.

Positive Class Instance :

The model output $g(z') = 5.038$

The base value $\phi_0 = -0.313$

Feature Contributions: Each feature's SHAP value (ϕ_i) represents its contribution to the deviation from the base value: ϕ_0

Additivity Property:

$$g(z') = -0.313 + \phi_{\text{dass}} + \phi_{\text{ist}} + \phi_{\text{lag}} + \phi_{\text{kreis}} + \phi_{\text{Strom}} + \phi_{\text{daran}} + \phi_{\text{die}} + \phi_{\text{ganze}} \\ + \phi_{\text{Temperatur}} + \phi_{\text{32 other features}}$$

$$g(z') = -0.313 + 0.26 + 0.24 + 0.22 + 0.21 + 0.21 + 0.21 + 0.19 + 0.18 \\ + 0.18 + 3.46$$

$$= 5.047 \simeq 5.038 \text{ (Model output)}$$

Negative Class Instance :

The model output $g(z') = -0.440861$

The base value $\phi_0 = -0.472$

Feature Contributions: Each feature's SHAP value (ϕ_i) represents its contribution to the deviation from the base value: ϕ_0

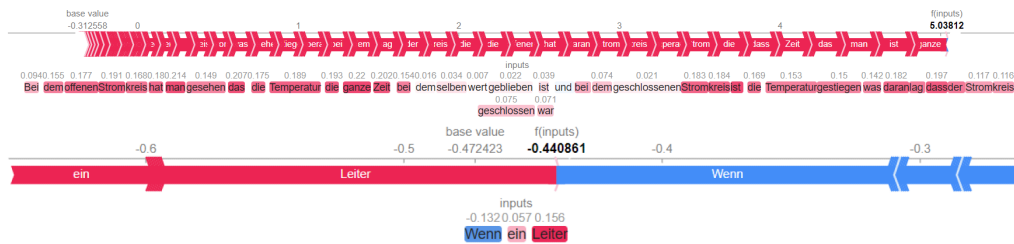


Figure 5.5: Analyzing Data: Text plot

Top: Response belongs to positive class, Bottom: Response belongs to negative class



Figure 5.6: Analyzing Data: Force plot

Top: Response belongs to positive class, Bottom: Response belongs to negative class

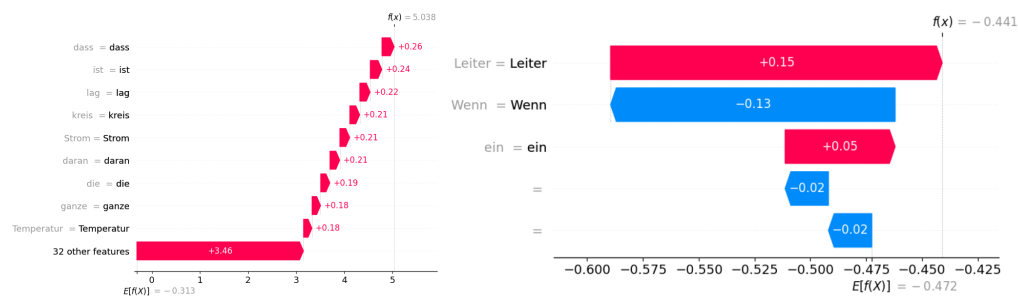


Figure 5.7: Analyzing Data: Waterfall plot

Left: Response belongs to positive class, Right: Response belongs to negative class

Additivity Property:

$$g(z') = -0.472 + \phi_{\text{Leiter}} - \phi_{\text{Wenn}} + \phi_{\text{ein}} - \phi_{=} + \phi_{=}$$

$$g(z') = -0.472 + 0.15 - 0.13 + 0.05 - 0.02 - 0.02$$

$$= -0.442 \simeq -0.441 \text{ (Model output)}$$

The plots provide a transparent and additive explanation of the model's output for this instance complying the SHAP's axiom.

- **Global Explainability:**

To determine the global explainability of the *Analyzing Data* skill, the top 60 features with positive SHAP values and the bottom 60 features with negative SHAP values in increasing order are calculated from the SHAP explainer. Figure 5.8 plots the feature distribution of the positive and the negative SHAP features. From the positive feature distribution and careful observation of the dataset, it is observed that the model has learned the highest contributing features like "emp"(2.00), "fluss"(1.875), etc, from the words like "temperature", "beeinflussen". Students frequently use these to analyze the data. Also, features like "Zeit"(< 1.875), "langsamer"(1.75), "beobachten"(< 1.25), leuchten(< 1.25), "schneller" (< 1.00), "hoher"(< 1.00), "steigen"(< 1.00) etc features are used in the responses by the students to compare data in tables. Also, the model has learned some domain-specific features like "Temperatur"($\simeq 1.0$), "Heiß"(> 0.75), "weiß" (< 0.75), "heiße"(<0.50) etc. as the responses are collected from the open-ended tasks based on modules of the transformation of the different forms of energy.

The lower plot of the figure 5.8 shows the features from the smallest negative SHAP value to the larger negative SHAP value. This gives an overall idea of what features are pushing away the model's decision to classify an instance as positive. From the plot, it is observed that features like "Rest", "C", "ergie", and "Beobachtung" have zero SHAP values, which means these features contribute to neither of the classes for the model's decision. Although these features are meaningful in explaining energy-related concepts in the responses, surprisingly, the explainer does not allocate any SHAP value to them. The plot shows the features with very low negative SHAP values, for instance, "Differenz"(< -0.05), "angeschlossen"(< -0.05), "Grad"(< -0.05), "Experiment"(< -0.05), "bearbeitet"(< -0.05) are used by the students to write responses for analyzing data. This explains why the magnitude of these features is almost near zero if not a positive SHAP value. The features with higher

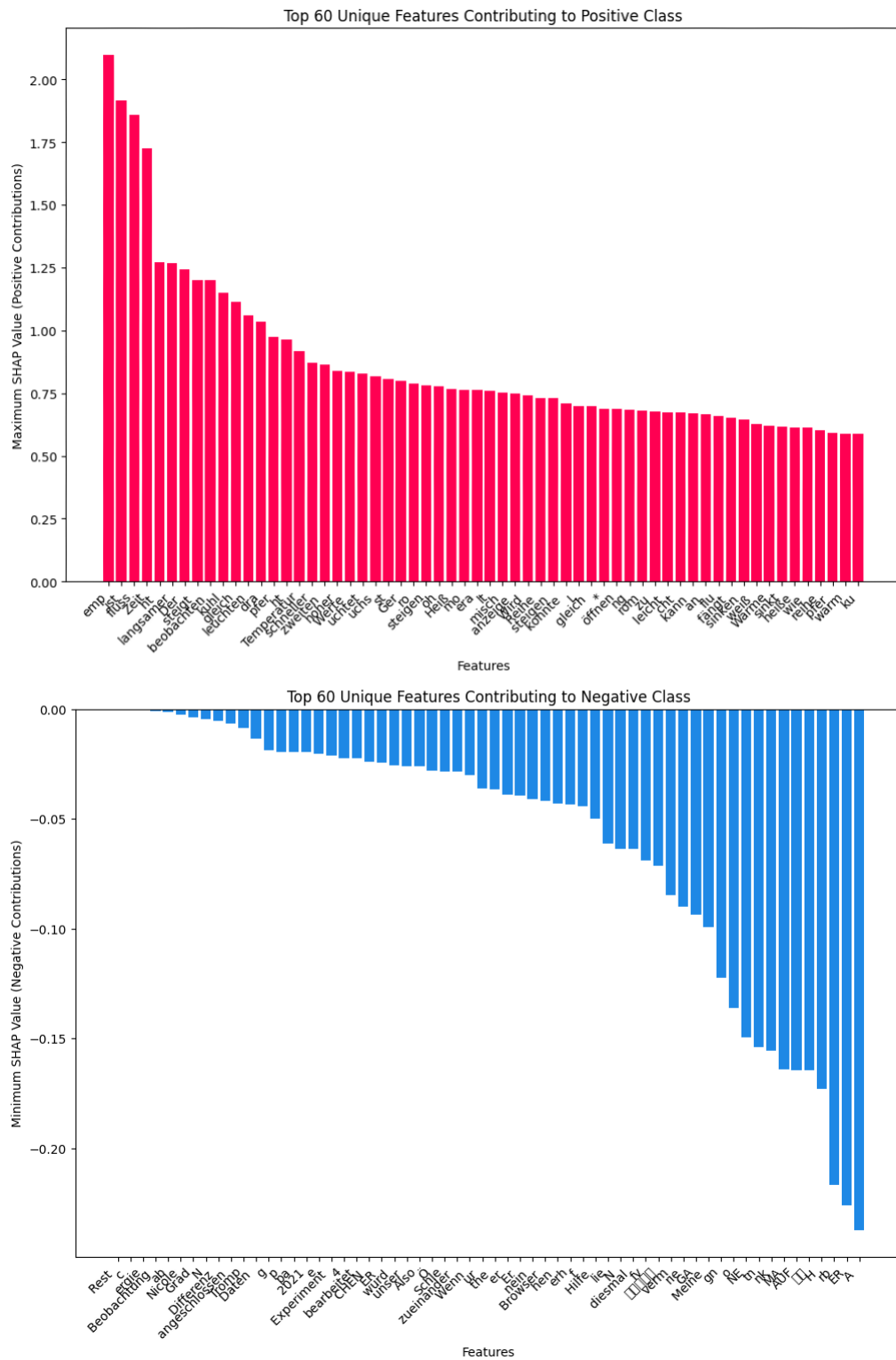


Figure 5.8: Analyzing Data: Global explanation
Top: Positive features, Bottom: Negative features

negative SHAP values are generalized features used in responses. However, for this skill, the model has learnt features like "Beobachtung" as a non-contributing feature, where the different forms of this same stem("beobachten"(≈ 1.25)) are considered features with large positive SHAP values, as described before.

As a global explanation, both plots show how the model has learnt importance at a dataset level and explain reasonably with few exceptions. Overall, the model has learned many features that can be used to compare data as highly contributing features. This global analysis by the SHAP offers the stakeholders a crucial understanding of the model's decision-making process and trust in this automatic skill identification system.

