QUD Empowered LLM Co-Writing Tool

University of California, Santa Cruz

Ting-Yu Chou | Yousuf Golding | Kiara LaRocca | Neng Wan
tchou11 / ygolding / klarocca / newan@ucsc.edu

Abstract

Large language models are increasingly used by students as part of their writing process, but the impact on learning and comprehension remains unclear. We present a QUD-based (Questions Under Discussion) feedback system designed to support student summary writing without diminishing active reasoning and skill development. Our system leverages the QUD framework from discourse linguistics to provide targeted, question-based feedback that guides students to identify missing content in their summaries rather than generating answers directly. The system operates in two phases: a teacher-side golden QUD generation process that creates diverse summaries and extracts semantically clustered questions, and a student-side analysis pipeline that compares student summaries against these golden questions using embedding-based similarity matching with LLM-as-a-judge verification. We conducted a validation study with 100 participants across 10 articles, evaluating approximately 118 system-generated questions. Results demonstrate a statistically significant correlation ($r = 0.28$, $p = 0.002$) between AI-predicted question frequency and human importance ratings, with 80% of articles showing positive correlations. These findings suggest that our frequency-based ranking approach is feasible for identifying important questions, though per-article variation indicates room for refinement. Our work contributes a novel operationalization of QUD theory for educational NLP and provides a foundation for question-based pedagogical feedback systems.

1 Introduction

Writing remains a cornerstone of critical thinking and learning in education, yet the rapid advancement of large language models (LLMs) has fundamentally shifted how students engage with this process. While tools like Grammarly and WordTune assist with clarity and correctness, they often bypass deeper cognitive engagement by providing ready-made solutions. Our project addresses this challenge by developing a QUD-powered application that scaffolds students' thinking without offering direct answers, prompting users to reflect, revise, and reason through their writing.

LLMs are now prevalent across nearly every level of education, but their role in post-secondary learning is especially critical. College and university students are typically not learning to write from scratch; they are learning to write in discipline-specific ways, such as forming clear arguments in philosophy, synthesizing sources in literature reviews, or articulating research questions. Rather than resisting the shift toward LLM usage, we propose to embrace it in a way that prioritizes cognitive engagement and learning.

We also acknowledge that many students may not wish to use an LLM to help them write; they may believe they are capable of writing perfectly and insist on letting the LLM do the work entirely. While usage of our tool may be voluntary at this current juncture, we do foresee a situation in which younger students may be the target audience of our tool. Schools, such as primary or middle, commonly have students write essays or do work in school-monitored computer labs, where access to LLMs could be restricted. In this case, we would propose our tool, as it stands or as a completely finished product, could replace unrestricted LLM usage and would allow students to use these modern tools while still learning how to do the work themselves.

1.1 Problem Statement

The core problem that we are addressing is how uninstructed LLM usage for writing tasks often bypasses the learning benefits that writing is meant to provide. Ideally, we would like to design and implement a tool that would look more or less like an LLM that would simply not give direct answers to or do the work for the student, but this is a topic with a lot of psychological and linguistic nuance which requires more research to do.

When a student asks ChatGPT to "help me write a summary," the model typically returns a complete summary, offering little pedagogical support. If we address this by prompting the LLM to generate step-by-step instructions for writing the summary, those instructions often become too lengthy to be practically useful. Posing questions, then, is an effective way to engage the mind and stimulate active processing, and the most realistic place to start for a 10 week long project.

1.2 Research Questions

Our project addresses the following research questions:

1. **RQ1:** Can QUD-based feedback effectively identify missing content in student summaries compared to the themes present in the source article?

2. **RQ2:** Does AI-predicted question importance (based on frequency across diverse summaries) correlate with human judgments of question importance?

3. **RQ3:** How does question-based feedback compare to vanilla LLM critique in terms of user experience, understanding, and sense of ownership?

2 Background and Related Work

2.1 Writing as a Learning Activity

A substantial body of research supports the idea that writing serves as a powerful vehicle for learning (Emig, 1977), provided it engages the writer in cognitively and meta-cognitively demanding tasks (Klein and van Dijk, 2019). The Knowledge Transforming Model (Scardamalia and Bereiter, 1991) suggests that learning occurs when writers integrate content knowledge with rhetorical goals, reshaping their understanding through the writing process. Similarly, Galbraith (1992) conceptualizes writing as a mechanism for cognitive offloading that enables iterative revision and refinement of thought.

The Self-Regulation View (Nückles et al., 2020) further argues that writing leads to learning when students engage in elaboration, organization, and self-monitoring. Empirical studies and meta-analyses have generally validated these theories, although with modest and variable effect sizes (van Dijk et al., 2022). Importantly, tasks that include explicit cognitive or metacognitive prompts consistently show larger learning gains (Berthold et al., 2007).

