A Feasibility Study on Generating Retrieval Practice Questions Using Large Language Models

Gabriel Diaz, Antonio Laverghetta Jr.
Quantitative Analysis Center, Wesleyan University

Introduction

- **Retrieval practice** is a well-established learning strategy that improves long-term retention by requiring students to actively recall previously learned material. (Roediger & Karpicke, 2006)
- Developing large banks of high-quality multiple-choice questions is **time-consuming and labor-intensive** for instructors.
- **Large language models (LLMs)** offer a promising approach for automatically generating educational assessment items (Latif N. Drasgow et al. , 2023).
- Although LLMs show promise for educational assessment items, ensuring the content validity and overall quality of generated questions remains an important challenge that often requires expert evaluation (Kasneci et al., 2023).
- This project investigates two complementary approaches to improving AI-assisted assessment development:
  - **Improve question generation** using retrieval-augmented generation (RAG) and different prompting strategies.
  - **Evaluate whether LLMs can assist expert reviewers** by automatically assessing the content validity of generated questions.

Methods

- Constructed a retrieval database from introductory psychology course materials to support Retrieval-Augmented Generation (RAG) (Lewis et al., 2020).
- Generated multiple-choice retrieval practice questions using multiple LLMs under four prompting conditions: **Zero-shot, Few-shot, Zero-shot + RAG, and Few-shot + RAG.**
- Evaluated question quality using an expert-reviewed content validity rubric and compared performance across prompting strategies, RAG, LLMs, and course topics.

Research Questions

- Does Retrieval-Augmented Generation (RAG) improve the content validity of LLM-generated retrieval practice questions?
- How do prompting strategies (zero-shot vs. few-shot) and different LLMs affect the quality of generated retrieval practice questions?

