Identifying Skills in Constructed Responses with Explainable NLP Models

Presenter: Smita Bhattacharya

M.Sc. in Data Science & Artificial Intelligence

Saarland University

Mentor: Sebastian Gombert

Leibniz Institute for Educational Research and Educational Information (DIPF)

Supervisors:

Prof. Dr. Vera Demberg

Saarland University

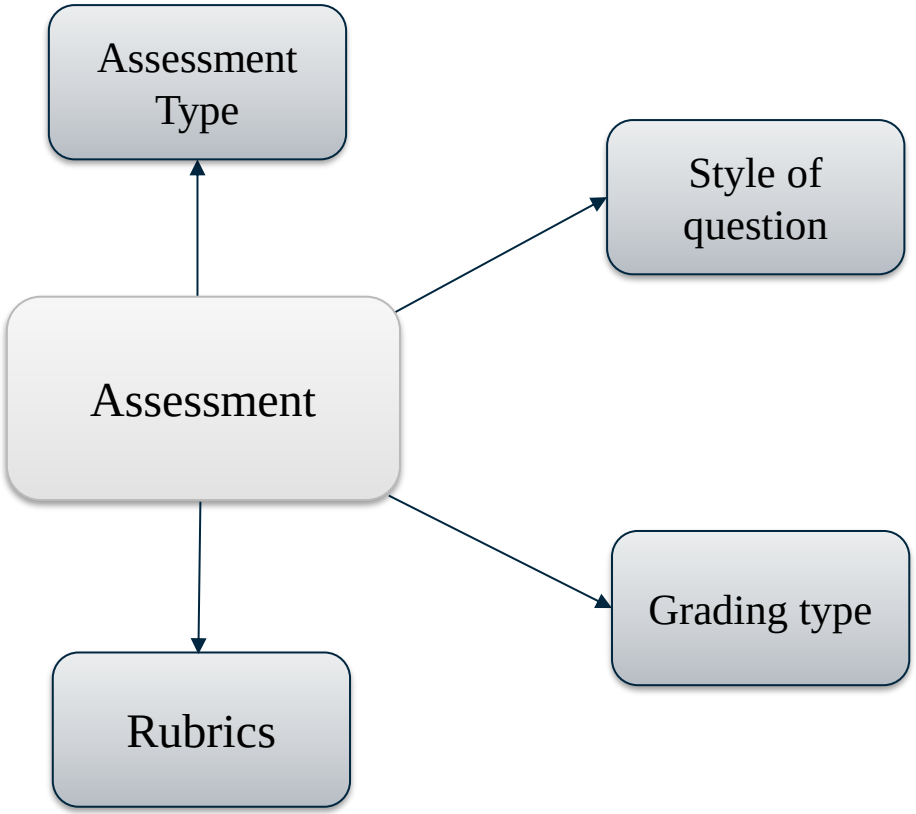
Prof. Dr. Hendrik Drachsler

Leibniz Institute for Educational Research and Educational Information (DIPF)

Agenda

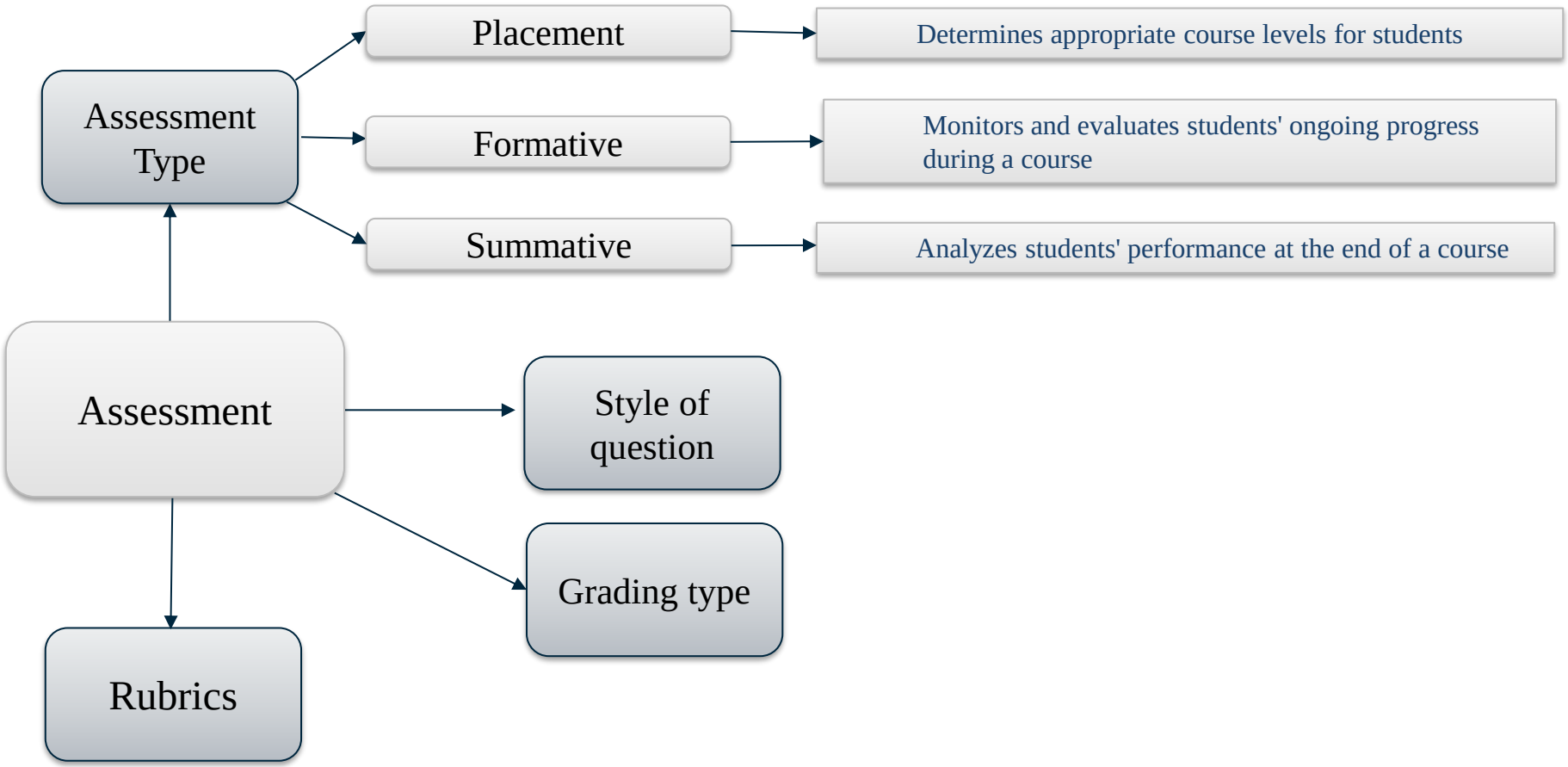
1. **Introduction**
2. Motivation
3. Related works
4. Datasets
5. RQ1, RQ2, RQ3: Methodology, results, discussion
6. RQ4: Methodology, results, discussion
7. RQ5: Methodology, results, discussion
8. Limitations
9. Future direction
10. Summary

Introduction



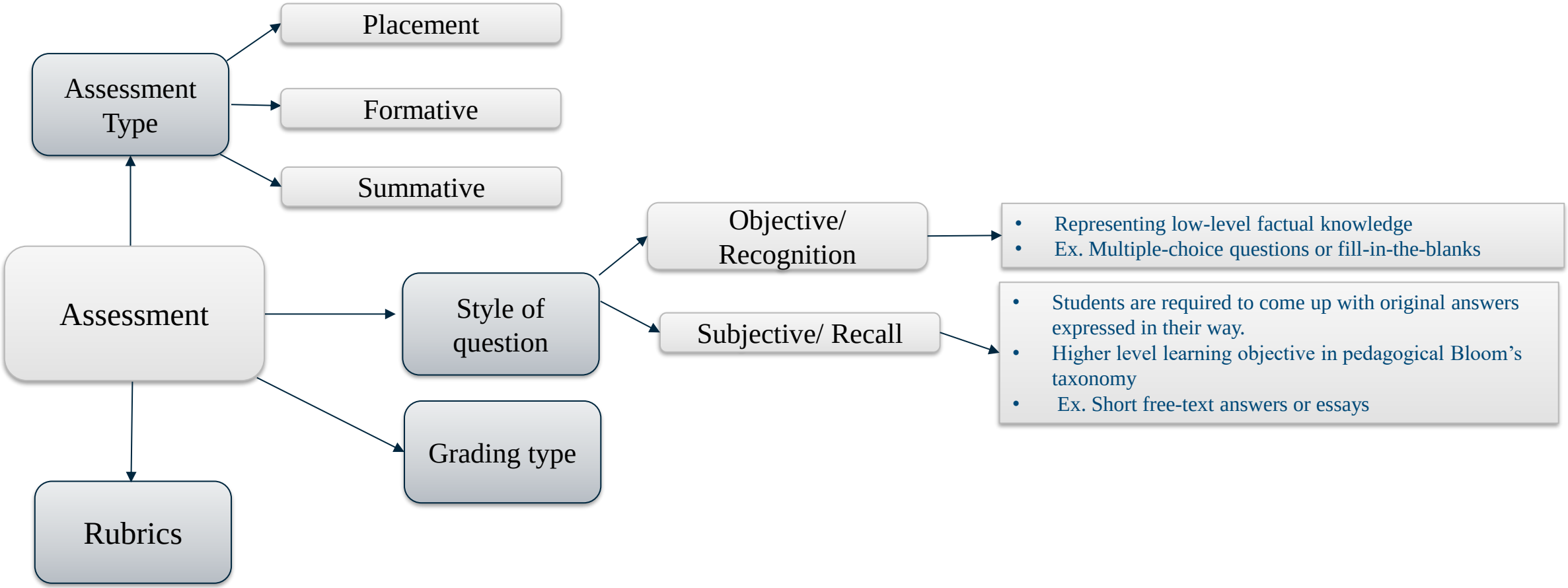
(Burrows et al., 2015), (Mertler, 2001)

Introduction



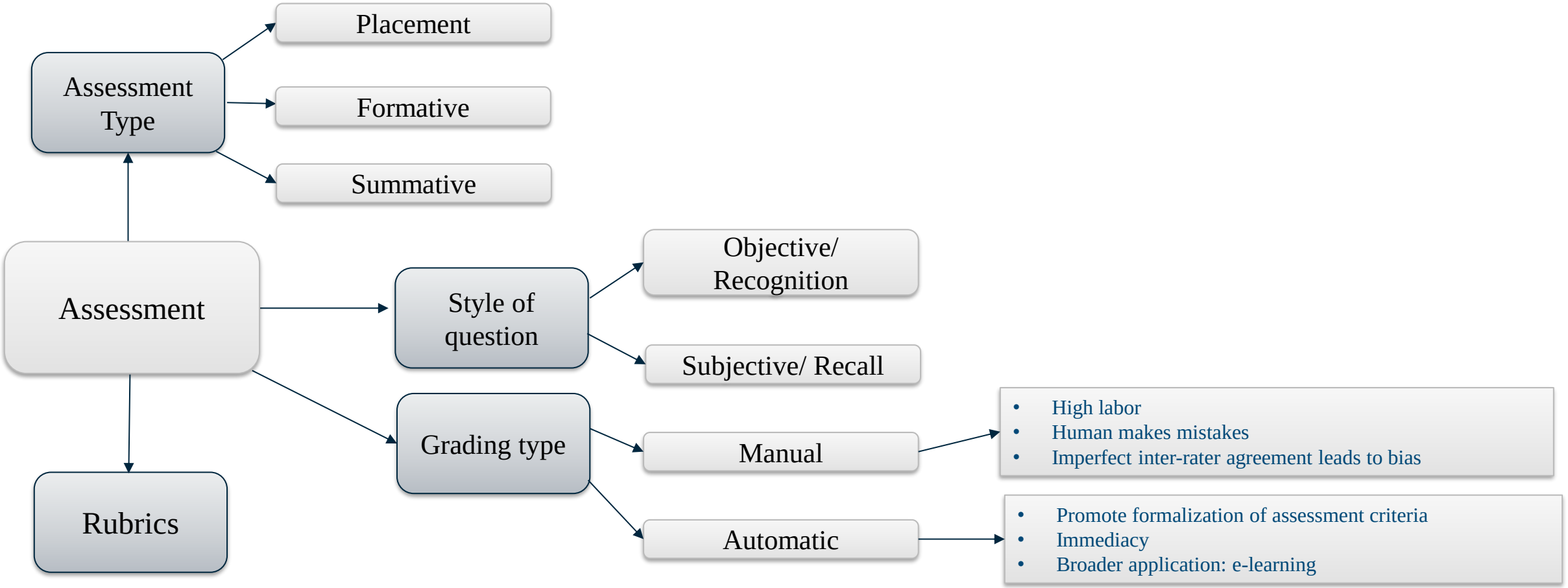
(Burrows et al., 2015), (Mertler, 2001)

Introduction



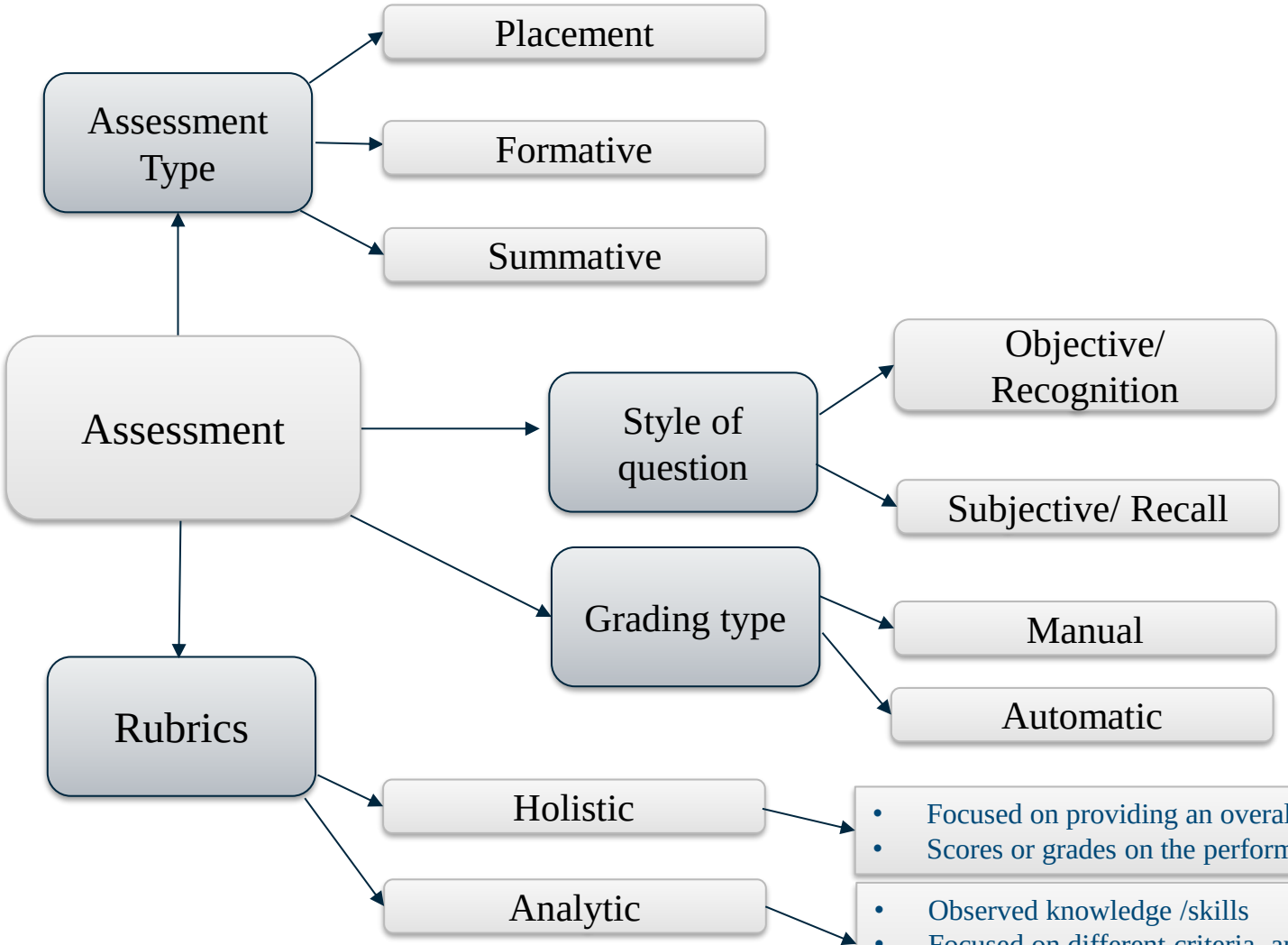
(Burrows et al., 2015), (Mertler, 2001)

Introduction



(Burrows et al., 2015), (Mertler, 2001)

Introduction



- Focused on providing an overall judgement of the work quality– Unidimensional
- Scores or grades on the performance of the student

- Observed knowledge /skills
- Focused on different criteria, and then the scores are summed up – Multi-dimensional
- This determines the individual student's strengths and weaknesses

(Burrows et al., 2015), (Mertler, 2001)

Introduction: Automatic short answer grading/ assessment

- Automatic short answer assessment (ASAA) or **automatic short answer grading (ASAG)**
  - **Application of NLP** in the **educational context**

- Burrows et al. (2015) conducted a comprehensive literature review focusing on ASAG
  - Predicting **holistic scores or grades**
  - **Discrete or continuous values** → Quality of a response

Introduction: ASAG example

Q. What do you observe when current flows through the conductor?

Explain your observation using energy terms.

