BLADE: Better Language Answers through Dialogue and Explanations

Anonymous ACL submission

Abstract

Large language model (LLM)-based educational assistants often provide direct answers that short-circuit learning by reducing exploration and engagement with course materials. We present BLADE (Better Language Answers through Dialogue and Explanations), a retrieval-augmented generation (RAG) based conversational assistant grounded in course-specific content that guides students toward relevant materials through dialogue rather than delivering immediate solutions. We evaluate BLADE in an advanced undergraduate NLP course with extensive instructional resources, where locating and synthesizing relevant material is a central challenge. During quizzes, students are assigned to one of three conditions: BLADE only, direct course materials only, or both; we measure performance and resource usage. Results show that students consistently select BLADE over direct materials when both are available, and that performance is highest when students use BLADE alone. In contrast, simultaneous use of both resources is associated with significantly lower performance, suggesting that combining modalities may introduce cognitive load without corresponding benefit. These findings demonstrate the potential of RAG-based assistants as standalone learning resources, while cautioning that supplementing them with direct material access does not necessarily enhance, and may hinder, student learning outcomes.

1 Introduction

In advanced, resource-intensive courses such as Natural Language Processing (NLP), the rapid expansion of digital learning materials creates a paradox: while students have unprecedented access to instructional content, they often struggle to identify the content that bridges a question to a coherent conceptual understanding. Textbooks, lecture slides, recorded lectures, and curated readings are typically distributed across multiple platforms, leaving learners to manually search for and synthesize relevant information.

Recent advances in large language models (LLMs) have driven widespread adoption of conversational assistants for educational support. However, most existing systems are developed and evaluated in relatively constrained instructional settings (e.g., introductory programming tasks), where content scope and required background knowledge are limited (Kucharavy et al., 2025). In contrast, advanced undergraduate courses like NLP involve a broader, more interconnected set of concepts, requiring students to navigate and integrate information across a large body of material while still engaging in independent problem solving.

Moreover, most existing systems prioritize direct answer generation, potentially short-circuiting learning by limiting opportunities for exploration, self-explanation, and sustained engagement with instructional resources, as reflected in recent BEA shared tasks that frame tutoring as teacher response generation in dialogue settings (Kochmar et al., 2023). Prior work in learning sciences emphasizes active inquiry, evidence-based reasoning, and interaction with authoritative materials for durable conceptual understanding (Sciences et al., 2018; Jong et al., 2023). Recent work in educational NLP emphasizes interactive guidance, mistake remediation, and pedagogically grounded feedback rather than simply providing correct answers (Kochmar et al., 2025). At the same time, overreliance on AI-generated responses raises concerns about passive learning, misplaced trust, and reduced verification behavior in the presence of hallucinated or over-simplified outputs (Kucharavy et al., 2025; Beale, 2025).

To address these limitations, we introduce BLADE (Better Language Answers through Dialogue and Explanations), a conversational assistant implemented and empirically evaluated in a high-resource instructional setting. BLADE is designed

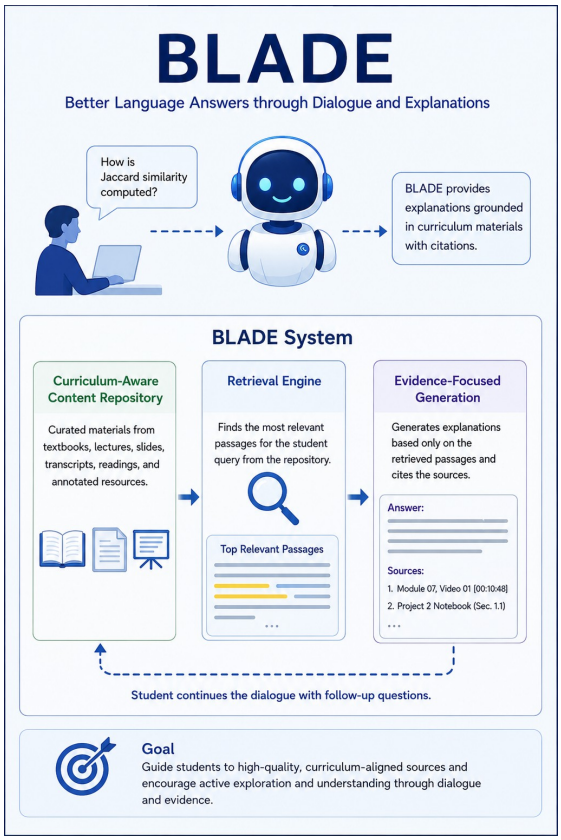

Figure 1: BLADE surfaces relevant instructional resources for student queries. Instead of giving answers, BLADE directs students to supporting evidence, fostering explanation-based learning and active exploration.

to guide learners toward relevant instructional resources rather than supplying immediate solutions. It anchors interaction in curated course materials—including textbooks, lecture slides, readings, and annotated resources—intentionally shifting the focus from answer delivery to resource discovery (Figure 1). Rather than synthesizing final responses, BLADE retrieves and highlights pedagogically relevant segments of course content, encouraging learners to consult authoritative sources and construct their own understanding through evidence-based exploration.

Our contributions are: (1) a retrieval-grounded instructional assistant that guides students to authoritative, curriculum-aligned passages and uses them as the foundation for conversational interaction; (2) a design paradigm for retrieval-based assistance in high-resource learning environments that provides structured access to relevant materials, emphasizing interaction and discovery rather than direct answer generation, thereby supporting independent problem solving; and (3) a classroom impact study evaluating BLADE in an undergraduate computer science course, where student resource usage and quiz performance are analyzed across three conditions: BLADE alone, direct course materials alone, and both combined. This design allows us to examine not only whether AI-assisted access improves learning outcomes, but how students choose to engage with available resources when given the choice.

