HTWK Hochschule für Technik, Wirtschaft und Kultur Leipzig

Einsatz von Large Language Models zur automatisierten Erstellung mathematischer Übungsaufgaben

Bachelor-Kolloquium von Ines Rohrbach

Prüfer: Herr Prof. Dr. Martin Grüttmüller

Herr M.Sc. Alexander Pögelt

Gliederung

- Motivation
- Zielstellung
- Grundlagen
  - Large Language Model
  - Prompt Engineering
  - Taxonomie
  - Question & Test Interoperability
- Implementierung
- Live Demo
- Methodik der Evaluation
- Evaluation
- Ergebnisse
  - Phase 1
  - Phase 2
  - Zusammenfassung
- Einschränkungen
- Fazit

Motivation

Punkte: 1 | Keine Antwort

Problematisch ist …

- [x] Zeitaufwendige Aufgabenerstellung
- [ ] „LLM-Generated Mathematics Items“ [1]
- [ ] tech4compKI [2]

Antwort abgeben

Zielstellung

Aufgabengenerierung

- Mathematisches Konzept
- Aufgabentyp
- Taxonomie
- weiterverwendbares Format

Aufgabenevaluation

- Taxonomie & Qualität
- Inwiefern stimmt die maschinelle Evaluation mit der menschlichen Wahrnehmung überein?

Large Language Model (LLM)

- Verwendetes Modell: GPT-4o von OpenAI [3]
- Framework: LangChain [4]

```
chain = prompts() | get_llm(model) | get_parser()
try:
    result = chain.invoke({
        "topic_information": topic_information,
        "cpd_information": cpd_information,
        "kd_information": kd_information,
        "topic": topic,
        "item_type": item_type,
        "examples": examples,
    }, config={"callbacks": [PromptLogger()]})
except Exception as e:
    result = correct_backslash_in_output(e)
log_prompt_output_pairs(is_prompt=False, text=result)
```

Prompt Engineering

```
prompt = """
Use the given information of the knowledge dimensions to classify the task
```

```
and explain why.
```

```
The output should be formatted in JSON following the schema below:
{"reason": "...", "decision": "..."}
```

```
Knowledge Dimensions:
...
```

```
Task:
...
Answer:
{"reason": "...", "decision": "..."}
Task:
...
Answer:
```

Prompt Engineering

```
prompt = """
Use the given information of the knowledge dimensions to classify the task
```

```
and explain why.
```

```
The output should be formatted in JSON following the schema below:
{"reason": "...", "decision": "..."}
```

```
Knowledge Dimensions:
...
```

```
Task:
...
Answer:
{"reason": "...", "decision": "..."}
Task:
...
Answer:
```

Prompt Engineering

```
prompt = """
Use the given information of the knowledge dimensions to classify the task
```

```
and explain why.
```

```
The output should be formatted in JSON following the schema below:
{"reason": "...", "decision": "..."}
```

```
Knowledge Dimensions:
...
```

```
Task:
...
Answer:
{"reason": "...", "decision": "..."}
Task:
...
Answer:
```

Prompt Engineering

```
Knowledge Dimensions:
...
```

```
Task:
...
Answer:
{"reason": "...", "decision": "..."}
Task:
...
Answer:
{"reason": "...", "decision": "..."}
```

```
Task:
Wählen Sie den Allquantor aus.
A) $$\\forall$$
B) $$\\exists$$
C) $$A$$
Answer:"""
```

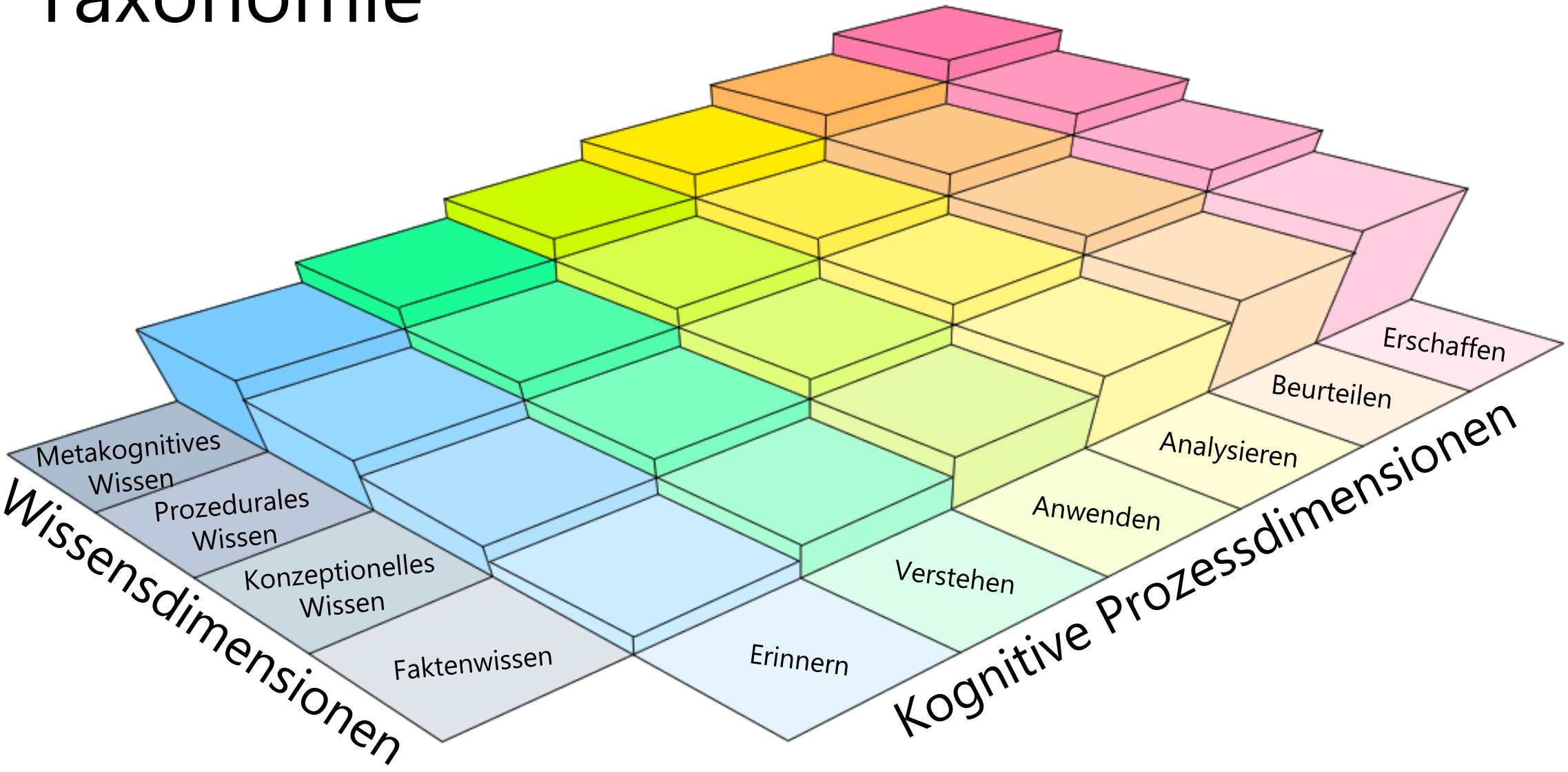
Prompt Engineering

```
Knowledge Dimensions:
...
```

```
Task:
...
Answer:
{"reason": "...", "decision": "..."}
Task:
...
Answer:
{"reason": "...", "decision": "..."}
```

```
Task:
Wählen Sie den Allquantor aus.
A) $$\\forall$$
B) $$\\exists$$
C) $$A$$
Answer:"""
```