Correct student response

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.

→ ASAG → Score/Grade→ Holistic Grading; Feedback→ Analytic Grading

Incorrect student response
The temperature remains the same

→ ASAG → Score/Grade→ Holistic Grading; Feedback→ Analytic Grading

Translated example of AFLEK (Gombert et al., 2023) dataset

Introduction:

- Online Learning Management Systems **(LMS)**, exemplified by **Moodle**, have gained significant importance in the post-COVID era.
- LMS emphasizes:
    - Formative assessments to provide feedback to students
    - Analytical grading is required
    - Different grading criteria :
        - Presence of domain knowledge (Gombert et al., 2023)
        - Development of different scientific skills(My thesis)

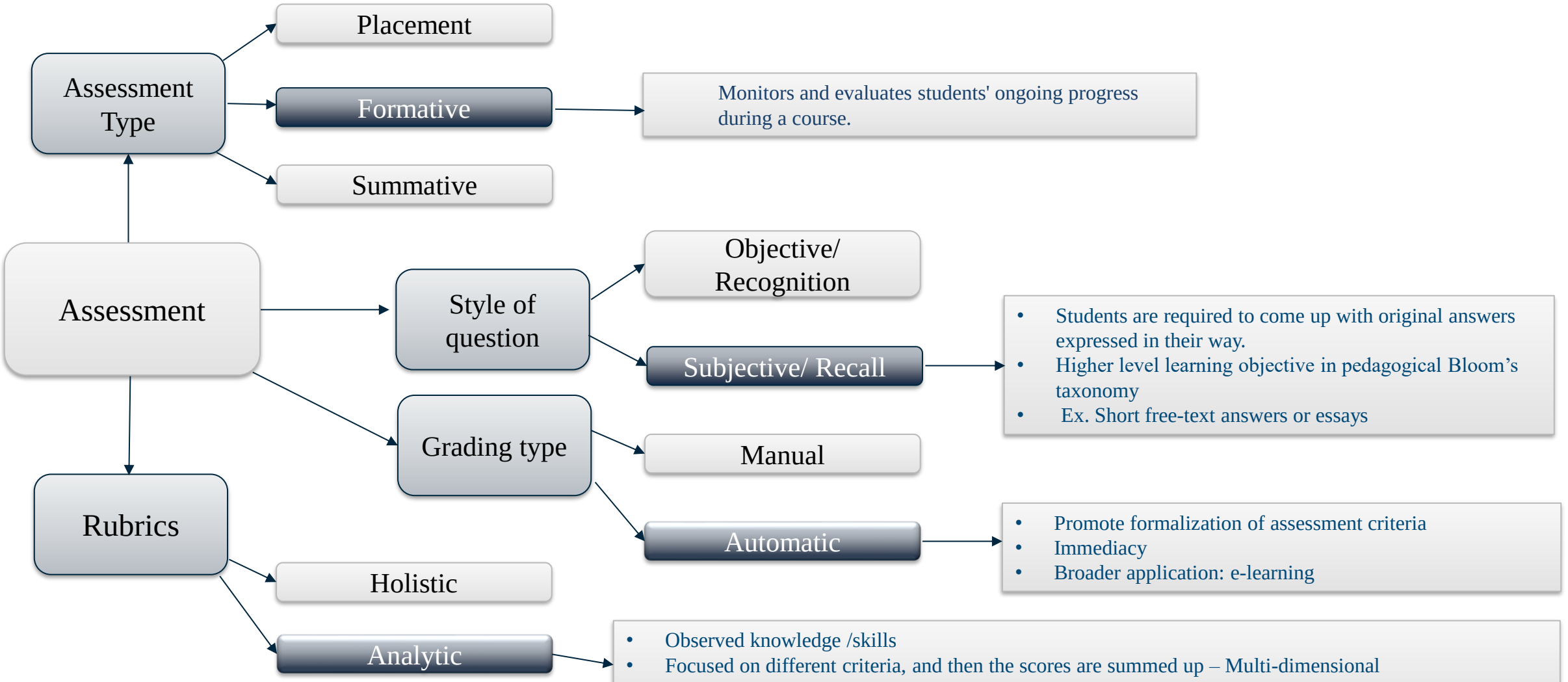
Agenda

1. Introduction
2. **Motivation**
3. Related works
4. Datasets
5. RQ1, RQ2, RQ3: Methodology, results, discussion
6. RQ4: Methodology, results, discussion
7. RQ5: Methodology, results, discussion
8. Limitations
9. Future direction
10. Summary

Motivation

- Analytical assessment by coding active-knowledge in German K12 students' responses(Gombert et al., 2023)
- STEM education requires:
  - Knowledge of a topic
  - Analytical skills for hands-on experiments
- This study aims to identify scientific skills in constructed responses
- Those skills are:
  - Constructing Explanations
  - Analyzing Data
  - Planning Investigations
- Lack of dataset → AFLEK short answer dataset (Gombert et al., 2023)
- Assessment is an inherently high-stakes scenario → where the system's explainability is essential for making its decision reliable to the stakeholders.

Introduction



(Burrows et al., 2015), (Mertler, 2001)

Agenda

1. Introduction
2. Motivation
3. **Related works**
4. Datasets
5. RQ1, RQ2, RQ3: Methodology, results, discussion
6. RQ4: Methodology, results, discussion
7. RQ5: Methodology, results, discussion
8. Limitations
9. Future direction
10. Summary

UNIVERSITÄT DES SAARLANDES

3. Related works

3.1 Automatic short answer grading

3.2 Different Scoring Mechanisms

3.3 Explainability

3.1 Related works for Automatic Short Answer Grading

- **Feature-based:** Identification of pre-defined concept lexicon (Callear et al., 2001)

- Pattern matching between different features of student response and reference answer(Bachman et al., 2002, Cutrone & Chang, 2010, Hahn & Meurers, 2012)

  → Score is assigned based on alignment quality

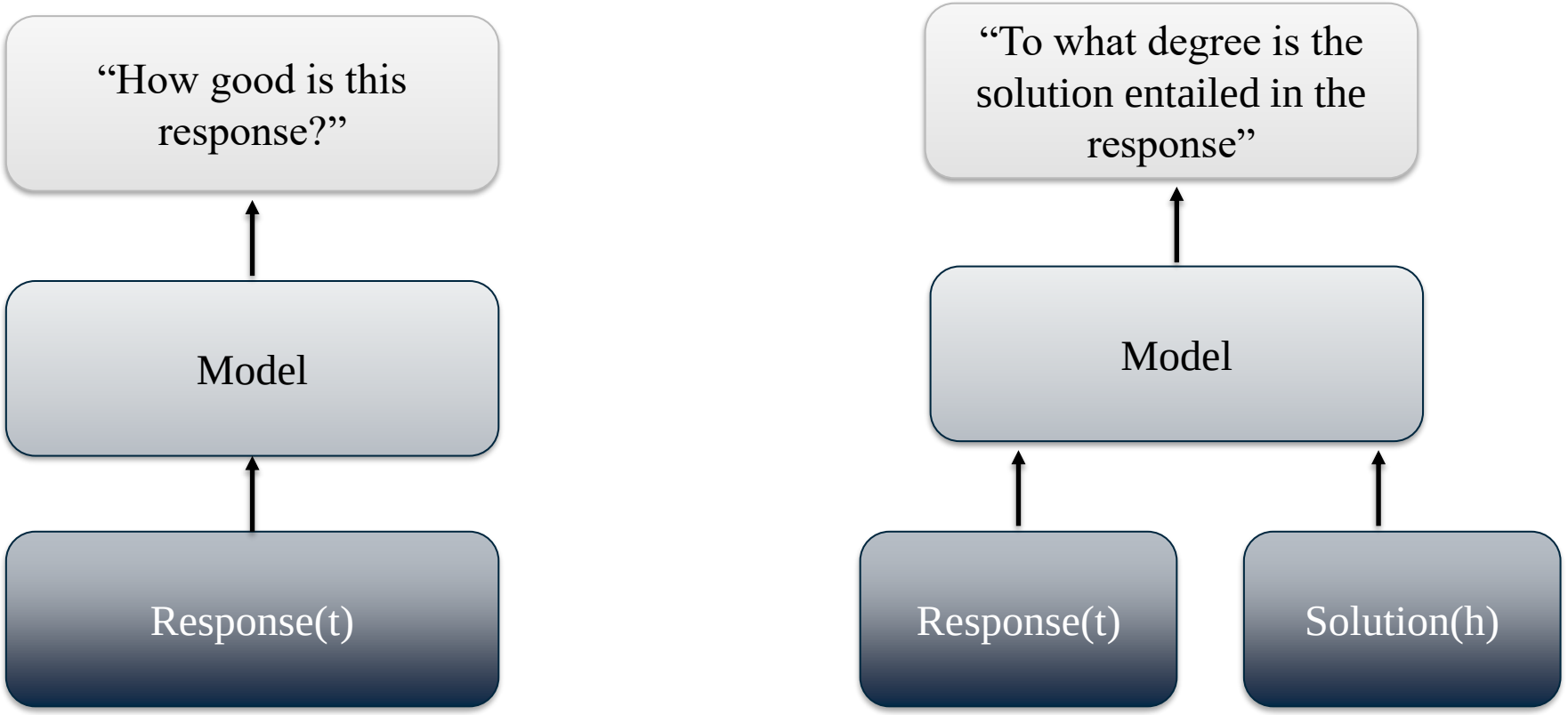
- **Supervised ML and Unsupervised ML:**

- KNN, Linear discriminant analysis are trained with lexical and semantic features (Meurers et al.,2011, Horbach et al., 2013, Crossley et al., 2016)

- LSTM model(Maharjan et al.,2018)

- Transformers-based architectures such as BERT(Devlin et al., 2019)
  - On English Dataset(Sung et al., 2019, Camus & Filighera, 2020)
  - On German Dataset (Gombert et al.,2023, Nath et al.,2023)

3.2 Different scoring mechanism



Instance-based vs. Entailment-based scoring
(Dzikovska et al., 2013, Burrows et al., 2015, Bexte et al., 2023)

3.3 Explainability

Glass box:

Blackbox:
- Lack of transparency of inner working
- Challenging to ensure it learns reliable patterns
- Transformer-based models
- Requires post hoc explanations

- Framework → PDR (Murdoch et al.,2019)
- Methods → Post hoc analysis

3.3 Explainability Framework & methods

PDR Framework(Murdoch et al.,2019**):**

- A structured approach to interpretability by considering these three key aspects

- **Predictive Accuracy:**
  - Refers to evaluating the model's fit quality using well-established machine-learning evaluation metrics

- **Descriptive Accuracy:**
  - Post-hoc methods are used to interpret whether a model's learned patterns align with coding guidelines
    - Methods: SHAP(**SH**apley **Additive** ex**P**lanations) (Lundberg et al., 2017)
    - Its estimation methods align the most with human intuition
  - Two types of post-hoc interpretability:
    - Local explanations/ prediction-level interpretability
    - Global explanations/dataset-level interpretability

- **Relevancy:**
  - Refers to stakeholders' requirements and their relevance to descriptive and predictive accuracy
  - High reliability of the predictive and descriptive models

Agenda

1. Introduction
2. Motivation
3. Related works
4. **Datasets**
5. RQ1, RQ2, RQ3: Methodology, results, discussion
6. RQ4: Methodology, results, discussion
7. RQ5: Methodology, results, discussion
8. Limitations
9. Future direction
10. Summary

5. Datasets: AFLEK Short Answer Dataset (Gombert et al., 2023)

- Analyse und Förderung von Lernverläufen zur Entwicklung von Kompetenzen
- Source of Data: Schleswig-Holstein Gemeinschaft Schule and Gymnasium
- Number of Moodle courses: 2 energy-related modules in Physics course
- Number of Periods: 6
- Duration: 45 minutes
- Presence of human-annotated evidence span
- Unit structure:
  - Started with a driving question about energy and associated phenomena that motivate the lessons
  - Each driving question is subdivided into three smaller questions:
    - Each sub-question assesses each scientific skill separately
      - Constructing Explanations
      - Analyzing Data
      - Planning Investigations

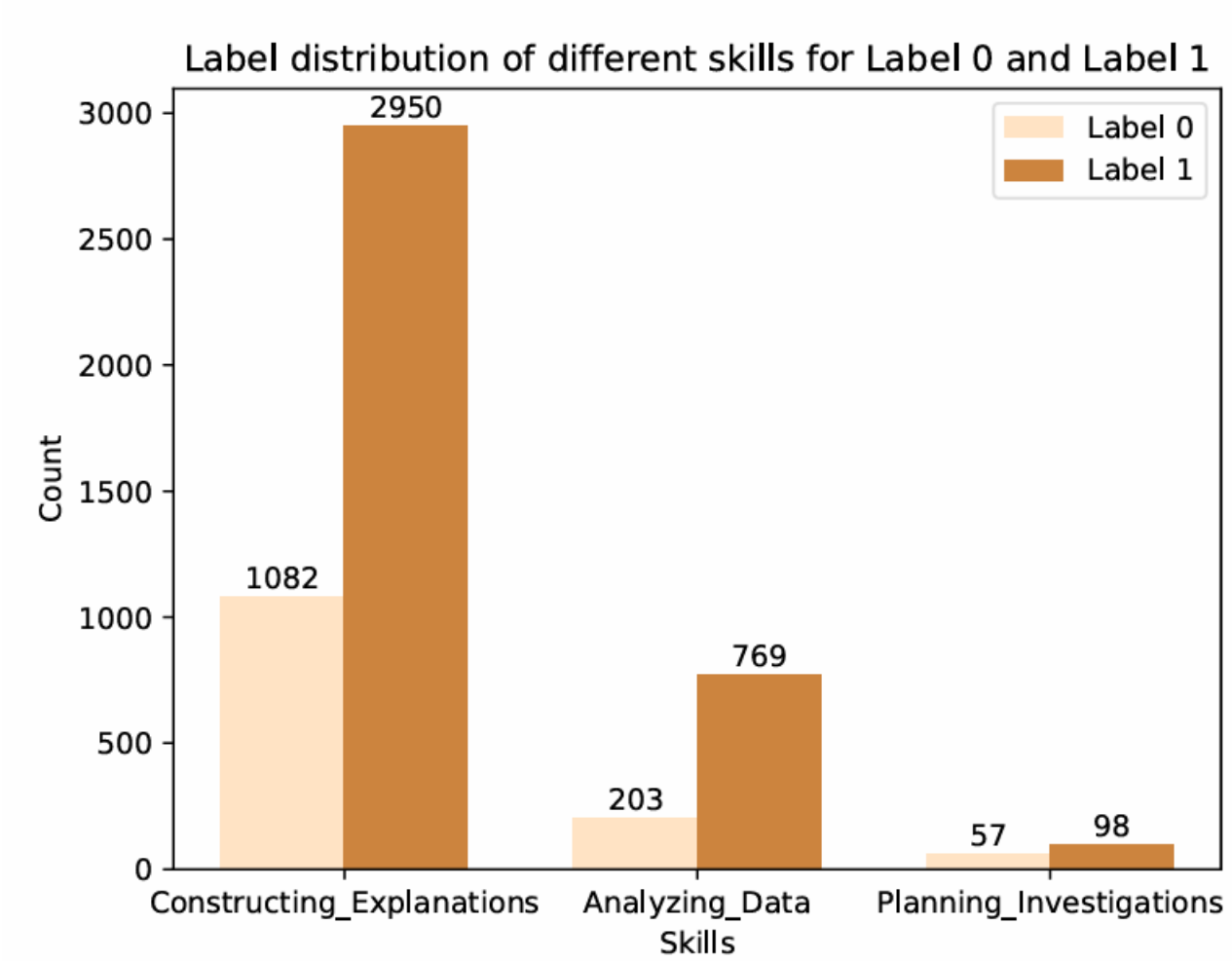
5. Datasets: CREG-TUE (Ott, 2014)

- The Corpus of Reading Comprehension Exercises in German (CREG-TUE)
- It was designed as a control group obtained by letting German native speakers answer a subset of the same CREG(Ott et al., 2012) dataset
- True label: “Extra Concept”, ” Correct”
- False label: “Missing Concept”, ” Blend”
- Secondary evaluation to validate our methods on an established benchmark for the task of short answer scoring.