5.2.1.3 Skill: Planning Investigations

As the table 5.1 illustrates, in terms of the F1- score, the finetuned GELECTRA_{large} model has the best predictive power for identifying this skill among the students' responses; we have used this model for the descriptive assessment as well in terms of the local and global explainability.

- **Local Explainability:**

The instances for the skill *Planning Investigations* in the table 5.3 show that the students are asked to investigate an experiment and note the temperature in a particular time interval in a provided table for both open and closed state of a circuit. The positive student's response properly demonstrates his/her investigation regarding his/her observation from the experiment, while the student failed to do so in the negative instance. The previous fig 4.1 shows that the skill *Planning Investigations* has very little data. Also, we have observed from the data that the responses are collected only from one experiment. The annotators have mostly classified an instance as positive if they observed the presence of the correct values in both open and closed circuit states. The exception to this leads to a negatively classified instance. Unlike other skills, the data for *Planning Investigations* is not annotated with the evidence span by the annotators. Due to this, we could not provide the evidence span for the instances. However, we have included a negative instance in the table 5.3 to illustrate this scenario in the various forms of local explainability.

The text plot 5.9 contains a text explanation of two text responses. The instance on the top is positive, while the negative one is at the bottom. Negative SHAP values cause the logit value to decline, while positive SHAP values push the prediction towards the positive class by increasing the logit value. The colour gradient of the red and blue text chunks represents the magnitude of the impact of the corresponding words. Darker red and blue impact the model's output, while the lighter portions contribute less.

The text plot for the positive instance reveals how each feature in the text instance contributes to the model's positive prediction. The model's base value ϕ_0 is -0.145564, which denotes the average logit output of the model when there is no input feature. This serves as a reference point from which the contributions of individual tokens are assessed. The model's output logit value $g(z') = 5.10443$, the sum of the base value ϕ_0 and the SHAP value contribution (ϕ_i) of each feature in the text. This plot shows that the student could analyze the open and closed circuit state temperature at a particular time interval. Although it is difficult to confirm without evidence span, but it can be observed that along with a few other textual features, the SHAP explainer mainly allocates the positive SHAP values to the data the student collected by investigating the experiment.

The model's base value ϕ_0 of the negative instance is 0.102672, and the model's output $g(z') = -1.13309$. Most of the features in the response have a negative SHAP value, indicating the model did not witness the features in this instance which could have pushed the model's prediction towards a positive class.

Figure 5.10 contains the force plot of the same positive(Top) and the negative instances(Bottom), which summarizes the overall features' contribution linearly. It also contains the direction to visualize how the positive and negative features push the model's prediction. The model output $g(z')$ for the positive instance is 5.10 and for the negative instance is -1.13 in the force plot those are slightly different than the output of the text plots(5.10443, -1.13309) due to the rounding of the value.

The waterfall plot 5.11 shows detailed additive contributions of the features step by step. This displays the proof of SHAP's Additivity axiom, as described before. The plot on the left explains the positive class instance, while the plot on the right explains the negative one. Here we will discuss both of the plots.

– Positive Class Instance :

The model output $g(z') = 5.104$

The base value $\phi_0 = -0.146$

Feature Contributions:

Each feature's SHAP value (ϕ_i) represents its contribution to the deviation from the base value: ϕ_0

Additivity Property:

The SHAP values for each feature sum up to the difference between the model output and the base value:



$$g(z') = -0.146 + \phi_8 + \phi_{20} - \phi_{=} - \phi_{PE} + \phi_{,} - \phi_{RA} + \phi_O + \phi_{VER} \\ + \phi_{31} + \phi_{99} \text{ other features}$$

$$g(z') = -0.146 + 1.28 + 1.26 - 1.24 - 1.14 + 1.06 - 1.02 + 0.91 + 0.88 \\ + 0.86 + 2.41$$

$$= 5.114 \simeq 5.104 \text{ (Model output)}$$

– **Negative Class Instance :**

The model output $g(z') = -1.133$

The base value $\phi_0 = 0.103$

Feature Contributions:

Each feature's SHAP value (ϕ_i) represents its contribution to the deviation from the base value: ϕ_0

Additivity Property: The SHAP values for each feature sum up to the difference between the model output and the base value:

$$g(z') = 0.103 - \phi_{RA} - \phi_{RA} - \phi_{IN} - \phi_T - \phi_T - \phi_T - \phi_T - \phi_{\circ} \\ - \phi_{=} - \phi_{99} \text{ other features}$$

$$g(z') = 0.103 - 0.07 - 0.07 - 0.06 - 0.06 - 0.06 - 0.06 - 0.06 - 0.06 \\ - 0.05 - 0.68$$

$$= -1.127 \simeq -1.133 \text{ (Model output)}$$

• **Global Explainability:**

To determine the global explainability of the *Planning Investigations* skill, the top 60 features with positive SHAP values and the bottom 60 features with negative SHAP values in increasing order are calculated from the SHAP explainer. Figure 5.12 plots the feature distribution of the positive and the negative SHAP features. From the positive feature distribution on the top of the figure and careful observation

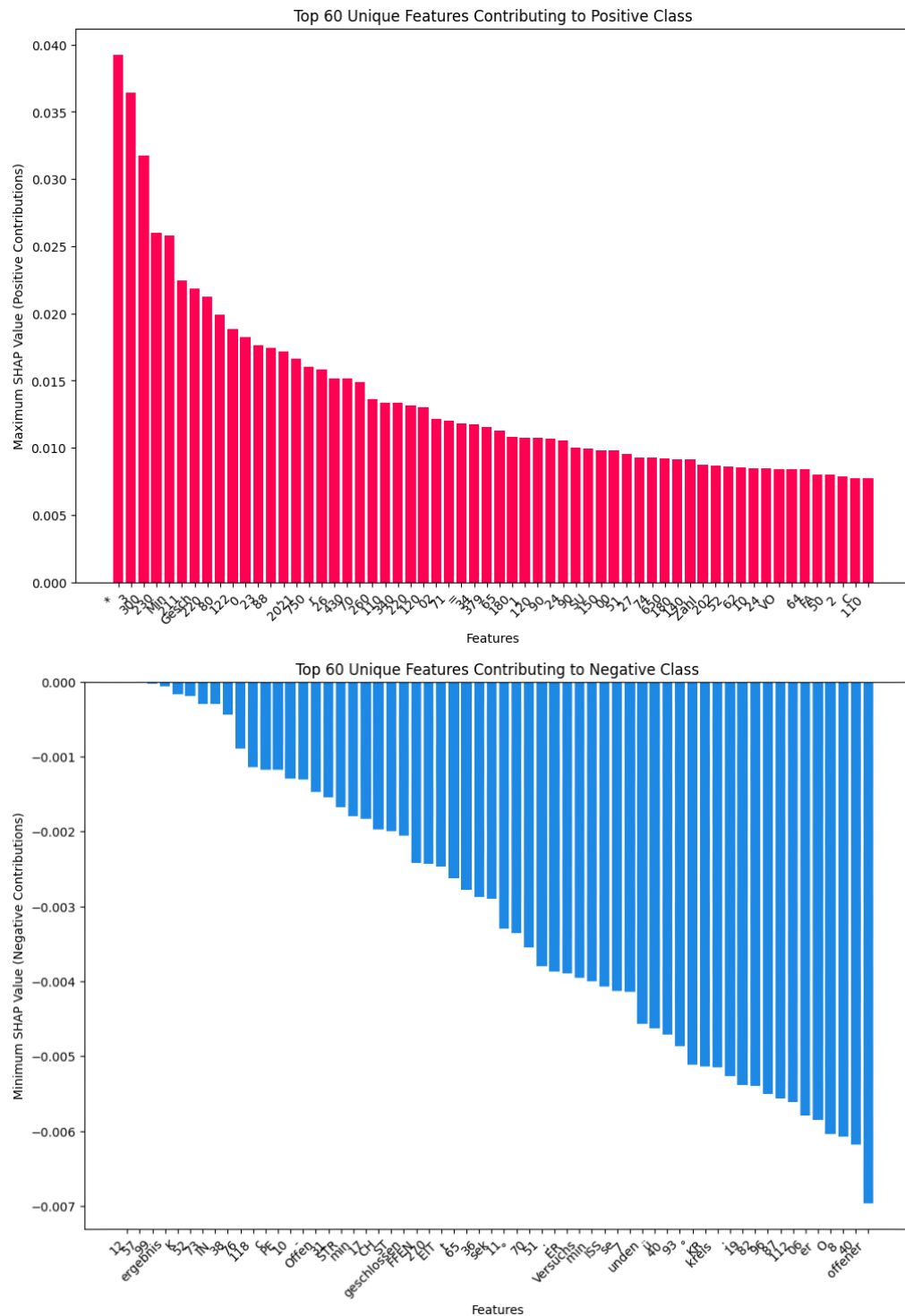


Figure 5.12: Planning Investigation: Global Explainability
Top: Positive features, Bottom: Negative features

of the dataset, it is observed that the model has learned mostly the numbers which are collected from the experiments and a few textual features as the highest contributing features. For instance, "3" ($\simeq 0.040$), "300" (> 0.035), "Min" (0.025), etc., which probably explains the fact that the model is trained with very few instances, and also they have a template-like structure. Only the numbers entered in this template structure make the instances different.

The plot at the bottom of the figure 5.12 shows the features from the smallest negative SHAP value to the larger negative SHAP value. This gives an overall idea of what features are pushing away the model's decision to classify an instance as positive. The plot shows that the pattern of the negative feature distribution is almost complementary to the positive feature distribution. The SHAP explainer allocated the negative SHAP values to the textual features, which are repetitive in nature in each instance due to their template-like structure and probably do not provide any meaningful information to the model to make its decision. Also, it shows few numerical features as negatively contributing features. For instance, "40" (-0.006), "93" (-0.005) etc. Three features, i.e., "12", "57", and "99" have zero SHAP values, implying no influence on the the model's decision.

For this skill, from the table 5.1, we can observe that there is a pattern of declined F1 score than the other skills. From the global explanations, the plots show a pattern of mostly the numerical features as the positive features and mostly the repetitive textual features as negative features. These occurrences probably indicate that the amount of the data is not big enough, as well as the quality of the data is not linguistically diverse enough to conclude whether the model has made decisions for the right reason. Without the evidence span, observation of this pattern is tentative and requires further substantiation.

