

# Fairness in Formative AI: Cognitive Complexity in Chatbot Questions Across Research Topics

**Alexandra Barry Colbert**  
College Board  
acolbert@collegeboard.org

**Karen D. Wang**  
San Jose State University  
karen.wang02@sjsu.edu

## Abstract

This study evaluates whether questions generated from a Socratic-style AI chatbot designed to support project-based AP courses maintain cognitive complexity parity when inputted with research topics of sensitive and non-sensitive nature. We present empirical findings indicating no significant conversational complexity differences, highlighting implications for equitable AI use in formative assessment.

## 1 Introduction

AI systems are increasingly deployed in educational settings for instructional and assessment support (U.S. Department of Education, 2023; Samala et al., 2025). These systems offer unprecedented opportunities for providing students with timely, individualized support at scale. Recent work demonstrates that AI systems can generate discussion prompts, assessment items, and lesson plans across various academic subjects (Scaria et al., 2024; Zheng et al., 2024; Maity et al., 2025). One promising direction is to adapt large language models (LLMs) as Socratic questioning chatbots. Specifically, AI-generated questions can prompt students to elaborate on their thinking, make connections between different topics, and develop critical thinking and metacognitive skills (Danry et al., 2023; Al-Hossami et al., 2024; Hung et al., 2024; Favero et al., 2024). This approach may help address concerns about AI tools diminishing critical thinking abilities through cognitive offloading (Gerlich, 2025), as Socratic questioning actively engages students in reflective reasoning rather than providing ready-made solutions. At the same time, successful implementations of AI systems in educational settings require careful examination of how these systems operate across varied content areas and student populations (Garcia Ramos and Wilson-Kennedy, 2024; Kaldaras et al., 2024). In the case of AI systems designed for Socratic dialogue, con-

cerns persist regarding potential algorithmic bias in the way LLM generates questions based on the varying content and perspectives found in student writing.

Existing research on LLM question generation in educational settings has focused on creating assessment items from structured course materials such as textbooks, lesson plans, and existing question banks (Tran et al., 2023; Chan et al., 2025; Shu et al., 2025). While these studies show promising results for automated question generation for STEM subjects, there remains limited investigation into how LLMs generate questions in response to student-produced writing, particularly essays addressing potentially controversial or sensitive topics. This research gap is significant because LLMs may have learned implicit associations during pre-training that influence their question generation patterns when encountering politically charged or socially sensitive content compared to neutral academic materials (Westwood et al., 2025; Bai et al., 2025). The absence of systematic investigation into how LLMs perform when generating questions based on student essays creates uncertainty about their reliability and fairness in real-world educational contexts when students write about various topics.

Our investigation in this study is guided by the following research question: to what extent does the nature of student essay topics influence the characteristics of LLM-generated, student-facing questions? Specifically, we are interested in whether students choosing potentially controversial or sensitive research topics may receive LLM-generated questions that differ in cognitive complexity from those posed to students with more neutral research topics. This study is part of a broader project at College Board's GenAI Studio to develop AI-powered educational tools in collaboration with AP students and educators, learning scientists, learning engineers, psychometricians, and ELA content experts.

An LLM-based, student-facing chatbot (“SAM”) is prompt-engineered to generate Socratic questions and support students’ deeper learning of the research process and critical thinking skills. Given the variety of interactions students might take with SAM, it is crucial to explore, prior to classroom deployment, equitable performance across different student inputs.

## 2 Methodology

### 2.1 Synthetic Student Work Samples

To ensure breadth of topic coverage and a balanced distribution of sensitive and non-sensitive essay topics, we first used an LLM to simulate student work samples based on 47 suggested themes in AP Seminar research tasks. In the AP Seminar course, to demonstrate research progress, students are required to submit stepwise evidence of their research process, including (1) an Annotated Bibliography or source chart including citations and summaries of source content, (2) an argument Outline which organizes and synthesizes sources, and (3) a completed research paper. For this study, samples were generated for the first two distinct work phases (Annotated Bibliography and Outline) and aim to represent progressing completion statuses of each phase (Severely Underdeveloped, First Draft, Final Draft). We did not generate completed papers as the chatbot is focused on providing formative support.

