Golden Retrievers: Fetching Expert Curriculum Knowledge to Enhance Pedagogical Agents

E. Margaret Perkoff[1], **Sean von Bayern[1]**, Zekun Zhao[2], Jon Z. Cai[1], Emily Doherty[1], Kristin Wright-Bettner[1], James Martin[1], Martha Palmer[1]

iSAT Strand 1, Theme 2

[1]University of Colorado Boulder, [2]University of California Santa Cruz

About iSAT

The NSF AI Institute for Student-AI Teaming (iSAT) develops and tests conversational AI for classroom environments; systems that we refer to as **pedagogical agents**. The iSAT Jigsaw Interactive Agent (JIA) is an AI assistant that provides collaboration and content support directly to students during jigsaw-style group activities.

Research Questions

- How do **AMR**, **TF-IDF**, and **LLMs** compare as methods for encoding documents for knowledge retrieval?
- How does **each retrieval method** effect the downstream task of knowledge-grounded response generation?

Our motivation for this work is finding an encoding method capable of converting **any set of curriculum documents** into a knowledge base for pedagogical agents with minimal demands on the teacher.

Datasets

- **Conversational Data:** a subset of **1,400 students utterances** from the Summer 2024 series of JIA lab studies, where groups of 2-3 students were recorded doing a jigsaw activity.
- **Knowledge Base:** a collection of **1,745 knowledge facts** taken from the assorted curriculum documents (slides, lesson plans, handouts, etc.) for the jigsaw activity mentioned above.

Prior to this project, **both datasets were annotated for AMR** using a rigorous two-pass human annotation process.

Experiments

For each student utterance, we retrieved a set of **5** knowledge facts using each of our **3** retrieval methods. Scoring was calculated using **SEMBLEU of n-grams** for AMR, **cosine similarity** for the others.

We then prompted a model to respond **four times** to each student utterance: once for each set of knowledge facts retrieved, and once without any knowledge as a baseline condition.

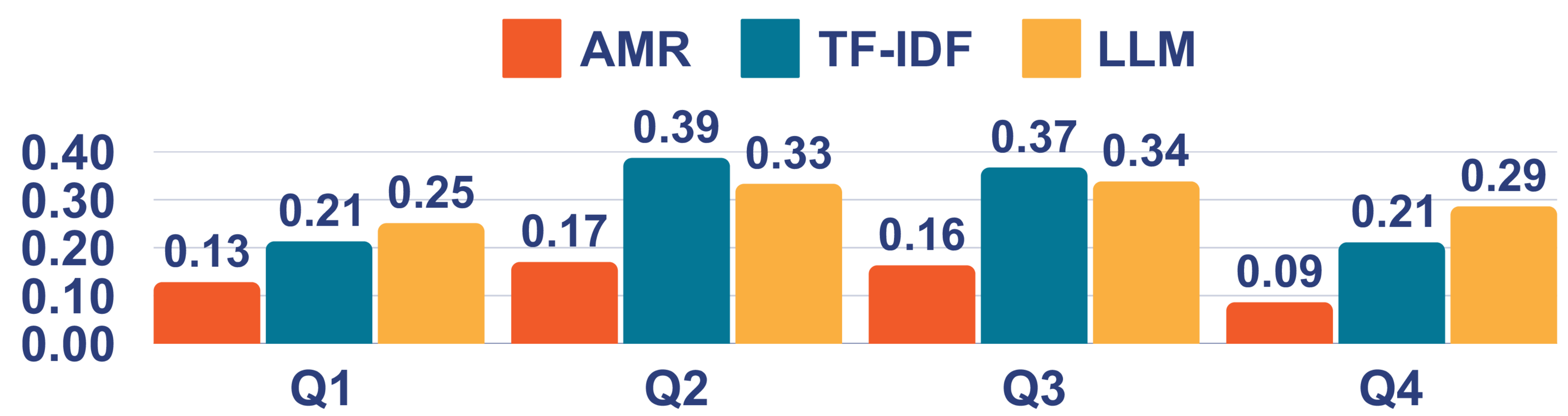

Table 1. Knowledge retrieval evaluated across all retrieval methods, broken down by worksheet question and measured in terms of **MRR**.