Taxonomie


Metakognitives Wissen
Prozedurales Wissen
Konzeptionelles Wissen
Faktenwissen
Wissensdimensionen
Erinnern
Verstehen
Anwenden
Analysieren
Beurteilen
Erschaffen
Kognitive Prozessdimensionen

Question & Test Interoperability (QTI) [5]

- XML-basierter Standard
- 18 Aufgabentypen
  - Single- & Multiple-Choice, Drag-and-Drop und Textbox

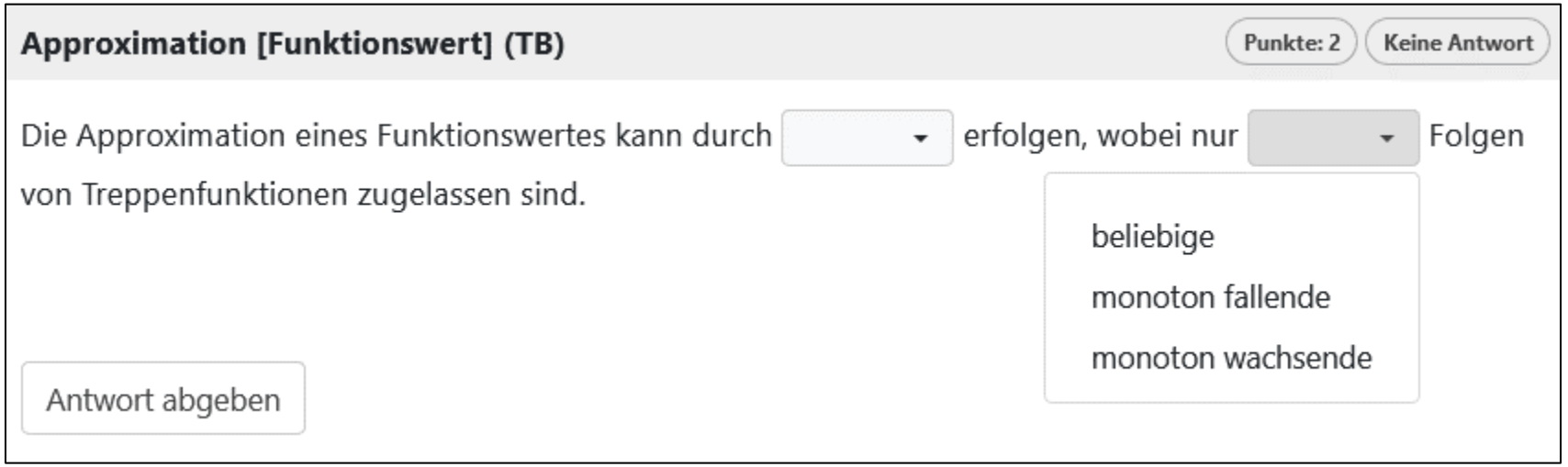


Abb. 2

Question & Test Interoperability (QTI) [5]