We used AWS Bedrock Flows to automate synthetic work sample generation through a set of chained prompts against Claude 3.5 Sonnet (see Figure 1). Given a sample theme from AP Seminar Course and Exam Description (CED) (e.g., Utopia) and topic sensitivity (“is controversial or sensitive in nature” or “is not controversial or sensitive in nature”), a topic was generated. Topic sensitivity was loosely defined in the prompt and left to the model’s interpretation. The generated topics were reviewed by a human reviewer to ensure the sensitivity distinctions were reasonable. Table 1 presents sample sensitive and non-sensitive topics based on five theme inputs.

Once a topic was generated, the flow branched to generate five subsequent work samples: Annotated Bibliography - Severely Underdeveloped, Annotated Bibliography - First Draft, Annotated Bibliography - Final Draft, Outline - First Draft, and Outline - Final Draft. The zero-shot prompts requested content “at the level of a 10th grader”. This process

generated 235 synthetic student work samples for sensitive topics and 235 for non-sensitive topics. Table 2 summarizes the distribution of synthetic work samples across work phase, development status, and topic sensitivity.

### 2.2 Chatbot Configuration and Question Generation

SAM is designed to help students reflect on their research process and provide teachers with evidence of student engagement with their research topics. The version of SAM tested in this study harnesses in-context learning capabilities of Claude 3.5 Sonnet. This foundational model was chosen for its Constitutional AI framework, which uses reinforcement learning to promote safe and fair chat experiences (Bai et al., 2022). SAM’s system prompt defines its role as a Socratic mentor and includes instructions related to conversation management, a rubric against which to support question generation, and safety measures to enforce AP-appropriate interactions. For example, in the following snippet of the Annotated Bibliography prompt, SAM is instructed to plan out its conversation with students by generating a starter question of specific types in its chain of thought.

```
Create a starter question for each of the following, in order. If elements of the work are not present in the work draft, do not skip questions, but adapt the questioning to let the student advocate for work that has not materialized yet or how they plan on approaching it.
```

```
Output the questions in <conversation_plan/> tags.
```

1. A summary question
2. A decision question
3. An evidence question
4. A speculative question
5. A reflection question

SAM’s line of questioning includes question types tailored to the work phase and span the breadth of the research experience (summary, decision, evidence, speculative, and reflection). The system prompt includes sample questions developed by learning scientists and learning engineering specialists, which serve as few-shot examples to guide SAM’s question generation (see Table 5). In order to balance personalization with standardization, SAM was encouraged to ground questions in the provided student work, but also was config-