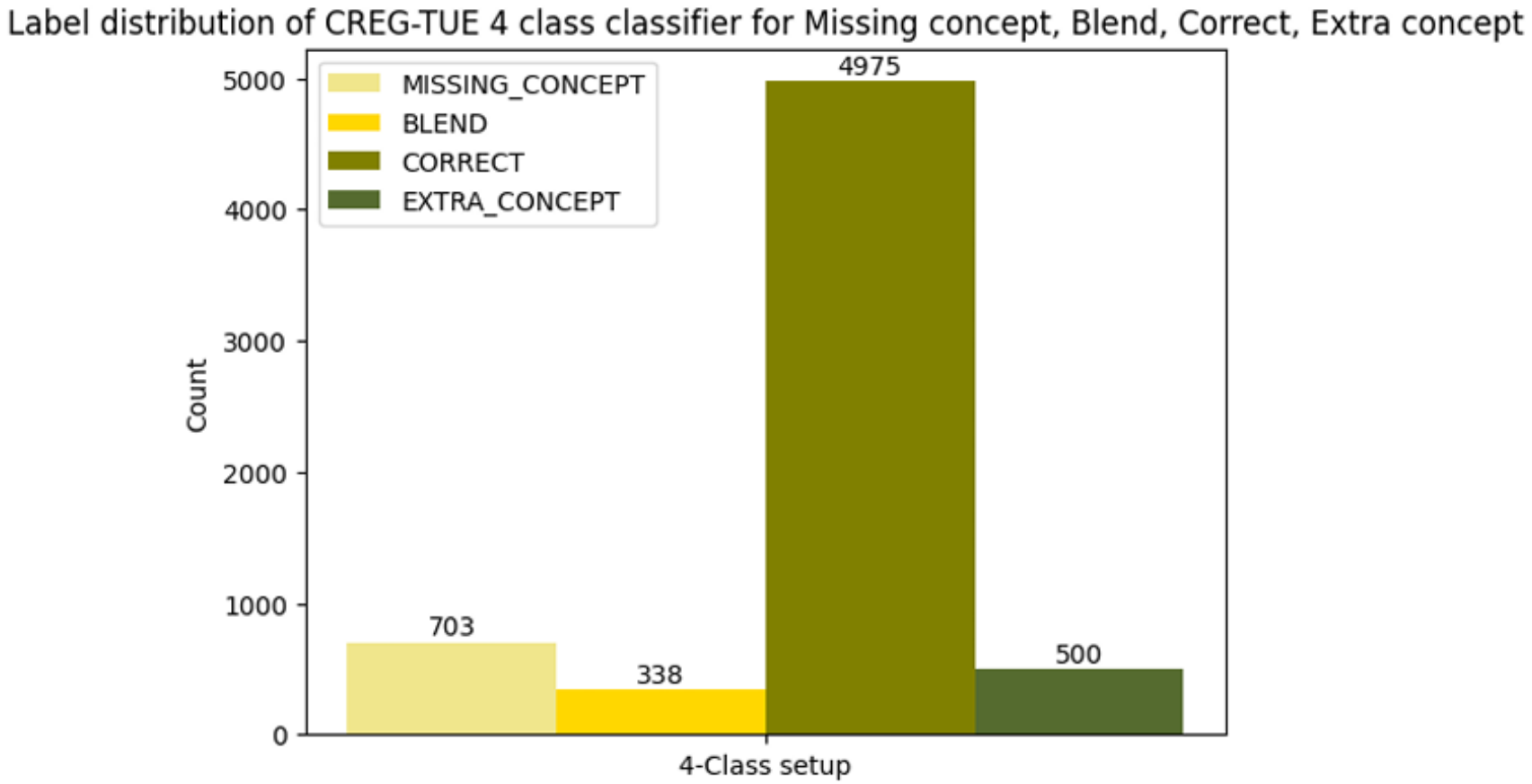
Descriptive statistics:

	AFLEK	CREG-TUE
No. of students	620	100
No. of questions	31	143
No. of reference answer	31	180
No. of responses	5159	6516
Label type	Binary	Binary Four class labels
Avg. number of responses per Student	8.1	65.16
Avg. Number of words per Response	25.11	7.01

AFLEK Dataset



CREG-TUE dataset



Agenda

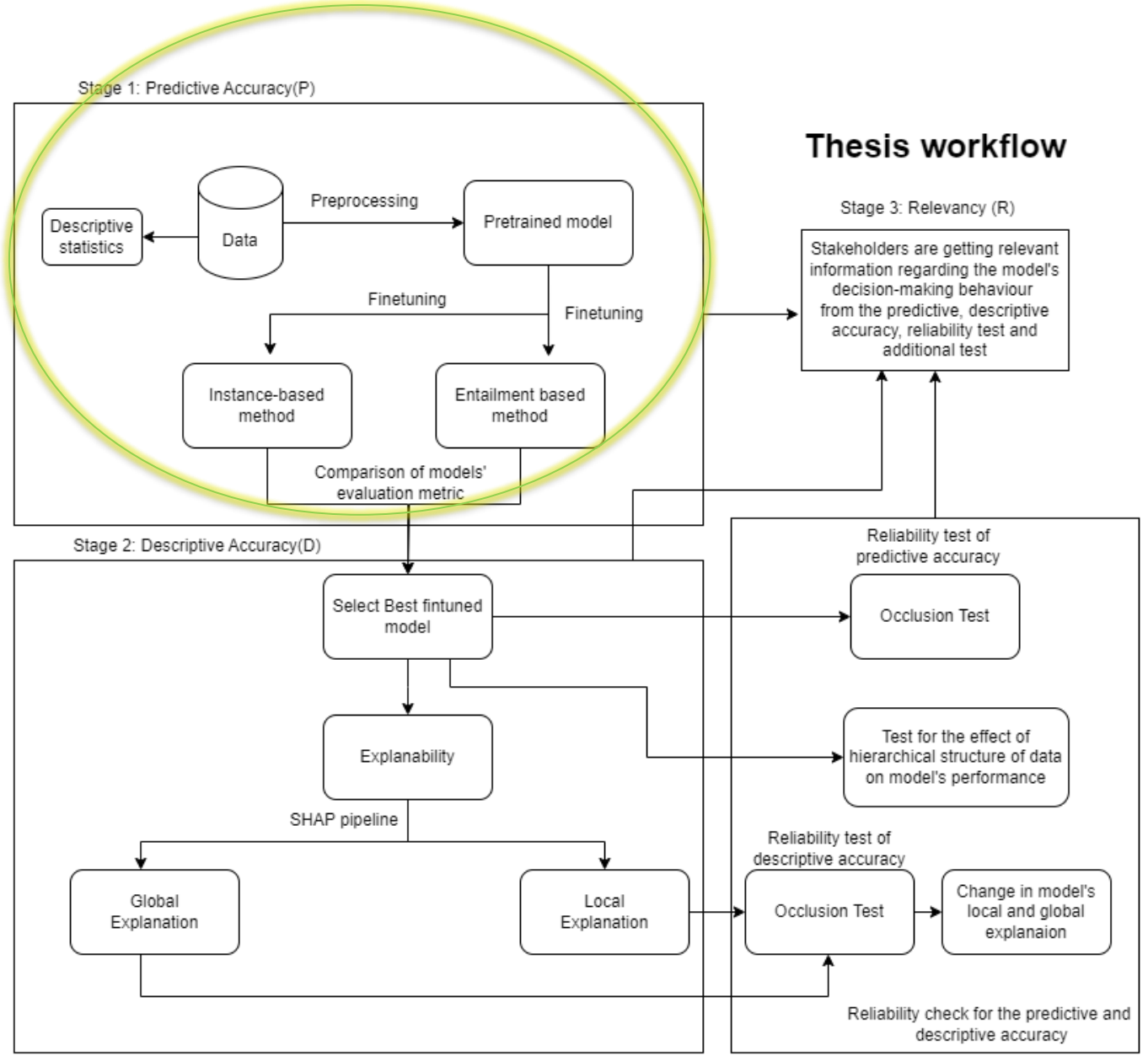
1. Introduction
2. Motivation
3. Related works
4. Datasets
5. **RQ1, RQ2, RQ3: Methodology, results, discussion**
6. RQ4: Methodology, results, discussion
7. RQ5: Methodology, results, discussion
8. Limitations
9. Future direction
10. Summary

- **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 apply to the domain-specific standard dataset (Secondary evaluation)?

  - Methodology
  - Result
  - Discussion

Workflow:

RQ1 & RQ2 & RQ3



Methodology:

- GermanBERT(base & large) (Chan et al., 2020), GermanELECTRA(base & large) (Chan et al., 2020) and XLM RoBERTa(base & large) (Conneau et al.,2020) models are used
- Scoring mechanisms: Instance-based and entailment-based
- Classification Type:
  - Binary for RQ1 and RQ2
  - 2-class setup(Binary) and 4-class setup for RQ3
- 5-fold cross-validation is used for training
- Loss function: Binary cross entropy
- Optimizer: AdamW(Loshchilov & Hutter, 2019)
- Evaluation metrics: Accuracy, Precision, Recall and F1 score

Results: AFLEK dataset

		Instance						Entailment					
	Skills	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: AFLEK dataset

		Instance						Entailment					
	Skills	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: AFLEK dataset

		Instance						Entailment					
	Skills	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: AFLEK dataset

		Instance						Entailment					
	Skills	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: AFLEK dataset

		Instance						Entailment					
	Skills	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: AFLEK dataset

		Instance						Entailment					
	Skills	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: AFLEK dataset

		Instance						Entailment					
	Skills	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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	95.7[illegible]	***1.0***	***1.0***	94.93	96.61	***1.0***
	P	92.[illegible]2	93.[illegible]4	***1.0***	9[illegible].14	9[illegible].83	74.[illegible]5	85.[illegible]7	95.3	**96.[illegible]5**	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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: AFLEK dataset

		Instance						Entailment					
	Skills	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**	**GBb**	**GEb**	**XRb**	**GBl**	**GEl**	**XRl**
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.[illegible]4	93.83	74.95	85.47	95.34	**96.25**	[illegible]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.[illegible]3	***89.78***	74.34	83.40	82.73	81.16	[illegible]**16**	87.12	65.78

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

Results: CREG-TUE

2 Class Setup												
	Instance						Entailment					
	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: CREG-TUE

2 Class Setup												
	Instance						Entailment					
	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: CREG-TUE

2 Class Setup												
	Instance						Entailment					
	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Results: CREG-TUE

2 Class Setup												
	Instance						Entailment					
	GBb	GEb	XRb	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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	GBl	GEl	XRl	GBb	GEb	XRb	GBl	GEl	XRl
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

- C: Constructing Explanations
- A: Analyzing Data
- P: Planning Investigations

- GBb, GBl: GBERT-base, GBERT-large
- GEb, GEl: GELECTRA-base, GELECTRA-large
- XRb, XRl: XLMRoBERTa-base, XLMRoBERTa-large

- **Bold:** Best model within a method
- ***Bold Italics:*** Best model across methods

Discussion

RQ1: To what extent can scientific skills be detected in students' free-text responses using different Transformers language models (predictive accuracy)?

- Addressed the first desideratum of the PDR framework
- Applied three different Transformers models(base and large) to solve the problem
- Instance vs entailment scoring mechanism
- Result shows the usage of Transformers architecture is plausible to automate the identification of the scientific skills in constructed responses efficiently

Discussion

RQ2: Does adaptation of entailment-based scoring improve over instance-based scoring?

- For the AFLEK dataset
  - In two out of three skills, the entailment-based GBERTlarge model achieved a better f1 score

- For CREG-TUE dataset:
  - 2-class setup: In entailment-based scoring, GBERTbase and GBERTlarge performed better
  - 4-class setup: All models in entailment-based scoring performed better(GBERTlarge having the best f1-score)

Discussion

RQ3: To what extent can the same models and the different scoring techniques apply to the domain-specific standard dataset (Secondary evaluation)?

- Shows a similar trend as the AFLEK dataset
- Additionally, the result highlights the increased difficulty for models to classify the responses in four class categories accurately(Drop in F1 score in 4-class setup)
- Entailment-based model works better with more complex classifier

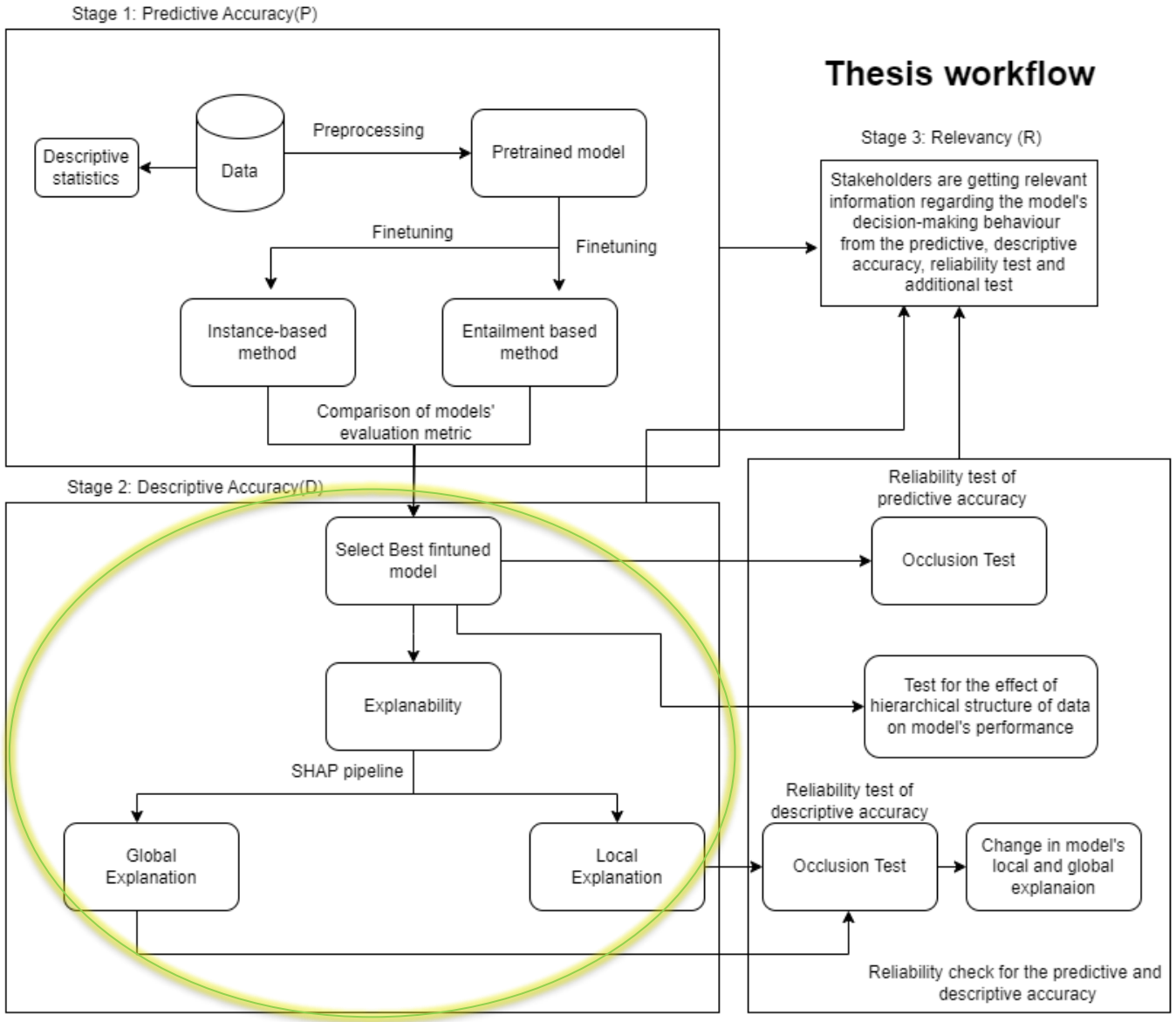
Agenda

1. Introduction
2. Motivation
3. Related works
4. Datasets
5. RQ1, RQ2, RQ3: Methodology, results, discussion
6. **RQ4: Methodology, results, discussion**
7. RQ5: Methodology, results, discussion
8. Limitations
9. Future direction
10. Summary

RQ4: To what extent do input words considered important by the models for their predictions match human-coded ones (descriptive accuracy)?