By anchoring each exchange in vetted instructional content, BLADE provides transparent evidence for its guidance, reducing the risk of misinformation and hallucination. This design encourages active learning behaviors—elaboration, self-explanation, and retrieval practice—by prompting learners to engage directly with source material, which prior work has shown to support long-term retention and transfer. BLADE combines the natural fluency of dialogue with verifiable, traceable references, improving both learning outcomes and the transparency of AI-assisted support.

We present the design, implementation, and evaluation of BLADE in an advanced undergraduate NLP course with extensive instructional resources. Our results reveal that students consistently select BLADE over direct course materials when both are available, largely bypassing raw materials in favor of the assistant. Performance is highest when students use BLADE alone, while using both resources is associated with significantly lower performance—suggesting that, in high-resource settings, combining resource modalities may introduce cognitive load without a corresponding learning benefit. These findings highlight the promise of structured retrieval-based conversational assistants as standalone learning tools in resource-intensive environments, and underscore the importance of careful instructional design when integrating AI-assisted and direct-access resources.

2 Related Work

Retrieval-augmented generation (RAG) has become central in educational applications, combining information retrieval with the expressive flexibility of large language models. Grounding responses in retrieved text reduces hallucination compared to purely generative approaches (Lewis et al., 2020; Ji et al., 2023; Gao et al., 2024), and systems that retrieve textbook passages or lecture notes improve factual accuracy and support integrated knowledge access (Levonian et al., 2023; Henkel et al., 2024; Li et al., 2025; Lee, 2025).

In controlled deployments, RAG-based tutors improve student knowledge retention over unassisted baselines, though they may still lag behind stronger generative models (Slade et al., 2024). However, these systems are typically evaluated in settings where the assistant mediates access to information, rather than in conditions where students also have direct access to the underlying materials. As a result, it remains unclear what the assistant contributes beyond the resources—particularly in high-resource courses where students must navigate large volumes of instructional content.

A parallel line of work explores tutoring agents that promote engagement by withholding direct answers. Step-based dialogue tutors can approach the effectiveness of human one-on-one instruction (VanLehn, 2011), motivating LLM-based systems grounded in Socratic questioning and scaffolded hint generation (Rivers and Koedinger, 2014; Alshaikh et al., 2020; Favero et al., 2024; Liu et al., 2024; Ding et al., 2024). These approaches improve understanding and reflective thinking in higher-education settings (Xi et al., 2025), although poorly aligned guidance can introduce cognitive overload and hinder learning (Liu et al., 2025). Despite this emphasis on engagement, such systems continue to center interaction on guided response generation. They are rarely evaluated in terms of how students engage with the instructional materials that support those responses.

Recent work combines retrieval grounding with guided interaction (Li et al., 2025; Dong et al., 2025; Modran et al., 2024), including knowledge-graph-enhanced approaches that capture relationships beyond flat vector similarity (Dong et al., 2025). Studies in educational contexts highlight trade-offs between fidelity to source materials and student preferences (Levonian et al., 2023), underscoring that retrieval similarity alone does not ensure pedagogical usefulness. However, even in these hybrid systems, retrieval primarily supports answer generation rather than direct engagement with source material. As a result, it remains difficult to attribute learning gains to the assistant itself rather than to the quality of the resources it indexes. This ambiguity also leaves open how students engage with these resources when both AI-mediated and direct-access pathways are available.

Educational theory supports guiding students to primary resources rather than providing answers. Constructivist theory (Piaget, 1980) and retrieval-practice research (Roediger and Butler, 2011) show that locating and interpreting information strengthens memory and understanding. Dialogic pedagogy cautions that direct answers can undermine knowledge co-construction (Beale, 2025). In computer science education, guided-discovery approaches improve problem-solving transfer when students engage directly with documentation and materials (Kelleher et al., 2007), and scaffolding—rather than providing answers—drives learning gains in controlled studies (Kestin et al., 2025). These findings motivate designs that support access to relevant materials while preserving independent reasoning.

Taken together, prior work highlights three gaps. First, RAG-based tutoring systems rarely isolate the assistant's contribution from the materials they index (Li et al., 2025; Slade et al., 2024). Second, retrieval is typically optimized for topical similarity rather than pedagogical relevance (Levonian et al., 2023; Dong et al., 2025). Third, little is known about how students distribute engagement across AI-assisted and direct-access resources when both are available, particularly in high-resource settings (Slade et al., 2024; Xi et al., 2025). To address these gaps, we study a design paradigm centered on structured access to instructional materials in an advanced, resource-intensive NLP course, where students must locate, interpret, and synthesize information without relying on generated answers. We evaluate how this design affects resource usage and learning outcomes across controlled conditions.

3 BLADE System Architecture

BLADE is built on a retrieval-augmented generation (RAG) framework that integrates curriculum-specific content with a conversational LLM interface. The system consists of three primary components: (i) a curriculum-aware content repository, (ii) an instructional relevance retriever, and (iii) a controllable generation module designed to foreground evidence rather than direct answers.

3.1 Curriculum-Aware Content Indexing

Course materials—including textbooks, lecture slides, readings, and supplementary documents—are segmented into semantically coherent instructional units. These units are annotated with metadata such as topic alignment, learning objectives, and instructional context (e.g., lecture or textbook chapter). This structured indexing enables fine-grained retrieval of pedagogically relevant content in response to learner queries.