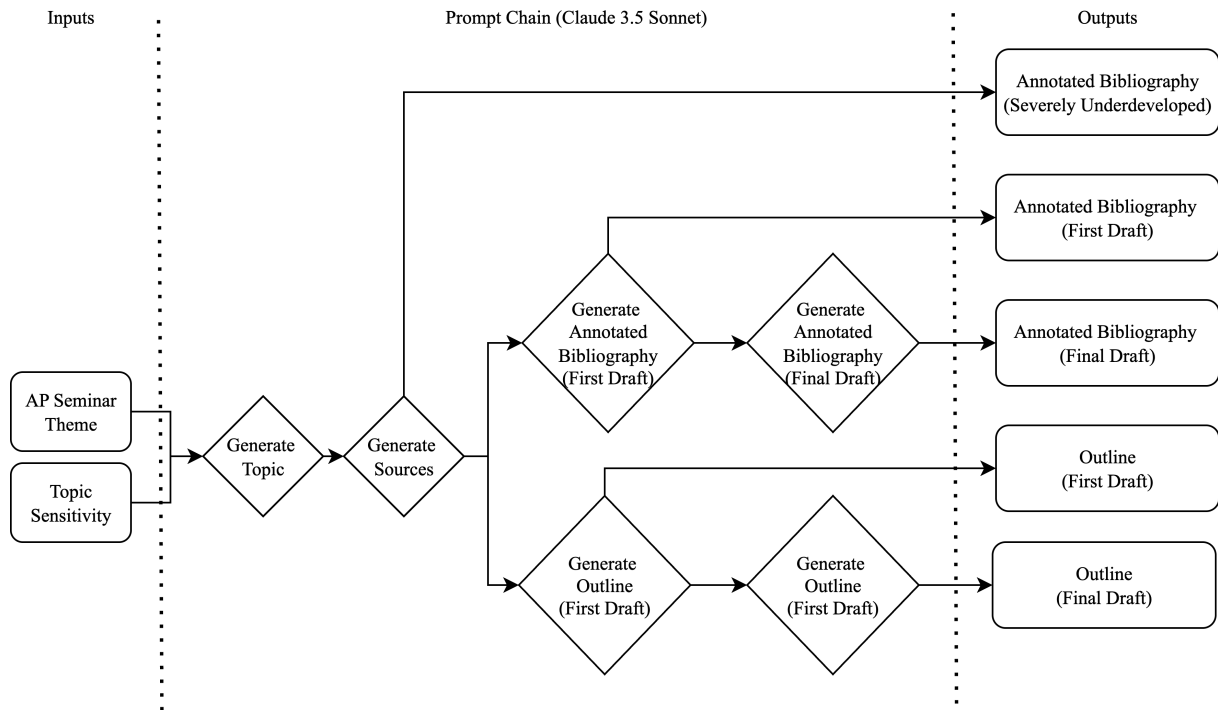


Figure 1: AWS Bedrock Flows automated generation of synthetic student work samples from input themes and topic sensitivities

Theme	Sensitive Topic	Non-Sensitive Topic
Evolution	<i>The role of eugenics in modern evolutionary theory</i>	<i>The evolutionary adaptations of migratory bird species</i>
Home	<i>The impact of home ownership on wealth inequality</i>	<i>The impact of smart home technology on energy efficiency</i>
Language	<i>The impact of gendered language on societal perceptions and biases</i>	<i>The impact of bilingual education on cognitive development in children</i>
Place	<i>The impact of gentrification on marginalized communities</i>	<i>The impact of urban green spaces on community well-being</i>
Rights and responsibilities	<i>The right to die</i>	<i>The role of individual rights and civic responsibilities in a democratic society</i>

Table 1: Sample Synthetic Topics by Theme and Topic Sensitivity

Work Phase	Development Status	Topic Sensitivity	
		Sensitive	Non-Sensitive
Annotated Bibliography	Severely Underdeveloped	47	47
	First Draft	47	47
	Final Draft	47	47
Outline	First Draft	47	47
	Final Draft	47	47
<b>Total</b>		<b>235</b>	<b>235</b>

Table 2: Synthetic Work Input Counts across Work Phase, Development Status, and Topic Sensitivity (n=470)

ured with a temperature of 0 for reproducibility. As a result, some of the questions generated by SAM closely matched the few-shot examples, while others referenced specific topics, sources, and arguments from the student work input.

Each of the 470 synthetic student work samples was submitted to SAM, which generated a conversation plan containing five starter questions for each sample. In total, this process produced 2,350 Socratic-style questions for subsequent coding and analysis.

### 2.3 Depth of Knowledge (DOK) Coding

Depth of Knowledge (DOK) was identified as a suitable framework for measuring cognitive complexity of SAM’s questions. DOK measures the complexity of thinking required to complete a task and focuses on context and depth of the task (Christopherson and Webb, 2024), making it appropriate for comparison of rigor across questions. An expert human rater labeled the DOK level of each unique SAM question, blinded to each output’s topic sensitivity condition. Table 3 presents sample SAM outputs labeled by their DOK level.

### 2.4 Statistical Analysis of Conversation DOK Levels

For samples that had four or more classifiable questions, Conversation DOK was calculated as the mean Depth of Knowledge value across questions in each conversation. To compare cognitive complexity between groups, an independent two-sample t-test was performed on mean Conversation DOK levels by topic sensitivity. To control for multiple comparisons, a Bonferroni correction was applied.

## 3 Results

Of the 2,350 Socratic-style questions produced by SAM, 1,062 were unique, as the same questions can be generated for multiple work samples. SAM’s most frequent questions across conversations are reported in Table 6.

