# Reasoning LLMs are Competent Courseware Reviewers

Dwyer, R. P. & Yaron, D. J.

Automated evaluation | Content validation | LLM

# Expanded Abstract

## Introduction and Approach

There is robust causal evidence that learning by doing is more effective than reading or watching video alone (Koedinger et al., 2018), so evidence-based online courseware centers student problem-solving with feedback. The REAL CHEM general chemistry courseware integrates over 1,000 questions into the text compared to just 176 in the traditional Open Stax Chemistry 2e textbook. This evidence-based design presents a challenge for course designers, however, as it is difficult for subject matter experts and student reviewers to identify and correct errors in this large quantity of targeted feedback. Scalable methods are needed to help authors review and improve large sets of activities efficiently.

Reasoning LLMs enable these questions and their feedback to be checked for errors accurately and inexpensively. We developed a system to convert each question to markdown and pass it to a reasoning LLM for review. To verify and improve the quality of the LLM's output, we built an automated evaluation system using an LLM-as-a-judge architecture. This iterative design-test-refine process represents a nested learning engineering cycle within the broader REAL CHEM development effort.

## Findings and Implications

For the automated evaluation system, we selected and identified errors (or verified no errors were present) in 60 representative REAL CHEM questions. The performance of several gpt-5 family models was measured against this set, with gpt-5-mini with medium reasoning effort and a carefully targeted prompt template achieving 88 percent accuracy. The results show that model size, reasoning depth, and prompt design all matter: limited reasoning reduces sensitivity to real errors, while targeted prompts significantly outperform generic ones. The automated evaluation system also enables iterative prompt improvement.

This approach offers a practical, low-cost method for improving large-scale instructional content. Modifications to this general approach have been used to assign alt text and learning objectives where those were missing. In REAL CHEM, the system has already improved hundreds of questions in the existing courseware.

# References

Koedinger, K. R., Scheines, R., & Schaldenbrand, P. (2018). Is the Doer Effect Robust Across Multiple Data Sets? International Educational Data Mining Society.