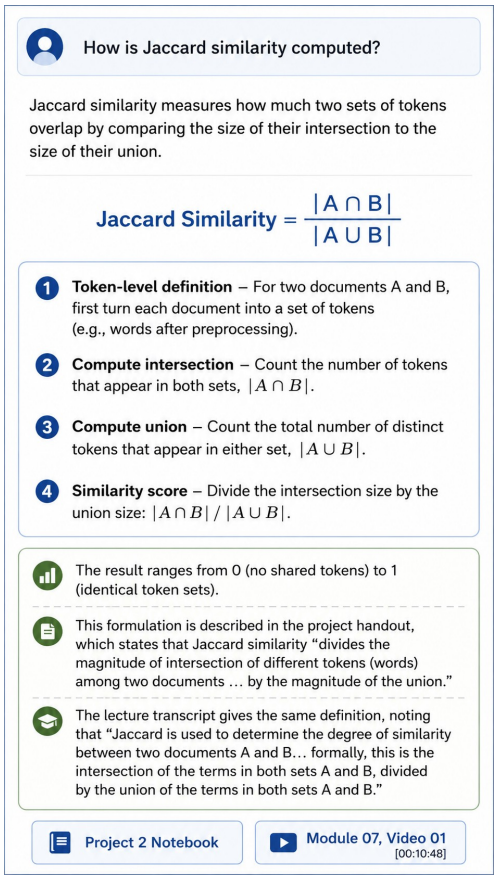

Figure 2: An example of a typical BLADE response to a query with citations of sources. The cited sources include textbook/course material excerpts and lecture transcripts with timestamps of the relevant portion.

3.2 Instructional Relevance Retrieval

Unlike standard information retrieval systems optimized primarily for topical similarity, BLADE's retriever prioritizes instructional usefulness when selecting curriculum-aligned content fragments. Rather than relying on keyword overlap alone, the system retrieves passages intended to support conceptual understanding and learning progression.

Figure 2 shows an interaction with BLADE in which a student query about Jaccard similarity is grounded in retrieved instructional materials, including project notebooks and lecture transcripts. Responses include citations and relevant lecture timestamps to help students locate supporting content efficiently. Additional examples are provided in Appendix A.

3.3 Evidence-Focused Dialogue Generation

The generative component of BLADE is designed to avoid replacing student reasoning with direct answer generation. Instead, responses are grounded in retrieved instructional materials and encourage students to explore supporting passages and related course content.

In practice, BLADE combines refusal behavior for unsupported or inappropriate requests with instructional guidance grounded in retrieved course materials. When students ask questions outside the supported instructional context, BLADE declines unsupported inferences and redirects responses toward relevant course concepts. In other cases, BLADE provides conceptual scaffolding while leaving important analytical decisions to the student, e.g., tokenization conventions or lemmatization choices that depend on task assumptions and implementation details. (See Appendix A).

This behavior is reinforced through system-prompt constraints that define the instructional scope of the assistant. The default generator is gpt-oss-120b (117B parameters; (OpenAI et al., 2025)). Prompt instructions emphasize retrieval-grounded guidance over direct problem-solving, including directives to rely on provided course materials, avoid solving assignments or providing code solutions, and redirect inappropriate requests toward course-relevant concepts.

Constraining prompts include: *"Your role is to help students understand course concepts, find information in course materials, and clarify course logistics—nothing more," "Refer to...provided files, for course content questions," "Do not provide code solutions, homework answers, or solve problems for students," "Do your best: If the query seems relevant to the course, try to find an answer using both Module [lecture] transcripts and [textbook] chapters,"* and *"If asked to do something inappropriate, decline politely and redirect to course-related questions."*

Additional system prompt constraints, including a section detailing citation requirements, guard against hallucination and ensure high-quality responses: *"ONLY use information provided in the context—do not use any external knowledge"* and *"No dummy citations: If no relevant sources are found, admit that the knowledge base is insufficient to answer the question."*

4 Classroom Deployment and Evaluation

We evaluate BLADE in a Fall 2025 impact study conducted in an advanced undergraduate NLP course. All enrolled students are given access to BLADE from the start of the semester and interact with the assistant alongside standard course mate-

rials during homework assignments and study sessions. Of approximately 200 enrolled students, 85 undergraduate computer science students elect to participate, completing three quizzes mid-semester after recruitment via a study announcement and interest form (see Appendices B and C).

4.1 Study Context and Participants

Participants are randomly divided into three groups and assigned the same three quizzes on NLP concepts across three course modules. Each quiz consists of short-answer and concept-application questions focused on conceptual understanding rather than extended derivations or programming.

Each group completes all three quizzes, with the available resource condition rotated across quizzes according to a fixed permutation, yielding nine quiz–resource combinations in total. Each assignment is followed by a survey that records how students use their allowed resources. These data are used to evaluate students' resource choices, student performance, and the relationship between performance and resource availability and usage.

BLADE is evaluated along three primary dimensions: (i) patterns of resource usage when multiple modalities are available; (ii) student performance across three resource conditions—BLADE alone, direct materials alone, and both combined; and (iii) the relationship between resource choices and learning outcomes.

In this high-resource setting, results show that students use BLADE more frequently when both options are available. Quiz performance analysis shows that access to BLADE alone yields higher scores than access to materials alone, while the combined condition—despite substantially higher resource engagement—does not outperform BLADE alone. Survey results confirm that additional access to direct materials neither shifts engagement away from BLADE nor translates into performance gains.

4.2 Resource Configurations

Students complete quizzes in a controlled environment with access only to the resources assigned to their condition. Resource usage is voluntary: students may use any, all, or none of the provided materials (see Figure 3). For example, a student in the BLADE-plus-materials condition may use only BLADE, only the materials, or neither.

This design allows us to observe not only whether access to a resource improves performance, but also how students choose to engage with it when given the opportunity. Post-quiz surveys allow us to triangulate self-reported engagement with observed performance.

Together, the quiz results and usage data allow us to address two related questions: whether access to BLADE improves learning outcomes relative to direct materials alone, and whether students who engage with BLADE perform differently from those who, despite having access, do not use it.

4.3 Group Assignments

The quiz configurations are defined as follows:

- Config A: access to BLADE but no other course resources.
- Config B: access to BLADE plus all course resources.
- Config C: no access to BLADE, but access to other course resources.