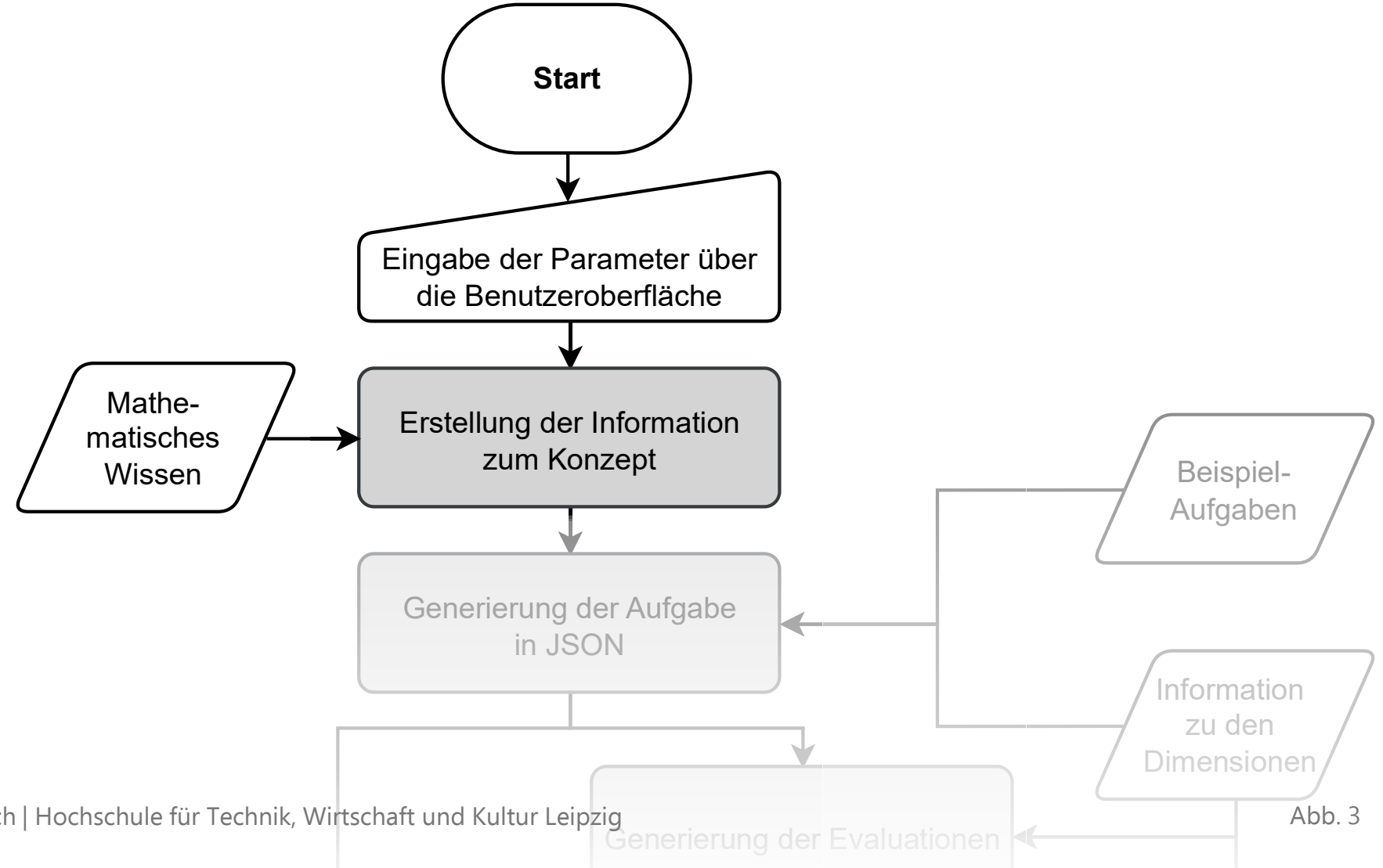
QTI (ca. 1500-2000 Token)

```
<assessmentItem xmlns="http://www.imsglobal.org/xsd/
imsqti_v2p1" xmlns:xsi="http://www.w3.org/2001/XMLSchema-
instance" xsi:schemaLocation="http://www.imsglobal.org/xsd/
imsqti_v2p1 http://www.imsglobal.org/xsd/qti/qtiv2p1/
imsqti_v2p1p1.xsd http://www.w3.org/1998/Math/MathML
http://www.w3.org/Math/XMLSchema/mathml2/mathml2.xsd"
identifier="ideb5f8a7c-ae00-41de-896c-8c0367e8e6d7"
title="Allquantor (SC)" adaptive="false"
timeDependent="false">
<responseDeclaration identifier="RESPONSE_1"
cardinality="single" baseType="identifier">
<correctResponse>
<value>$$\forall$$</value>
</correctResponse>
</responseDeclaration>
<outcomeDeclaration identifier="SCORE" cardinality="single"
baseType="float">
<defaultValue>
<value>0</value>
</defaultValue>
</outcomeDeclaration>
<outcomeDeclaration identifier="MAXSCORE"
cardinality="single" baseType="float">
<defaultValue>
<value>0</value>
</defaultValue>
<outcomeDeclaration identifier="FEEDBACKBASIC"
```

JSON (ca. 200-400 Token)

```
{
  "question_type": "single-choice",
  "question": {
    "title": "Allquantor (SC)",
    "question_text": "Wählen Sie den Allquantor aus.",
    "options": [
      {
         "text": "$$\\forall$$", "is_correct": true,
         "feedback": "Sieht gut aus!"
      },
      {
         "text": "$$\\exists$$",
         "is_correct": false,
         "feedback": "Das ist der falsche Quantor. [...]"
      },
      {
         "text": "$$A$$",
         "is_correct": false,
         "feedback": "Das ist ein normales A in einer
Mathematik-Umgebung. [...]"
      }
    ]
  }
}
```