2.2 Questions Under Discussion (QUD) Framework

Questions Under Discussion is a framework from discourse linguistics that views text as a series of implicit questions and answers (Roberts, 2012). Every segment of discourse addresses some question, whether explicitly stated or not. This framework has been used primarily in theoretical linguistics to understand discourse coherence and information structure.

The QUD framework specifies several principles for well-formed questions: language quality (the question must be grammatical and make sense), answer compatibility (the answer sentence must actually answer the question), givenness (the question should not introduce new concepts beyond anchor and context), and relevance (the reader could reasonably ask this question after seeing the preceding context).

Recent work has begun applying QUD to NLP tasks. Wu et al. (2023) introduced QUDEVAL for evaluating QUD discourse parsing, and Namuduri et al. (2025) proposed QUDsim for quantifying discourse similarities in LLM-generated text. Our work extends this line of research by modifying QUD for educational feedback.

2.3 Summarization Evaluation

Traditional summarization evaluation metrics like ROUGE (Lin, 2004) measure surface-level similarity through n-gram overlap but fail to capture whether a student genuinely understood the main themes. A student might use completely different words yet perfectly capture the key ideas, or conversely, might copy phrases verbatim without real comprehension.

The Pyramid method (Nenkova and Passonneau, 2004) addresses this by decomposing reference summaries into Semantic Content Units (SCUs), weighting them by frequency across multiple human model summaries, and scoring system summaries by how many of these SCUs they cover. Gao et al. (2019) developed automated approaches to pyramid evaluation, and more recently, Zhang et al. (2025) proposed QAPyramid, which decomposes summaries into question-answer pairs for fine-grained content selection evaluation.

Our approach shares the insight that frequency across multiple summaries indicates importance, but we apply this to questions rather than content units, enabling actionable feedback that prompts student thinking.

2.4 LLMs in Education

The rise of LLMs like ChatGPT has introduced new dynamics into how students approach writing. These tools are widely accessible and capable of producing fluent text, enabling students to outsource elements of the writing process that were traditionally sites of learning. Some students use LLMs to brainstorm arguments or generate drafts, suggesting potential for co-writing that could extend cognitive engagement

However, the concern remains that direct answer generation may bypass the cognitive work that produces learning. Since LLMs will most likely be around for the forseeable future, research must be done into how to use them in an educational setting without compromising the quality of education received. Our system is a good first step: by using questions as the medium of feedback, we can preserve the student's role in reasoning through the content.

3 System Design

3.1 Design Philosophy

The QUD Writing Advisor is built around a two-phase pipeline designed with a clear separation between expensive one-time setup and lightweight repeated analysis. This architecture makes the system practical for classroom deployment, where one article might be used by dozens of students.

Several key principles guided the system design. First, we prioritized **reusability**—the expensive process of analyzing an article and generating "golden questions" should only happen once, with those results used for all subsequent student analyses. Second, we aimed for **semantic understanding** rather than surface matching, using embeddings and similarity measures that capture meaning rather than exact words. Third, we wanted **actionable feedback**—not just telling students they missed something, but showing them exactly which questions they missed and where in the source article to look.

3.2 Phase 1: Golden QUD Generation (Teacher-Side)

The first phase generates diverse summaries of the source article and extracts questions from each. This diversity is crucial because different viewpoints emphasize different aspects of the same material.

3.2.1 Diverse Summary Generation

To create diversity, we systematically vary five categories: persona (the role of the writer, e.g., science editor, policy analyst, educator), depth (introductory vs. intermediate vs. advanced), style (explanatory vs. narrative vs. analytical), structure (single paragraph vs. multi-paragraph vs. sectioned), and length (150-250 words vs. 200-300 vs. 250-350 words). We also vary the GPT decoding parameters across four batches, using different temperature and penalty settings to encourage lexical and structural diversity.

The system initially generates 24 summaries (4 batches of 6), then filters out near-duplicates using cosine similarity on embeddings. Any summaries with similarity above 0.92 are considered duplicates and removed. From the remaining diverse summaries, we select 20 using a max-min diversity algorithm that iteratively selects summaries to maximize the minimum pairwise distance.

3.2.2 QUD Extraction Pipeline

Each of the 20 selected summaries undergoes QUD extraction through a multi-step process:

1. **Sentence Numbering:** The text is tokenized into numbered sentences using NLTK.