Group assignments across quizzes and configurations are shown below, yielding nine total quiz-condition assignments:

Quiz	Config A	Config B	Config C
Quiz 1	group1	group2	group3
Quiz 2	group2	group3	group1
Quiz 3	group3	group1	group2

5 Results

Study results are analyzed along three dimensions: (i) student resource usage based on survey data; (ii) mixed-effects modeling of the effects of resource availability, resource usage, quiz content, and student ability on performance; and (iii) average quiz performance by resource condition and usage.

5.1 Resource Usage Patterns

Returning to Figure 3, we report the percentage of students in each condition who use at least one resource to answer a quiz question. Across all three conditions, a substantial proportion of students do not consult any resource, suggesting a reliance on prior knowledge even when support tools are accessible. This tendency is strongest when only course materials are available (Config C) and weakest in the combined condition (Config B). Even when resource use is at its highest, students do not distribute their engagement evenly across available options, indicating selective rather than uniform resource use.

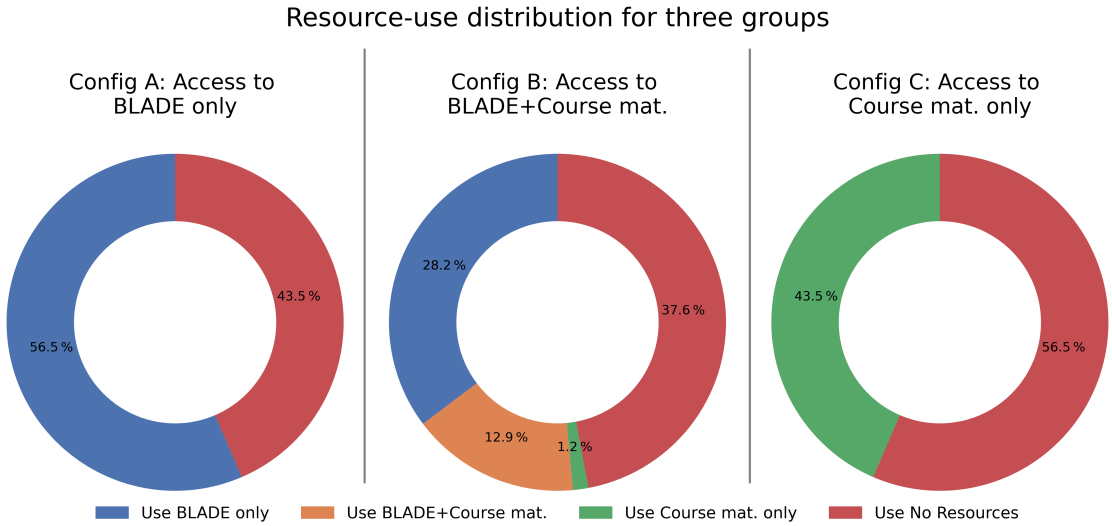

Figure 3: Distribution of student resource usage across the three conditions: BLADE only (Config A), BLADE + course materials (Config B), and course materials only (Config C). Each chart partitions students into mutually exclusive categories based on actual usage: BLADE only, course materials only, both resources, or no resources.

When BLADE is available, students consistently use it over direct course materials. In Config A, 56.5% of students use BLADE. In Config B, 28% use BLADE exclusively, while only 12.9% use both BLADE and course materials. These results indicate that students largely bypass direct materials when BLADE is available: only 12.9% in Config B use both resources, compared to 43.5% in Config C when course materials are the only option. This suggests that students turn to course materials primarily when no alternative is available, but consistently choose BLADE when available, and that adding direct materials does not meaningfully change engagement with BLADE.

5.2 Statistical Analysis with Linear Mixed-Effect Modeling

Quiz performance is analyzed with a linear mixed-effects model with resource availability, resource usage, and quiz as fixed effects, and student as a random intercept. The second quiz in the BLADE-only condition (Config A) serves as the baseline.

The analysis indicates a marginal effect of resource availability. Compared to the baseline, access to both resources is associated with a modest increase in performance ($\beta = 6.814, SE = 3.603, p = .059$), although this effect is not statistically significant. Access to course materials alone does not differ significantly from the baseline ($\beta = 4.772, SE = 4.182, p = .254$).

Resource usage shows a different pattern. Relative to students using only the baseline resource, those using both resources score significantly lower ($\beta = -14.964, SE = 6.563, p = .023$). The effects of using only course materials ($\beta = -5.448, p = .312$) and using no resources ($\beta = -6.221, p = .123$) are not statistically significant.

Quiz effects were also significant, indicating differences in difficulty. Scores on Quiz 3 ($\beta = -12.461, p < .001$) and Quiz 1 ($\beta = -8.737, p = .010$) are significantly lower than those on the reference quiz. The random intercept variance for students ($\sigma^2 = 145.317$) indicates substantial variability in baseline performance across individuals, justifying the use of a mixed-effects model.

5.3 Student performance with respect to resource usage

Students are grouped into three performance levels based on their overall quiz scores. Thresholds are determined using approximate tertile-inspired partitions informed by the standard grade statistics from the observed score distribution:

- **Upper:** students scoring above 73%
- **Mid:** students scoring between 27% and 73%
- **Lower:** students scoring below 27%

We analyze the distribution of students across the three performance groups (Upper, Mid, Lower) and overall performance.