5.3 Reliability

As mentioned in the background chapter, we have used the occlusion study for error analysis and increase the reliability of the predictive and descriptive accuracy of our model. The human-coded evidence spans are used to examine whether the models made predictions for the equivalent human logic. For this reason, the human-coded evidence spans of each instance in the dataset are "MASKED" to generate a new version of the dataset to conduct this study, i.e. the occluded dataset. In our AFLEK dataset, we only have the marked evidence spans for the skill *Constructing Explanations* and *Analyzing data*. For this reason, unfortunately we could not perform the occlusion study on the *Planning Investigation* skill. We will discuss the reliability of the predictive and the descriptive accuracy skill-wise.

5.3.1 Constructing Explanations:

Occlusion study for predictive accuracy: If the predictive performance of the model trained on the regular non-occluded dataset, which has learnt the linguistic signals match with the human-coded evidence span, should drop substantially if the occluded version of the dataset is passed. Table 5.4 shows the result of the model's predictive performance in the occlusion study. All the evaluation metrics drastically dropped from the model trained on the original data. The drop in precision suggests that the model is making larger false positive errors by wrongly identifying negative instances as positive. Also, the critical parts being occluded, the model with decreased recall is failing to identify a considerable number of true positive instance which increase the number of false negatives. Figure 5.13 illustrates the statistical significance of the performance difference of the *GBERT large* model with original data and the same model with occluded data using the *GLMER* model. Here we have picked the model trained with original data as the baseline for the *GLMER* model.

Figure 5.13 illustrates the result of the *GLMER* model. We will focus here on the fixed and the random effects. From the fixed effect, we can see that the model performance with original data has log odds of 2.7944 for correct prediction with high statistical significance ($\Pr(> |z|) < 2e - 16 (***)$), indicating strong baseline performance. The log odds of the model performance for correct prediction decrease by 3.3839 with high statistical significance ($\Pr(> |z|) < 2e - 16 (***)$) if we use the occluded data. The standard error is .1105 for the model performance with original data whereas it is 0.1070 for the occluded data.

Regarding the student-wise random effect, there is a variance of 2.13296 with a standard deviation of 1.4605. This means some students perform consistently better or worse than others, irrespective of the model type(*GBERT large* with original or occluded data). The variance of 0.05984 among questions with a standard deviation of 0.2446 suggests

Constructing Explanation				
GBERT large	Accuracy	Precision	Recall	F1 Score
Original Data	85.36	86.48	95.00	90.48
Masked Data	45.93	72.99	33.81	46.21
Analyzing Data				
GBERT large	Accuracy	Precision	Recall	F1 Score
Original Data	83.54	84.99	96.35	90.24
Masked Data	61.26	70.46	80.98	75.36

Table 5.4: Occlusion study for the skill *Constructing Explanations and Analyzing Data*

```

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial ( logit )
Formula: correct ~ model + (1 | question_id) + (1 | student_id)
Data: df

      AIC      BIC   logLik deviance df.resid
 6969.6   6997.5  -3480.8   6961.6     8060

Scaled residuals:
    Min       1Q   Median       3Q      Max
-15.5552  -0.4017   0.1940   0.3734   2.4854

Random effects:
Groups      Name      Variance Std.Dev.
student_id (Intercept) 2.13296  1.4605
question_id (Intercept) 0.05984  0.2446
Number of obs: 8064, groups: student_id, 615; question_id, 22

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.7944      0.1105   25.28 <2e-16 ***
modelGBERT_large_masked -3.3839      0.1070  -31.61 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
              (Intr)
mdlGBERT_l_ -0.449

```

Figure 5.13: GLEMR result for skill *Constructing Explanations* in occlusion study

Constructing Explanations		
	Original Text	Translated Text
Sample Question	Notiere die anschließend deine Beobachtungen unter dem Video.	Then note your observations below the video.
Solution	In dem Video ist ein Laptop zu sehen , auf dem ein Spiel läuft . Nach einiger Zeit erscheint die Meldung , dass eine Überhitzung erkannt wurde und der Laptop gleich in einen Standby Modus wechselt . Dabei ertönt ein lautes wiederkehrendes Signal . Nach Ablauf weniger Sekunden geht der Bildschirm aus und das Signal hört auch auf .	The video shows a laptop running a game. After a while, a message appears saying that overheating has been detected and the laptop will go into standby mode. A loud, recurring signal is heard. After a few seconds, the screen goes off and the signal stops.
Response	Zuerst wird ein Videospiel gespielt und dann kommt die Benachrichtigung , dass einer Überhitzung erkannt wurde . Anschließend schaltet sich der Laptop automatisch aus .	First a video game is played and then a notification appears that overheating has been detected . The laptop then switches off automatically.
Occluded response	Zuerst wird ein Videospiel gespielt und dann kommt die [MASK] , [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] Anschließend schaltet sich der Laptop automatisch aus [MASK]	First a video game is played and then comes the [MASK] , [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] Then the laptop switches off automatically [MASK]
Analyzing Data		
	Original Text	Translated Text
Sample Question	1. Was beobachtest du wenn der Leiter stromdurchflossen ist? 2. Erkläre deine Beobachtung mithilfe von Energiebegriffen. 3. Was bedeutet deine Beobachtung für den heidf werdenden Laptop?	1. What do you observe when current flows through the conductor? 2. Explain your observation using energy terms. 3. What does your observation mean for the laptop that is heating up?
Solution	Wenn der Leiter stromdurchflossen ist , leuchtet die Lampe und die Temperatur T (in °C) des Drahtes steigt mit der Zeit t (in s) an . Die Höchsttemperatur liegt bei x°C.Wenn ein Leiter stromdurchflossen ist , wird elektrische Energie transportiert . Ein Teil dieser elektrischen Energie wird in thermische Energie umgewandelt . Diese thermische Energie wird durch die erhöhte Temperatur deutlich .	When current flows through the conductor, the lamp lights up and the temperature T (in °C) of the wire increases over time t (in s). The maximum temperature is x°C. When current flows through a conductor, electrical energy is transported. A part of this electrical energy is converted into thermal energy. This thermal energy is evident from the increased temperature.
Response	Bei dem offenen Stromkreis hat man gesehen das die Temperatur die ganze Zeit bei dem selben wert geblieben ist und bei dem geschlossenen Stromkreis ist die Temperatur gestiegen was daran lag dass der Stromkreis geschlossen war	With the open circuit, it was seen that the temperature remained at the same value the whole time and with the closed circuit, the temperature rose , which was due to the fact that the circuit was closed.
Occluded response	[MASK] [MASK] offenen Stromkreis hat man gesehen [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] und [MASK] [MASK] geschlossenen Stromkreis [MASK] [MASK] [MASK] [MASK] was daran lag dass der Stromkreis geschlossen war	[MASK] [MASK] open circuit was seen [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] and [MASK] [MASK] closed circuit [MASK] [MASK] [MASK] [MASK] which was because the circuit was closed

Table 5.5: Student's response instance for the skill *Constructing Explanations, Analyzing Data* in Occlusion study

that some questions are more difficult to answer, which affects the likelihood of the correct prediction by both models. The significant negative estimate of the performance of the GBERT base model on occluded data indicates it performs worse when we use the occluded data where the evidence span is masked. Conversely, it suggests there is a significant overlap of the features considered important by the model matches with the human-coded evidence span.

Occlusion study for descriptive accuracy:

Eventually, suppose we pass this occluded version of a particular instance of the dataset to the SHAP explainer. In that case, the SHAP value should also decrease if those evidence spans have the most impact on the model's decision. Table 5.5 shows an example of an occluded instance passed through the SHAP explainer. The table contains the original German and translated English versions for the same. In the occluded response, the human-coded evidence span is MASKED.

- **Global Explainability:**

To determine the global effect of the masked evidence span in the data for the *Constructing Explanations* skill, the top 60 positive SHAP features and the 60 negative SHAP features in increasing order are calculated from the SHAP explainer. Figure 5.14 plots the feature distribution of the positive and the negative SHAP features. If we compare the highest feature value of the positive class in figure 5.4 for original data and figure 5.14 for occluded data, a drop of global SHAP logit output of almost $\simeq 4$ can be seen. For the negative feature, the range of the smallest top negative SHAP value decreases by nearly 0.001. This likely indicates that the occluded data influenced a decreased importance of the most critical features in the model's prediction, resulting in a less confident model. For instance, in the feature "Richtungen", the SHAP explainer assigns a SHAP value between 7 and 8, where in the original case the value was almost equal to 9. This shows that the marginal contribution of the features to the model's prediction has declined due to absence of the same supporting context (Evidence span) to meaningfully contribute to the model's prediction globally.

The lower plot of the figure 5.14 shows the features from the smallest negative SHAP value to the larger negative SHAP value. This gives an overall idea of what features are pushing away the model's decision to classify an instance as positive. From the plot, it can be observed that these negative features are very generalised and provide very little domain-specific information.