Workflow:

RQ4



Methodology

Descriptive Accuracy:

- Best-performed model from the previous stage
  - Constructing Explanations: GBERT-large
  - Analyzing Data: GBERT-large
  - Planning Investigations: GELECTRA-large

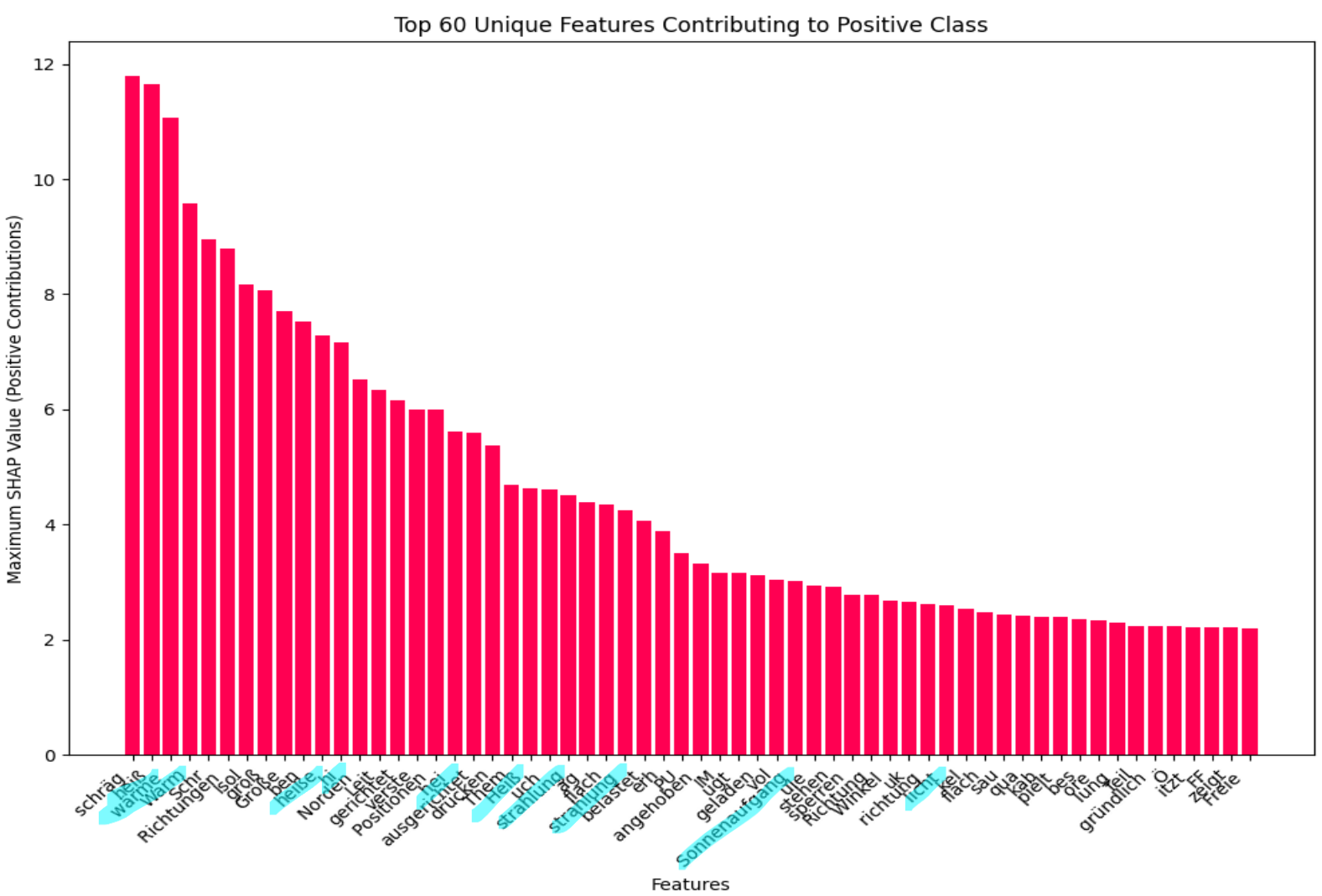
- Explanation model: SHAP
  - Local Explainability: Text plot
  - Global Explainability: Top 60 positive and negative SHAP features plotted in summary plot

Analysis

- Global explanation :
  - High SHAP value features:
    - Domain-specific energy-related term
    - Transformation/direction-related terms
    - Comparative terms
  - Negative SHAP value features: Mostly general words

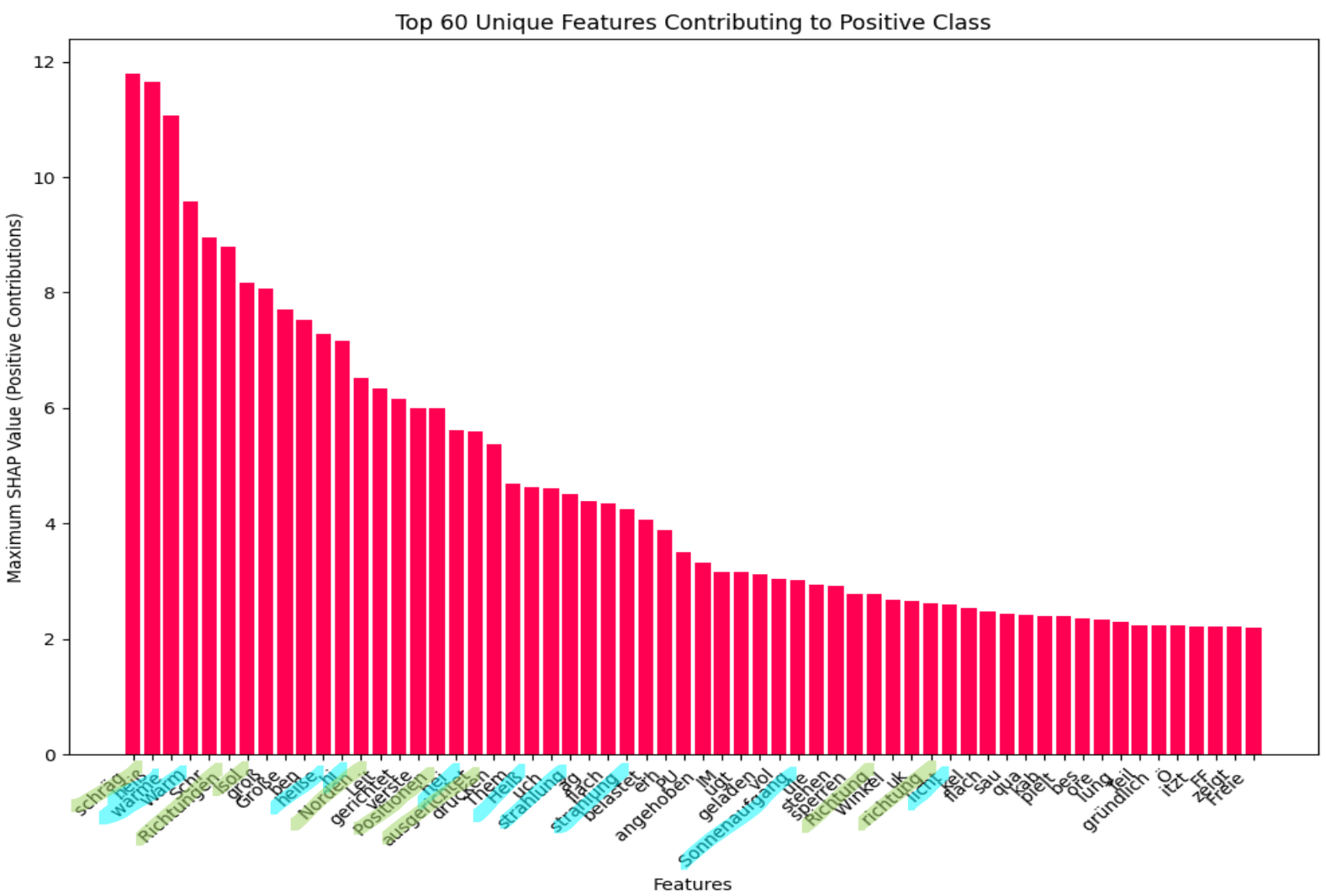
- Local explanation:
  - Considerable overlap between the human-annotated evidence span and the words provided positive SHAP value by the model(With few exceptions)

Constructing Explanations: Positive Global Explainability



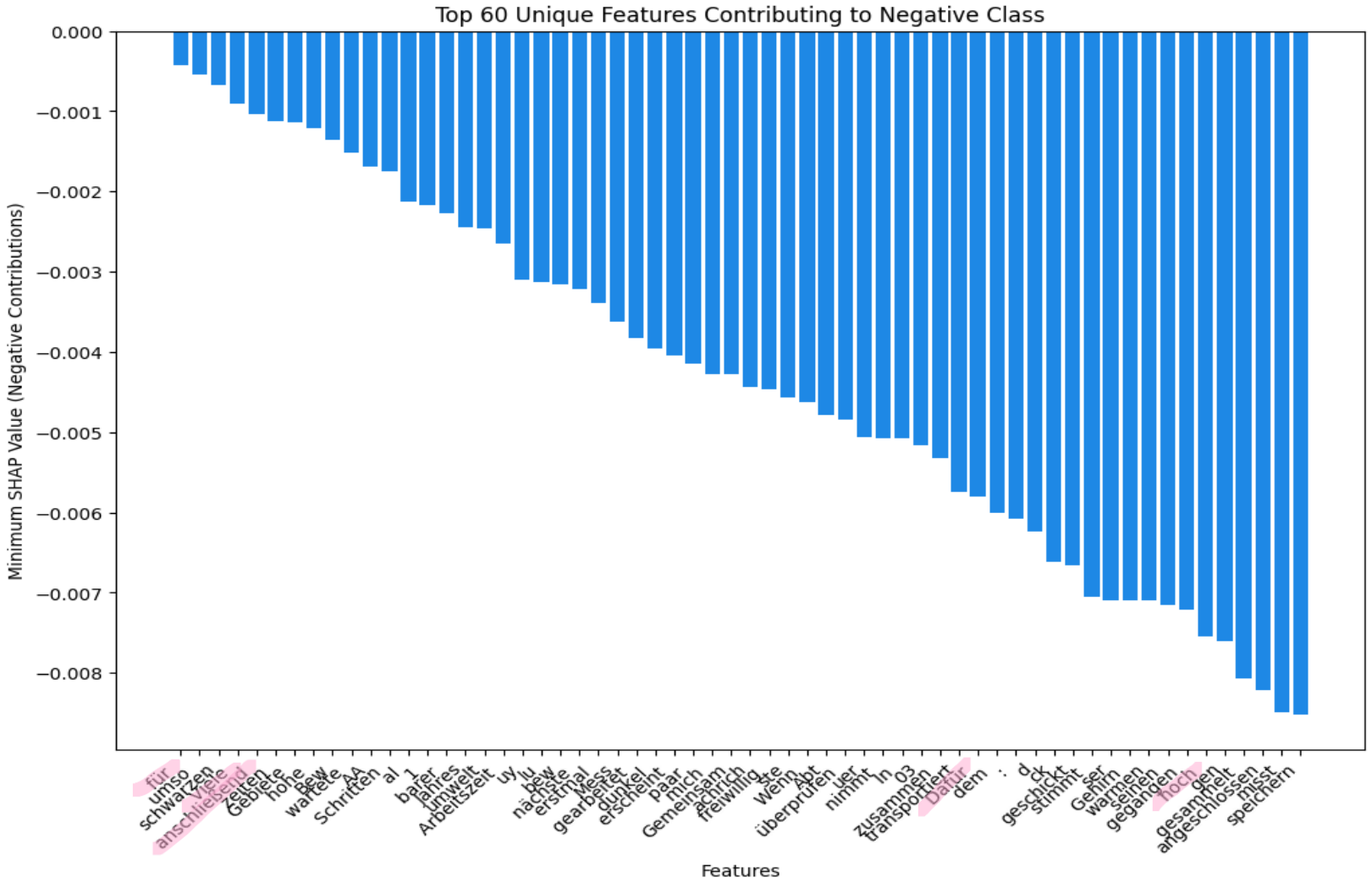
- Responses collected from modules on the transformation of the different forms of energy
- Features received higher positive SHAP value
- Energy-related features/words
- "strahlung" , "heiß" ,"hi", "hei" ,"heiße" ,"Heiß" ,"warme", "warm" ,"Sonnenaufgang" ,"licht"

Constructing Explanations: Positive Global Explainability



- Responses collected from modules on the transformation of the different forms of energy
- Features received higher positive SHAP value
- Energy-related features/words
- "strahlung" , "heiß" ,"hi", "hei" ,"heiße" ,"Heiß" ,"warme", "warm" ,"Sonnenaufgang" ,"licht"
- Direction-related features/words
- "schräg" , "Richtungen" , "Isol" , "Norden", "positionen", "ausgerichtet" , "richtung" etc

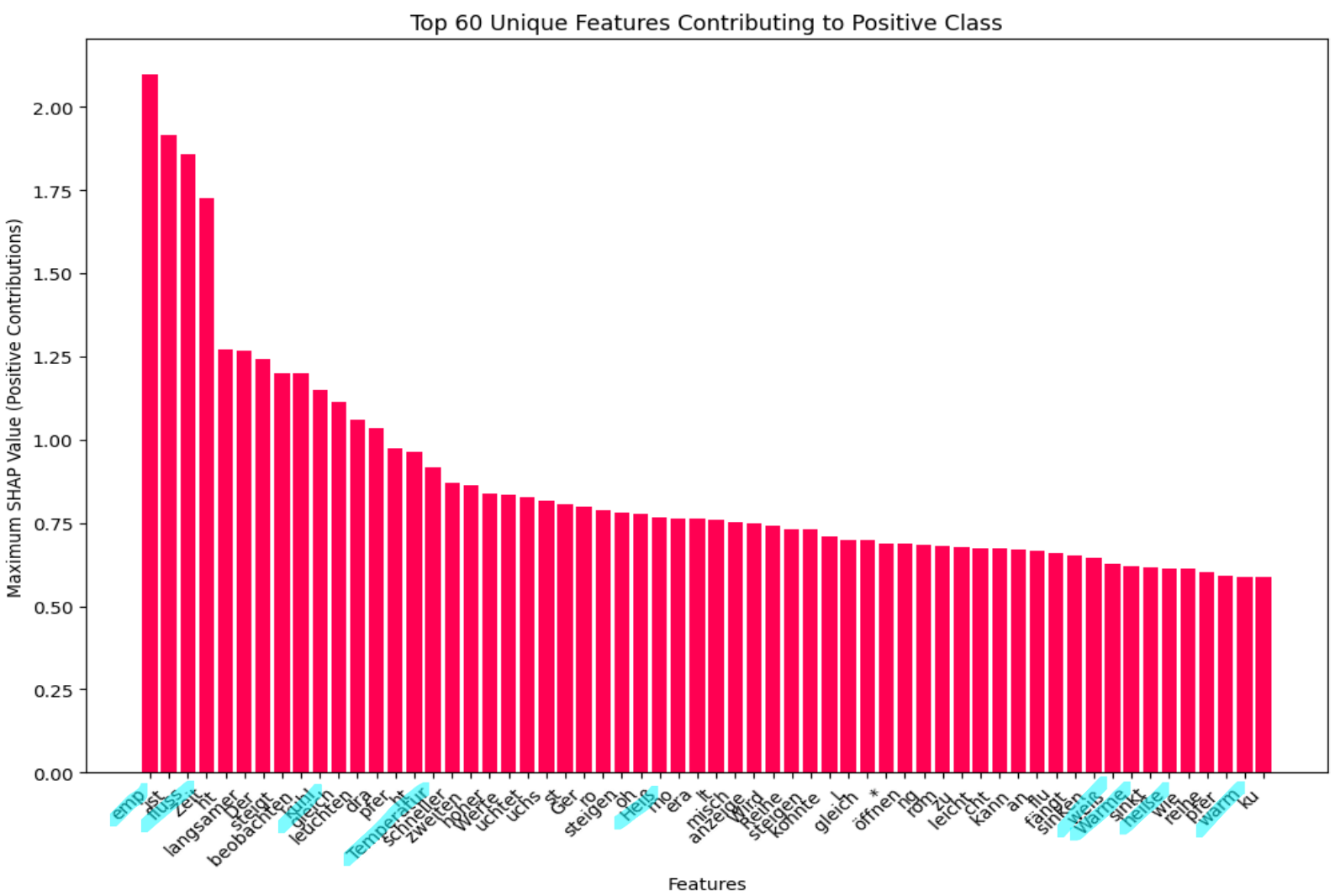
Constructing Explanations: Negative Global Explainability



Negative features(Exception)

- Negative features are very generalized and provide very little domain-specific information
- Features received negative SHAP value
- Description of a situation or can be used to construct certain explanations
- "für", "viele" , "anschließend" , "erstmal" , " Dafur",, "hoch"