### 3.1 Question DOK by Work Phase and Topic Sensitivity

Figure 2 compares the distribution of DOK labels across work phases for SAM’s questions. 5.9% of Annotated Bibliography questions and 33.4% of Outline questions were unclassifiable. One reason for the high volume of unclassifiable questions in

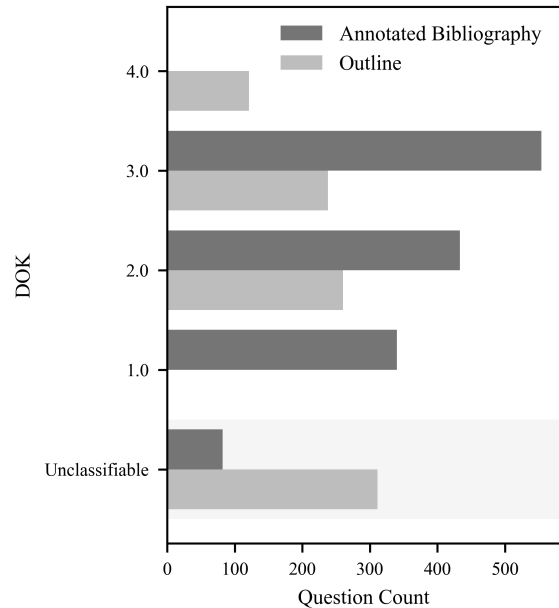


Figure 2: Question DOK (n=2,350) by Work Phase

the Outline phase was that many of SAM’s questions followed a sample question in the prompt, which elicited "yes"/"no" responses that prevented accurate cognitive complexity classification. Given that one third of Outline questions were unclassifiable, including multiple unclassifiable questions per conversation, conversation-level DOK could not be reliably calculated for Outline conversations. For this reason, we focus the DOK comparison for the questions generated for Annotated Bibliography work samples only.

Figure 3 visualizes question DOK distribution across the 282 Annotated Bibliography conversations, comparing those generated from (a) sensitive and (b) non-sensitive topics. Read vertically, each subplot illustrates the progression of conversational DOK over time. The conversations typically start with a question of low cognitive complexity (DOK of 1) and end with a question of high cognitive complexity (DOK of 3). Visual comparison of these conversation trajectories reveals no substantial differences in DOK patterns between sensitive and non-sensitive topics across all five questions.

### 3.2 Conversation DOK by Topic Sensitivity

We calculated conversation-level DOK scores for all 282 Annotated Bibliography conversation plans to quantitatively assess whether topic sensitivity influenced the cognitive complexity level of LLM-generated questions. Figure 4 displays the distribution of conversation-level DOK scores across sen-

DOK Level	SAM Question	Question Type	Work Phase
Level 1 (Recall)	<i>What is the main focus of your research based on these sources?</i>	Summary	Annotated Bibliography
Level 2 (Skills and Concepts)	<i>What is your current research topic, and how did it evolve as you developed the outline?</i>	Decision	Outline
Level 3 (Strategic Thinking)	<i>Which of these sources do you find most influential in shaping your understanding of entrepreneurial opportunities and small business success?</i>	Evidence	Annotated Bibliography
Level 4 (Extended Thinking)	<i>Reflecting on your research and outline development process, what have you learned about conducting research on such a complex topic?</i>	Reflection	Outline
Unclassifiable	<i>Based on your research so far, can you think of any potential alternatives to the death penalty that might address the concerns raised by both supporters and opponents?</i>	Speculative	Annotated Bibliography

Table 3: Sample SAM Questions by DOK Level

	Sensitive	Non-sensitive
Conversation Count	141	141
Mean	2.152	2.153
Standard Deviation	0.303	0.299

Table 4: Descriptive Statistics of Conversation DOK by Topic Sensitivity (n=282)

sitive and non-sensitive topic groups, with descriptive statistics presented in Table 4. Levene’s test confirmed homogeneity of variances,  $F(1, 280) = 0.010$ ,  $p = 0.919$ , satisfying the assumptions for parametric testing. An independent samples t-test comparing group means revealed no significant difference in conversation-level DOK between sensitive and non-sensitive topic,  $t(280) = -0.020$ ,  $p = 0.984$ . This result suggests that SAM generated questions with equivalent levels of cognitive challenge, regardless of topic sensitivity.