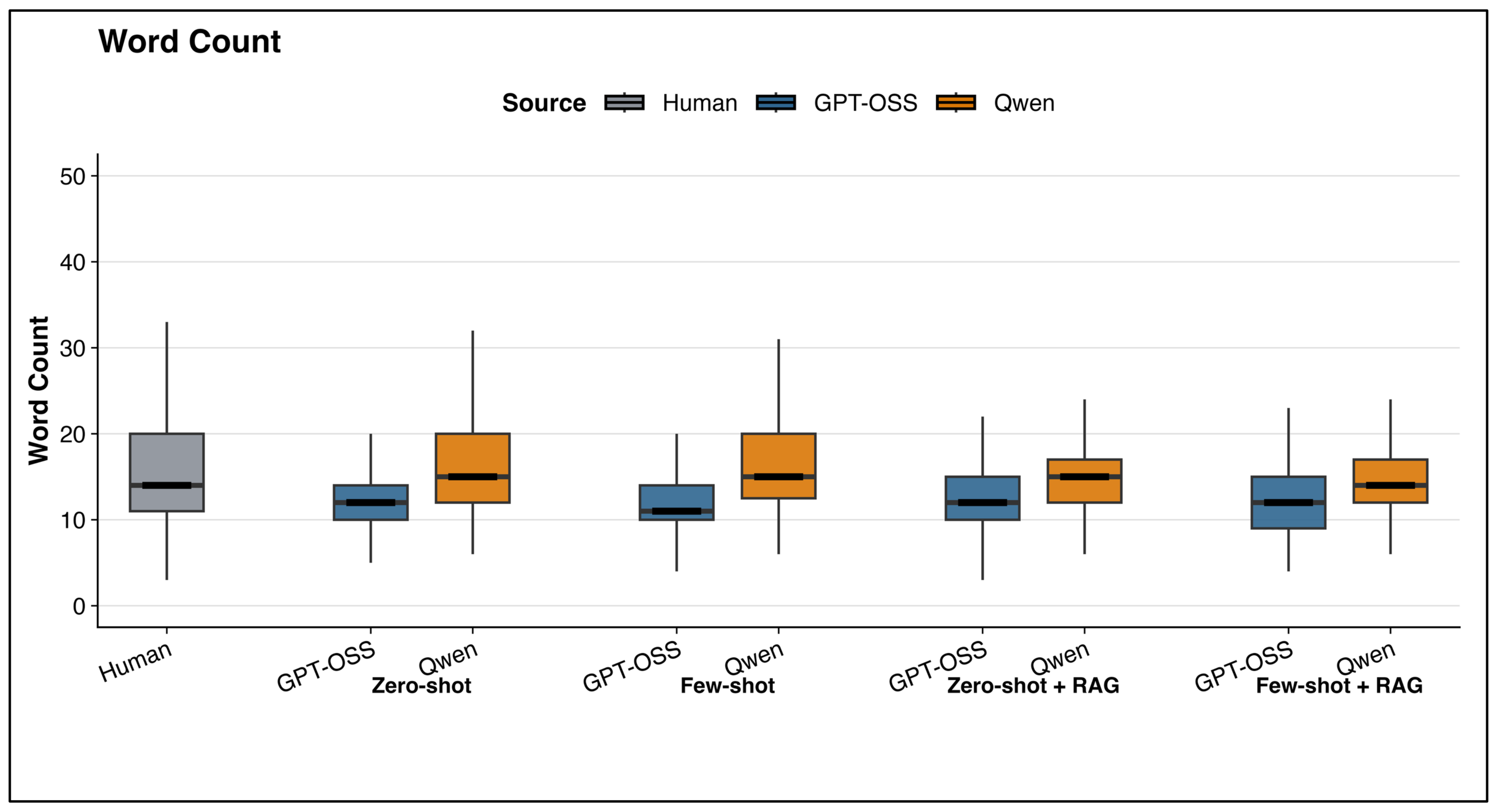

Figure 1: Word count remained consistent across models and prompting strategies

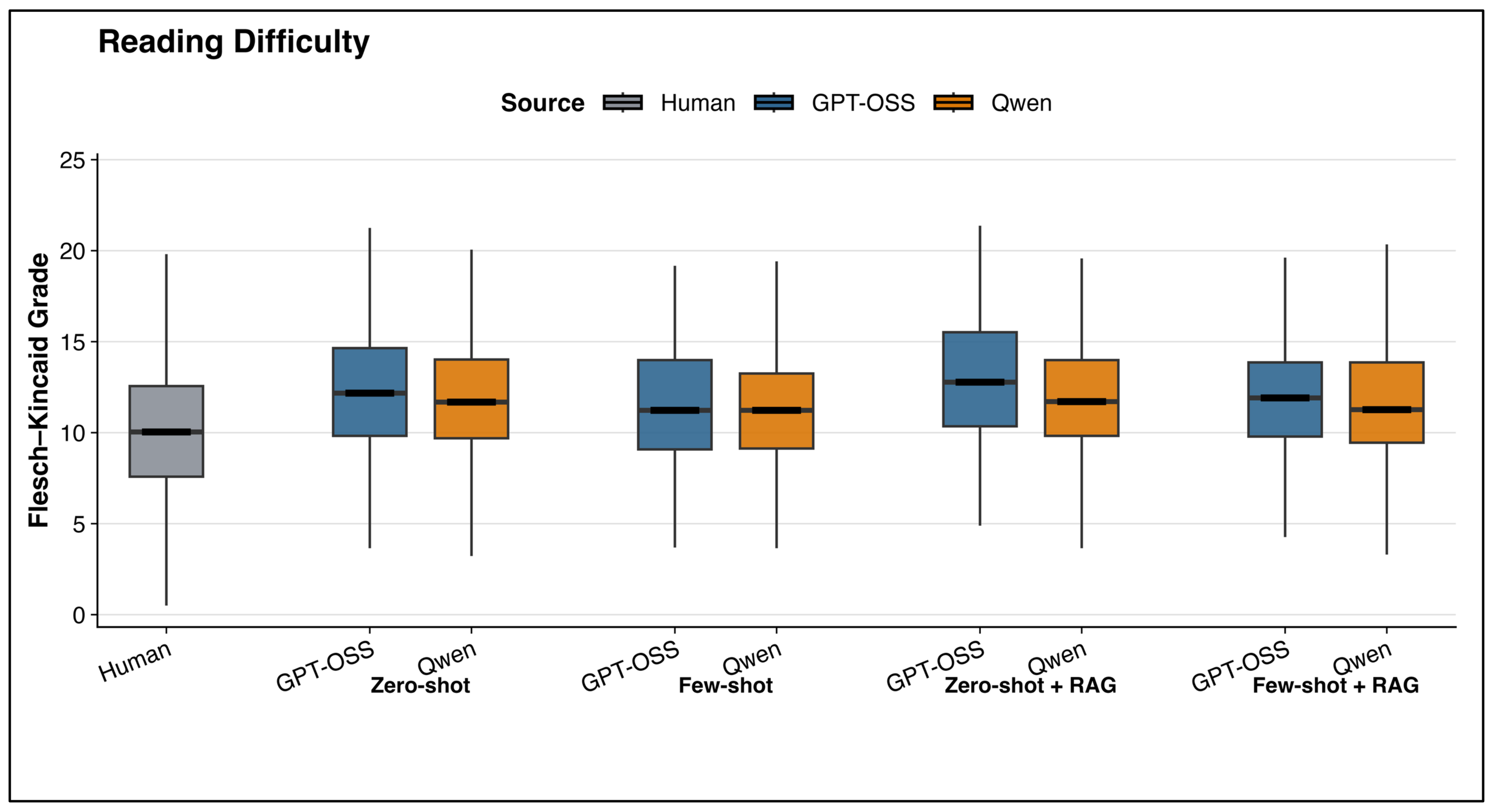

Figure 2: Reading difficulty remained consistent across models and prompting strategies