Implementierung



Implementierung

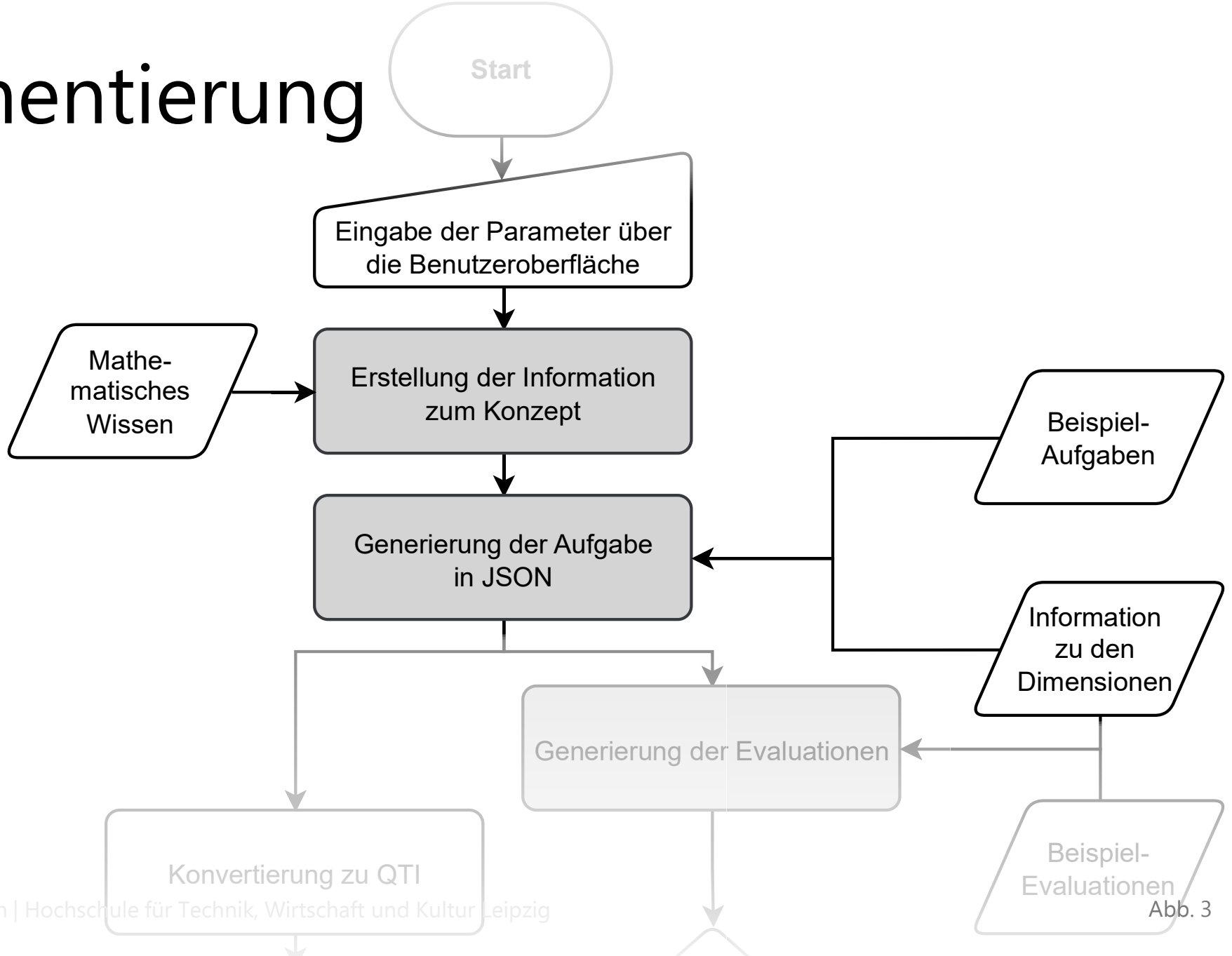
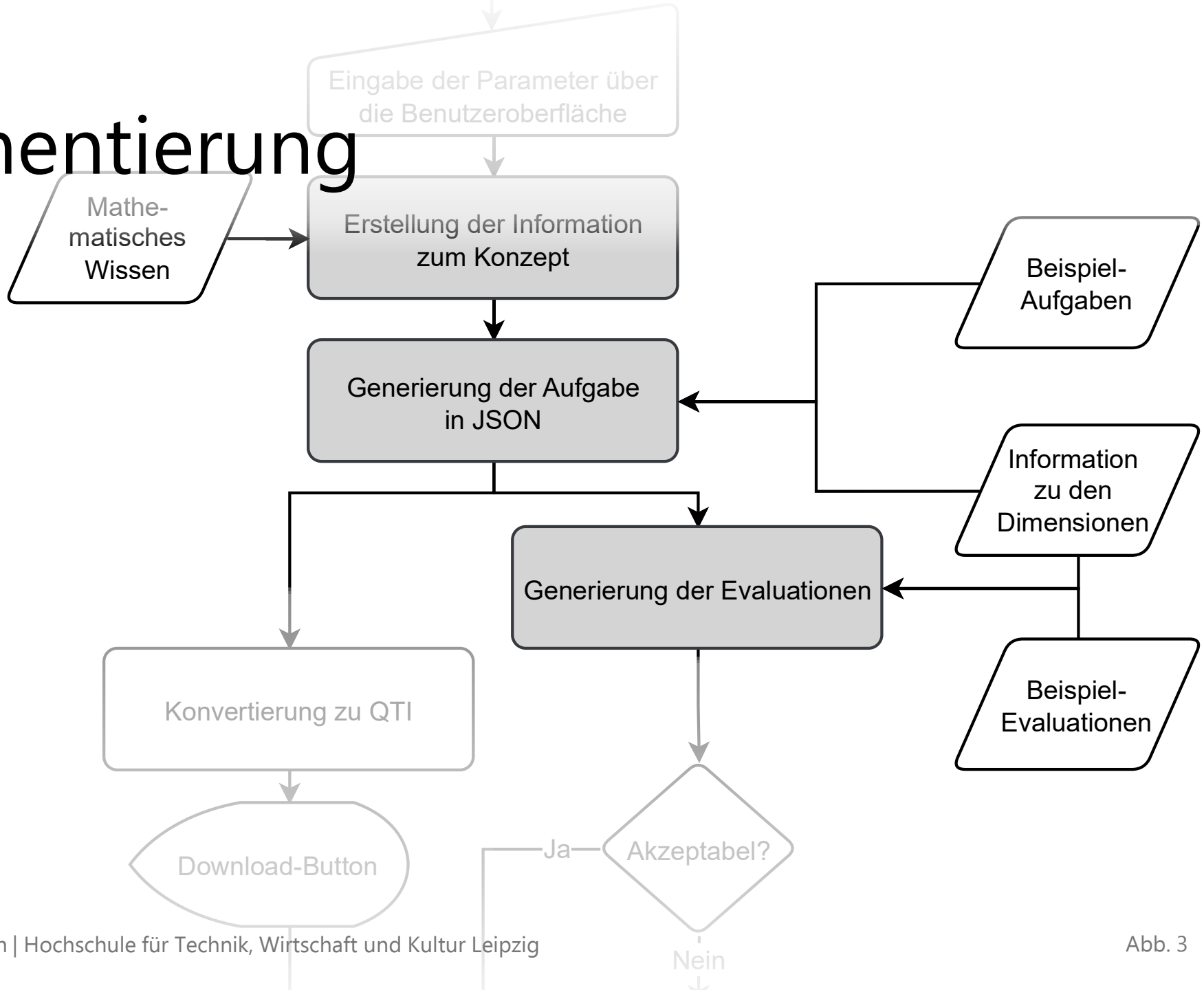
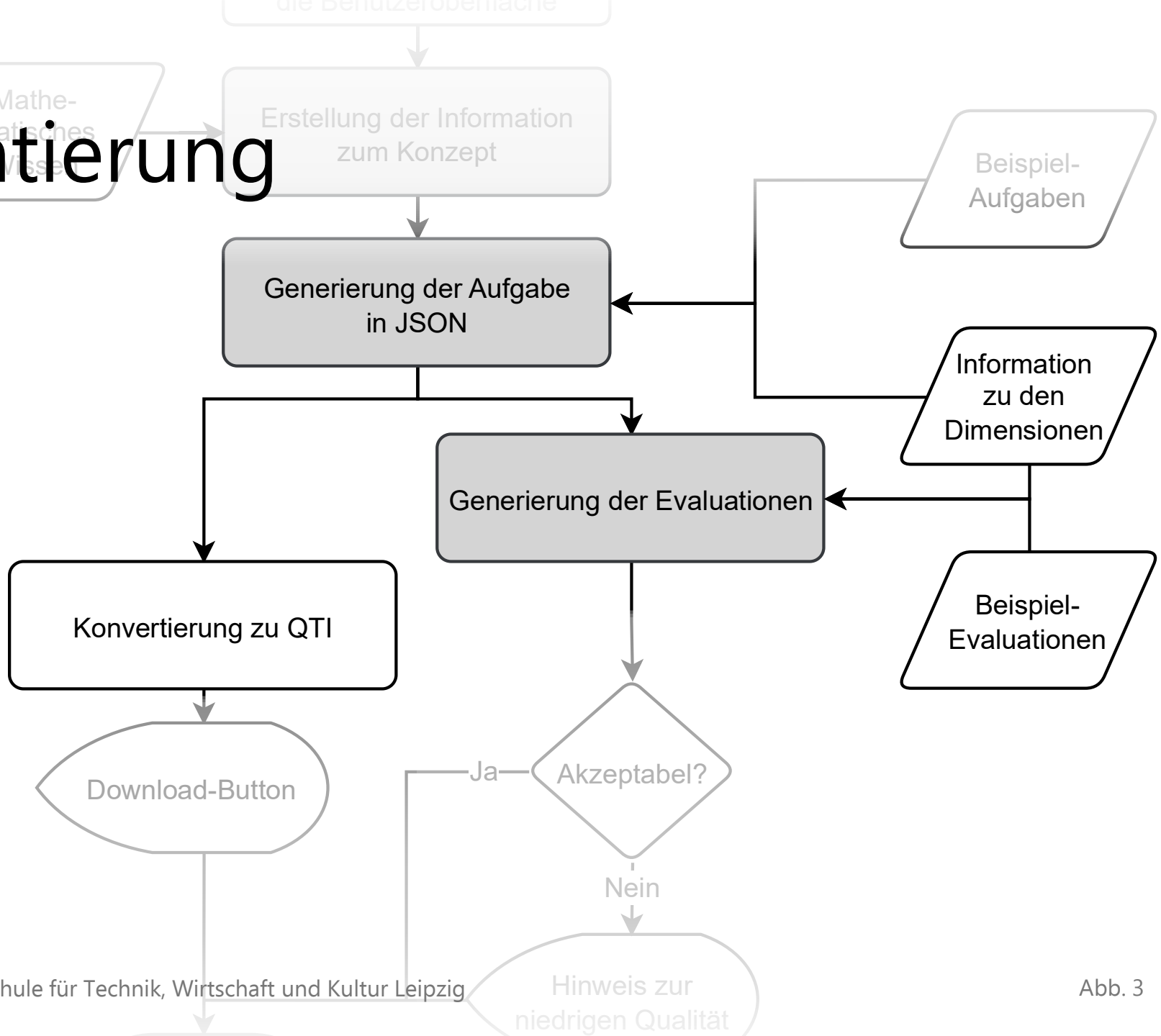