### 3.3 A Note about Potential Bias in Questions

Even though the expert labeler was not specifically tasked with identifying bias in the questions generated by SAM, they did surface five outputs that exhibited nuanced bias. These five questions followed a consistent pattern: *How might your research change if you were to include more sources from [an ethnic/religious/political group] perspectives?*

All five instances were speculative questions, likely generated because AP instructional materials on synthesizing perspectives were incorporated into SAM’s system prompt. These outputs present a dual interpretation challenge. On the one hand, such questions could be interpreted as problematically inviting students to presume that all members of a particular group share the same view of an issue. On the other hand, such questions might have value in encouraging students to include perspectives from communities that they had not previously considered.

While these five outputs were generated from sensitive topics, the data is not sufficient to establish a systematic relationship between topic sensitivity and problematic question generation. Fisher’s exact test revealed no significant association after multiple comparison correction ( $p = 0.030 > 0.00625$ , OR = 11.1, 95% CI: 0.62–202). Future research conducting targeted bias assessment at the individual question level is needed to determine whether topic sensitivity, question type, or other variables introduce biased questioning patterns.

## 4 Discussion

This study examined whether cognitive complexity of LLM-generated questions differs based on the sensitivity of student research topics. Our analysis

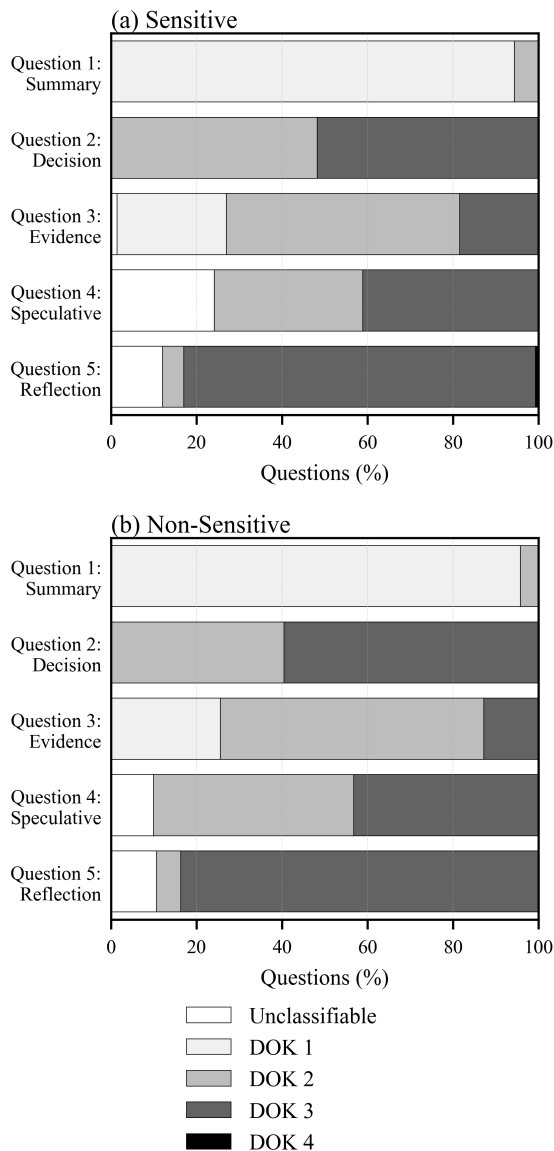


Figure 3: Question DOK (n=2,350) of Annotated Bibliography conversations by Topic Sensitivity