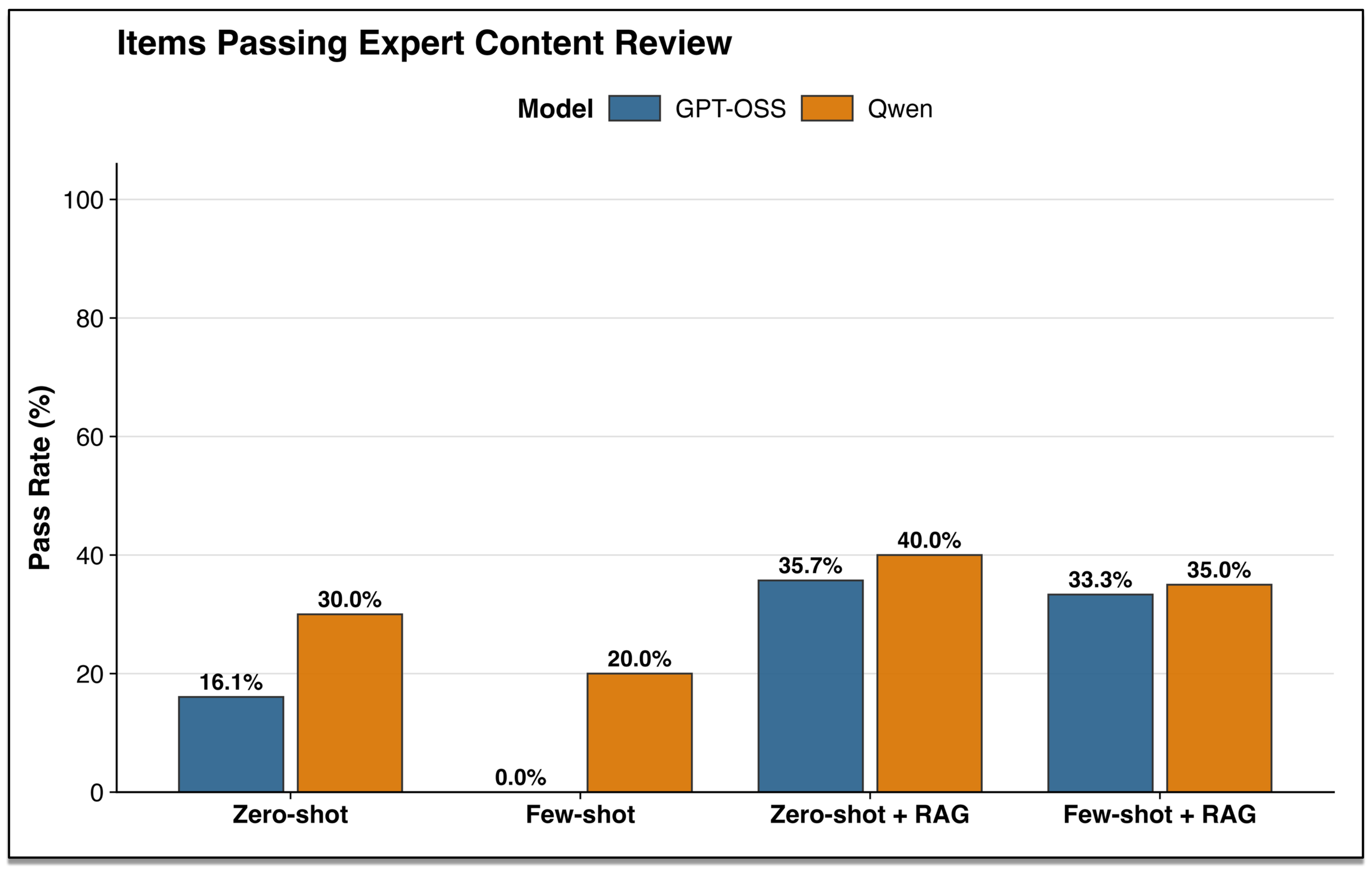

Figure 3: RAG increased the proportion of questions passing expert content review across models and prompting strategies

Results

- **Question length remained consistent** across models and prompting strategies, indicating similar question lengths.
- **Reading difficulty remained comparable** across all generation conditions, suggesting consistent readability regardless of model or prompting strategy.
- **Retrieval-Augmented Generation (RAG) substantially increased expert content review pass rates** for both GPT-OSS and Qwen compared to non-RAG approaches.
- **Qwen generally achieved higher expert content review pass rates than GPT-OSS** across comparable prompting strategies.

Discussion

- **Grounding LLM generation with course materials through RAG consistently improved question quality,** demonstrating the value of retrieval-augmented generation for educational content creation.
- **Similar word counts and reading difficulty across conditions suggest that improvements in expert ratings were driven by content quality rather than differences in question length or complexity.**
- **Both GPT-OSS and Qwen demonstrated the potential to support scalable retrieval practice question generation, particularly when combined with RAG.**
- **Future work will complete the planned LLM-based evaluation framework, validate its agreement with expert reviewers, and evaluate generated questions across additional courses and student learning outcomes.**

References:

Kasneci, E., et al. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, 103, 102274.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Advances in Neural Information Processing Systems, 33, 9459–9474.

Roediger, H. L., III, & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. Psychological Science, 17(3), 249–255.

Wang, Z., Drasgow, F., et al. (2023). Large Language Models for Education: Opportunities, Challenges, and Future Directions.