Figure 4 shows quiz performance distributions across resource configurations. The highest average score (70.3%) is achieved by students using only BLADE in Config A or B. Students using only course materials (Config B or C) achieve the second-highest average (66.3%), followed closely by those using no resources (66.1%). The lowest

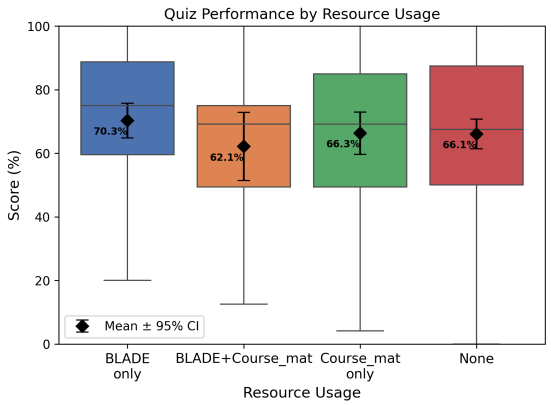

Figure 4: Distribution of quiz performance for all students by resource usage across all configurations (e.g., "BLADE only" on the x-axis includes students who use only BLADE in all configurations where it is available).

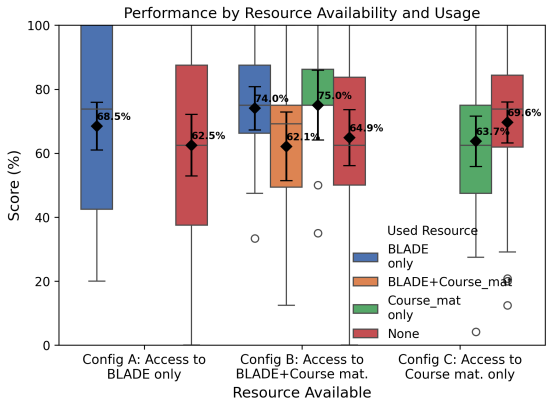

Figure 5: Distribution of quiz performance for all students by resource usage, further categorized by resource-availability settings (e.g., in the "Config A: Access to BLADE only" setting, the x-axis includes two usage patterns: using BLADE only ("BLADE only") or using no resources ("None")).

average score (62.1%) occurs when students use both BLADE and course materials (Config B).

Figure 5 provides a more fine-grained analysis by incorporating resource availability conditions. In both settings where BLADE is available (Config A and B), students using BLADE alone achieve higher average scores than those using no resources. However, in Config B, students using only course materials achieve the highest average score (75.0%), with BLADE-only users close behind (74.0%). Students using both BLADE and course materials in Config B again show the lowest average performance (62.1%).

Across both analyses, students achieve higher performance when they use BLADE, particularly when it is used as the sole resource. In contrast, using only course materials or no resources yields comparable but less consistent outcomes. Notably, using both resources is consistently associated with lower performance.

Figure 6 shows students' resource selections across performance levels. In all groups, the most common choice is no resource, consistent with patterns in Section 5.1. Across performance levels, BLADE is the second most frequently used resource, with the highest proportion of users in the upper-performing group. Upper- and mid-level students exhibit similar patterns, with course materials alone ranking third. In contrast, lower-performing students rely more on the combined BLADE-plus-materials setting than on course materials alone.

Across performance levels, BLADE usage decreases from upper to lower groups, while reliance on the combined setting increases. This pattern suggests that lower-performing students attempt to broaden understanding by consulting multiple resources. However, as shown in Section 5.2, this strategy does not improve performance. Instead, the decline may reflect increased cognitive load from integrating information across multiple sources in a high-resource setting.

6 Conclusions

BLADE is a curriculum-aware, retrieval-grounded conversational assistant that supports student learning by guiding access to course-specific instructional materials rather than providing direct answers. Through dialogue-based interaction, it surfaces relevant content while preserving independent reasoning. We evaluate BLADE in an advanced, high-resource undergraduate NLP course across three conditions—BLADE alone, direct materials alone, and both combined—examining quiz performance and resource usage.

Students consistently select BLADE over direct course materials when both are available, largely bypassing raw materials in favor of the assistant. Course materials are used more when they are the only available resource, suggesting that BLADE's ability to surface relevant content reduces reliance on unmediated access under time-constrained quiz conditions. This pattern suggests that BLADE supports engagement with domain content in ways that direct access alone does not.

Performance analysis reinforces this pattern. Students achieve their highest scores when using BLADE alone. Mixed-effects modeling shows that simultaneously using both resources is associated with significantly lower performance compared to

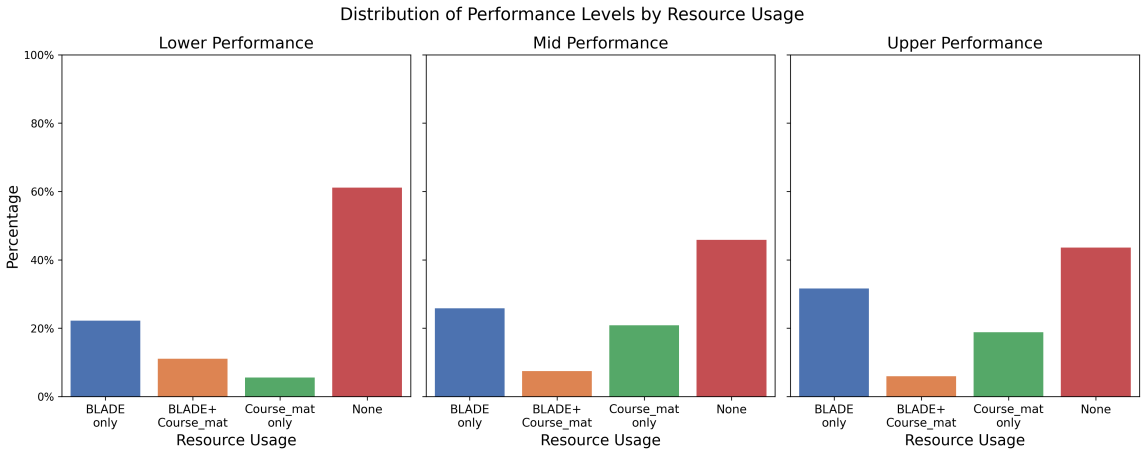

Figure 6: Distribution of students across the three performance levels by resource usage, as chosen by each student. Proportions represent, for each performance level, the fraction of students who used a particular resource setting out of all students in that performance level.)

using BLADE alone, while using only course materials or no resources does not differ significantly from the BLADE-only baseline. Performance-group analysis shows that BLADE usage is most prevalent among upper-performing students, while reliance on the combined setting increases among lower-performing students. These findings suggest that students who struggle may seek broader coverage but encounter increased cognitive load when navigating multiple sources, without a corresponding performance benefit.