Analyzing Data: Positive Global Explainability



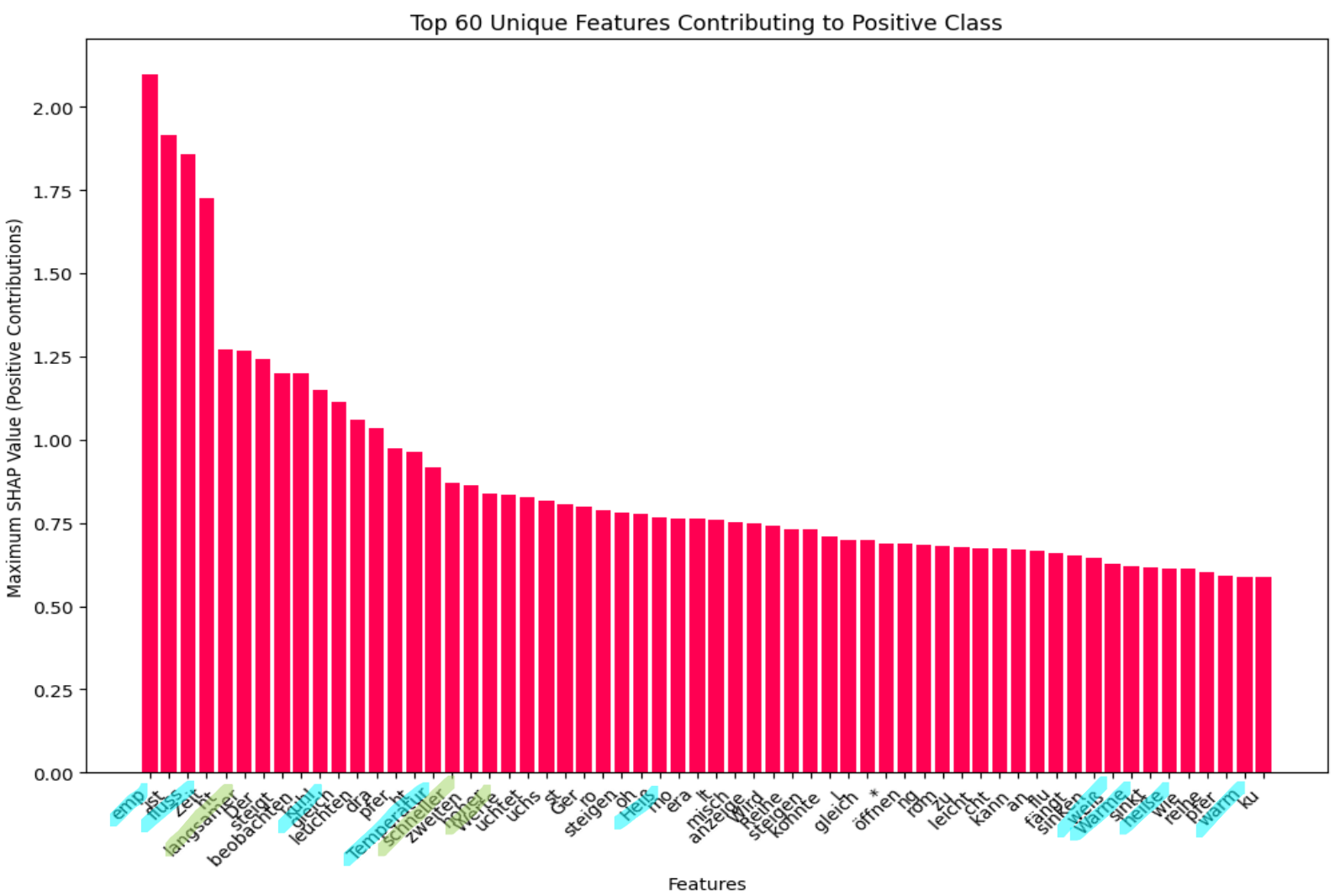
Domain Specific features

Comparative features

Features used for general analysis

- Responses collected from modules on the transformation of the different forms of energy
- Features received higher positive SHAP value
- Domain-specific energy relaed words
- "**temp**erature", " beein**fluss**en“, "Temperatur"(≃ 1.0), "Heiß"(> 0.75), "weiß" (< 0.75), "heiße"(<0.50)

Analyzing Data: Positive Global Explainability



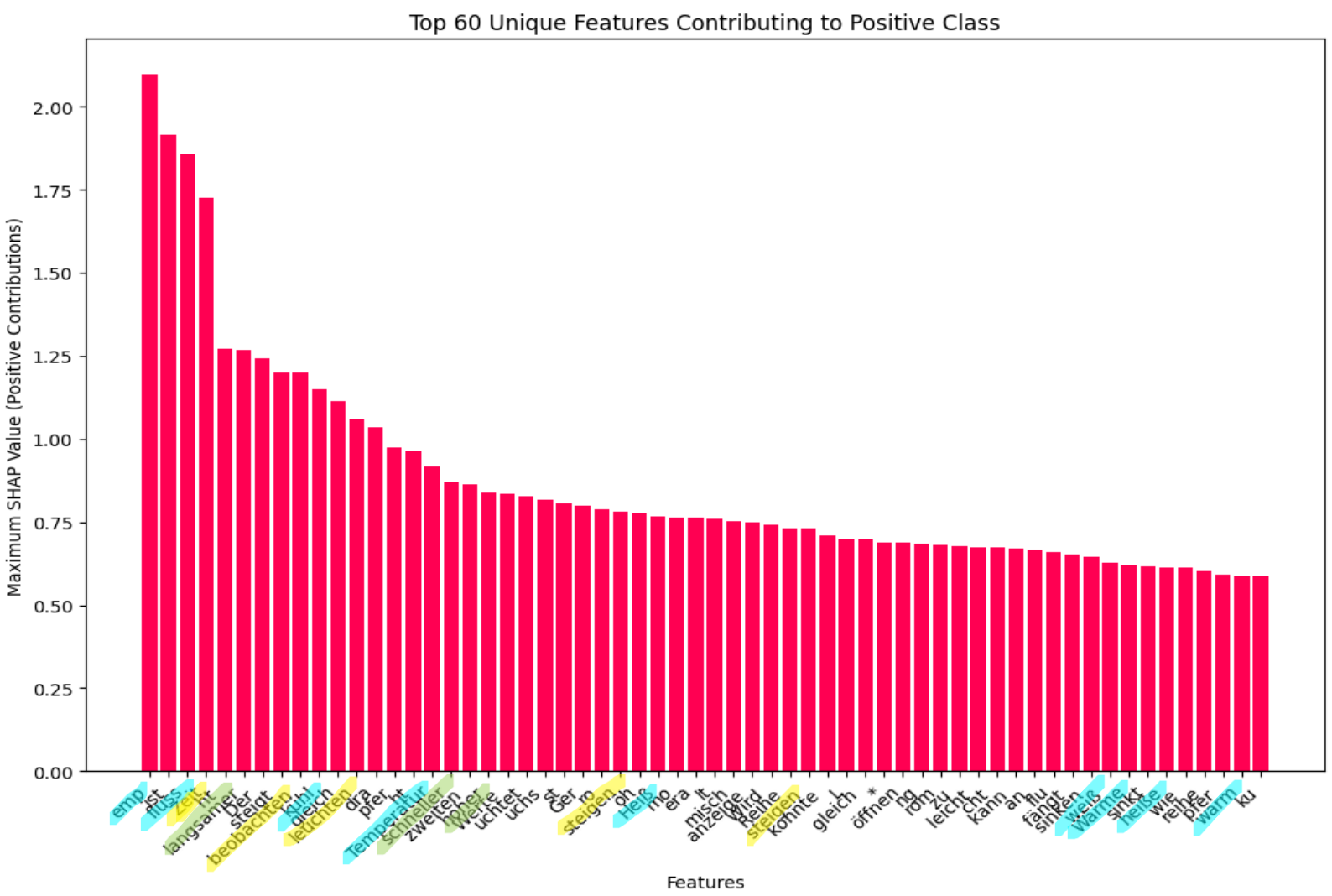
Domain Specific features

Comparative features

Features used for general analysis

- Responses collected from modules on the transformation of the different forms of energy
- Features received higher positive SHAP value
- Domain-specific energy relaed words
- "**temp**erature", " beein**fluss**en“, "Temperatur"(≃ 1.0), "Heiß"(> 0.75), "weiß" (< 0.75), "heiße"(<0.50)
- Comparative words: "langsamer"(1.75), "schneller" (< 1.00), "hoher"(< 1.00)

Analyzing Data: Positive Global Explainability



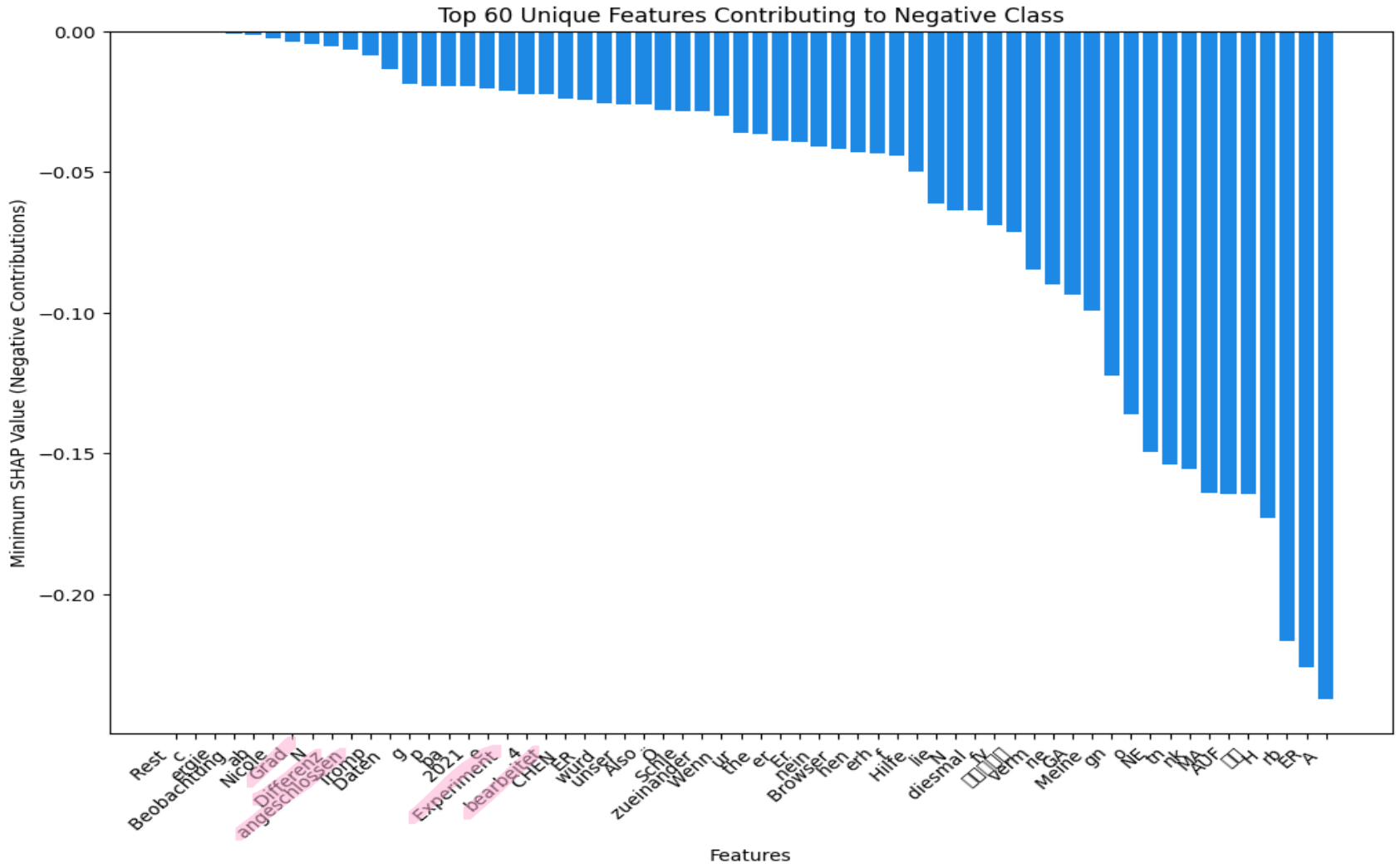
Domain Specific features

Comparative features

Features used for general analysis

- Responses collected from modules on the transformation of the different forms of energy
- Features received higher positive SHAP value
- Domain-specific energy relaed words
- "**temp**erature", " beein**fluss**en“, "Temperatur"(≃ 1.0), "Heiß"(> 0.75), "weiß" (< 0.75), "heiße"(<0.50)
- Comparative words: "langsamer"(1.75), "schneller" (< 1.00), "hoher"(< 1.00)
- Words used for analysis:"Zeit"(< 1.875), "beobachten"(< 1.25), leuchten(< 1.25), "steigen"(< 1.00)

Analyzing Data: Negative Global Explainability

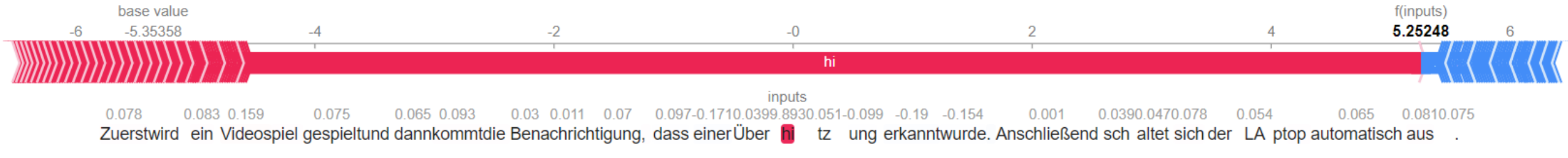


- Zero SHAP value features: "Rest", "C", "ergie", and "Beobachtung"
- Low negative SHAP value features: "Differenz"(<-0.05), "angeschlossen"(<-0.05), "Grad"(<-0.05), "Experiment"(<-0.05), "bearbeitet"(<-0.05)
- Highly negative SHAP value features are very generalized

Constructing Explanations: Local Explainability

Q. Notiere die anschließend deine Beobachtungen unter dem Video.

→ Zuerst wird ein Videospiel gespielt und dann kommt die **Benachrichtigung, dass einer Überhitzung erkannt wurde.** Anschließend schaltet sich der LAptop automatisch aus.



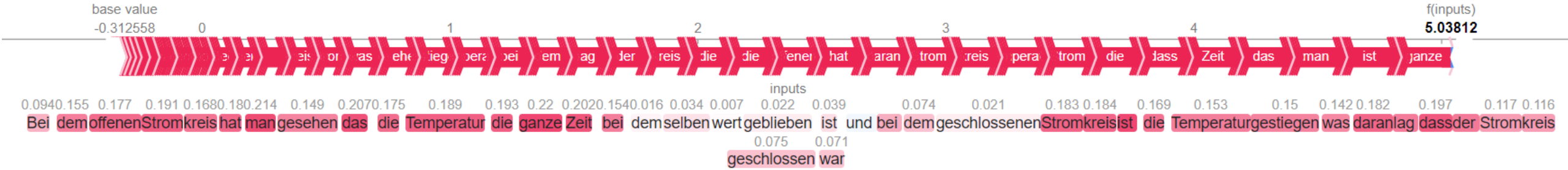
Text plot of positively classified instance

- Highly contributing features: "hi"(9.893), "Über"(0.039), "tz"(0.051) of the word Überhitzung,"Benach"(0.03), "richtigung"(0.07), "dass"(0.097)
- Exception: "einer"(-0.17) "ung"(-0.099)
- Important features outside evidence span: "Zuerst", "dann"

Analyzing Data: Local Explainability

Q. Was beobachtest du wenn der Leiter stromdurchflossen ist? Erkläre deine Beobachtung mithilfe von Energiebegriffen.

→ **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



Text plot of positively classified instance

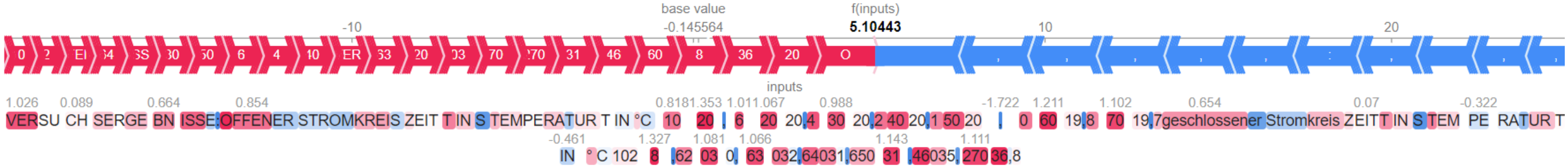
- Highly contributing features: "Temperatur"(0.193), "die"(0.22), "ganze"(0.202), "Zeit"(0.2020), "gestiegen"(0.15)

Planning Investigations: Local Explainability

Q. Notiert euch die Werte in der unten stehenden Tabelle.

→ VERSUCHSERGEBNISSE:OFFENER STROMKREIS ZEIT T IN S TEM PERATUR T IN °C **10 20,6 20 20,4 30 20,2 40 20,1 50 20,0 60 19,8 70 19,7**
geschlossener Stromkreis ZEIT T IN S TEMPERATUR T IN ° C **1028,62030,63032,64031,65031,46035,27036,8**

