| | |
|---|---|
| **Titel** | Engaging Exam Evaluations: Making use of Learning Materials for LLM-based QA Assessments |
| **Workshopleitung** | Andrea Palmini and Tunc Yilmaz |
| **Teilnehmendenanzahl** | 32 |
| **Datum** | 05.03.2025 |
| **Uhrzeit** | 13:00-14:00 |
| **Struktur der Veranstaltung** (Vortrag, Gruppenarbeit, Diskussion…?) | Präsentation und Diskussion |

| | |
|---|---|
| **Zusammenfassung** | The workshop began with an overview of the ministry funded IMPACT Project, which centers around implementing AI-based tools for different use cases in learning and teaching. Panellists introduced the details of a particular subproject under IMPACT, that concerns the development of short essay type exam evaluation tools. A brief summary of the earlier experiences relating to automated exam evaluations has been shared with the participants.<br><br>The session continued with providing theoretical about Large Language Models (LLMs), as the most recent exam evaluation applications developed by the team made use of open source LLMs. Description, use-cases and architecture of such models have been shared with those in the audience who were interested in the technical infrastructure.<br><br>Further technical background, mostly relating to the RAG skeleton used by the evaluation tool employed by the presenters, has been sketched in the third part of the workshop. Key terms, such as transformer models, vector stores, and embeddings were touched upon, leading to an overall depiction of the full RAG infrastructure that supports the main application behind the scenes.<br><br>The workshop later dwelled on certain topics of concern, mostly referring to highly controversial points of discussion such as ethics, privacy, bias, computational costs and hallucinations. Before continuing with a real-life showcase of their exam evaluation tools, the presenters elaborated on some of these highly debated focus points, and how their implementation might remedy some, if not all, of the concerns mentioned.<br><br>At the last section of the workshop before the concluding discussion session, the features of the AI-based exam evaluation tool has been presented, with a walk-through of capabilities and limitations based on several perspectives and real life scenarios. |

| | |
|---|---|
| **Diskussion** | The discussion over the panel chat as well as in the dedicated session mostly revolved around how to make the best out of the practical application. A large portion of the participants who voiced their opinions agreed that: (1) real-life exam evaluations require batch processing of answers, (2) such tools can be a potential at the hands of the human evaluator, who can form a layer of reviewing mechanism for auto-generated content, and (3) defining a fixed blueprint solution and assigning suggestive grades (percentages) to student answers would definitely be a feature that many evaluators seek in real life use-cases.<br><br>The participants and the workshop hosts agreed on the fact that generating a blueprint solution, primarily based on the information in the learning material documents rather than the pretrained information of the LLM would ensure a stronger baseline for answer assessments. Once this baseline is assured, consistent evaluation of exams can be executed for larger number of students with assigned grades. For further improvements, participants also mentioned a mechanism to apply further training for refining the grades through internal comparison of answer-grade pairs. |

| | |
|---|---|
| **Take-aways** | Participants directed their main focus on the practical implementation and real life use cases of the showcased application, reaching an overall consensus on the following principles:<br><br>- Real life examinations often involve hundreds of students. Therefore speeding up and scaling the evaluation process is crucial. Technical aspects, such as how the blueprint solution is generated, as well as what kind of an output data format is being used, all contribute significantly to the overall experience of assessing exams.<br>- Evaluators focus on the practicality of such AI-based tools, if locally available models are employed, and an additional layer of grade assessment (such as peer to peer answer comparisons, human reviews, etc.) is at all times sustained.<br>- Flexibility given to evaluators through prompting, dissecting the evaluation reasoning into parts, or by any other means is also highly appreciated. |