- **Local Explainability:**

The plot at the bottom of the fig 5.15 shows the occluded text explanation by the SHAP explainer for the same instance explained before. For easier comparability, both the original and the occluded version of the instance is demonstrated. While

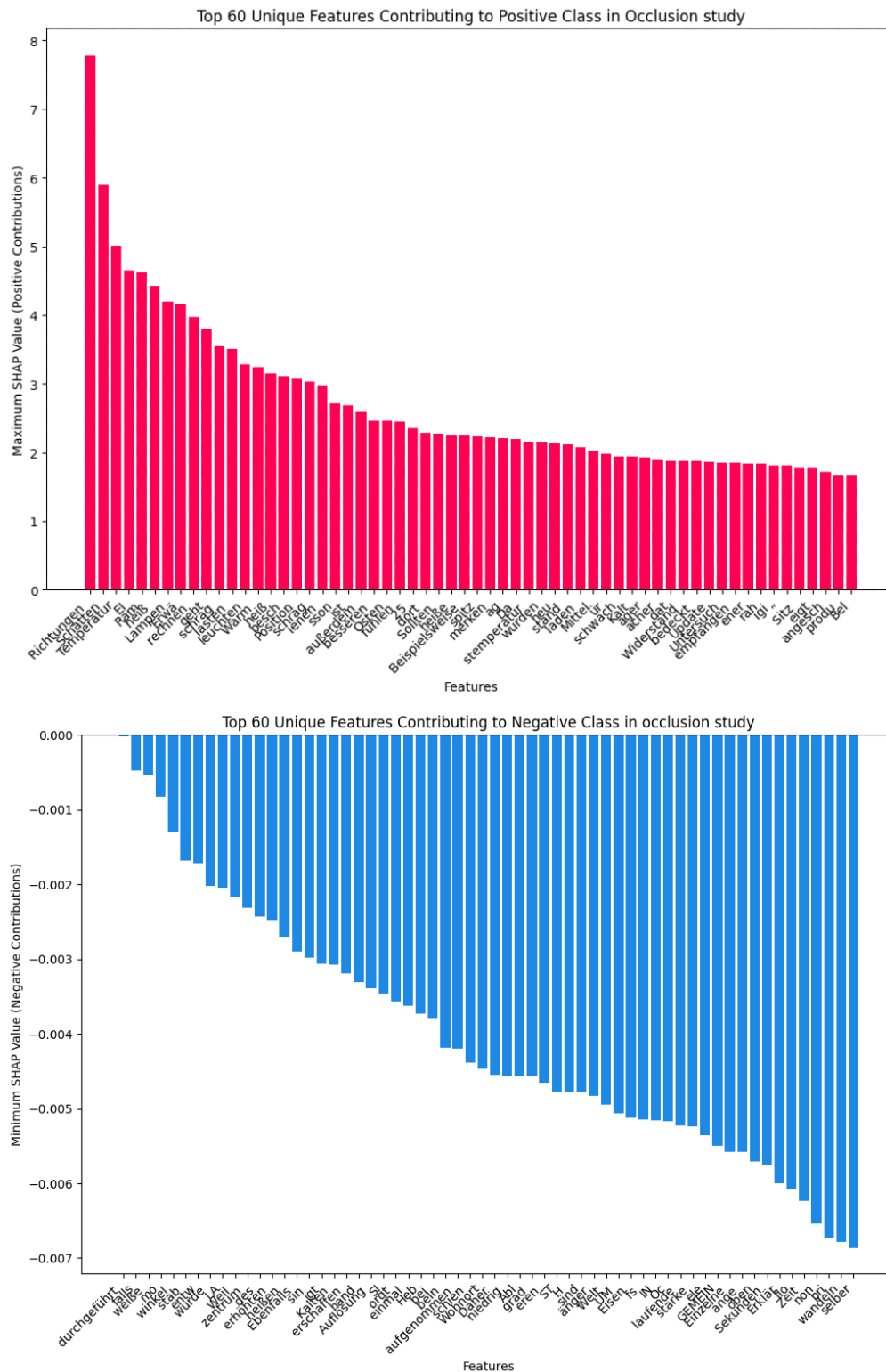


Figure 5.14: Constructing Explanation: Global Explainability in Occlusion study
Top: Positive features, Bottom: Negative features

the base value ϕ_0 of the model remains almost the same, the final model output $g(z')$ is -5.77837, whereas the model output $g(z')$ for the regular instance is 5.25248. The SHAP values for the occluded instance have drastically decreased compared to the unmasked instance, leading to the classification of the instance as a negative instance. Because of the masking, the SHAP values for the masked instance are less informative, this implies that the MASKED portions contain crucial information, if revealed, would offer a more accurate and stronger explanation.

The force plot at the bottom of the fig 5.16 illustrates the occluded explanation of the same instance. For easier comparability, both the original and the occluded version of the instance is demonstrated. Like in the occluded text plot, the model output in the force plot has also changed from 5.25 to -5.78, implying that the masking of the evidence span in the instance alters the model's decision and leads to false negatives from true positives.

The waterfall plot at the right side of the fig 5.17 shows the occluded explanation of the same instance. For easier comparability, both the original and the occluded version of the instance is demonstrated. The model's output drops significantly from $g(z') = 5.252$ to $g(z') = -5.778$, indicating a drastic reduction in the model's confidence in the positive class prediction for this instance. Here, we can see the marginal contribution of the features to the model's prediction has also declined. For example, the feature "Videos", "kommt" etc. Taking one feature in consideration, "Videos", it can be seen how the SHAP value is getting changed.

- **Original Text (Left Plot):** The term "Videos" has a more nuanced context in the original text, with words that support the positive class evidence.

$$\phi_{\text{Videos}}^{\text{original}} = +0.15$$

- **Masked Text (Right Plot):** In the occluded instance, as the evidence span is MASKED, the word "Videos" no longer has the same supportive context to contribute meaningfully to the model's prediction, leading to a negative SHAP value.

$$\phi_{\text{Videos}}^{\text{masked}} = -0.09$$

In the SHAP framework, the contribution of a feature is determined by the difference in model output with and without the feature, averaged over all possible coalitions of features. Let's denote the set of features as S and the feature of interest as i .

The SHAP value for feature i is computed as equation 3:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (5.1)$$

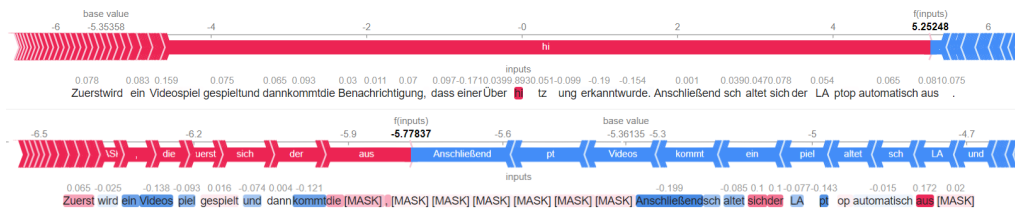


Figure 5.15: Constructing Explanation: Text plot
Top: Original Response, Bottom: Occluded response

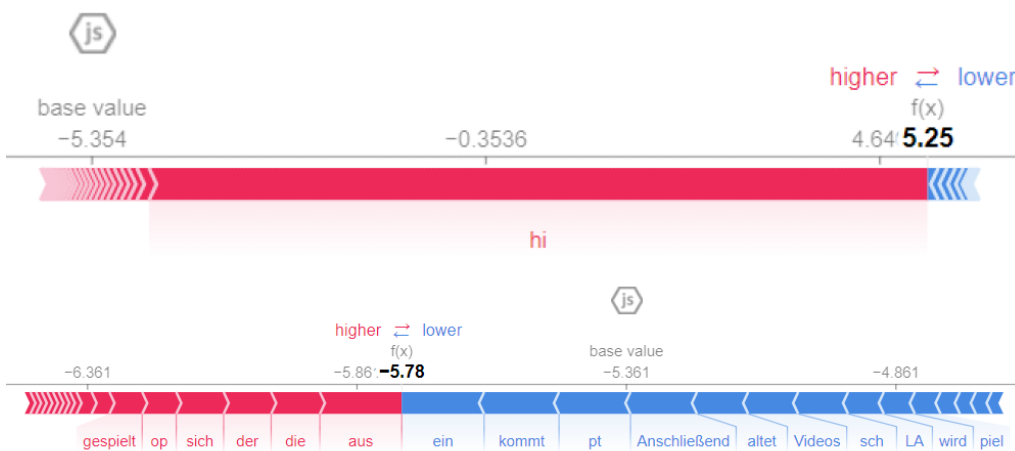


Figure 5.16: Constructing Explanation: Force plot
Top: Original Response, Bottom: Occluded response

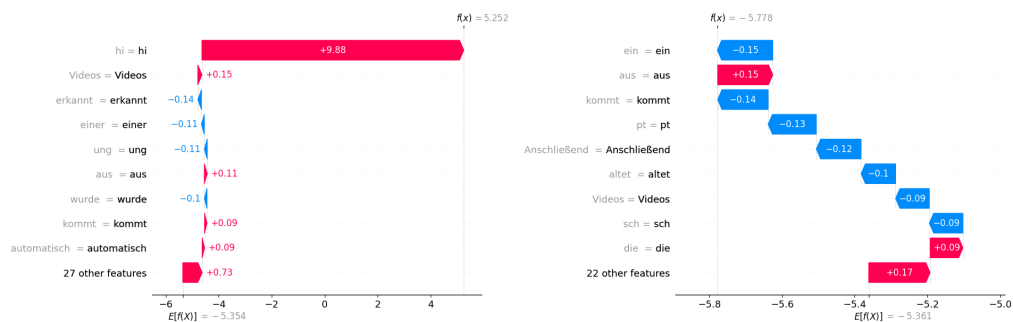


Figure 5.17: Constructing Explanation: Waterfall plot
Left: Original Response, Right: Occluded response

- F : The set of all features.
- S : A subset of features excluding i .
- $f_{S \cup \{i\}}(x_{S \cup \{i\}})$: Model output with feature i included.
- $f_S(x_S)$: Model output without feature i .

In the masked text, the context around "Videos" is altered, leading to different values of $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ and $f_S(x_S)$. The absence of crucial context words means $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$ is smaller, hence the drastic reduction in the SHAP value.

5.3.2 Analyzing Data

Occlusion study for predictive accuracy:

If the predictive performance of the model trained on the regular non-occluded dataset, which has learnt the linguistic signals match with the human-coded evidence span, it should drop substantially if the occluded version of the dataset is passed. Table 5.4 shows the result of the model's predictive performance in the occlusion study. All the evaluation metrics drastically dropped from the model trained on the original data. The drop in precision suggests that the model is making larger false positive errors by wrongly identifying negative instances as positive. Also, the critical parts being occluded, the model with decreased recall is failing to identify a considerable number of true positive instance which increase the number of false negatives. Figure 5.18 illustrates the statistical significance of the performance difference of the *GBERT large* model with original data and the same model with occluded data using the *GLMER* model. Here, we have picked the model trained with original data as the baseline for the *GLMER* model.