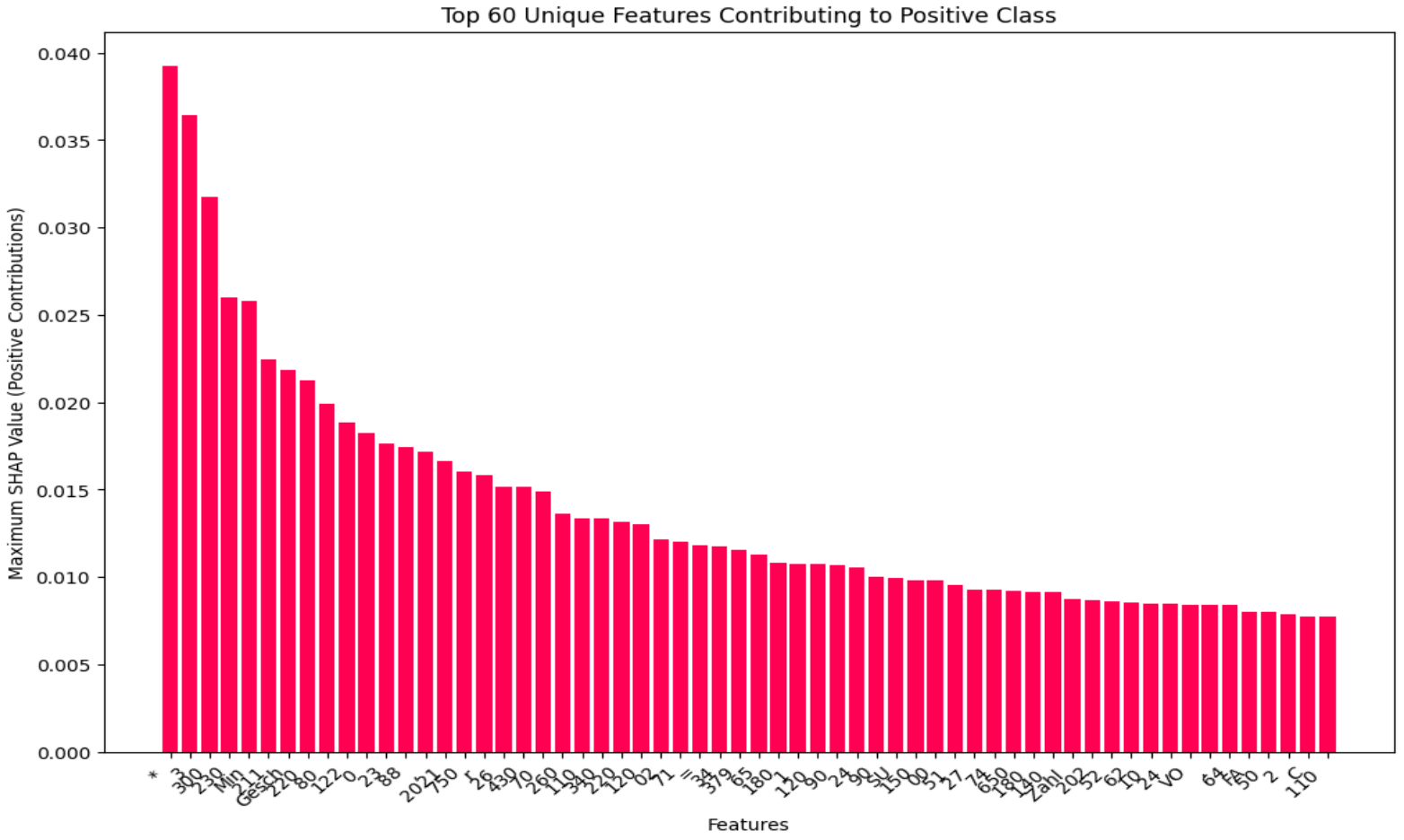
→ Evidence span is not present for this skill
→ Only the presence of the numbers makes an instance positive



Text plot of positively classified instance

- 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, 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.

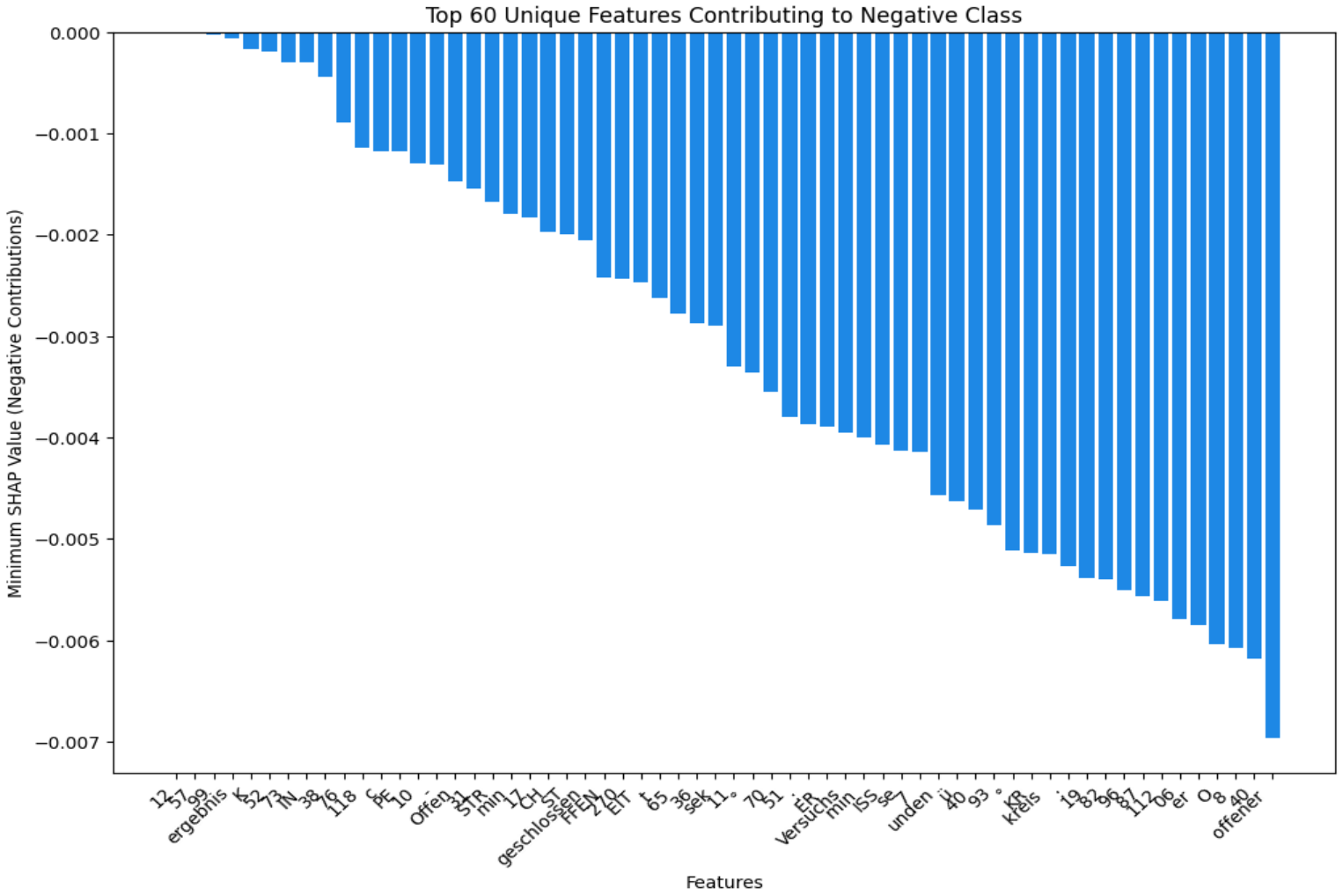
Planning Investigations: Global positive Explainability



- High positive SHAP features are mostly numbers which students have collected from the experiments
- Few textual features:
- “Min”, “Zahl” etc.

- This pattern probably explains
  - Lower number of instances for training
  - Answers come from a single question
  - Template-like structure of the answers
  - Only the numbers entered in this template structure make the instances positive. Or else negative

Planning Investigations: Global negative Explainability



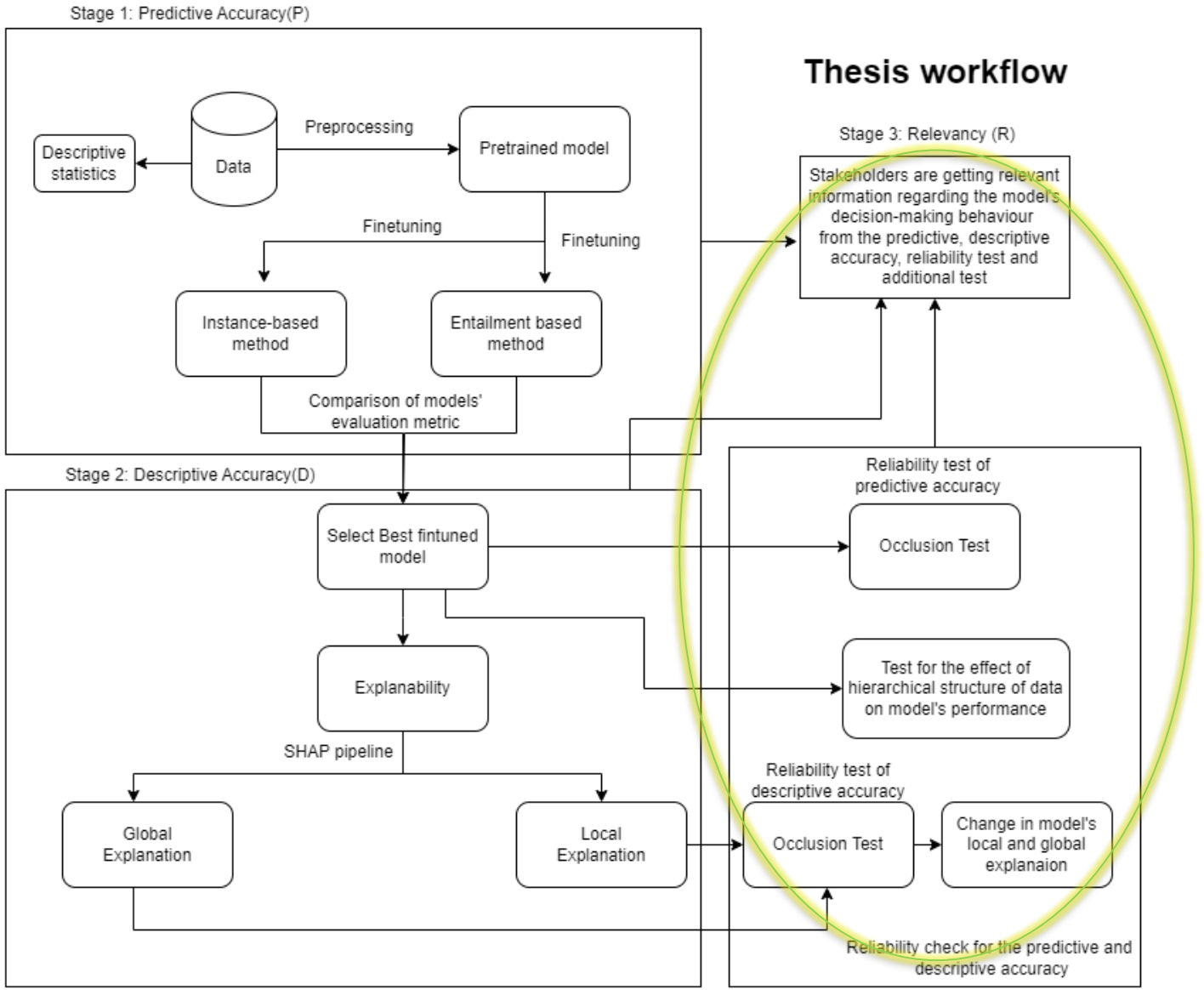
- Template-like structure of the responses leads to the textual features as negative features
- Model does not learn any new and important information from the repetitive textual features

Agenda

1. Introduction
2. Motivation
3. Related works
4. Datasets
5. RQ1, RQ2, RQ3: Methodology, results, discussion
6. RQ4: Methodology, results, discussion
7. **RQ5: Methodology, results, discussion**
8. Limitations
9. Future direction
10. Summary

RQ5: To what extent the interpretation of these models' decision-making behavior is relevant to the stakeholders (Relevancy)?

Workflow:



RQ5

Methodology:

- Relevancy: Highly reliable predictive and descriptive accuracy of the model
- Reason: Elevate the trustworthiness of the stakeholders
- Method: Occlusion study (Zeiler and Fergus, 2014)
- Occluded dataset: MASKED human-annotated evidence span
  - Measures the changes in the model prediction
  - No change in model behaviour → Features not important
  - Degradation in the model's prediction and confidence of the explainer → Impactful feature

Occlusion study for Predictive accuracy:

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

- Lower precision: Larger false positive errors (Wrong identification of negative instance as positive)
- Lower recall: Larger number of false negative(Wrong identification of true positive as false negative)

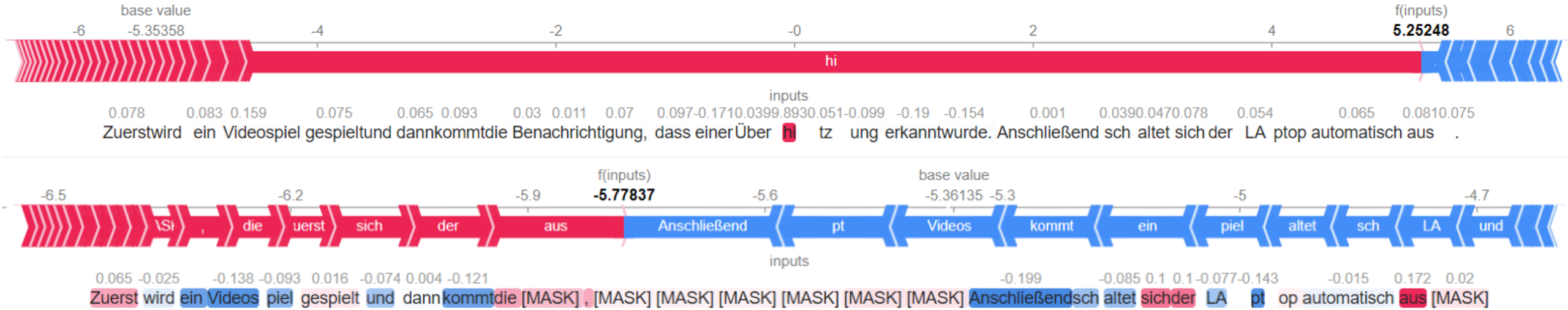
Occlusion study for descriptive accuracy:

Two scenarios are observed:

- Change in the explainer model's behaviour:
  - Local explainability changed in 2 ways
- Change in the average marginal contribution of the positive features which are out of the evidence span

Occlusion study for descriptive accuracy:

Scenario 1: Complete misclassification



Skill: Constructing Explanations
Top: Original text plot, **Bottom:** Occluded text plot

Occlusion study for descriptive accuracy:

Scenario 2: Right classification with lower explainer confidence

Skill: Analyzing Data

Top: Original text plot, **Bottom:** Occluded text plot

Change in average marginal contribution of features out of evidence span:

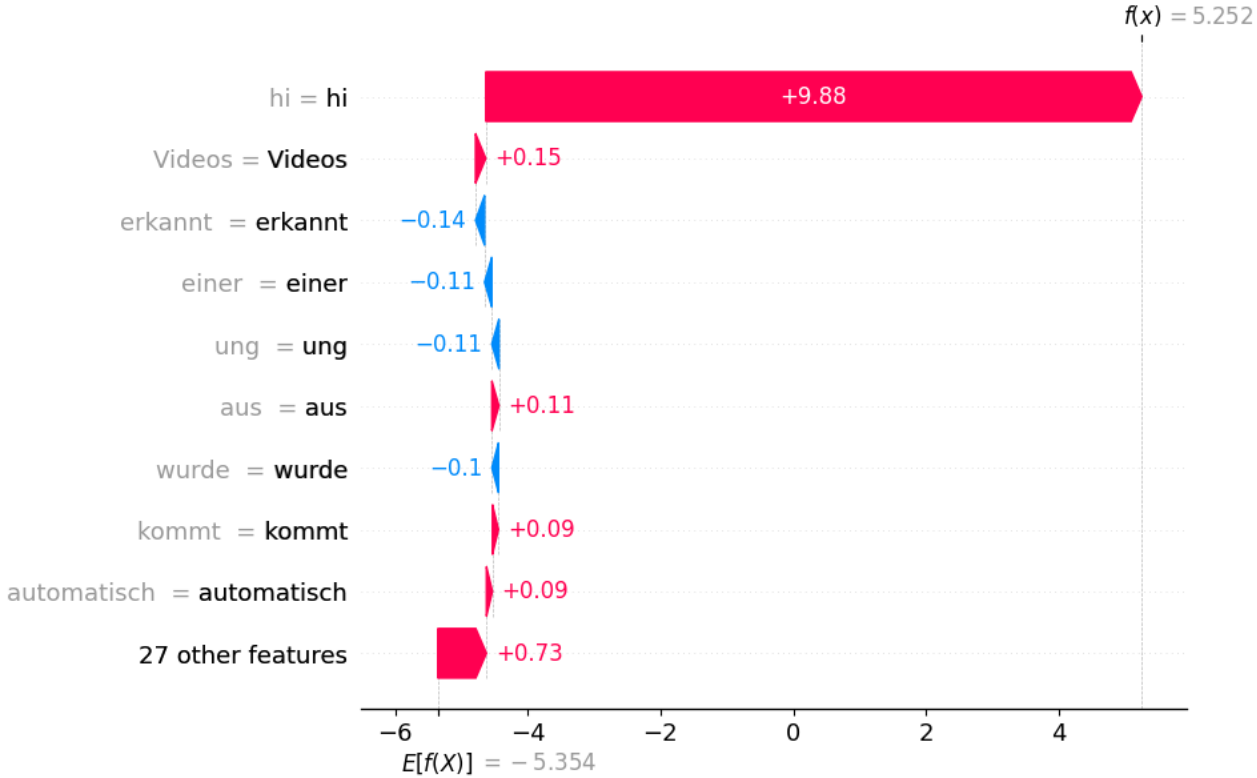
SHAP value calculation:

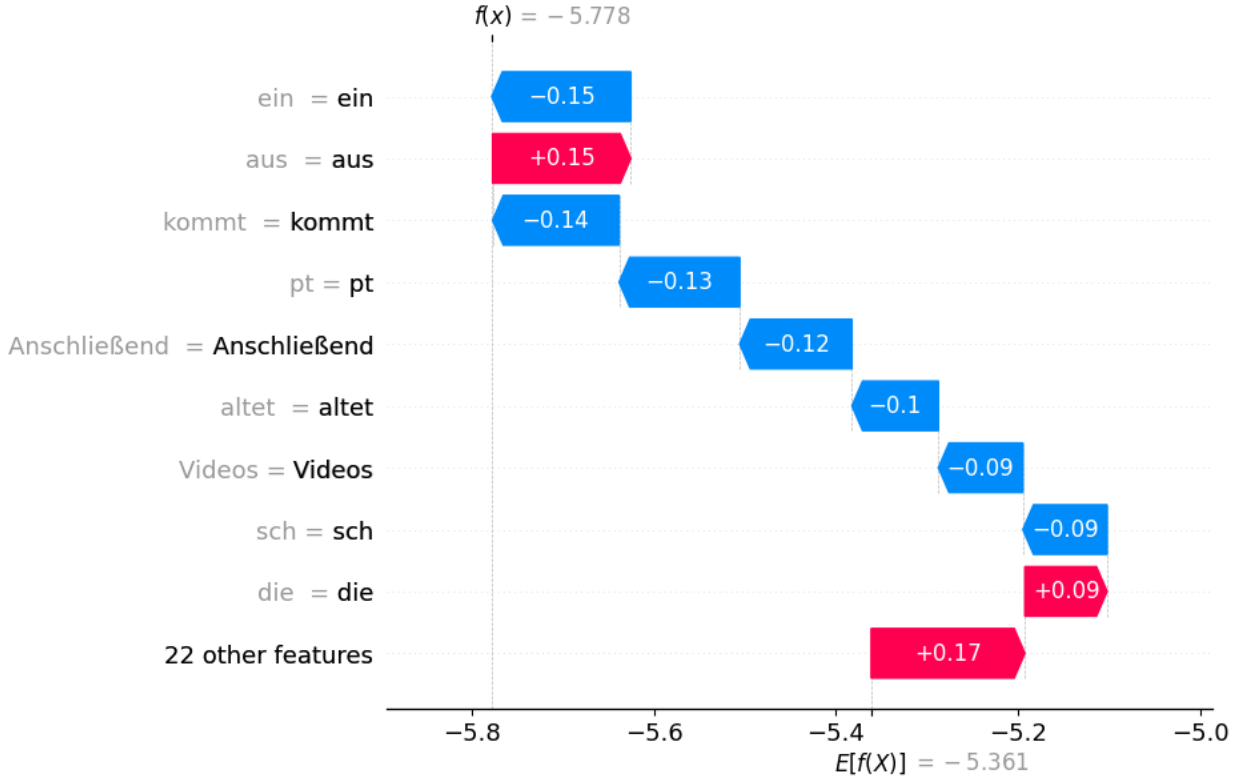
$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}}(x_{S \cup \{i\}}) - f_S(x_S) \right]$$

- 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.

Change in the average marginal contribution of features out of evidence span:

Skill: Constructing Explanations

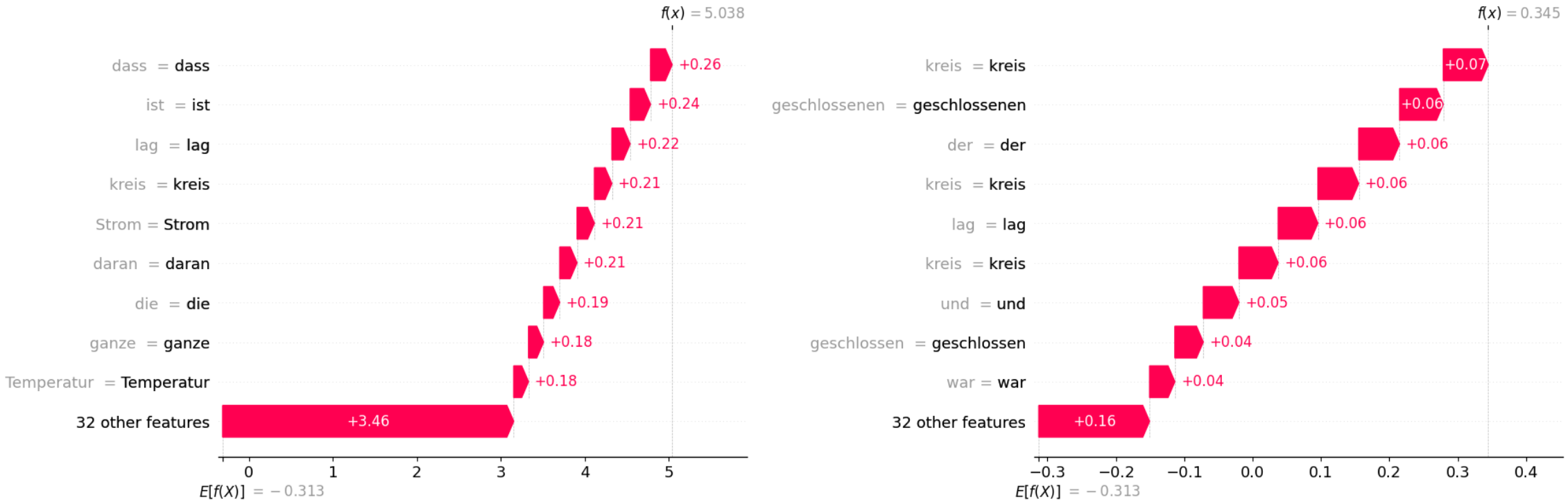




Top: Original waterfall plot, **Bottom:** Occluded waterfall plot

Change in the average marginal contribution of features out of evidence span:

Skill: Analyzing Data



Top: Original waterfall plot, **Bottom:** Occluded waterfall plot

Analysis using waterfall plot:

- The model's output drops from g(z′) = 5.252 to g(z′) = 0.345
- Declined marginal feature contribution in the model's prediction.
  - "kreis"
  - ϕ(original kreis) = +0.21
  - ϕ(masked kreis) = +0.06

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

Additional Analysis

- Hierarchical structure of the data is observed
- Three different Generalized mixed effect models (Bates et al., 2015) using R (R Core Team, 2024) with increasing random effect for each skill
- Applied ANOVA on these models to understand if there is any significant random effect on the model's performance
- H0: There is no significant random effect on the model's performance
- H1: There is significant random effect on the model's performance
- Significant level(α) = 0.05
- Student model: Pr(>Chisq): <2e-16***
- Reject the null hypothesis and accept the alternative hypothesis.
- This suggests that the model is more likely to be correct in predicting the responses of different students

Discussion

RQ5: To what extent the interpretation of these models' decision-making behavior is relevant to the stakeholders (Relevancy)?

- The predictive accuracy is quite reasonable
- Descriptive accuracy suggests
    - The models mostly learn for the right reason, with few exceptions
    - Learnt impactful features by the model match the human-annotated evidence span
- Occlusion study shows the reliability of predictive & descriptive accuracy
- Student-wise random effect showed for different students, the model might be more likely to predict the label correctly.
- Stakeholders are well aware of the pros and cons of the models' behaviours, which eventually helps in making further decisions

Agenda

1. Introduction
2. Motivation
3. Related works
4. Research questions
5. Datasets
6. RQ1, RQ2, RQ3: Methodology, results, discussion
7. RQ4: Methodology, results, discussion
8. RQ5: Methodology, results, discussion
9. **Limitations**
10. Future direction
11. Summary

Limitation

- The data comes from a narrow domain
  - Generalizability could be restricted.
- The impact of data augmentation was not explored
- Negative instance consideration
  - Check for false negatives
- The model is more likely to be correct in predicting the responses of different students

Agenda

1. Introduction
2. Motivation
3. Related works
4. Research questions
5. Datasets
6. RQ1, RQ2, RQ3: Methodology, results, discussion
7. RQ4: Methodology, results, discussion
8. RQ5: Methodology, results, discussion
9. Limitations
10. **Future direction**
11. Summary

Future direction

- Data augmentation to check if model performance improves
- Cross-skill training within the domain
  - Provides insight into low-resource scenario
- In real life, skills are transferrable to different STEM education scenarios.
  - Testing with different domain data(Chemistry or any other course) to understand the transferability of the model’s learning

Agenda

1. Introduction
2. Motivation
3. Related works
4. Research questions
5. Datasets
6. RQ1, RQ2, RQ3: Methodology, results, discussion
7. RQ4: Methodology, results, discussion
8. RQ5: Methodology, results, discussion
9. Limitations
10. Future direction
11. **Summary**

Summary

1. Explainable analytical skill assessment system

2. Employed PDR framework

3. Adapting the larger models and entailment-based scoring, especially for higher classification complexity, provides enhanced performance

4. Explainer model using SHAP

5. Global explanation highlights
   - Importance of domain-specific features, words related to describe energy related phenomenon(Comparison, direction)
   - Negative features are general feature
6. Local explanation
   - Model learnt mostly for the right reason
   - Automatically identified important features match with human-annotated evidence span with few exceptions

7. Results of the occlusion study showed higher reliability of the predictive and explainer models

8. GLMER model shows a student-wise random effect on model prediction

9. Stakeholders are well informed regarding the pros and cons of the model's behaviour (Relevancy)

UNIVERSITÄT
DES
SAARLANDES

Thanks for your attention !!!

Acknowledgement

- **Mentor, Reviewers:**
  - Prof. Dr. Vera Demberg
  - Prof. Dr. Hendrik Drachsler
  - Sebastian Gombert
- **Dataset:**
  - Researchers of Leibniz Institute for Science and Mathematics Education, Kiel: AFLEK
  - Prof. Dr. Walt Detmar Meurers: CREG-TUE

References:

- 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
- Mertler, C. A. (2001). Designing scoring rubrics for your classroom. Practical assessment, research, and evaluation, 7(1), 25.
- Callear, D. H., Jerrams-Smith, J., & Soh, V. (2001). Caa of short non-mcq answers.
- Bachman, L. F., Carr, N., Kamei, G., Kim, M., Pan, M. J., Salvador, C., & Sawaki, Y. (2002). Areliable approach to automatic assessment of short answer free responses. In Coling 2002: The 17th international conference on computational linguistics: Project notes.
- 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).
- 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)
- 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)
- 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.
- 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).
- 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.
- 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).

References:

- 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).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).
- 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.
- Nath, S., Parsaeifard, B., & Werlen, E. (2023). Automated short answer grading using bert on german datasets.
- 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).
- 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).
- Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Interpretable ma chine learning: definitions, methods, and applications. arXiv preprint arXiv:1901.04592.
- Lundberg, S. M., &Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in neural information processing systems, 30.
- 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
- Chan, B., Schweter, S., & Möller, T. (2020). German’s next language model. arXiv preprint arXiv:2010.10906.
- 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

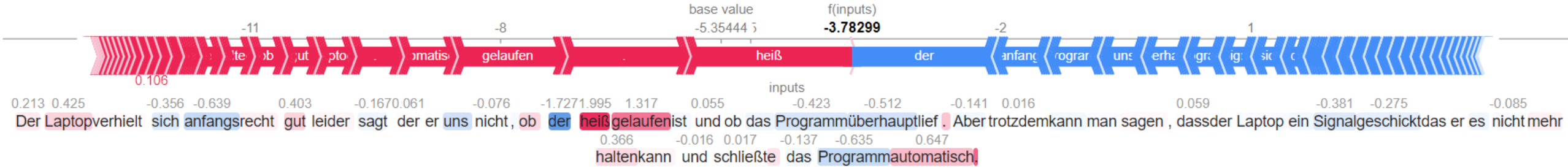
References:

- Loshchilov, I., & Hutter, F. (2019). Decoupled weight decay regularization. In International conference on learning representations. Retrieved from https://openreview.net/forum?id=Bkg6RiCqY7
- 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
- R Core Team. (2024). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from https://www.R-project .org/

UNIVERSITÄT DES SAARLANDES

Additional Slides

Negative Instance: Constructing Explanations



3.1 Related works for Automatic Short Answer Assessment

- The methods adopted for solving ASAG can be categorized as
  - **Feature based**
    - → Different features are created and found the similarity in different ways
  - **Supervised ML and Unsupervised ML**
    - → Different hand-crafted and semi-hand-crafted features are created to train different machine learning algos
    - → Without hand-crafted features, deep learning algorithms have demonstrated impressive results

Feature-based ASAG

Callear et al., 2001

- Automated Text Marker
- Smaller concept lexicons are collected from reference answers to create the keyword lexicons
- Pattern matching between response and predefined keyword lexicon
- Presence of lexicon is regarded as correct

Bachman et al., 2002

- Web-based language assessment system replaced P&P system
- Extracting important elements from the reference answer and student response by tagging and parsing them
- Pattern matching between tagged and parsed responses and reference answer
- Interactively proposes the assessment to the user.

Cutroe & Chang., 2010

- IE and IR are used to get the canonical form of the reference answer and the student response
- These two canonical forms are matched for exact matching.
- Synonyms for the unmatched words from WordNet, as well as their POS and relative word positions, are determined to get the accuracy of the matching.

Hahn & Meurers, 2012

- Surface representation: syntax-semantic interface representation
- Rules are used to map the interface to Lexical resource semantics representation
- LRS of response, reference answer and questions are aligned to determine the focused Information structure
- Numerical scores are given based on potential alignment quality

Supervised and unsupervised ASAG

Meurers et al.,2011

- Instead of surface-based info, more linguistically informed features are derived from student responses and reference answers.
- Features like overlapping keywords, tokens, chunks, dependency triples, matching type, lemma, and synonyms
- These features are used to train KNN to evaluate the similarity

Horbach et al., 2013

- highlights the importance of including reading text in addition to the reference answer and student response
- Linguistic features like tokens, chunks, dependency triples of reference answers, student responses and reading text are derived
- These features are used to train KNN to determine the similarity

Crossley et al., 2016

- The hints used during the assessment serve as a reference, rather than a regular reference answer.
- Different linguistic features like psycho-linguistic norms such as word information indices are derived along with regular linguistic features
- Linear discriminant analysis models are trained to score the different assessments in the chemistry domain.

Marvaniya et al., 2018

- Unsupervised way of generating scoring rubrics
- Correct answer provides additional info than the reference answer
- Clustering is used to get the representative student answers for a particular grade category to use as reference answer
- Lexical overlap, sentence-embedding-based similarity are calculated between question, set of representative answers and the response.

Supervised and unsupervised ASAG

Maharjan et al.,2018

- Feature engineering is time-consuming.
- Semantic similarity approaches in ASAG assign different scores to semantically different but correct answers.
- Context is important as the response can capture long-term dependency between a student's answer and the previous context.
- Problem definition, tutor question, and Student responses are concatenated and tri-letter word encoding is used to train the LSTM model for the ASAG task.

Sung et al., 2019

- Transformers-based architectures such as BERT are used to finetune for ASAG task.
- English datasets such as MNLI (Williams et al., 2018), and other psychology domain datasets are used to show the utility of BERT in cross-domain scenario

Camus & Filighera, 2020

- Transformer-based models are used for ASAG tasks and have shown improved transfer learning performance of these models based on knowledge distillation
- English datasets such as MNLI, and sciEntsBank (Dzikovska et al., 2013) are used

Gombert et al. 2023, Nath et al. (2023)

- Showed comparative performance study among different feature-based and transformer models specifically pretrained and fine-tined with German language data in ASAG.
- German datasets such as CREG(Ott et al., 2012), CSSAG(Pado &Kiefer (2015), AFLEK (Gombert et al. 2023) are used.

3.2 Related Works: Trusted learning analytics

Greller & Drachsler, 2012

Learning Analytics (LA)

• Aim: to support both learners and teachers

• Utilizes computational and data-driven methods for assessment

• A potential drawback of data-driven approaches in learning analytics is the presence of different biases in training sets, which can lead to models learning unintended shortcuts rather than precise regularizations.