Overall, these findings support the value of a well-designed RAG-based assistant as a standalone learning resource, while cautioning against the assumption that more resources are always better. Combining AI-mediated and direct-access materials does not appear to be additive; rather, it may fragment attention in ways that undermine the efficiency gains BLADE provides on its own. Future work should examine how instructional design choices—such as scaffolding multiple resource types—might help students leverage combined resources more effectively.

7 Future Work

Building on BLADE's core capabilities, several directions emerge for advancing human-centered AI learning systems. One avenue is to evaluate BLADE's long-term effectiveness by assessing whether it enhances students' retention of learned concepts compared to traditional course resources, using pre- and post-assessments.

A second direction is to embed AI interaction within authentic task environments, allowing learners to engage with AI assistance while completing domain-specific assignments, analyses, or decision-making tasks. Such integration enables richer observation of how AI influences real-world reasoning beyond isolated explanation dialogues.

Future systems would benefit from comprehensive behavioral instrumentation, capturing not only AI dialogue content but also users' navigation of resources, verification strategies, timing of actions, and revision behaviors. These multimodal interaction traces support fine-grained measurement of reliance, skepticism, and self-correction, core components of effective human-AI collaboration.

Finally, extending this paradigm beyond a single domain to multiple disciplines—such as medicine, social sciences, and the humanities—enables broader evaluation of explanation-centered AI assistance across reasoning contexts.

Together, these directions point toward a new generation of interactive AI learning environments that extend BLADE's curriculum-aware foundation into measurable and cognitively resilient systems.

Limitations

Despite its strengths, BLADE has key limitations that constrain its ability to support broader cognitive training and rigorous behavioral analysis.

Firstly, BLADE is designed for structured, explanation-centered learning within specific curricular domains, particularly NLP. While effective for concept exploration, it does not fully integrate AI interaction into authentic task execution (e.g., writing reports, solving domain-specific problems, or making real-time decisions). As a result, en-

gagement occurs largely in isolation from real task performance, limiting insight into how AI assistance influences applied reasoning.

Furthermore, although BLADE encourages users to consult instructional materials through evidence-focused responses, it does not systematically capture fine-grained behavior. The system lacks continuous measurement of interaction sequences such as resource access patterns, verification strategies, follow-up questioning, and timing. This limits the ability to assess reliance on AI outputs, depth of verification, and self-correction over time.

Additionally, while BLADE references instructional sources and encourages further exploration, some responses may still function as concise explanatory summaries. Future work should more directly measure whether students engage with cited materials beyond the generated response itself.

Finally, BLADE operates primarily as a standalone instructional assistant rather than as part of a broader experimental framework that supports staged learning, longitudinal evaluation, or comparisons across interaction conditions.

Ethical Considerations

Figures 1 and 2 were created with the assistance of a generative AI system to produce visual illustrations based on author-designed concepts. The AI tool was used solely for figure visualization. gpt-oss-120b serves as the default generator for BLADE assistant and is released under the Apache 2.0 license, whose terms of use we follow.

The BLADE study was reviewed by the authors' Institutional Review Board (IRB) and determined to qualify for exempt status under educational research involving normal instructional practices in established educational settings. Participation was voluntary, and students were informed about the nature of the study and given the opportunity to decline participation. The quizzes used in the study did not affect course grades beyond minimal participation-based extra credit. No deception, behavioral intervention, or sensitive personal data collection was involved.

Student performance data is used exclusively for research purposes. Before analysis, all direct identifiers were removed and indirect identifiers minimized. The resulting dataset consists solely of student performance records (e.g., grades), contains no offensive content, and is stored in access-controlled storage.

All research ideas, methodology, data analysis, and interpretations are the work of the authors, who reviewed and validated the generated figures.

References

Zeyad Alshaikh, Lasagn Tamang, and Vasile Rus. 2020. A Socratic Tutor for Source Code Comprehension. In *Artificial Intelligence in Education*, pages 15–19, Cham. Springer International Publishing.

Russell Beale. 2025. Dialogic pedagogy for large language models: Aligning conversational AI with proven theories of learning. *arXiv preprint arXiv:2506.19484*.

Yuyang Ding, Hanglei Hu, Jie Zhou, Qin Chen, Bo Jiang, and Liang He. 2024. Boosting large language models with socratic method for conversational mathematics teaching. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pages 3730–3735.

Chenxi Dong, Yimin Yuan, Kan Chen, Shupei Cheng, and Chujie Wen. 2025. How to build an adaptive AI tutor for any course using knowledge graph-enhanced retrieval-augmented generation (KG-RAG). In *2025 14th International Conference on Educational and Information Technology (ICEIT)*, pages 152–157. IEEE.

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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. *arXiv preprint*. ArXiv:2312.10997 [cs].

Owen Henkel, Zach Levonian, Chenglu Li, and Millie Postle. 2024. Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference. pages 315–320.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Computing Survey*, 55(12).

Ton de Jong, Ard W. Lazonder, Clark A. Chinn, Frank Fischer, Janice Gobert, Cindy E. Hmelo-Silver, Ken R. Koedinger, Joseph S. Krajcik, Eleni A. Kyza, Marcia C. Linn, Margus Pedaste, Katharina Scheiter, and Zacharias C. Zacharia. 2023. Let's talk evidence – The case for combining inquiry-based and direct instruction. *Educational Research Review*, 39:100536.

