Retrieve-augmented Generation to Improve Math Question-Answering

Trade-offs Between Groundedness and Human Preference

Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, Wanli Xing
Presenter: Zach (Digital Harbor Foundation) <zach@digitalharbor.org> @zwlevonian@hci.social

Large Language Models for Math Question-Answering

<table>
<thead>
<tr>
<th>Conceptual Understanding</th>
<th>Learning gains???</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tutor Q &amp; A</td>
<td></td>
</tr>
<tr>
<td>Procedural Understanding</td>
<td></td>
</tr>
</tbody>
</table>

Retrieval-augmented Generation for Correctness & Groundedness

- Implemented a RAG system
- Asked GPT-3.5 student questions from Math Nation
- Retrieved textbook sections from OpenStax (prealgebra)
- Evaluated via survey
- Code & data available!

![Diagram of RAG system]

What we found: Humans prefer LLM responses to students' conceptual math questions when created with retrieval-augmented generation and "just the right amount" of prompting guidance.

Survey Results

- Prompt guidance affects groundedness
- Raters preferred low guidance (n=144)
- Too much guidance is possible!

Preferred responses (ranked)

<table>
<thead>
<tr>
<th>Prompt Guidance</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low</td>
<td>63.8%</td>
</tr>
<tr>
<td>Low</td>
<td>57.5%</td>
</tr>
<tr>
<td>None</td>
<td>51.7%</td>
</tr>
</tbody>
</table>

Rori: a chatbot math tutor

- Accessed via WhatsApp
- Used in classrooms and at home
- Mostly in West Africa, esp. Ghana

![Image of students using Rori]

Learning Engineering Virtual Institute

Partners: University of Utah, University of Florida, University of Oxford