Figure 5.18 illustrates the result of the *GLMER* model. We will focus here on the fixed and the random effects. From the fixed effect, we can see that the model performance with original data has log odds of 2.2799 for correct prediction with high statistical significance ($\Pr(> |z|) < 2e - 16 (***)$), indicating strong baseline performance. The log odds of the model performance for correct prediction decrease by -0.4827 with high statistical significance ($\Pr(> |z|) 0.00147 (**)$) if we use the occluded data. The standard error is 0.2002 for the model performance with original data and 0.1518 for the occluded data.

The student-wise random effect has a variance of 3.42759 with a standard deviation of 1.8514. This means some students perform consistently better or worse than others, irrespective of the model type (*GBERT large* with original or occluded data). The variance of 0.02439 among questions with a standard deviation of 0.1562 suggests that some questions are more difficult to answer, which affects the likelihood of the correct prediction by both models. The significant negative estimate of the performance of the *GBERT large* model on occluded data indicates it performs worse when we use the occluded data where the evidence span is masked even after accounting for the variability across students and questions. Conversely, it suggests that there is a significant overlap of the

```

Generalized linear mixed model fit by maximum likelihood (Laplace
Approximation) [glmerMod]
Family: binomial ( logit )
Formula: correct ~ model + (1 | question_id) + (1 | student_id)
Data: df

      AIC      BIC   logLik deviance df.resid
1857.6   1879.8   -924.8   1849.6     1940

Scaled residuals:
    Min       1Q   Median       3Q      Max
-4.1719  0.1614  0.2647  0.3881  2.1695

Random effects:
Groups      Name      Variance Std.Dev.
student_id (Intercept) 3.42759  1.8514
question_id (Intercept) 0.02439  0.1562
Number of obs: 1944, groups: student_id, 431; question_id, 6

Fixed effects:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)      2.2799    0.2002   11.39 < 2e-16 ***
modelGBERT_large_masked -0.4827    0.1518   -3.18  0.00147 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Correlation of Fixed Effects:
      (Intr)
mdlGBERT_l_ -0.338

```

Figure 5.18: GLEMR result for skill *Analyzing Data* in occlusion study

features considered important by the model that matches the human-coded evidence span.

Occlusion study for descriptive accuracy:

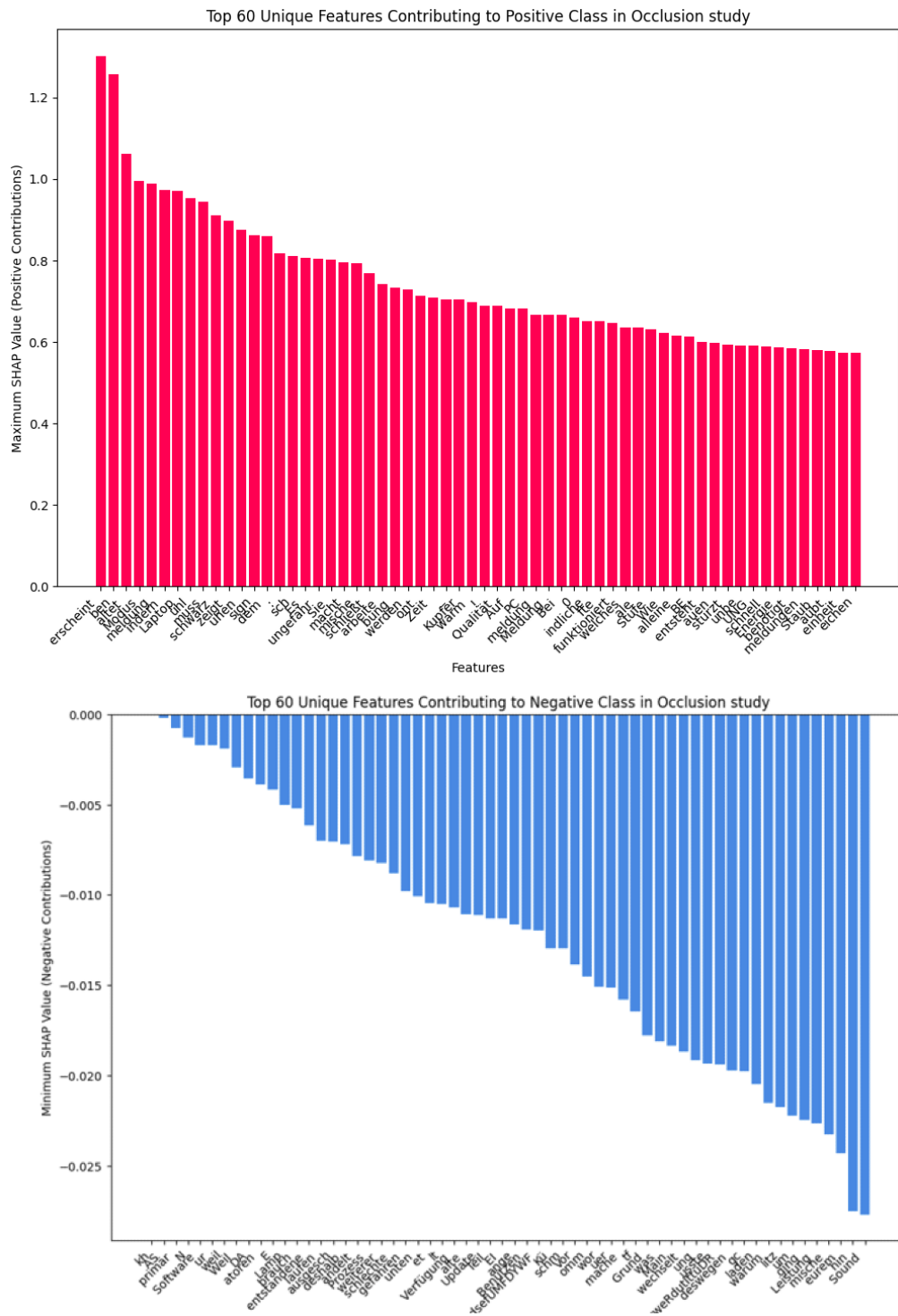
- Global Explainability:

To determine the global effect of the masked evidence span in the data for the *Analyzing Data* skill, the top 60 positive SHAP features and the 60 negative SHAP features in increasing order are calculated from the SHAP explainer. Figure 5.19 shows the plots of the feature distribution of both positive and negative ones. Let's compare the highest feature value of the positive class in figure 5.8 for original data and figure 5.19 for occluded data. A drop of global SHAP logit output of almost $\simeq 1.1$. For the negative feature, the range of the smallest top negative SHAP value decreases by nearly 0.175. This likely indicates that the occluded data influenced a decreased importance of the most critical features in the model's prediction, resulting in a less confident model like previous skill. For instance, the SHAP explainer assigns a global feature importance of "Zeit" in the original case is 1.75 where the occluded model assigns a decreased global importance of almost $\simeq 0.8$. The lower plot of the figure 5.19 shows the features from the smallest negative SHAP value to the larger negative SHAP value. This gives an overall idea of what features are pushing away the model's decision to classify an instance as positive. From the plot, it can be observed that these negative features are very generalised and provide very little domain-specific information.

- Local Explainability:

Table 5.5 shows an example of an occluded instance passed through the SHAP explainer. The plot at the bottom of the fig 5.20 shows the occluded text explanation by the SHAP explainer for the same instance explained before. For easier comparability, both the original and the occluded version of the instance is demonstrated. While the base value of the model remains the same, the final model output $g(z')$ is 0.344639, whereas the model output $g(z')$ for the regular instance is 5.038. The SHAP values for the occluded instance have drastically decreased in magnitude compared to the unmasked instance, indicating a weaker but still positive influence. Because of the masking, the SHAP values for the masked instance are less informative, even if the model still predicts a positive class instance for the presence of other unmasked features in the instance. This implies that the MASKED portions, contain crucial information, if revealed, would offer a more accurate and stronger explanation.

The force plot at the bottom of the fig 5.21 illustrates the occluded explanation of the same instance. For easier comparability, both the original and the occluded version of the instance is demonstrated. Like in the occluded text plot, the model



output $g(z')$ in the force plot has also decreased from 5.04 to 0.34, implying that the masking of the evidence span in the instance is decreasing the model's confidence to classify the instance belonging to the positive class.

The waterfall plot at the right side of the fig 5.22 shows the occluded explanation of the same instance. For easier comparability, both the original and the occluded version of the instance is demonstrated. The model's output drops significantly from $g(z') = 5.038$ to $g(z') = 0.345$, indicating a drastic reduction in the model's confidence in the positive class prediction. Here, we can see the marginal contribution of the features to the model's prediction has also declined. For example, the feature "kreis"

- **Original Text (Left Plot):** The term "kreis" has a more nuanced context in the original text, with words that support the positive class evidence.

$$\phi_{\text{kreis}}^{\text{original}} = 0.21$$

- **Masked Text (Right Plot):** In the occluded instance, as the evidence span is MASKED, the word "kreis" no longer has the same supportive context to contribute meaningfully to the model's prediction, leading to a reduced SHAP value.

$$\phi_{\text{kreis}}^{\text{masked}} = 0.06$$

In the SHAP framework, the contribution of a feature is determined by the difference in model output with and without the feature, averaged over all possible coalitions of features. Let's denote the set of features as S and the feature of interest as i .

The SHAP value for feature i is computed as equation 3:

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)] \quad (5.2)$$

- F : The set of all features.
- S : A subset of features excluding i .
- $f_{S \cup \{i\}}(x_{S \cup \{i\}})$: Model output with feature i included.
- $f_S(x_S)$: Model output without feature i .

In the masked text, the context around "kreis" is altered, leading to different values of $f_{S \cup \{i\}}(x_{S \cup \{i\}})$ and $f_S(x_S)$. The absence of crucial context words means $f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S)$ is smaller, thus reducing the SHAP value.