Abb. 3

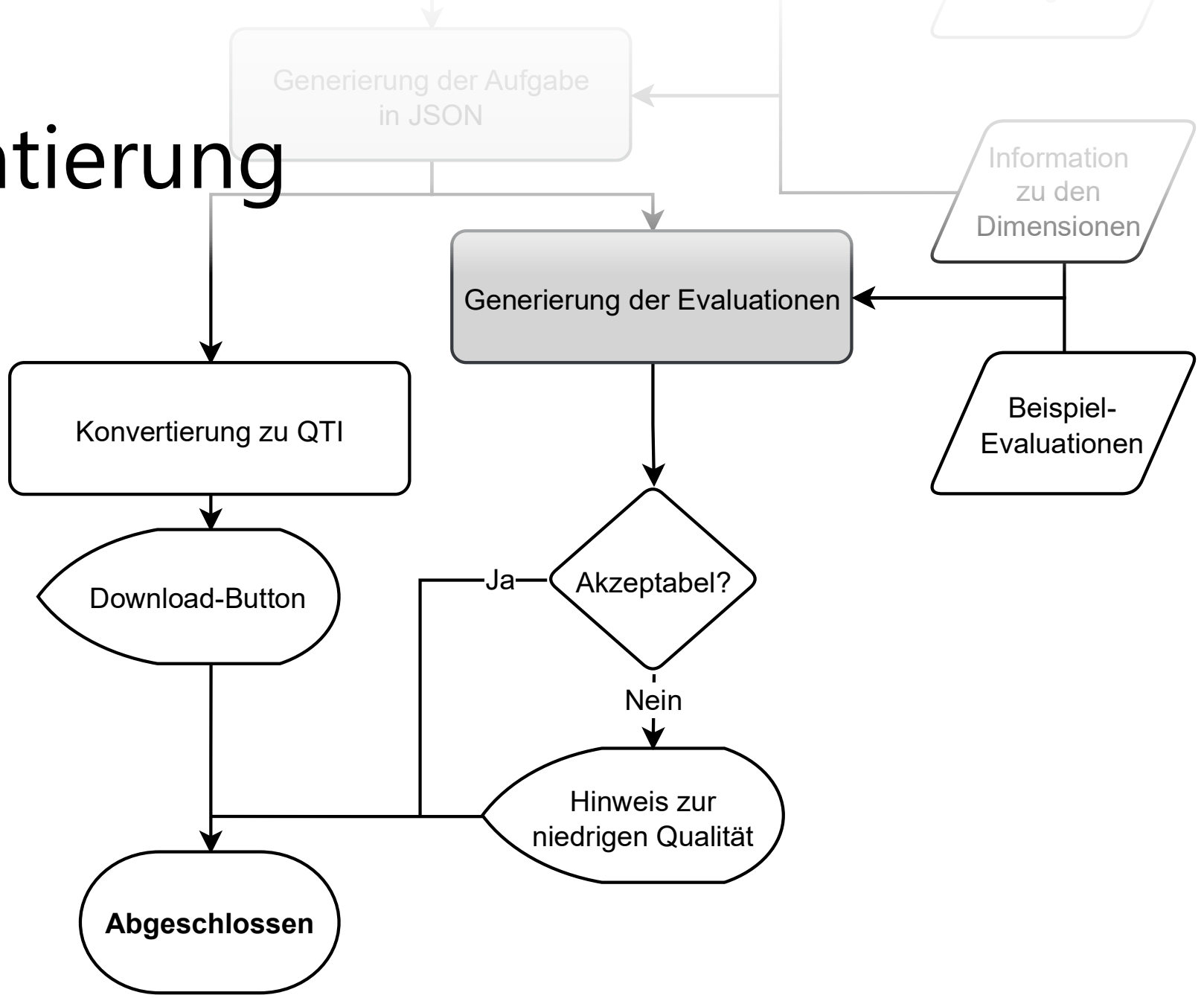
Implementierung



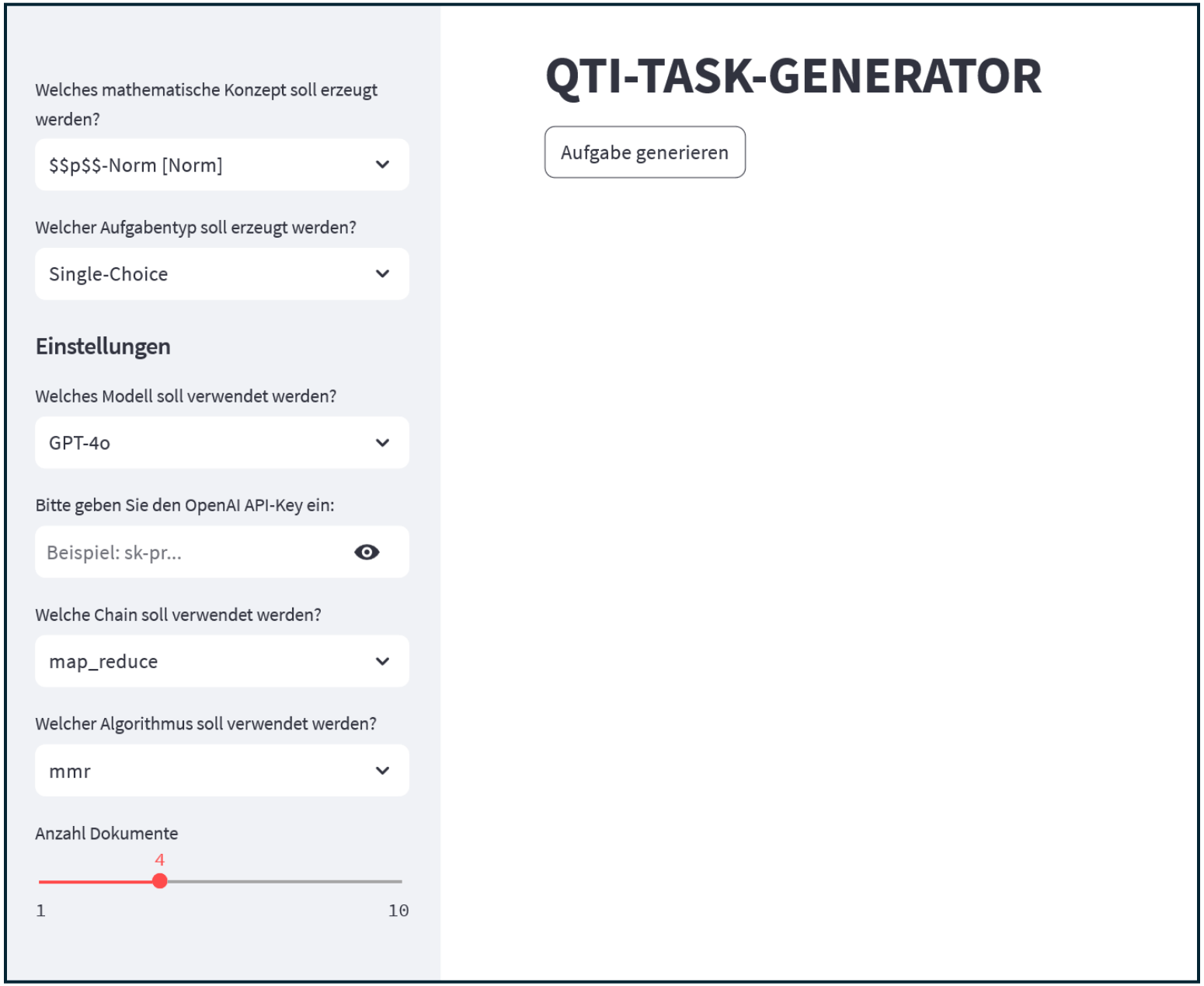
Implementierung



Implementierung



Live Demo

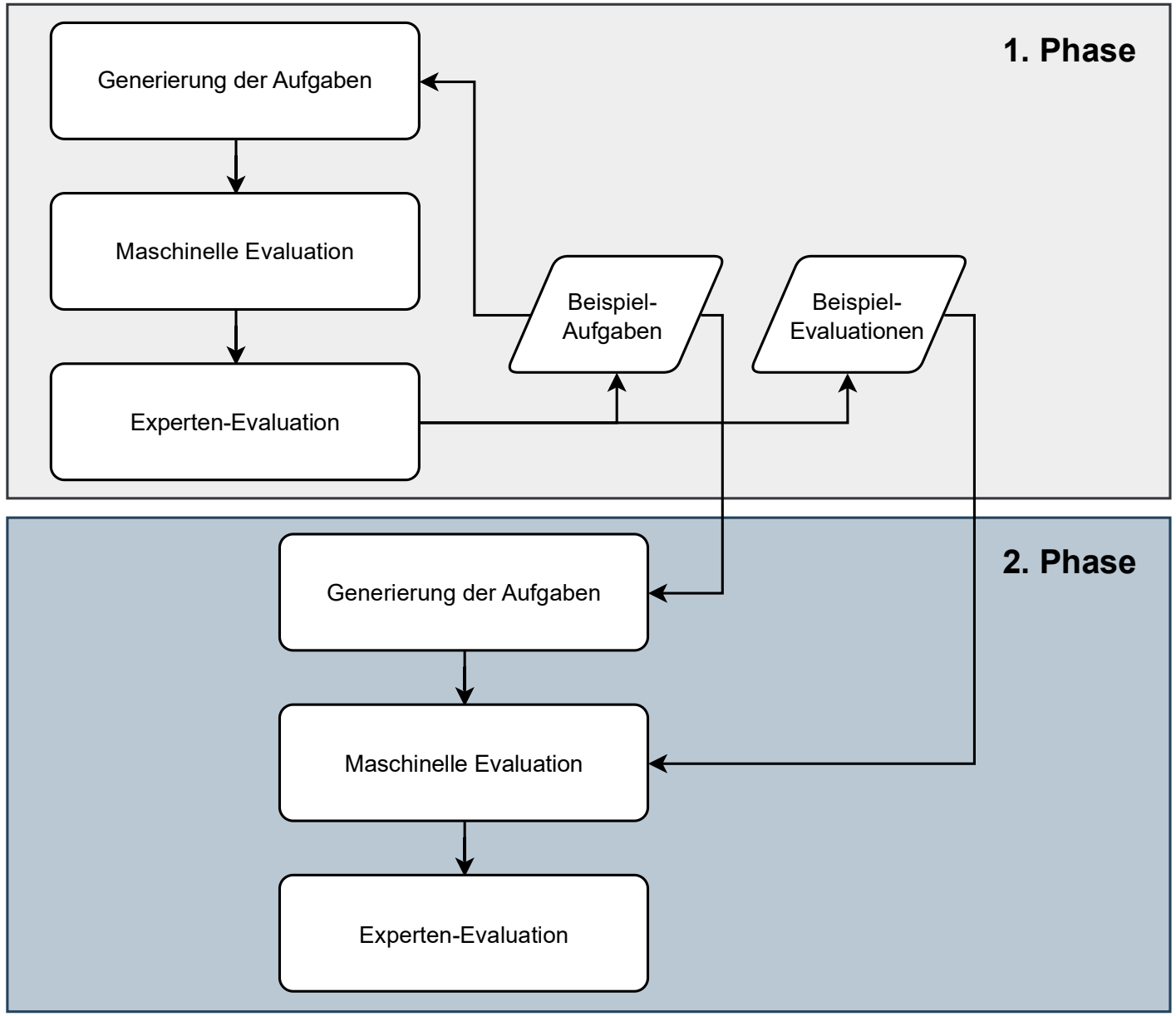

Welches mathematische Konzept soll erzeugt werden?
$$p$$-Norm [Norm]
Welcher Aufgabentyp soll erzeugt werden?
Single-Choice
Einstellungen
Welches Modell soll verwendet werden?
GPT-4o
Bitte geben Sie den OpenAI API-Key ein:
Beispiel: sk-pr...
Welche Chain soll verwendet werden?
map_reduce
Welcher Algorithmus soll verwendet werden?
mmr
Anzahl Dokumente
4
1
10
QTI-TASK-GENERATOR
Aufgabe generieren