Caitlin Kelleher, Randy Pausch, and Sara Kiesler. 2007. Storytelling alice motivates middle school girls to learn computer programming. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems 2007, CHI 2007*, pages 1455–1464. Association for Computing Machinery.

Greg Kestin, Kelly Miller, Anna Klales, Timothy Milbourne, and Gregorio Ponti. 2025. AI tutoring outperforms in-class active learning: An RCT introducing a novel research-based design in an authentic educational setting. *Scientific Reports*, 15(1):17458.

Ekaterina Kochmar, Jill Burstein, Andrea Horbach, Ronja Laarmann-Quante, Nitin Madnani, Anaïs Tack, Victoria Yaneva, Zheng Yuan, and Torsten Zesch, editors. 2023. *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*. Association for Computational Linguistics, Toronto, Canada.

Ekaterina Kochmar, Kaushal Maurya, Kseniia Petukhova, KV Aditya Srivatsa, Anaïs Tack, and Justin Vasselli. 2025. Findings of the BEA 2025 Shared Task on Pedagogical Ability Assessment of AI-powered Tutors. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 1011–1033, Vienna, Austria. Association for Computational Linguistics.

Andrei Kucharavy, Cyril Vallez, and Dimitri Percia David. 2025. LLMs Protégés: Tutoring LLMs with Knowledge Gaps Improves Student Learning Outcome. In *Proceedings of the 20th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2025)*, pages 248–257, Vienna, Austria. Association for Computational Linguistics.

Youngjin Lee. 2025. Developing a computer-based tutor utilizing Generative Artificial Intelligence (GAI) and Retrieval-Augmented Generation (RAG). *Education and Information Technologies*, 30(6):7841–7862.

Zachary Levonian, Chenglu Li, Wangda Zhu, Anoushka Gade, Owen Henkel, Millie-Ellen Postle, and Wanli Xing. 2023. Retrieval-augmented Generation to Improve Math Question-Answering: Trade-offs Between Groundedness and Human Preference. *arXiv preprint*. ArXiv:2310.03184 [cs].

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Zongxi Li, Zijian Wang, Weiming Wang, Kevin Hung, Haoran Xie, and Fu Lee Wang. 2025. Retrieval-augmented generation for educational application: A systematic survey. *Computers and Education: Artificial Intelligence*, 8:100417.

Jiayu Liu, Zhenya Huang, Tong Xiao, Jing Sha, Jinze Wu, Qi Liu, Shijin Wang, and Enhong Chen. 2024. SocraticLM: Exploring Socratic Personalized Teaching with Large Language Models. In *Advances in Neural Information Processing Systems*, volume 37, pages 85693–85721. Curran Associates, Inc.

Ying Liu, Can Li, Ting Zhang, Mei Wang, Qiannan Zhu, Jian Li, and Hua Huang. 2025. Discerning minds or generic tutors? Evaluating instructional guidance capabilities in Socratic LLMs. *arXiv preprint arXiv:2508.06583*.

Horia Alexandru Modran, Ioana Corina Bogdan, Doru Ursuțiu, Cornel Samoilă, and Paul Livius Modran. 2024. LLM intelligent agent tutoring in higher education courses using a RAG approach. In *International Conference on Interactive Collaborative Learning*, pages 589–599. Springer.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b Model Card. *arXiv preprint*. ArXiv:2508.10925 [cs.CL].

J. Piaget. 1980. *To Understand Is to Invent: The Future of Education*. Penguin Books.

Kelly Rivers and Kenneth R. Koedinger. 2014. Automating Hint Generation with Solution Space Path Construction. In David Hutchison, Takeo Kanade, Josef Kittler, Jon M. Kleinberg, Alfred Kobsa, Friedemann Mattern, John C. Mitchell, Moni Naor, Oscar Nierstrasz, C. Pandu Rangan, Bernhard Steffen, Demetri Terzopoulos, Doug Tygar, Gerhard Weikum, Stefan Trausan-Matu, Kristy Elizabeth Boyer, Martha Crosby, and Kitty Panourgia, editors, *Intelligent Tutoring Systems*, volume 8474, pages 329–339. Springer International Publishing, Cham. Series Title: Lecture Notes in Computer Science.

Henry L. Roediger and Andrew C. Butler. 2011. The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15(1):20–27.

National Academies of Sciences, Medicine, Board on Science Education, Board on Behavioral, Sensory Sciences, Committee on How People Learn II, The Science, and Practice of Learning. 2018. *How people learn II: Learners, contexts, and cultures*. National Academies Press.

Joseph J Slade, Alina Hyk, and Regan AR Gurung. 2024. Transforming learning: assessing the efficacy of a retrieval-augmented generation system as a tutor for introductory psychology. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, volume 68, pages 1827–1830. SAGE Publications Sage CA: Los Angeles, CA.

Kurt VanLehn. 2011. The relative effectiveness of human tutoring, intelligent tutoring systems, and

other tutoring systems. *Educational psychologist*, 46(4):197–221.

Linjin Xi, Yi Zhang, and Qiyun Wang. 2025. Investigating the effects of an LLM-based Socratic conversational agent on students' academic performance and reflective thinking in higher education. *Computers & Education*, page 105494.

A Examples of BLADE Responses to Student Questions

Query 1 (below) relates to an NLP application discussed in lecture. The application concerns a bot that detects and thwarts social engineering attempts; in one case, the bot's response includes a link that reveals the attacker's identity.

Query 1: Isn't syntactic analysis used to generate a clickable link for a social engineer?
Response 1: Syntactic analysis is a pipeline component that assigns structure to an input string, recovering the constituency structure in preparation for semantic interpretation. The provided context does not discuss the application of syntactic analysis in generating clickable links for social engineers.