Figure 5.20: Analyzing Data: Text plot
Top: Original Response, Bottom: Occluded response



Figure 5.21: Analyzing Data: Force plot
Top: Original Response, Bottom: Occluded response

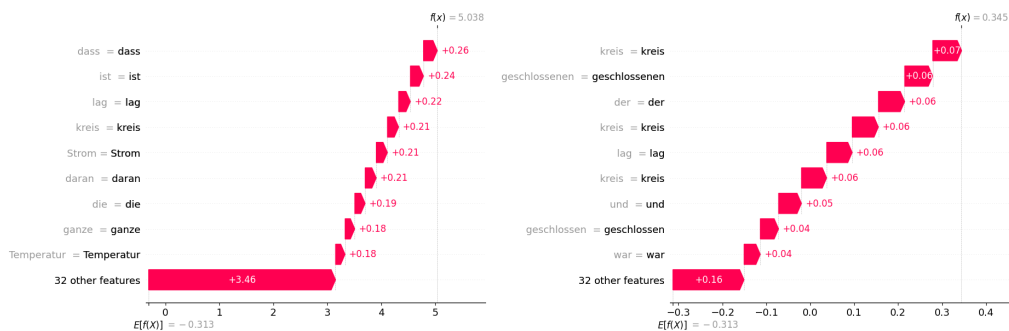


Figure 5.22: Analyzing Data: Waterfall plot
Left: Original Response, Right: Occluded response

5.4 Relevancy

Apart from the predictive, descriptive accuracy, and reliability assessment of the predictive and descriptive accuracy of the model, in this section, we evaluated the other possible effects which might be relevant for the stakeholder to know regarding the models' decision-making behaviour. For this reason, we further conducted additional analysis. AFLEK dataset contains multiple responses (multiple observations) from each student (each subject) and multiple responses (multiple observations) from each question (each item). This implies that individual responses are not completely independent from one another, but they are grouped under students and questions. As a further study, we collected the predictions of the top model one from the instance-based and another from the entailment-based method for each skill based on their predictive performance (For computational limitation, we restricted this study only to the top two models). Furthermore, we have implemented three *generalised linear mixed effect models (GLMER)* with increasing random effects for each skill. We used *ANOVA* on these three models to understand if there is any significant random effect which may have influenced the models' performance.

- **Constructing Explanations:** The best-performing model from the instance-based scoring is the *GBERT base* model, where the *GBERT large* model has the best F1 score from the entailment scoring. We collected the predictions of these models and passed them through three different *GLMER* models. Finally, *ANOVA* is used on these three models.

Figure 5.23 illustrates the *ANOVA* output of the three *GLMER* models for the Constructing explanation skill. The *question_model* includes the random intercept for the question to check if, for different questions, the models might be more likely to be correct in predicting the label. The *student_model*, along with various questions, accounts for the variability of different students on the models' performance by including additional random intercepts for students. Lastly, the *full_model* is the most complex one. It includes a random slope for models under students to check whether some models might have an easier time predicting the responses of specific students being correct.

The *student_model* has significantly lower AIC, BIC and deviance and higher log-likelihood than the *question_model* ($\Pr(> \text{Chi}^2) < 2e - 16 (***)$). This indicates that adding the student random effect significantly improves the model fit to the predictions of the *GBERT base* and *large* models. Similarly, the *full_model* has no significant improvement over the *student_model*, as indicated by the non-significant Chi-square value.

This indicates that for different students, the model is more likely to be correct in predicting the responses.

- **Analyzing Data:** The best-performing model from the instance-based scoring is the *GBERT base* model, where the *GBERT large* model has the best F1 score from the entailment scoring. We collected the predictions of these models and passed them through three different *GLMER* models. Finally, *ANOVA* is used on these three models.

Figure 5.24 shows the *GLMER* model output for the *Analyzing data* skill. illustrates the *ANOVA* output of the three *GLMER* models for the Constructing explanation skill. The *question_model* includes the random intercept for the question to check if, for different questions, the models might be more likely to be correct in predicting the label. The *student_model*, along with various questions, accounts for the variability of different students on the models' performance by including additional random intercepts for students. Lastly, the *full_model* is the most complex one. It includes a random slope for models under students to check whether some models might have an easier time predicting the responses of specific students being correct.

The *student_model* has significantly lower AIC, BIC and deviance and higher log-likelihood than the *question_model* ($\Pr(> \text{Chi}^2) < 2e - 16 (***)$). This indicates that adding the student random effect significantly improves the model fit to the predictions of the *GBERT base* and *large* models. Similarly, the *full_model* has no significant improvement over the *student_model*, as indicated by the non-significant Chi-square value.

This indicates that for different students, the model is more likely to be correct in predicting the responses.

- **Planning Investigations:**

The best-performing model from the instance-based scoring is the *GELECTRA large* model, whereas the *GBERT large* model has the best F1 score from the entailment-based scoring. As for the *Planning Investigation* skill responses come only from one question, we have implemented one *generalized linear model (GLM)* and another *GLMER* model. We collected the predictions of the *GELECTRA large* and *GBERT large* models and passed them through two different models. Finally, *ANOVA* is used on these two models.

The *base_model* includes a linear relation between the response and the *student_model* consisting of the random intercept for the students to check if, for different students, the models might be more likely to be correct in predicting the label. The residual degrees of freedom for the *student_model* is one less due to the inclusion of the random effect than the *null_model*. There is no change in residual deviance and a very small change in the deviance after adding the random effect for the students to the *student_model*, and a non-significant high chi-square value implies there is no effect on the *GELECTRA large* and *GBERT large* models of the


```

boundary (singular) fit: see help('isSingular')
Data: df
Models:
question_model: correct ~ model + (1 | question_id)
student_model: correct ~ model + (1 | question_id) + (1 | student_id)
full_model: correct ~ model + (1 | question_id) + (1 + model | student_id)
      npar    AIC    BIC logLik deviance   Chisq Df Pr(>Chisq)
question_model    3 6206.4 6227.4 -3100.2   6200.4
student_model     4 5791.2 5819.2 -2891.6   5783.2 417.2538  1    <2e-16 ***
full_model        6 5793.4 5835.4 -2890.7   5781.4   1.7833  2      0.41
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 5.23: Comparison of different GLMER models for *Constructing Explanations*

```

Data: df
Models:
question_model: correct ~ model + (1 | question_id)
student_model: correct ~ model + (1 | question_id) + (1 | student_id)
full_model: correct ~ model + (1 | question_id) + (1 + model | student_id)
      npar    AIC    BIC logLik deviance   Chisq Df Pr(>Chisq)
question_model    3 1791.8 1808.6 -892.92   1785.8
student_model     4 1616.2 1638.5 -804.08   1608.2 177.6769  1    <2e-16 ***
full_model        6 1620.1 1653.5 -804.03   1608.1   0.1039  2      0.9494
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Figure 5.24: Comparison of different GLMER models for *Analyzing Data*

```

Analysis of Deviance Table

Model: binomial, link: logit

Response: correct

Terms added sequentially (first to last)

      Df    Deviance Resid. Df Resid. Dev Pr(>Chi)
NULL              254      280.59
model            1 0.00055593    253      280.59  0.9812

```

Figure 5.25: Comparison of GLM and GLEMR models for skill *Planning Investigations*

student-wise variance in the data.

Our result suggests using state-of-the-art models is appropriate for identifying scientific skills in constructed responses. We also evaluated the instance and entailment-based scoring. This suggests that considering the global performance of the model, the entailment-based models work better than the instance-based ones. Providing greater context through a sample solution to the models, as in the entailment-based scoring, has improved their performance in most cases, with a few exceptions. On the other hand, additional random effect analysis shows that for *Constructing Explanations* and *Analyzing data* skills, there is a significant random effect student-wise, indicating that student-specific variability significantly influences the predictions. This shows that our models are sensitive to individual student differences, and it is essential to account for enhanced reliability in the future. However, this mixed effect analysis is limited by the number of models we have examined due to resource challenges and needs further investigation.

Chapter 6

Discussion

In developing our explainable analytical skill assessment system, we employed an interpretable predictive, descriptive and relevant framework(PDR) to maintain interpretability from various perspectives. The first desideratum of this framework is predictive accuracy, which estimates the connection between the underlying data and the model, a key aspect we addressed in RQ1. We utilized various techniques to build our skill assessment system, commonly used for automatic short-answer grading. Three different Transformers models were implemented for all three skills, including their base and large variants. We followed two distinct scoring methods to implement these models: instance-based and entailment-based scoring. Regardless of the scoring techniques and skills, most models(29 out of 36 cases) achieved an F1 score higher than 86.00, demonstrating the effectiveness of our system. Notably, the *GBERT large* model in the entailment-based scoring outperformed compared to the instance-based scoring in two out of three skills. To response to RQ1 (To what extent can scientific skills be detected in students' free-text responses using different Transformers models?) and RQ2 (Does adopting the entailment-based scoring improve over the instance-based scoring?), we can say that the usage of Transformers architecture is plausible to automate the identification of the scientific skills in constructed responses efficiently and the overall trend suggests that prioritizing large models(e.g.*GBERT-large*, *GELECTRA-large*) and leveraging the entailment-based scoring whenever possible can enhance the possibility of predicting precise class compared to instance-based scoring.

We also evaluated identical architectures and scoring methods using the CREG-TUE dataset (a key aspect to address the RQ2) to compare our approach to similar short-answer scoring datasets from this domain to contextualize our work better. While the instance-and entailment-based scoring achieves similar results for the two-class setup,

using the entailment-based scoring leads to notably higher results in the four-class setup. F1 scores dropped considerably in the four-class setup, with all models falling below 84.00. This highlights the increased difficulty for models to classify the responses in four-class categories accurately. In this dataset, *GBERT-large* emerged as the best-performing model in a four-class setup, while in a two-class setup, the *GBERT base* model in the entailment-based scoring performed the best. The impact of data size was more evident in AFLEK. Skills with less data (e.g., Planning Investigations) seemed to benefit more from large models. The CREG-TUE experiment demonstrates the prominent effect of classifier complexity. The four-class setup proved more challenging for all models than the two-class setup in AFLEK, where skills have varying difficulty levels. Therefore, in response to RQ3, it can be stated that the approaches used for the AFLEK dataset translate well for the CREG-TUE dataset. Irrespective of the classifier complexity, the entailment-based scoring method works better than the instance-based one.