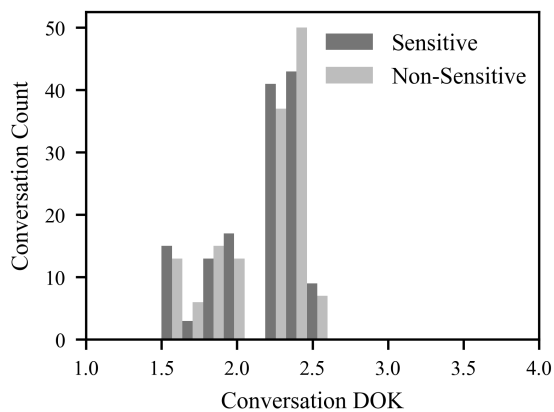


Figure 4: Conversation DOK (n=282) distribution by Topic Sensitivity

of 282 conversation plans, each consisting of questions generated by an LLM-based Socratic chatbot (SAM) in response to synthetic student Annotated Bibliography work samples, revealed comparable cognitive complexity patterns regardless of topic sensitivity. These results suggest that, under the tested configurations, SAM’s prompt engineering successfully maintains equitable question generation across both sensitive and non-sensitive topics. This finding is encouraging for developers and educators implementing LLM-based formative feedback, especially in courses where students explore controversial or sensitive topics. At the same time, our results also underscore the importance of systematic psychometric evaluation in verifying LLM performance on student-facing tasks such as question generation, especially prior to classroom deployment.

#### 4.1 Limitations

There are several limitations that affect the generalizability of our findings. First, our analysis relied on synthetic student work samples generated by LLMs rather than authentic student work. While this approach enabled controlled comparison across topic sensitivity categories, synthetic samples may not accurately reflect the style, structure, and linguistic diversity present in real student work. However, this limitation may be less pronounced for Annotated Bibliography work samples, as this assignment follows a relatively standardized format that constrain the range of possible outputs. Second, this study relies heavily on the performance of a single LLM (Claude 3.5 Sonnet) for both synthetic work generation and question generation. As the reasoning capabilities of LLMs continue to improve, it is important to re-evaluate these findings with newer models and configurations. Third, our DOK comparison focused on one stage of the AP Seminar research process (Annotated Bibliography) and examined only "starter" questions. In real classroom implementations, SAM can ask up to three follow-up questions per starter question, designed to draw more information from students based on rubric criteria and personalized to specific student responses. Due to their dynamic, response-dependent nature, follow-up questions are challenging to anticipate without authentic student interactions. However, since follow-up questions are expected to comprise more than 50% of actual conversation content, future conversational DOK measures must incorporate them to provide complete

evaluation.

Future research should address these limitations by examining authentic student work interactions with the chatbot across multiple phases of the research process and incorporating follow-up questions analysis. Additionally, investigations should extend beyond topic sensitivity to examine other student-related variables that may influence the rigor and equity of support provided by the LLM-based chatbot, such as language proficiency, socioeconomic status, and cultural background.

## Acknowledgments

The authors thank Julie Miles and Garrett Ziegler for their content expertise and contributions to question coding, without which this work would not have been possible. The authors also thank the greater College Board GenAI Studio team for their collaboration on the development of SAM.