Methodik

64 Aufgaben

300 Aufgaben

80 Aufgaben



Evaluation

Taxonomie

- Kognitive Prozessdimension
- Wissensdimension

Qualitätsmerkmale

- Ausführbarkeit
- Aufgabenstellung
- Auswertung
- Feedbackzweige
- Kontext

Bewertungsmetriken

$$\text{Accuracy} = \frac{TP+TN}{TP+FP+TN+FN}$$

$$\text{Precision} = \frac{TP}{TP+FP}$$

$$\text{Recall} = \frac{TP}{TP+FN}$$

$$\text{F1-Score} = 2 \cdot \frac{Precision \cdot Recall}{Precision+Recall}$$

[6]

Ereignisse	Akzeptabel	Unakzeptabel
Akzeptabel	**True Positive**	**False Negative**
Unakzeptabel	**False Positive**	**True Negative**

Experten Evaluation

Maschinelle Evaluation

Ergebnisse – Phase 1

Taxonomie

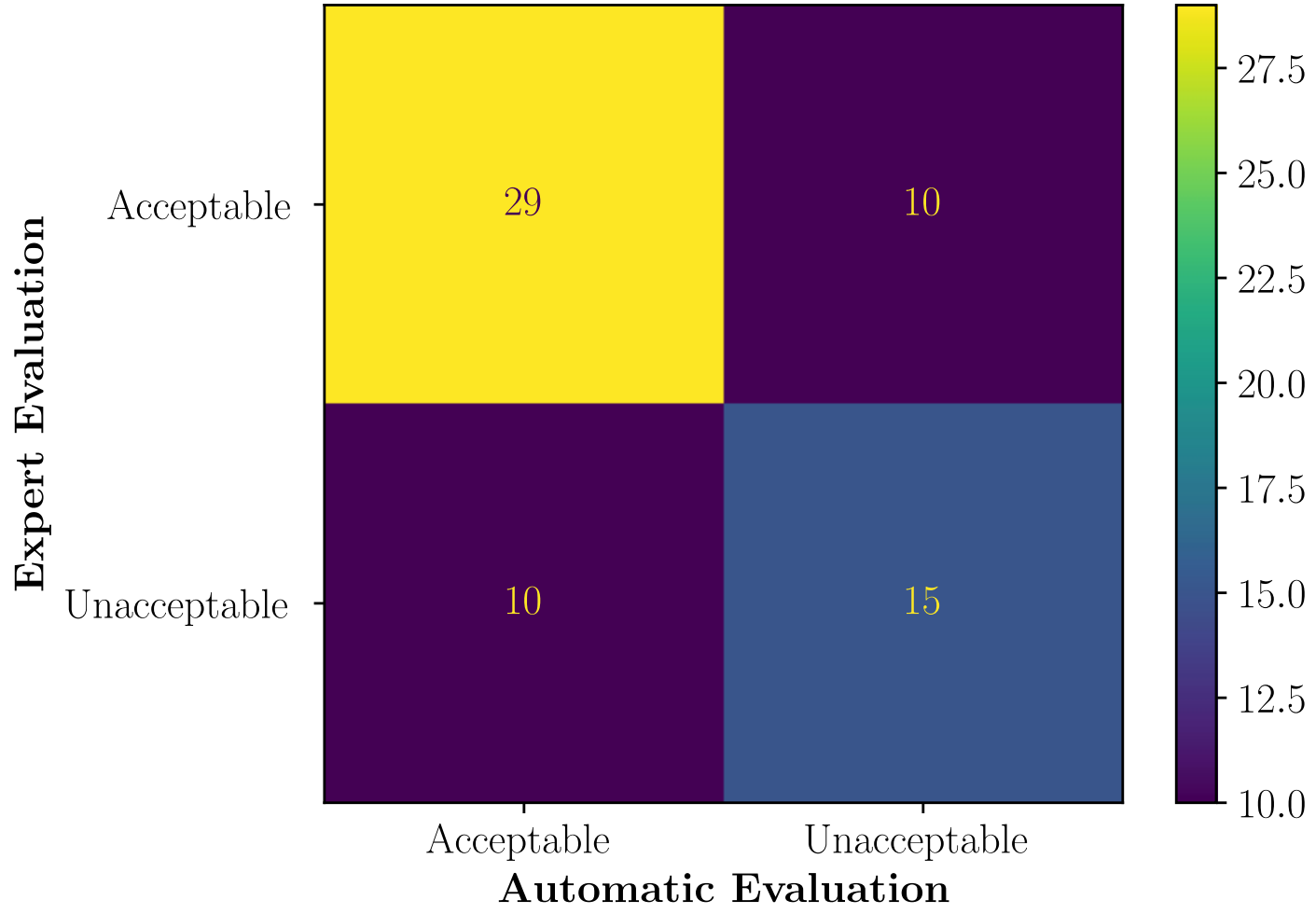
- Wissensdimension
  - eine Abweichung
- Kognitive Prozessdimension

	Anzahl
Erinnern	39
Verstehen	14
Anwenden	-
Analysieren	11
Beurteilen	-
Erschaffen	-