Drachsler and Greller, 2016

Trusted Learning Analytics:

• Addresses issues of data privacy

• The problem of asymmetrical power relationships in learning scenarios.

Slade and Tait, 2019

Ethical guidelines for LA

stress the importance of sound models

• Free from algorithmic bias,
• Transparency
• Clarity for end-users.

3.3 Explainability

Murdoch et al. (2019)

Growing awareness of interpretability, ML models are grouped as glass box and black box models.

Glass box:
- Provide insights into the inner workings
- Regression and tree-based models

Blackbox:
- Lack of transparency of inner working
- Challenging to ensure it learns reliable patterns
- Transformer-based models
- Requires post hoc explanations

A wide range of techniques with equally wide results are categorized as interpretations

- Great deal of misunderstanding on the concept of interpretability

- Framework and methods
- Framework → PDR
- Methods → Post hoc analysis

3.3 Explainability Framework

Murdoch et al. (2019)

PDR Framework :

- A structured approach to interpretability by considering these three key aspects

- **Predictive Accuracy:**
  - Refers to evaluating the model's fit quality using well-established machine learning evaluation methods.

- **Descriptive Accuracy:**
  - Post-hoc methods are used to interpret whether a model's learned patterns align with coding guidelines
  - Two types of post-hoc interpretability :
    - Local explanations/ prediction-level interpretability
    - Global explanations/dataset-level interpretability

- **Relevancy:**
  - Refers to stakeholders' requirements and their relevance to descriptive and predictive accuracy
  - High reliability of the predictive and descriptive models

3.3 Explainability: Methods

(Bastings &Filippova, 2020)

- Gradient-based
- Propagation-based: LRP
- Occlusion-based

And

- Additive feature attribution methods

(Chefer et al., 2021)

- Combination of gradient-based and propagation-based methods: Transformers explainability
- ASAG application(Gombert et.al, 2023)

(Li et al., 2016)

- Gradient-based:
- Get model gradients using backpropagation to determine the feature importance

(Zeiler and Fergus, 2014)

- Occlusion-based methods: Occludes or eliminates a part of the whole input features and measures how that changes the model prediction
- also be used to evaluate the reliability of other explainability methods

(Binder et al., 2016)

- Layer-wise relevance propagation(LRP): pickups only the most important portion of the input relevant to the output using custom backward passes to calculate relevance score in various layers

(Lundberg et al., 2017)

- Additive Feature attribution methods:
- The unique solution to this class in game theory
- SHAP is the game theory method, and its estimation methods align the most with human intuition

3.1 Feature-based ASAG

- Identification of pre-defined concept lexicon (Callear et al., 2001)

- Pattern matching between student response and reference answer
- Score assignment based on quality of alignment

    - Tagged and parsed form (Bachman et al., 2002)

    - Canonical forms by using techniques of IE &IR (Cutrone & Chang, 2010),

    - Lexical resource semantics representation (Hahn & Meurers, 2012)

3.1 Supervised and unsupervised ASAG

- Hand-crafted and semi-handcrafted lexical and semantic feature similarity:
  - Overlapping keywords, tokens, chunks, dependency triples, matching type, lemma, synonyms, psycholinguistic norms
  - Trained KNN, Linear discriminant analysis(Meurers et al.,2011, Horbach et al., 2013, Crossley et al., 2016)

- Different styles of reference answers are also considered
  - Standard reference answers (Meurers et al.,2011)
  - Reference answer + reading text (Horbach et al., 2013)
  - Hints presented in different stages of assessments(Crossley et al., 2016)
  - Clustering correct answers from a grade category to create representational reference answers (Marvaniya et al., 2018)

3.1 Supervised and unsupervised ASAG

- Feature engineering is time-consuming
- Semantic similarity approaches assign different scores to semantically different but correct answers
  - Disparity in assessment
- Importance of context is considered
  - long-term dependency between a student's answer and the previous context
- LSTM model(Maharjan et al.,2018)
- Transformers-based architectures such as BERT(Devlin et al., 2019)
  - On English Dataset(Sung et al., 2019, Camus & Filighera, 2020)
  - On German Dataset (Gombert et al.,2023, Nath et al.,2023)

3.3 Explainability: Methods

- **Gradient-based**(Li et al., 2016)**:** Get model gradients using backpropagation to determine the feature importance

- **Propagation-based**(Binder et al., 2016) **Layer-wise relevance propagation:** Pickups only the most important portion of the input relevant to the output using custom backward passes to calculate relevance score in various layers

- **Occlusion-based**(Zeiler and Fergus, 2014)**:** Occludes or eliminates a part of the whole input features and measures how that changes the model prediction

- **Additive feature attribution methods:** SHAP(**SH**apley **Additive** ex**P**lanations)
  - SHAP is the game theory method, and its estimation methods align the most with human intuition (Lundberg et al., 2017)

Additional Analysis

- Hierarchical structure of the data is observed
  - One student(each subject) provided multiple responses (multiple observations)
  - One question (each item) has many students' responses (multiple observations)
  - Individual responses are not completely independent of one another
  - Grouped under students and questions
- Three different Generalized mixed effect models (Bates et al., 2015) using R (R Core Team, 2024) with increasing random effect for each skill
- Applied ANOVA on these models to understand if there is any significant random effect on the model's performance

Results

Models	npar	AIC	BIC	loglik	deviance	Chisq	Df	Pr(>Chisq)
Question	3	6206.4	6227.4	-3100.2	6200.4			
Student	4	5791.2	5819.2	-2891.6	5783.2	417.25	1	<2e-16***
Full	6	5793.4	5835.4	-2890.7	5781.4	2.78	2	0.9494

- H0: There is no significant random effect on the model's performance
- H1: There is significant random effect on the model's performance
- Significant level(α) = 0.05
- Reject the null hypothesis and accept the alternative hypothesis.
- Suggests that there is a significant student-wise random effect on the model's performance

question_model

question_model: correct~ model +(1 | question_id)

- Random intercept by question here means that for different questions, the model might be more likely to be correct in predicting the label.
- Modelling the different random intercepts per items/ questions

student_model

student_model: correct~ model +(1 | question_id) + (1|student_id)

- Random intercept by students here means that for different students, the model might be more likely to be correct in predicting the label.
- Modelling the different random intercepts per items/ questions and per subject/student

full_model

full_model: correct~ model +(1 | question_id) + (1+ model | student_id)

- random slope for model under student means that some models might have an easier time predicting whether the responses of specific students are correct
- Modelling the different random intercepts per items/ questions, per subject/student and random slope for different model

Occlusion study for Predictive accuracy:

Two scenarios are observed:

- Global explanation: The range of the positive and negative SHAP values decreased
  - Occluded data influenced a decreased importance of the most critical features in the model's prediction

- Change in the explainer model's behaviour:
  - Local explainability changed in 2 ways
- Change in the average marginal contribution of the positive features which are out of the evidence span

Analysis using waterfall plot:

- The model's output drops from $g(z') = 5.038$ to $g(z') = -5.778$
- Declined marginal feature contribution in the model's prediction.
  - "Videos", "kommt"
  - ϕ(original Videos) = +0.15
  - ϕ(masked Videos) = −0.09

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

5. 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 apply 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 is the interpretation of these models' decision-making behavior relevant to the stakeholders (Relevancy)?

Discussion

- RQ1: To what extent can scientific skills be detected in students' free-text responses using different Transformers language models (predictive accuracy)?

•Explainable analytical skill assessment system
• Employed PDR framework: Maintaining interpretability from various perspectives
• 3 different transformer models and their base and large variation
• Instance vs entailment scoring mechanism
• The usage of Transformers architecture is plausible to automate the identification of the scientific skills in constructed responses efficiently

Discussion

- RQ2: Does adaptation of entailment-based scoring improve over instance-based scoring?
  - For the AFLEK dataset in two out of three skills, the entailment-based scoring model achieved a better f1 score
  - For CREG-TUE dataset:
    - Entailment-based model in a 2-class setup is slightly better than instance-based
    - with increased classifier complexity, the entailment-based model outperforms instance-based model with a high margin.

Discussion

- **RQ3:** To what extent can the same models and the different scoring techniques apply to the domain-specific standard dataset (Secondary evaluation)?
- To contexualize our work better in the ASAG domain, CREG-TUE is used.
- This highlights the increased difficulty for models to classify the responses in four class categories accurately.
- Entailment-based model works better with more complex classifier

Discussion

RQ4: To what extent do input words considered important by the models for their predictions match human-coded ones (descriptive accuracy)?

- SHAP explainer as an explanation model to explain the approximate behaviour of our Transformers models
- Global explanation :
  - High SHAP value featues: Domain-specific energy-related terms
  - Negative SHAP value features: Mostly general words
- Local explanation:
  - considerable overlap between the human-annotated evidence span and the words provided positive SHAP value by the model
- 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.

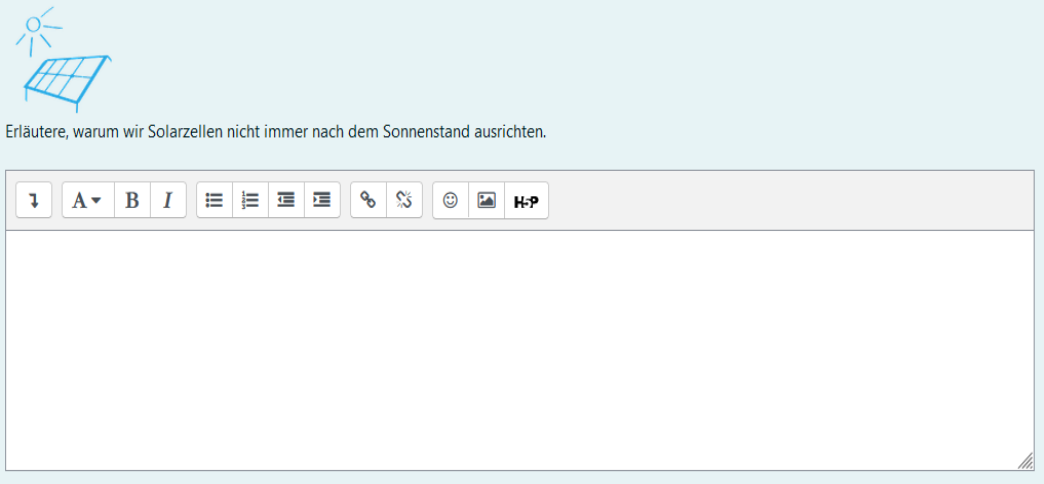
Discussion

RQ5: To what extent the interpretation of these models' decision-making behavior is relevant to the stakeholders (Relevancy)?

- Predictive accuracy of models are quite reasonable
- Descriptive accuracy suggests
  - The models mostly learn for the right reason with few exceptions
  - their reason for making decisions also matches with the human-annotated evidence span
- Occlusion study shows the reliability of predictive & descriptive accuracy
- 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

Example Question of Constructing Explanations

Q: Explain why we don't always align solar cells with the position of the sun.



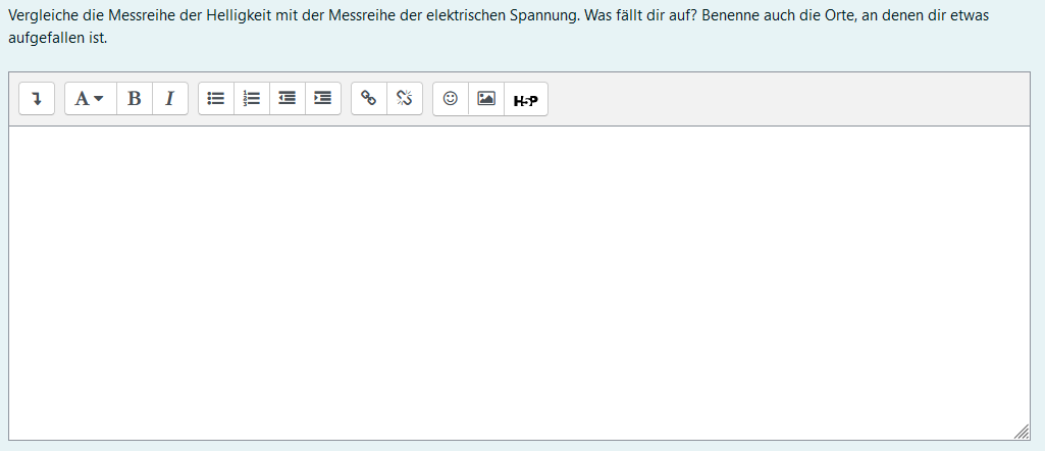
- Goal:
  - Determine if deviations from normative ideas result from a lack of task comprehension or a deficit in normative understanding.
- Reasons:
  - This label shall help us to conclude whether the students understand the tasks and their prompts.
- Evidence:
  - Students use the information given to them as reasons for ideas used in their answers.

Example Question of Analyzing Data

Q: 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.



- Goal:
  - Trace whether students analyze and interpret the data that they obtained from experiments.
- Reason:
  - This label shall help us to draw conclusions on whether the students are trained to analyze and interpret data.
- Evidence:
  - Students describe data and evaluate their ideas about the observed phenomenon based on that data.

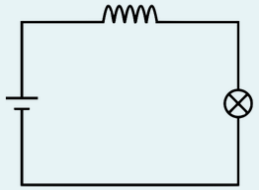
Note: As the automated computation of these answers is quite challenging, we limit ourselves to check whether the students work on task-relevant evidence – like their data or possible error sources in the experiments.

Erläutere, warum wir Solarzellen nicht immer nach dem Sonnenstand ausrichten.

Vergleiche die Messreihe der Helligkeit mit der Messreihe der elektrischen Spannung. Was fällt dir auf? Benenne auch die Orte, an denen dir etwas aufgefallen ist.

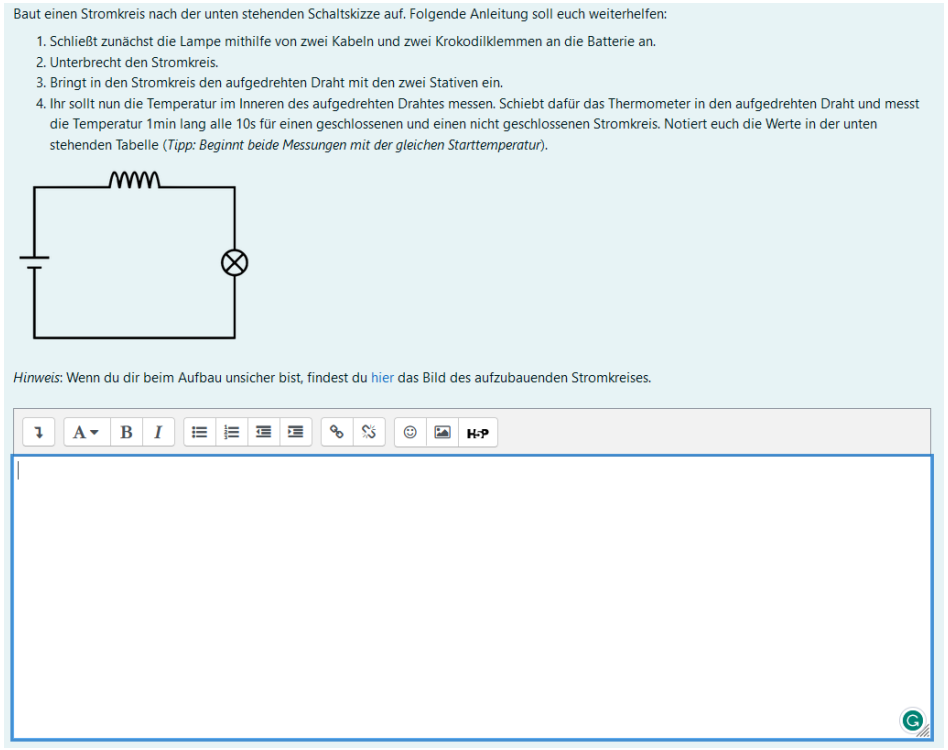
Baut einen Stromkreis nach der unten stehenden Schaltskizze auf. Folgende Anleitung soll euch weiterhelfen:

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.

Example Question of Planning Investigations



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.

SHAP approach

- SHAP(Shapley Additive exPlanations) is used for transformers explainability
- A pretrained GBERT base model is fed to the SHAP text explainability pipeline for the text explanation for a single instance.
- A positive classified response and a negative classified response for Constructing Explanation skills are examined separately.
- Text explainability works as follows:

Example examined for the positive class	Text chunk
Red	The text chunk has increased the probability of that instance belonging to the positive class
Blue	The instance has decreased the probability of that instance belonging to the positive class
Gradient	Magnitude of influence

Example examined for the negative class	Text chunk
Red	The text chunk has increased the probability of that instance belonging to the negative class
Blue	The instance has decreased the probability of that instance belonging to the negative class
Gradient	Magnitude of influence