## References

- Erfan Al-Hossami, Razvan Bunescu, Justin Smith, and Ryan Teehan. 2024. [Can language models employ the socratic method? experiments with code debugging](#). In *Proceedings of the 55th ACM Technical Symposium on Computer Science Education V. 1*, pages 53–59.
- Xuechunzi Bai, Angelina Wang, Iliia Sucholutsky, and Thomas L Griffiths. 2025. [Explicitly unbiased large language models still form biased associations](#). *Proceedings of the National Academy of Sciences*, 122(8):e2416228122.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, and 1 others. 2022. [Constitutional ai: Harmlessness from ai feedback](#). 2022. *arXiv preprint arXiv:2212.08073*, 8(3).
- K. W. Chan, F. Ali, J. Park, K. S. B. Sham, E. Y. T. Tan, F. W. C. Chong, L. Y. Ng, J. H. Lim, P. B. Ooi, L. K. Mun, B. K. S. Gan, T. S. Hew, N. W. Er, and G. K. Sze. 2025. [Automatic item generation in various stem subjects using large language model prompting](#). *Computers and Education: Artificial Intelligence*, 100344(8).
- S. Christopherson and N. Webb. 2024. [Complexity and difficulty in a coherent standards-based education system](#). <https://www.webbalign.org/difficulty-and-complexity>.
- Valdemar Danry, Pat Pataranutaporn, Yaoli Mao, and Pattie Maes. 2023. [Don't just tell me, ask me: Ai systems that intelligently frame explanations as questions improve human logical discernment accuracy over causal ai explanations](#). In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.
- Lucile Favero, Juan Antonio Pérez-Ortiz, Tanja Käser, and Nuria Oliver. 2024. [Enhancing critical thinking in education by means of a socratic chatbot](#). In *International Workshop on AI in Education and Educational Research*, pages 17–32. Springer.
- Jennifer Garcia Ramos and Zakiya Wilson-Kennedy. 2024. [Promoting equity and addressing concerns in teaching and learning with artificial intelligence](#). In *Frontiers in Education*, volume 9, page 1487882. Frontiers Media SA.
- Michael Gerlich. 2025. [Ai tools in society: Impacts on cognitive offloading and the future of critical thinking](#). *Societies*, 15(1):6.
- Jui-Tse Hung, Christopher Cui, Diana M Popescu, Saurabh Chatterjee, and Thad Starner. 2024. [Socratic mind: Scalable oral assessment powered by ai](#). In *Proceedings of the Eleventh ACM Conference on Learning@ Scale*, pages 340–345.
- Leonora Kaldaras, Hope O Akaeze, and Mark D Reckase. 2024. [Developing valid assessments in the era of generative artificial intelligence](#). In *Frontiers in education*, volume 9, page 1399377. Frontiers Media SA.
- S. Maity, A. Deroy, and S. Sarkar. 2025. [Can large language models meet the challenge of generating school-level questions?](#) *Computers and Education: Artificial Intelligence*, 100370(8).
- Agariadne Dwinggo Samala, Soha Rawas, Tianchong Wang, Janet Marie Reed, Jinhee Kim, Natalie-Jane Howard, and Myriam Ertz. 2025. [Unveiling the landscape of generative artificial intelligence in education: a comprehensive taxonomy of applications, challenges, and future prospects](#). *Education and Information Technologies*, 30(3):3239–3278.
- Nicy Scaria, Suma Dharani Chenna, and Deepak Subramani. 2024. [Automated educational question generation at different bloom's skill levels using large language models: Strategies and evaluation](#). In *International Conference on Artificial Intelligence in Education*, pages 165–179. Springer.
- C. Shu, N. Yao, Y. Chen, V. Wijeratne, L. Ma, J. Loo, S. Khan, M. Tariq, B. G. Zapirain, and A. Abuelmaatti. 2025. [Ai-assisted multiple-choice questions generation with multimodal large language models in engineering higher education](#). In *2025 IEEE Global Engineering Education Conference (EDUCON)*, pages 1–9. IEEE.
- A. Tran, K. Angelikas, E. Rama, C. Okechukwu, D. H. Smith, and S. MacNeil. 2023. [Generating multiple choice questions for computing courses using large language models](#). In *2023 IEEE Frontiers in Education Conference (FIE)*, pages 1–8. IEEE.

Office of Educational Technology U.S. Department of Education. 2023. [Artificial intelligence and the future of teaching and learning: Insights and recommendations](#). Technical report, U.S. Department of Education, Washington, DC. “Examples Are Not Endorsements”, public domain; Prepared by Miguel A. Cardona; Roberto J. Rodríguez; Kristina Ishmael.

Sean J Westwood, Justin Grimmer, and Andrew B Hall. 2025. [Measuring perceived slant in large language models through user evaluations](#). Technical Report 4262, Stanford Graduate School of Business Faculty Working Papers.

Ying Zheng, Xueyi Li, Yaying Huang, Qianru Liang, Teng Guo, Mingliang Hou, Boyu Gao, Mi Tian, Zita Liu, and Weiqi Luo. 2024. [Automatic lesson plan generation via large language models with self-critique prompting](#). In *International Conference on Artificial Intelligence in Education*, pages 163–178. Springer.

## A Sample Synthetic Work Input and SAM Outputs

Presented below is a sample synthetic Annotated Bibliography (First Draft) for a sensitive topic, accompanied by SAM’s generated conversation plan.

### A.1 Sample Sensitive Annotated Bibliography

*Annotated Bibliography*

1. Harrington, J. (2019). *The Moral Case for Capital Punishment*. *Journal of Criminal Justice Ethics*, 18(2), 45-67.

*Annotation: This article presents a philosophical argument in favor of the death penalty, contending that it is a morally justified form of punishment for the most heinous crimes. The author examines various ethical frameworks, including retributivism and utilitarianism, to build a comprehensive case for capital punishment.*

2. Wilkins, E. (2019). *Executing the Innocent: A Systemic Failure of the Death Penalty*. *Harvard Law Review*, 132(4), 789-832.