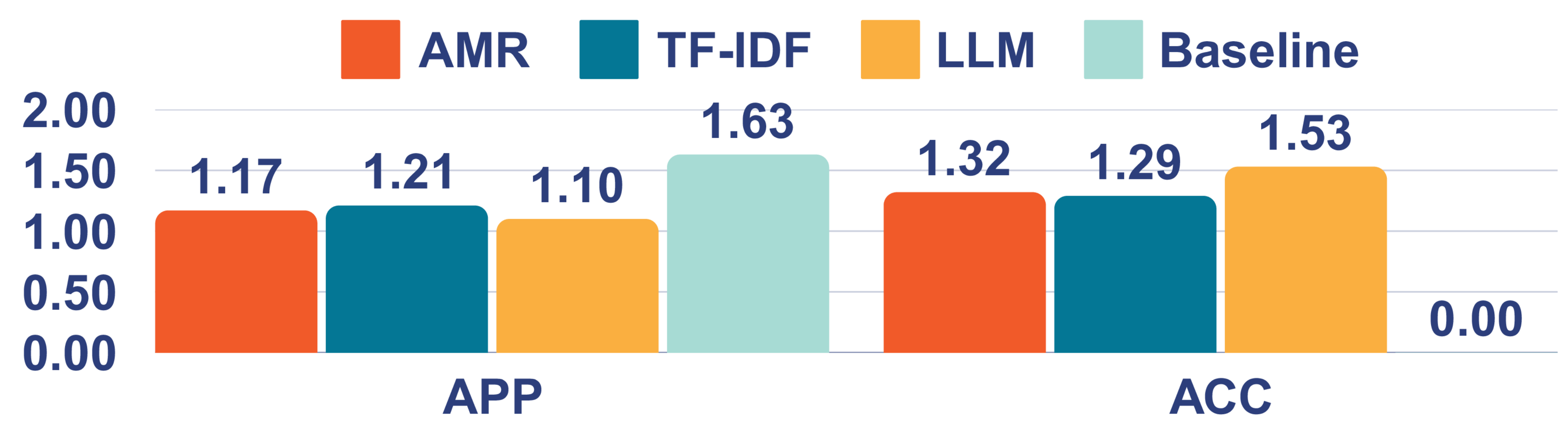

Table 2. Agent responses evaluated across all generation methods, measured in terms of **APP** and **ACC**. Baseline has no basis for ACC.

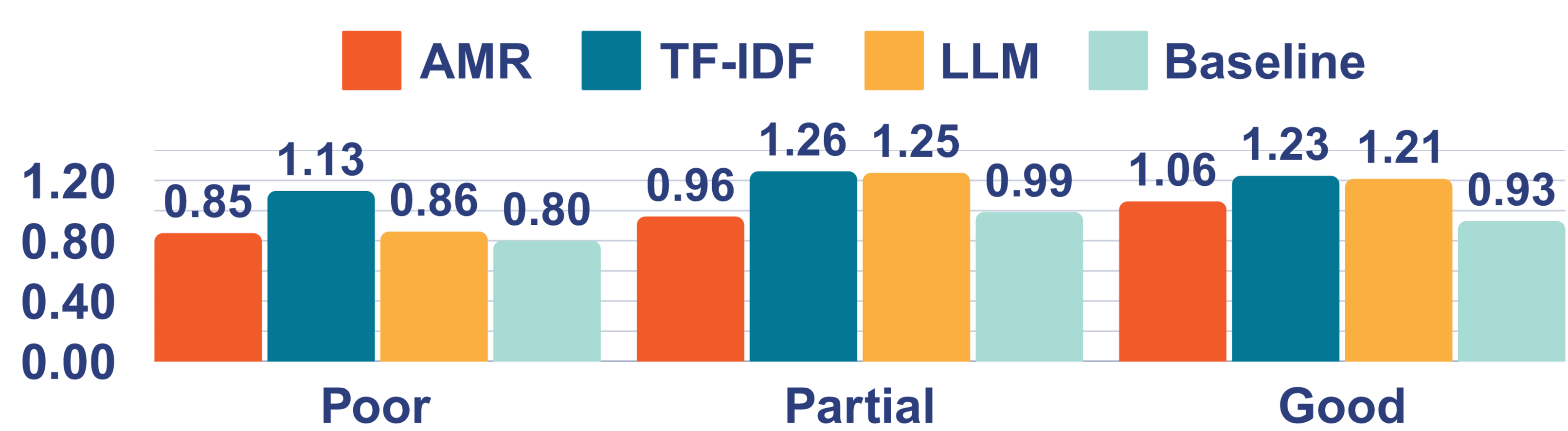

Table 3. Agent responses evaluated for **LOC**, broken down by **LOU**. Ideally, Control would inverse Understanding; but that is not the case.

Conclusions

We were surprised to find that **TF-IDF** performed as well if not better than **LLM embeddings** in both retrieval and generation tasks; also, that both methods outperformed **AMR**, despite being lower-effort and more readily automated. And we found a new problem for future work: generating responses with the consideration that Control should be **inversely** proportionate to Understanding.

Evaluation Metrics

We randomly chose **160** student utterances (plus **3** sets of knowledge and **4** generated responses **for each**) for evaluation based on:

- **Mean Reciprocal Rank (MRR):** Measures the quality of retrieved knowledge by picking the **first relevant fact** from a list and using the reciprocal of its rank as a score, i.e. **3rd = ⅓ = 0.33 points**.
- **Level of Understanding (LOU):** Measures apparent student understanding in terms of what is needed to complete the task.
- **Level of Control (LOC):** Measures how directly support is given to a student, e.g. "giving away the answer" would be high control.
- **Appropriateness (APP):** Measures how relevant a generated response is to a student utterance within conversational context.
- **Accuracy (ACC):** Measures how faithfully a generated response uses the provided knowledge (regardless of knowledge quality.)

After double-annotation of **15 different labels** per sampled utterance, our two pairs of annotators produced a total of **4,800 data points**.

Response Generation

We created a **minimalist template** that limits instructional boilerplate as much as possible, and we leveraged a **guidance grammar** to enforce additional controls without adding tokens to the prompt itself.

Prompt Template

A group of students are working together to answer the following question: **{QUESTION}**

The following is a transcript of their recent conversation:
{PRIOR UTTERANCES} x 5

You possess some knowledge that may be useful to them:
{KNOWLEDGE FACT} x 5

Use your knowledge to formulate a one-sentence hint that will help them make progress.

Guidance JSON

```
"properties":{
    "hint":{
        "title":"Hint",
        "type":"string"
    },
    "rationale":{
        "title":"Rationale",
        "type":"string"
    }
},
"required":[
    "hint",
    "rationale"
],
"type":"object"
```

iSAT

NSF AI Institute for Student-AI Teaming

This research was supported by the NSF National AI Institute for Student-AI Teaming (iSAT) under grant DRL 2019805. The opinions expressed are those of the authors and do not represent views of the NSF.