The second desideratum of the PDR framework is descriptive accuracy, which assesses whether the model learns to decide for the right reason (a key aspect of RQ4). For this assessment, we implemented the SHAP explainer as an explanation model to explain the approximate behaviour of our Transformers models. The SHAP explainer played a crucial role in our work, providing a detailed understanding of how our Transformers models make decisions. We picked the best-performing Transformers models for each skill to fit in the SHAP explainer pipeline to get the local and global explanations. For all three skills, the global explanation of the Transformers models demonstrates that the models have learnt mostly domain-specific words as highly important words for models' decisions with few exceptions. The SHAP explainer has assigned a higher SHAP value for those energy-related, domain-specific words than the other general words. Furthermore, the local explanations suggest considerable overlap between the human-annotated evidence span and the words provided positive SHAP value by the model. Additionally, the waterfall plots illustrate the additivity property of SHAP to ensure the correctness of our explainer model. Moreover, to assess the reliability of our models' predictive and descriptive accuracy, we performed an occlusion test. We have generated the occluded version of the dataset by masking the human-annotated evidence span. We kept the length of the responses the same as the responses in the original version of the dataset. Afterwards, we tested the best-performing Transformers models with the occluded responses for two skills (For the third skill, we did not have the marked evidence span) and observed a huge decrease in their predictive performance. For assessing the reliability of the the SHAP explanation model's descriptive accuracy, we used the occluded version of the dataset to illustrate the global and local explanations. For the global explanations, the result shows that for both the skills, the range of the positive and the negative SHAP values are degraded notably, specifically the positive SHAP range. The SHAP explainer assigned a lower positive SHAP value to the positive features present in the dataset even after occlusion. This implies that the models' confidence

declined in taking a decision when the evidence span is occluded. In local explainability, we observed two scenarios; one is a positive instance that gets misclassified as a negative instance. Another scenario illustrates that even if the positive instance is classified as positive, the SHAP explainer output is considerably low likely because of the missing evidence span. To respond to the RQ4 (To what extent do the input words that are considered important by the models for their predictions match human-coded ones?), it can be stated that all the above occurrences suggest the input words or evidence span considered important by the models for their predictions match human-coded one as the models learnt the human-annotated evidence span as an important for models' decisions.

In terms of an explainable analytical skill assessment system, the third desideratum relevancy suggests whether the stakeholder gets the relevant interpretation from the assessment system in terms of predictive and descriptive accuracy. The result suggests the predictive accuracy of our models is quite reasonable and the descriptive accuracy illustrates that the models mostly learn for the right reason with few exceptions and their reason for making decisions also matches with the human-annotated evidence span. The occlusion study shows the reliability of our models' predictive and descriptive accuracy. Furthermore, our additional analysis suggests that the model is sensitive to student-specific performance. Considering every aspect, we can say that the stakeholders can get relevant information regarding the models' decision-making mechanism. As a result, to address the last research question of our thesis (RQ5: To what extent the interpretation of these model's decision-making behaviour are relevant to the stakeholders), we can say that the stakeholders are well aware of the pros and cons of the models' behaviours which is quite relevant for them which eventually help them to take further decisions.

Chapter 7

Conclusion, Limitations and Future Directions, Ethical Considerations

In this chapter, we will discuss the conclusion of our research. Then we will elaborate on the limitation of this work and the probable future direction of it. Additionally, we will finish this chapter by discussing ethical consideration.

7.1 Conclusion

In this thesis, we experimented with identifying three scientific skills in constructed responses from German middle school students using explainable models. We have adopted a standardized explainable framework to ensure the interpretability of our implementation. To instantiate each module of this framework, we have adopted the current state-of-the-art performing architecture in this domain and the explainable method which aligns the most with human intuition. As the first desideratum of this framework, We have evaluated six Transformers-encoder language models and achieved satisfactory results. Our results suggest that using the monolingual dataset, *BERT* and *ELECTRA*-like Transformers language models is appropriate for identifying scientific skills in constructed responses. We also evaluated whether instance- or entailment-based scoring works better when classifying the responses. Providing greater context through a sample solution to the models, as in the entailment-based scoring, has improved their performance in most cases, with a few exceptions. To compare our approach to similar datasets from this automatic short answer scoring domain, we also evaluated identical architectures for the well-known dataset CREG-TUE. Here, it was confirmed, as well, that

using entailment-based scoring can be regarded as the optimal approach. To instantiate the second desideratum for this framework, we have further analyzed the explainability of our model’s decision to ensure the descriptive accuracy of our models. For all the skills, the result shows that the models learn mostly domain-specific features at the dataset level globally, which is quite logical. To investigate it further, we performed local explainability and showed that the models’ local explanation considerably matches the human evidence span. To ensure the robustness of the explainability and to increase the reliability and trustworthiness, we further performed an error analysis based on the occlusion test. We showed that the model’s prediction drastically degrades if we mask the human evidence span, and the explainer also loses confidence in classifying instances correctly. This indicates that masked portions contain crucial information; if revealed, they offer a more accurate and robust explanation. Following the final desideratum of the framework relevancy to provide the stakeholders with all the relevant perspectives of the model’s decision-making behaviour (Predictive, descriptive accuracy and reliability), we additionally examined the effect of the hierarchical structures on these models’ performance. Our models are observed to be sensitive to student-wise variability, which will necessitate further investigations in the future. The work presented in this thesis is available in the following link: <https://github.com/Smita1908/MastersThesis>

7.2 Limitations and Future Directions

- The work is limited by focusing on two specific German datasets (AFLEK and CREG-TUE). The data we use to evaluate whether Transformers language models are appropriate for identifying scientific skills comes from a narrow domain. The generalizability of the findings might be restricted if applied to significantly different reasoning skills or assessment tasks.
- Furthermore, the impact of varying the training data size or using different data augmentation techniques was not explored, which could influence the overall conclusions.
- To mitigate the effect of the student-wise variability towards the model’s prediction, a future study can be conducted to cluster the similar responses of the students and can analyse the general pattern of each cluster to understand this occurrence.

7.3 Ethical Considerations

It is important to remember that these powerful models should not be used in isolation when applied in practical contexts. Ideally, teachers should double-check the results to ensure their correctness and consider the context of each student’s work. Therefore, our

system is best suited as a supplementary tool and not for high-stakes assessments where a single score holds significant weight.

References

- Anders, C. J., Weber, L., Neumann, D., Samek, W., Müller, K.-R., & Lapuschkin, S. (2022). Finding and removing clever hans: using explanation methods to debug and improve deep models. *Information Fusion*, 77, 261–295.
- Bachman, L. F., Carr, N., Kamei, G., Kim, M., Pan, M. J., Salvador, C., & Sawaki, Y. (2002). A reliable approach to automatic assessment of short answer free responses. In *Coling 2002: The 17th international conference on computational linguistics: Project notes*.
- Bastings, J., & Filippova, K. (2020, November). The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? In A. Alishahi, Y. Belinkov, G. Chrupała, D. Hupkes, Y. Pinter, & H. Sajjad (Eds.), *Proceedings of the third blackboxnlp workshop on analyzing and interpreting neural networks for nlp* (pp. 149–155). Online: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/2020.blackboxnlp-1.14> doi: 10.18653/v1/2020.blackboxnlp-1.14
- Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi: 10.18637/jss.v067.i01
- Belle, V., & Papantonis, I. (2021). Principles and practice of explainable machine learning. *Frontiers in big Data*, 4, 688969.
- Bexte, M., Horbach, A., & Zesch, T. (2023). Similarity-based content scoring-a more classroom-suitable alternative to instance-based scoring? In *Findings of the association for computational linguistics: Acl 2023* (pp. 1892–1903).
- Binder, A., Bach, S., Montavon, G., Müller, K.-R., & Samek, W. (2016). Layer-wise relevance propagation for deep neural network architectures. In *Information science and applications (icisa) 2016* (pp. 913–922).
- Burrows, S., Gurevych, I., & Stein, B. (2015). The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25, 60–117.
- Callear, D. H., Jerrams-Smith, J., & Soh, V. (2001). Caa of short non-mcq answers.
- Camus, L., & Filighera, A. (2020). Investigating transformers for automatic short answer grading. In *Artificial intelligence in education: 21st international conference, aied 2020, ifrane, morocco, july 6–10, 2020, proceedings, part ii 21* (pp. 43–48).

- Chan, B., Schweter, S., & Möller, T. (2020). German's next language model. *arXiv preprint arXiv:2010.10906*.
- Chefer, H., Gur, S., & Wolf, L. (2021). Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 782–791).
- Clark, K., Luong, M.-T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. In *International conference on learning representations*. Retrieved from <https://openreview.net/forum?id=r1xMH1BtvB>
- Crossley, S., Kyle, K., Davenport, J., & McNamara, D. S. (2016). Automatic assessment of constructed response data in a chemistry tutor. *International Educational Data Mining Society*.
- Cutrone, L. A., & Chang, M. (2010). Automarking: automatic assessment of open questions. In *2010 10th IEEE International Conference on Advanced Learning Technologies* (pp. 143–147).
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.
- Dougiamas, M., & Taylor, P. (2003). Moodle: Using learning communities to create an open source course management system. In *Edmedia+ innovate learning* (pp. 171–178).
- Drachsler, H., & Greller, W. (2016). Privacy and analytics: it's a delicate issue a checklist for trusted learning analytics. In *Proceedings of the sixth international conference on learning analytics & knowledge* (pp. 89–98).
- Dzikovska, M. O., Nielsen, R., Brew, C., Leacock, C., Giampiccolo, D., Bentivogli, L., ... Dang, H. T. (2013). Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. In *Second joint conference on lexical and computational semantics (*SEM), volume 2: Proceedings of the seventh international workshop on semantic evaluation (semeval 2013)* (pp. 263–274).
- Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., ... Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable nlp models. *Journal of Computer Assisted Learning*, 39(3), 767–786.
- Greller, W., & Drachsler, H. (2012). Translating learning into numbers: A generic framework for learning analytics. *Journal of Educational Technology & Society*, 15(3), 42–57.
- Hahn, M., & Meurers, D. (2012). Evaluating the meaning of answers to reading comprehension questions: A semantics-based approach. In *Proceedings of the seventh workshop on building educational applications using nlp* (pp. 326–336).