Qualitätsmerkmale

- ∅ Accuracy: 0.74
  - Kontext: 0.45

Ergebnisse – Phase 1



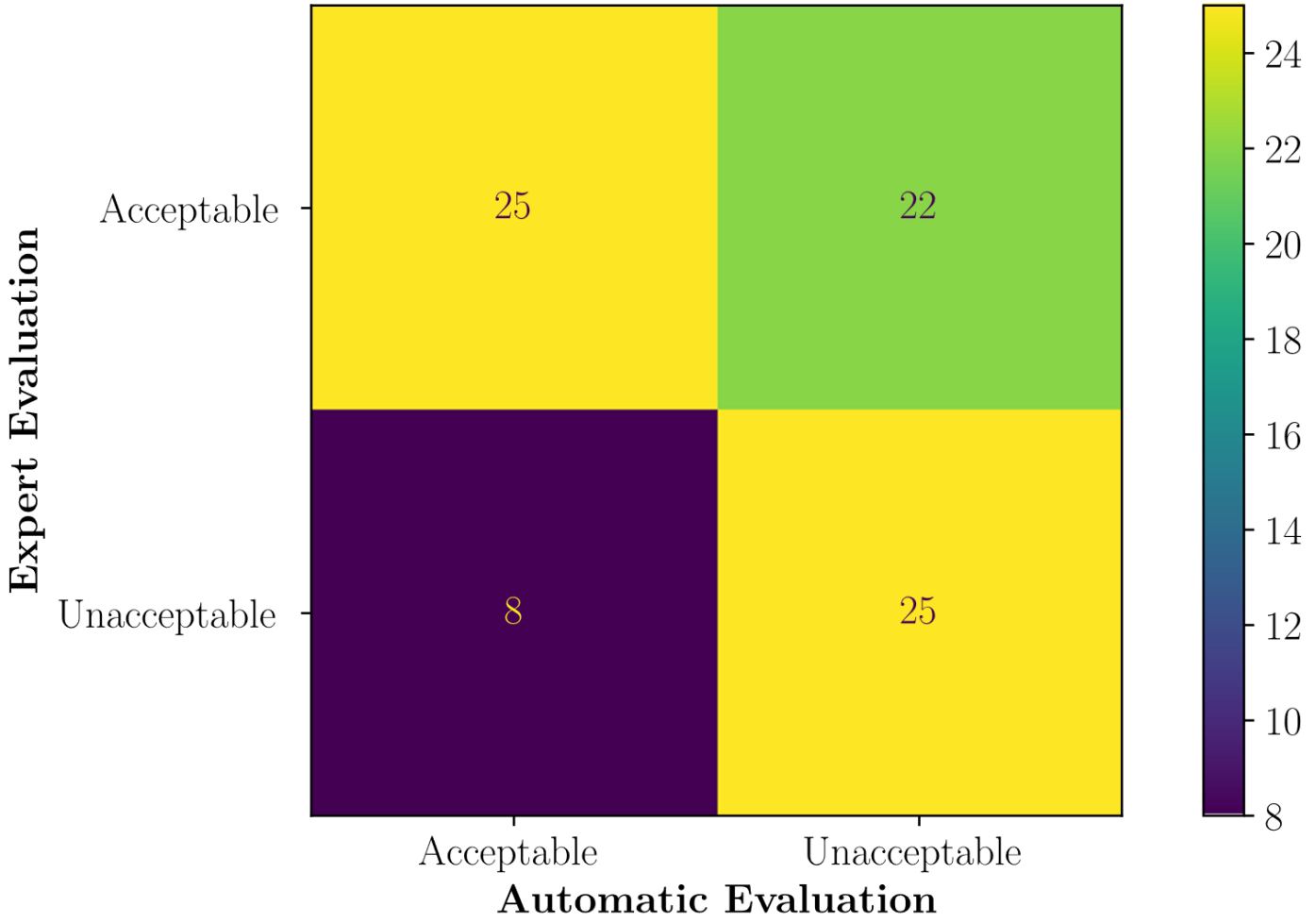
Ergebnisse – Phase 2

Taxonomie

- F1-Score: 0.96

Qualitätsmerkmale

- ∅ Accuracy: 0.78
  - Kontext: 0.53



Ergebnisse – Zusammenfassung

Aufgabengenerierung

- 60 % akzeptabel
- Manche Aufgaben bereits verwendet
- Feedback:
  - Kontrolle notwendig
  - Entlastend

Maschinelle Evaluation

- 63 % Übereinstimmung
- Precision: 0.76
- Potenzial

Einschränkungen

LLMs

- Determinismus
- Ausfallbeständigkeit

Menschliche Evaluation

- Allgemeingültigkeit

Fazit

Generierung & Evaluation von Übungsaufgaben

- Technologie: Large Language Models
- Eingabe: Mathematisches Konzept und Aufgabentyp
- Ausgabe: Aufgaben im QTI-Format
- Evaluation: Taxonomie und Qualität

Überprüfung der Leistung

- 144 generierte Aufgaben menschlich evaluiert
- Aufgaben: 60 % akzeptabel
- Maschinelle Evaluation: 63 % übereinstimmend

Ausblick

- LLMs anpassen
- Prompt Engineering

Quellen

Abbildungen

Literatur

[1] R. Meissner et al. *LLM-Generated Competence-Based E-Assessment Items for Higher Education Mathematics: Methodology and Evaluation*. 2023. Vorveröffentlichung

[2] *tech4compKI*. Wi-Ho - Wissenschafts- und Hochschulforschung. URL: https://www.wihoforschung.de/wihoforschung/de/bmbf-projektfoerderung/foerderlinien/forschung-zur-digitalen-hochschulbildung/zweite-foerderlinie-zur-digitalen-hochschulbildung/zweite-foerderphase/tech4compki/tech4compki.html (besucht am 20. 05. 2024)

[3] *OpenAI Developer Platform*. OpenAI Platform. URL: https://platform.openai. com/docs/overview (besucht am 08. 07. 2024)

[4] *LangChain*. URL: https://www.langchain.com/ (besucht am 29. 07. 2024)

[5] *Question & Test Interoperability®.* 1EdTech. URL: https://www.1edtech.org/standards/qti (besucht am 30. 05. 2024)

[6] A. Tharwat. „Classification Assessment Methods". In: *Applied Computing and Informatics* 17.1 (1. Jan. 2020), S. 168–192. issn: 2210-8327. doi: 10.1016/j.aci. 2018.08.003. URL: https://doi.org/10.1016/j.aci.2018.08.003 (besucht am 28. 08. 2024)

Einsatz von LLMs zur automatisierten Erstellung mathematischer Übungsaufgaben

Generierung & Evaluation von Übungsaufgaben

- Technologie: Large Language Models
- Eingabe: Mathematisches Konzept und Aufgabentyp
- Ausgabe: Aufgaben im QTI-Format
- Evaluation: Taxonomie und Qualität

Überprüfung der Leistung

- 144 generierte Aufgaben menschlich evaluiert
- Aufgaben: 60 % akzeptabel
- Maschinelle Evaluation: 63 % übereinstimmend

Ausblick

- LLMs anpassen
- Prompt Engineering