*Annotation: This comprehensive review of the death penalty system in the United States highlights the issue of wrongful convictions and executions. The author analyzes various factors, such as eyewitness misidentification, false confessions, and ineffective legal representation, that contribute to the execution of innocent individuals.*

3. Chambers, R. (2020). *The Deterrent Effect of the Death Penalty: A Review of the Evidence*. *Criminology and Public Policy*, 15(1), 171-204.

*Annotation: This article examines the empirical evidence on the deterrent effect of the death*

*penalty. The author critically evaluates the existing research and concludes that the deterrent effect of capital punishment is inconclusive, challenging the common argument that the death penalty deters crime.*

4. Linden, S. (2018). *Cruel and Unusual: The Psychological Toll of Executions on Prison Staff*. *Journal of Correctional Health Care*, 22(3), 256-271.

*Annotation: This study explores the psychological impact of carrying out executions on prison staff. The author presents findings from interviews and surveys, highlighting the emotional and mental health challenges faced by those involved in the execution process, and the need for better support and counseling for these individuals.*

5. Malik, A. (2021). *The Death Penalty and Human Rights: A Global Perspective*. *International Journal of Human Rights*, 27(2), 345-378.

*Annotation: This article provides a global analysis of the death penalty from a human rights perspective. The author examines the international legal framework, the practices of various countries, and the ongoing debates surrounding the compatibility of capital punishment with fundamental human rights.*

### A.2 SAM Output

<conversation\_plan>

1. Summary question: What is your research topic or question regarding capital punishment?

2. Decision question: How did you go about selecting these specific sources for your annotated bibliography?

3. Evidence question: Which of these sources do you find most compelling or influential in shaping your understanding of the death penalty, and why?

4. Speculative question: Based on your research so far, can you think of any potential alternatives to the death penalty that might address the concerns raised by both supporters and opponents?

5. Reflection question: How has your understanding of the death penalty evolved as you’ve researched and created this annotated bibliography?

</conversation\_plan>

## B Additional LLM Inputs and Outputs

Presented below are additional tables referenced in the text, including sample few-shot example questions and the most frequent SAM questions.



Question Type	Question Type Description	Work Phase	Few-Shot Example
Summary	Asks students to summarize their research topic	Annotated Bibliography	<i>What is your research topic?</i>
Decision	Asks students to explain their decision rationale for choosing a specific topic or including specific arguments or evidence	Annotated Bibliography	<i>How did you go about choosing your topic and narrowing down your focus?</i>
		Outline	<i>Describe for me the perspective in X source. How are you thinking of using that in section Y?</i>
Evidence	Asks students to recall sources used to supporting their conclusion/claims	Annotated Bibliography	<i>Which of the sources in your annotated bibliography were most influential in shaping your conclusion about [your research topic]?</i>
Speculative	Asks students to provide reasoning for observations or further explore related concepts	Annotated Bibliography	<i>Can you think of any evidence or data that would change your conclusion about [your research topic]?</i>
Reflection	Asks students to reflect on their research learning experience	Annotated Bibliography	<i>How did your understanding of [the topic] evolve as you researched and created your annotated bibliography?</i>
		Outline	<i>If you could improve one section, which would it be and what would you do?</i>

Table 5: Sample Few-Shot Example Questions in SAM’s Prompt by Question Type

Rank	Question	Work Phase	Question Type	DOK	Count
1	<i>Are there any sections or elements of your outline that you would like to improve or expand on?</i>	Outline	Reflection	Unclassifiable	125
2	<i>Reflecting on your research and outline development process, what have you learned about conducting research?</i>	Outline	Reflection	4	100
3	<i>How did you decide which perspectives or arguments to include in your outline?</i>	Outline	Decision	3	91
4	<i>What is the main focus of your research based on these sources?</i>	Annotated Bibliography	Summary	1	86
5	<i>What is your current research topic, and how did it evolve as you developed the outline?</i>	Outline	Reflection	2	79

Table 6: Most Frequent SAM Questions