- Horbach, A., Palmer, A., & Pinkal, M. (2013). Using the text to evaluate short answers for reading comprehension exercises. In *Second joint conference on lexical and computational semantics (* sem), volume 1: Proceedings of the main conference and the shared task: Semantic textual similarity* (pp. 286–295).
- Korman, D. Z., Mack, E., Jett, J., & Renear, A. H. (2018). Defining textual entailment. *Journal of the Association for Information Science and Technology*, 69(6), 763–772.
- Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., . . . Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.
- Livingston, S. A. (2009). Constructed-response test questions: Why we use them; how we score them. r&d connections. number 11. *Educational Testing Service*.
- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In *International conference on learning representations*. Retrieved from <https://openreview.net/forum?id=Bkg6RiCqY7>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. *Advances in neural information processing systems*, 30.
- Maharjan, N., Gautam, D., & Rus, V. (2018). Assessing free student answers in tutorial dialogues using lstm models. In *Artificial intelligence in education: 19th international conference, aied 2018, london, uk, june 27–30, 2018, proceedings, part ii* 19 (pp. 193–198).
- Marvaniya, S., Saha, S., Dhamecha, T. I., Foltz, P., Sindhgatta, R., & Sengupta, B. (2018). Creating scoring rubric from representative student answers for improved short answer grading. In *Proceedings of the 27th acm international conference on information and knowledge management* (pp. 993–1002).
- Mertler, C. A. (2001). Designing scoring rubrics for your classroom. *Practical assessment, research, and evaluation*, 7(1), 25.
- Meurers, D., Ziai, R., Ott, N., & Kopp, J. (2011). Evaluating answers to reading comprehension questions in context: Results for german and the role of information structure. In *Proceedings of the textinfer 2011 workshop on textual entailment* (pp. 1–9).
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable machine learning: definitions, methods, and applications. *arXiv preprint arXiv:1901.04592*.
- Nath, S., Parsaeifard, B., & Werlen, E. (2023). Automated short answer grading using bert on german datasets.
- Ott, N. (2014). *Creg-tue data set documentation*. (CREG-TUE Data Set Documentation)
- Ott, N., Ziai, R., & Meurers, D. (2012). Creation and analysis of a reading comprehension exercise corpus. *Multilingual corpora and multilingual corpus analysis*, 14, 47.

- Pado, U., & Kiefer, C. (2015). Short answer grading: When sorting helps and when it doesn't. In *Proceedings of the fourth workshop on nlp for computer-assisted language learning* (pp. 42–50).
- Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for conceptualizing and evaluating the validity of instructionally relevant assessments. *Educational Psychologist*, 51(1), 59–81.
- Poulton, A., & Eliens, S. (2021). Explaining transformer-based models for automatic short answer grading. In *Proceedings of the 5th international conference on digital technology in education* (pp. 110–116).
- R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Riordan, B., Horbach, A., Cahill, A., Zesch, T., & Lee, C. (2017). Investigating neural architectures for short answer scoring. In *Proceedings of the 12th workshop on innovative use of nlp for building educational applications* (pp. 159–168).
- Sawyer, R. K. (2005). *The cambridge handbook of the learning sciences*. Cambridge University Press.
- Slade, S., & Tait, A. (2019). Global guidelines: Ethics in learning analytics.
- Sun, X., Yang, D., Li, X., Zhang, T., Meng, Y., Qiu, H., ... Li, J. (2021). Interpreting deep learning models in natural language processing: A review. *arXiv preprint arXiv:2110.10470*.
- Sung, C., Dhamecha, T. I., & Mukhi, N. (2019). Improving short answer grading using transformer-based pre-training. In *Artificial intelligence in education: 20th international conference, aied 2019, chicago, il, usa, june 25-29, 2019, proceedings, part i 20* (pp. 469–481).
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). Attention is all you need. *advances in neural information processing systems*, 30(2017).
- Williams, A., Nangia, N., & Bowman, S. (2018, June). A broad-coverage challenge corpus for sentence understanding through inference. In M. Walker, H. Ji, & A. Stent (Eds.), *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1112–1122). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from <https://aclanthology.org/N18-1101> doi: 10.18653/v1/N18-1101

List of Figures

3.1	Overview of different data science life cycle stages(black text) where interpretability is important. Figure from the paper (Murdoch et al., 2019)	11
3.2	Instance-based vs. Entailment-based scoring (Burrows et al., 2015; Dzikovska et al., 2013)	13
3.3	Transformer architecture (Vaswani et al., 2017)	15
3.4	Overview of BERT pre-training process (Devlin et al., 2018)	17
3.5	Overview of ELECTRA pre-training process: Replaced token detection (Clark et al., 2020)	17
4.1	Label distributions for the three skills. <i>Label 0</i> refers to cases where a skill could not be detected in a response, while <i>Label 1</i> refers to cases where a skill could be detected.	26
4.2	Example question for assessing Constructing Explanations Skill. Question Translation: Explain why we don't always align solar cell with the position of the sun	28
4.3	Example question for assessing Analyzing Data Skill. Question Translation: Compare the series of measurements for brightness with the series of measurements for electrical voltage. What do you notice? Also name the places where you noticed something.	28

4.4	Example question for assessing Planning Investigation Skill. Translated question: "Build a circuit according to the circuit diagram below. The following instructions should help you: 1. First connect the lamp to the battery using two cables and two alligator clips. 2. Break the circuit. 3. Insert the twisted wire with the two tripods into the circuit. 4. You should now measure the temperature inside the untwisted wire. To do this, push the thermometer into the untwisted wire and measure the temperature every 10 seconds for 1 minute for a closed and non-closed circuit. Write down the values in the table below (tip: start both measurements with the same starting temperature).Note: If you are unsure about how to set it up, you can find the picture of the circuit to be set up here.	29
4.5	Label distributions for binary and four class classification of CREG-TUE dataset. For binary classifier, <i>True</i> refers to cases where a student could provide information in response properly, while <i>False</i> refers to cases where a student could not. For the four class classifiers, the labels explain the amount of information provided in the responses.	31
4.6	Thesis workflow	32
5.1	Constructing Explanation: Text plot Top: Response belongs to positive class, Bottom: Response belongs to negative class	44
5.2	Constructing Explanation: Force plot Top: Response belongs to positive class, Bottom: Response belongs to negative class	44
5.3	Constructing Explanation: Waterfall plot Left: Response belongs to positive class, Right: Response belongs to negative class	44
5.4	Constructing Explanation: Global explanation Top: Positive features, Bottom: Negative features	47
5.5	Analyzing Data: Text plot Top: Response belongs to positive class, Bottom: Response belongs to negative class	50
5.6	Analyzing Data: Force plot Top: Response belongs to positive class, Bottom: Response belongs to negative class	50
5.7	Analyzing Data: Waterfall plot Left: Response belongs to positive class, Right: Response belongs to negative class	50

5.8 Analyzing Data: Global explanation Top: Positive features, Bottom: Negative features	52
5.9 Planning Investigation: Text plot Top: Positive Response, Bottom: Negative response	55
5.10 Planning Investigation: Force plot	55
5.11 Planning Investigation: Waterfall plot Left: Original Response, Right: Occluded response	55
5.12 Planning Investigation: Global Explainability Top: Positive features, Bottom: Negative features	57
5.13 GLEMR result for skill <i>Constructing Explanations</i> in occlusion study	60
5.14 Constructing Explanation: Global Explainability in Occlusion study Top: Positive features, Bottom: Negative features	63
5.15 Constructing Explanation: Text plot Top: Original Response, Bottom: Occluded response	65
5.16 Constructing Explanation: Force plot Top: Original Response, Bottom: Occluded response	65
5.17 Constructing Explanation: Waterfall plot Left: Original Response, Right: Occluded response	65
5.18 GLEMR result for skill <i>Analyzing Data</i> in occlusion study	67
5.19 Analyzing Data: Global Explainability in Occlusion study Top: Positive features, Bottom: Negative features	69
5.20 Analyzing Data: Text plot Top: Original Response, Bottom: Occluded response	71
5.21 Analyzing Data: Force plot Top: Original Response, Bottom: Occluded response	71
5.22 Analyzing Data: Waterfall plot Left: Original Response, Right: Occluded response	71

5.23	Comparison of different GLMER models for <i>Constructing Explanations</i> . . .	74
5.24	Comparison of different GLMER models for <i>Analyzing Data</i>	74
5.25	Comparison of GLM and GLEMR models for skill <i>Planning Investigations</i> .	74

List of Tables

3.1	Confusion matrix illustrating every scenario that might occur in an experiment	19
4.1	This rubric lists various codes used for coding the AFLEK data for scientific skills	25
4.2	AFLEK length Summary	25
4.3	A sample example of CREG-TUE dataset with sample text, question, reference answer, and various students' responses	30
4.4	CREG dataset details	31
5.1	Results for the AFLEK dataset. The abbreviated model names such as GBb , GEb , XRb , GBI , GEI , XRI , correspond to <i>GBERTbase</i> , <i>GELECTRAbase</i> , <i>XLM-RoBERTabase</i> , <i>GBERTlarge</i> , <i>GELECTRALarge</i> <i>XLM-RoBERTalarge</i> respectively. In terms of skills, C corresponds to <i>Constructing Explanations</i> , A corresponds to <i>Analyzing Data</i> , and P corresponds to <i>Planning Investigations</i> . Bold indicates the best model within a particular method and the bold italic indicates the best model among the methods. The green colour indicates the larger version of the model has performed better than its base counterpart and for the red colour, it is vice versa.	38
5.2	Results for the two- and four-class evaluation setups of the CREG-TUE dataset. The abbreviated model names such as GBb , GEb , XRb , GBI , GEI , XRI , correspond to <i>GBERTbase</i> , <i>GELECTRAbase</i> , <i>XLM-RoBERTabase</i> , <i>GBERTlarge</i> , <i>GELECTRALarge</i> <i>XLM-RoBERTalarge</i> respectively. In terms of Class Setup, 2 Class Setup corresponds to <i>Binary Classification</i> , and 4 Class Setup corresponds to <i>4-Class Classification</i>	40
5.3	Student's response instance for the skill <i>Constructing Explanations</i> , <i>Analyzing Data</i> and <i>Planning Investigations</i>	42
5.4	Occlusion study for the skill <i>Constructing Explanations</i> and <i>Analyzing Data</i>	60
5.5	Student's response instance for the skill <i>Constructing Explanations</i> , <i>Analyzing Data</i> in Occlusion study	61