

## HUMAN-AI LEARNING EVALUATION

# Evaluating Socratic Dialogue Scaffolding by Fine-Tuned LLMs in Undergraduate Chemistry Tutoring

Elena Rostova, Douglas Jenkins, Tri Wulandari

1. Dept. of Chemistry Education, Universitas Tadulako; 2. Center for Human-AI Interaction; 3. Educational Technologies Institute.

Received: 12 August 2026 • Revised: 28 September 2026 • Accepted: 10 October 2026 • Published Online: 16 October 2026

### Abstract

*A double-blind randomized control trial comparing open-ended AI dialogue versus traditional static text feedback. The Socratic agent generated statistically higher conceptual transfer scores ( $d = 0.72$ ) on complex stoichiometry exams.*

**Keywords:** Large Language Models, Socratic Tutoring, Chemistry Education, Dialogue Scaffolding, Generative AI

## 1. Introduction

Generative large language models risk inducing passive cognitive reliance if they merely output direct answers to student homework problems. Fine-tuning models to adhere strictly to Socratic inquiry protocols preserves productive struggle while delivering high-touch pedagogical guidance.

## 2. Method

A 7B parameter open-weights language model was fine-tuned on 12,000 verified expert Socratic chemistry tutoring transcripts. A double-blind RCT ( $N=240$  chemistry undergraduates) compared the Socratic agent against a standard solution-providing conversational agent.

## 3. Results and Discussion

Students in the Socratic AI cohort achieved significantly higher scores on novel conceptual transfer stoichiometry problems (mean = 84.2 vs. 71.5,  $t(238) = 5.64$ ,  $p < 0.001$ ,  $d = 0.72$ ) and exhibited greater metacognitive self-explanation during think-aloud protocols.

## 4. Conclusion

Pedagogically constrained LLMs that refuse to furnish immediate answers can simulate expert Socratic one-on-one human tutoring, driving profound qualitative gains in conceptual mastery.

## Declarations

**Funding:** This study was supported by institutional research development grants.

**Conflicts of Interest:** The authors declare that they have no competing financial or personal interests.

**Data Availability:** The dataset and analysis matrices associated with this study are available in the public data repository.

**Author Contributions:** All authors contributed equally to conceptualization, investigation, formal analysis, and manuscript preparation.

## References

- Rostova, E., & Fitrasari, D. (2025). Guardrailing generative AI for Socratic inquiry in undergraduate STEM courses. *Computers and Education: Artificial Intelligence*, 7, 100234.
- Graesser, A. C., & Person, N. K. (2022). Question asking during tutoring. *American Educational Research Journal*, 31(1), 104–137.