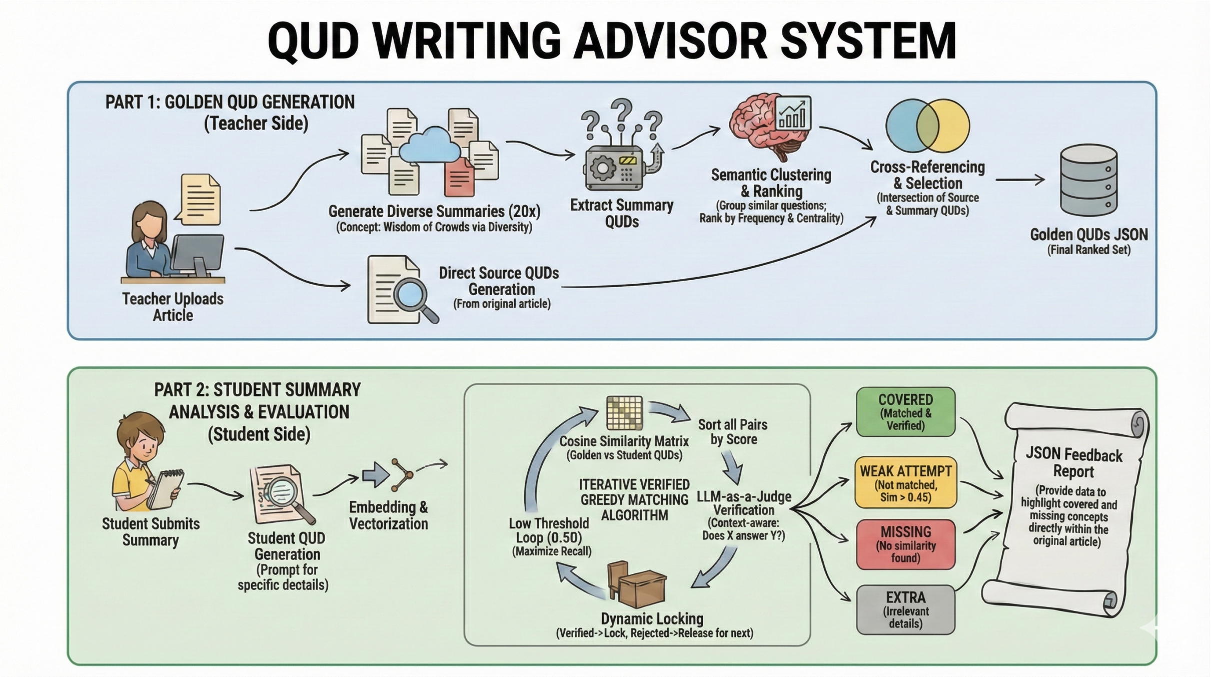
QUD WRITING ADVISOR SYSTEM
PART 1: GOLDEN QUD GENERATION (Teacher Side)
Teacher Uploads Article
Generate Diverse Summaries (20x)
(Concept: Wisdom of Crowds via Diversity)
Extract Summary QUDs
Semantic Clustering & Ranking
(Group similar questions; Rank by Frequency & Centrality)
Cross-Referencing & Selection
(Intersection of Source & Summary QUDs)
Golden QUDs JSON
(Final Ranked Set)
Direct Source QUDs Generation
(From original article)
PART 2: STUDENT SUMMARY ANALYSIS & EVALUATION (Student Side)
Student Submits Summary
Student QUD Generation
(Prompt for specific dectails)
Embedding & Vectorization
Cosine Similarity Matrix
(Golden vs Student QUDs)
Sort all Pairs by Score
ITERATIVE VERIFIED GREEDY MATCHING ALGORITHM
LLM-as-a-Judge Verification
(Context-aware: Does X answer Y?)
Low Threshold Loop (0.50)
(Maximize Recall)
Dynamic Locking
(Verified->Lock, Rejected->Release for next)
COVERED
(Matched & Verified)
WEAK ATTEMPT
(Not matched, Sim > 0.45)
MISSING
(No similarity found)
EXTRA
(Irrelevant details)
JSON Feedback Report
(Provide data to highlight covered and missing concepts directly within the original article)

Figure 1: Overview of the QUD Writing Advisor system architecture. **Part 1 (Teacher Side):** Golden QUD Generation processes the source article by generating 20 diverse summaries, extracting QUDs, performing semantic clustering and ranking, then cross-referencing with source QUDs to produce a final ranked set of golden questions. **Part 2 (Student Side):** Student Summary Analysis compares student-generated QUDs against golden QUDs using an iterative verified greedy matching algorithm with LLM-as-a-judge verification, classifying each question as Covered, Weak Attempt, Missing, or Extra.