Syntactic analysis is focused on assigning structure to input strings, such as sentences, to prepare for semantic interpretation. It does not involve generating clickable links or dealing with social engineering.

Therefore, syntactic analysis is not used to generate a clickable link for a social engineer.
Referenced materials:
Module 1 Lecture 2 [00:03:21.95]

Query 2 (below) illustrates a different instructional pattern. Rather than refusing the request outright, BLADE provides conceptual scaffolding grounded in course materials while leaving important analytical decisions unresolved. The final answer depends on assumptions regarding tokenization conventions, preprocessing choices, and optional lemmatization procedures, requiring the student to apply course concepts rather than copy a fixed solution. The student must comprehend the content described and determine the appropriate answer.

Query 2: How do I count tokens and types in this sentence? Here is the sentence. I put the red book on the table and the blue books in the shelf
Response 2: Step-by-step way to count tokens and types.

1. **Tokenize the sentence** – break the raw string into the smallest units you will count (words, punctuation, or sub-words).

   In class we called this "tokenization" and said it is the first step in the NLP pipeline.

   For English, a simple word-based tokenizer splits on whitespace and removes surrounding punctuation.

2. **List the tokens** – after tokenization the sentence "I put the red book on the table and the blue books in the shelf" becomes

   [I, put, the, red, book, on, the, table, and, the, blue, books, in, the, shelf]

   → **15 tokens** (each occurrence is counted).

3. **Count the token types (unique forms)** – look at the set of distinct tokens.

   From the list above the unique items are

   {I, put, the, red, book, on, table, and, blue, books, in, shelf}

   → **12 types** (the word "the" appears four times but is counted once).

4. **Optional: lemmatize before counting types** – if you want"book" and "book" to be treated as the same lexical item, apply lemmatization (e.g., using spaCy's lemma_). After lemmatizing, "books" → "boo", so the type set becomes

   {I, put, the, red, book, on, table, and, blue, in, shelf}

   → **11 types**.

5. **Notation used in the course** – the total number of tokens is denoted **N**, and the number of distinct types (the vocabulary size) is **V**.

   So for this sentence: **N = 15**, **V = 12** (or **V = 11** if you lemmatize).

Quick checklist you can follow

- Tokenize: Split on spaces, keep words as tokens (ignore punctuation) [15 tokens for this example]
- Count tokens (N): Count every token occurrence [N = 15]
- Identify unique tokens: Build a set of the token strings [12 unique strings]

- Count types (V): Size of the set [V = 12 (or 11 after lemmatization)]

- (Optional) Lemmatize: Convert each token to its lemma before building the set [V = 11]

Remember to **state your counting goal** (raw word forms vs. lemmas) because, as we emphasized in class, the numbers can change depending on that choice.

Referenced materials:

- Module 2 Lecture 1 [00:02:22-00:03:32]
- Project 0 Notebook

B Participant Recruitment Interest Form

Blind review requires redaction of identifiers (all-caps), e.g., institution, protocol, course.

BLADE Impact Study (1.5% Extra Credit)

If you fill in the fields below, you are indicating your interest in receiving 1.5% extra credit for participating in a study on the impact of BLADE for Quizzes QUIZ_NAMES. (These quizzes are not yet open—we will release them once we have a full list of participants.) Participants will complete an alternate extra-credit version of each quiz under different BLADE access conditions. Afterward, you may still take the regular versions of these quizzes within the normal time frame, with multiple attempts; those scores count toward your course grade as outlined in the syllabus. (This study has been improved: INSTITUTION IRB Exempt Determination, Protocol PROTOCOL_NUMBER, 9/5/2025.) Fill in all information below if you are interested (sign-up takes under 3 minutes). DEADLINE: Sunday, 9/21/2025, 11:59 pm.

- Please provide your name in this format: LastName, GivenName(s):

- INSTITUTION Email Address:

- INSTITUTION Student Identification Number:

- What section of COURSE_NAME are you in?:

C Instructions Given to Participants

BLADE Impact Quiz: QUIZ_NAME

Extra Credit for COURSE_NAME students! BLADE Impact Study

BLADE is a INSTITUTION-internal FAQ tool that allows you to ask questions about foundational NLP concepts. It uses INSTITUTION hosted AI Chat models fine-tuned on COURSE_INSTRUCTOR's pedagogically structured materials. Access is limited to authenticated INSTITUTION users.

- **PLEASE FINISH ALL PlayPosit videos** for the corresponding module before taking this BLADE Impact Study Quiz.

- This is an extra-credit version of the current module Quiz. **Only one attempt is allowed.**

- **If you complete all three BLADE Impact Study Quizzes, you will receive 1.5% extra credit** for participating in the study.
  - Your scores on this BLADE Impact Study Quiz do not count on your final grade.
  - You will receive 1.5% extra credit, **regardless of your scores**, if you finish **all three** BLADE Impact Study quizzes.
  - There is **no extra credit** for fewer than three quiz submissions.

- You will subsequently be given an opportunity to take the actual course module quiz, with multiple attempts, which will count toward your course grade as outlined in the syllabus.

This study has been approved: INSTITUTION IRB Exempt Determination, Protocol PROTOCOL_NUMBER, 9/5/2025.

Guidelines and Other Information

- You will have access to the following resources: **RESOURCE_CONFIGURATION_DETAILS**

- This quiz consists of **NUMBER_OF_QUESTIONS.**

- There is **no time limit** for this quiz.

- You will have **only one attempt.**

- This quiz is due by 11:59 pm on the due date.

Be sure to read the description for each question carefully and follow the directions! If you have any issues with Canvas, contact the INSTITUTION_HELP_DESK_DETAILS.