

Learning-By-Explaining with Generative AI: A Pilot Implementation in Introductory Biology

Ulrich, P. N. , McNaughton, K. , Salami, A. , & McCarthy, K. S.

Design Implementation Framework

Generative AI

Learning-by-explaining

STEM

Explanation activities promote learning but are difficult to scale. This project applied the Design Implementation Framework (DIF) in a nested Learning Engineering cycle to investigate how generative AI can support learning-by-explaining in introductory biology. In an experimental study, students were randomly assigned to view a brief video about explaining or an extended version that also modeled how to use ChatGPT to craft and improve explanations. The extended instruction resulted in longer exchanges and more causal language in the chats. Findings illustrate DIF's value for guiding AI integration and instruction that help students to leverage AI as a tool for active learning.

Introduction

Explanation is a powerful strategy for promoting conceptual understanding. However, students struggle to produce high-quality explanations without feedback. Instructors can scaffold this process, but the realities of high-enrollment courses make offering explanation activities and feedback challenging. Generative AI presents a promising path for scale but introduces its own set of barriers. On one hand, generative AI-based chatbots can foster active engagement and personalized support. On the other hand, casual use may yield cognitive disengagement in which students let the bot “do the thinking for them”. With this in mind, this work applies learning engineering to explore generative AI–supported explanation tasks in an introductory biology course.

Within this broader learning engineering effort, we employed the Design Implementation Framework (DIF; McCarthy et al., 2020) as a nested cycle (Craig et al., 2025). DIF includes five stages: defining and evaluating the problem, ideation, user experience/design, experimental evaluation, and implementation and feedback. Although similar to other frameworks, DIF emphasizes experimentation and co-design, which align with learning engineering’s focus on human-centered design (Thai et al., 2023). This project reflects a partnership across a learning sciences lab, biology instructor and the university Center for Excellence in Teaching, Learning and Online Education (CETLOE).

Method

Explanation activities were developed and implemented in two sections of introductory biology for non-majors. Courses met in-person, but the explanation activities were conducted asynchronously as homework.

Following DIF, we first defined the problem of using generative AI to support explanation strategies in large-enrollment courses and ideated solutions guided by the ICAP framework (Chi & Wylie, 2014) to promote constructive and interactive engagement. Although we considered creating a fully customized chatbot, our prior experiences suggested students tend to default to the general ChatGPT interface. Thus, we focused on designing instructional supports to help students use ChatGPT effectively and transfer these skills to new contexts. In the user experience/design phase, we developed a ChatGPT bot that collected basic student information and exported transcripts in .json format. The bot was piloted and refined with the CETLOE support team and research assistants.

We ran an implementation in Week 2 of the semester in which students completed a learning-by-explaining activity on transfer of energy in food chains (i.e., Trophic Pyramids and the Rule of 10), with minimal guidance on writing explanations or using the bot. Usability data indicated the bot was easy to use ($M = 3.51/4.0$). Notably, analysis showed that 58 of 100 students did no more than paste the prompt into the bot and export bot responses without further engagement.

In parallel, we developed instructional materials. We created two short “Learning By Explaining” videos and corresponding reference documents: a 2-minute overview of effective explanation writing (control) and a 5-minute version that also described productive ChatGPT use (treatment). In Week 6, students were randomly assigned to treatment or control groups and worked with the bot to write an explanation of diffusion and osmosis. They submitted their ChatGPT transcript (.txt) and final explanation (.docx) and completed a brief survey on usability and perceived learning. This design allowed us to experimentally compare how different instructional supports affected engagement with the bot.

Results

Results reflect the experimental implementation in Week 6. One hundred thirteen (113) students completed the activity.

Table 1.
Means and Standard Deviations for Transcript Features as a Function of Instruction Type

| | Control M (SD) | Treatment M (SD) | t-test | p |
|-----------------------|-------------------|---------------------|--------|--------|
| # of User Turns | 5.11 (2.68) | 6.42 (3.44) | 2.27 | .03 |
| Total User Word Count | 130.54 (79.32) | 234.77 (183.06) | 4.02 | < .001 |
| # of Questions | .94 (1.67) | 1.84 (2.71) | 2.14 | .03 |
| # of Causal Words | 2.79 (1.99) | 4.30 (3.31) | 2.98 | < .001 |

Compared to the control group ($n = 52$), students in the treatment condition ($n = 61$) produced more turns and higher overall word counts, indicating sustained and active participation. They also asked more questions, suggesting a shift toward exploratory and information-seeking behavior, and used more causal language, reflecting greater engagement in reasoning and cause-effect thinking (Table 1).

Sixty one (61) students completed the post-activity survey. Students in the treatment condition ($n = 29$) gave an average rating of 3.68/4.00 ($SD = .61$) while students in the control condition ($n = 32$) gave an average rating of 3.49 ($SD = .81$). While these means are in the predicted direction, the difference failed to reach significance, $t(57.33) = 1.03$, $p = .31$.

Discussion

These preliminary findings suggest that brief instruction can encourage more constructive and interactive engagement with generative AI. We are scoring the transcripts to understand how the manipulation affected student engagement with the chatbot and how this relates to explanation quality and exam performance.

This work also provided insights into refinements we can make prior to implementation in the spring. For example, given possible ceiling effects in our survey data, we plan to increase the range of the Likert scale (e.g., 7-points) to increase sensitivity. We are using data from this cycle as feedback for refinement of our spring implementation, illustrating DIF in action.

Acknowledgments

This work was funded by Georgia State University's Center for Excellence in Teaching, Learning and Online Education (CETLOE Catalyst 2025-2026).

References

- Chi, M. T., & Wylie, R. (2014). The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychologist*, 49(4), 219-243.
- Craig, S. D., Avancha, K., [...] & Goldberg, B. (2025). Using a Nested Learning Engineering Methodology to Develop a Team Dynamic Measurement Framework for a Virtual Training Environment. In International Consortium for Innovation and Collaboration in Learning Engineering (ICICLE) 2024 Conference Proceedings: Solving for Complexity at Scale (pp. 115-132). <https://doi.org/10.59668/2109.21735>

- McCarthy, K. S., Watanabe, M., & McNamara, D. S. (2020). The design implementation framework: Guiding principles for the redesign of a reading comprehension intelligent tutoring system. In M. Schmidt, A. Tawfik, Y. Earnshaw, & I. Jahnke (Eds.), *Learner and user experience research: An introduction for the field of learning design & technology*. EdTech Books. https://edtechbooks.org/ux/9_the_design_impleme
- Thai, K. P., Craig, S. D., Goodell, J., Lis., J., Schoenherr, J. R., & Kolodner J. (2023). Learning Engineering is Human-Centered. In J Goodell (Ed.), *The Learning Engineering Toolkit* (pp. 83-124). Routledge.