2. **Segmentation:** A GPT model redraws paragraph boundaries so that each segment addresses a single atomic topic.

3. **Decontextualization (Optional):** Specific names, places, dates, and details are abstracted away, leaving only general conceptual content. This forces question generation to focus on high-level themes rather than specific facts.

4. **Question Generation:** Each segment is analyzed to generate its implicit questions. The prompt asks for "unique, high-level, abstract questions that capture main themes when answered."

Across 20 summaries, this process typically yields 100-200 questions total per article.

3.2.3 Semantic Clustering and Ranking

The extracted questions are embedded using OpenAI's text-embedding-3-small model and clustered using agglomerative clustering with a distance threshold of 0.4. Questions that appear frequently across multiple summaries are ranked higher, following the insight that importance correlates with how often a theme is addressed across diverse perspectives.

Each cluster is assigned a representative question (the one closest to the cluster centroid), and clusters are ranked by a composite score combining:

- **Frequency:** How many summaries generated questions in this cluster

- **Centrality:** How connected the cluster is to other question clusters

Questions scoring above a threshold (default: 1.1) are selected as "golden questions." These are then cross-referenced with questions extracted directly from the original article to ensure grounding in the source material.

3.3 Phase 2: Student Summary Analysis

When a student submits a summary for evaluation, the system extracts QUDs from their summary using the same pipeline described above. It then compares the student's questions against the golden questions using cosine similarity on embeddings.

3.3.1 Iterative Verified Greedy Matching

Rather than simple threshold-based matching, we employ an iterative verified greedy matching algorithm:

1. Compute cosine similarity matrix between all golden QUDs and student QUDs
2. Sort all pairs by similarity score (highest first)
3. For pairs above a low threshold (default: 0.5), use an LLM-as-a-judge to verify whether the student's question semantically covers the golden question
4. Lock verified matches and exclude those questions from further matching
5. Continue until no more matches above threshold remain

This approach maximizes recall by catching semantic matches that might be missed by embedding similarity alone, while the LLM verification step improves precision.

3.3.2 Feedback Classification

Each golden question is classified into one of four categories:

- **Covered:** The student's summary addresses this question (similarity ≥ 0.5 and LLM-verified)
- **Weak Attempt:** The student's summary partially addresses this (similarity 0.4-0.5)
- **Missing:** The student's summary does not address this question
- **Extra:** The student raised questions not in the golden set (potentially indicating irrelevant content)

The final output provides the list of missing questions along with source context from the original article, enabling students to locate the relevant information for revision.

4 Implementation

4.1 Technical Stack

The system is implemented in Python with the following key components:

- **Backend:** FastAPI for REST API endpoints
- **LLM Integration:** OpenAI GPT-4o for diverse summary generation with structured outputs, GPT-4o-mini for QUD extraction and analysis
- **Embeddings:** OpenAI text-embedding-3-small for semantic similarity
- **Frontend:** React with deployment on Cloudflare Pages
- **Database:** Neon PostgreSQL for study data persistence

4.2 User Interfaces

We developed multiple interfaces to support different use cases:

1. **Command-Line Interface:** For batch processing of articles and summaries
2. **REST API:** For programmatic integration with endpoints for article processing and summary analysis
3. **Google Docs Add-on:** For in-document feedback, allowing students to receive feedback directly in their writing environment
4. **Web-Based Study Platform:** Custom interface for conducting user studies with participant code management and data collection

4.3 Key Configuration Parameters

Table 1 summarizes the key configuration parameters used in the system.

Table 1: System Configuration Parameters

Parameter	**Default**	**Description**
N_BATCHES	4	Summary generation batches
TARGET_K	20	Diverse summaries to keep
distance_threshold	0.4	Clustering threshold
SCORE_THRESHOLD	1.1	Min. golden question score
similarity_threshold	0.48	Missing question threshold

4.4 Cost and Performance

The system is designed to be cost-effective for classroom deployment:

- **Article Setup (Phase 1-2):** Approximately $0.15 per article, 5-10 minutes processing time
- **Student Analysis (Phase 3):** Approximately $0.0007 per summary, 10-30 seconds processing time

Because of the longer processing time for the phase 1-2 setup, teachers would be expected to upload articles in advance rather than the system being used on demand.

5 Evaluation and Results

5.1 Evaluation Framework

Our evaluation strategy focuses on validating the core components of the system. We identified eight potential evaluations for each phase of the pipeline, organized below. Due to time constraints, only Study 1 was completed during the capstone period.

5.1.1 Phase 1: Summary and Question Generation

E1.1 Summary Diversity: How diverse are the 20 generated summaries? This could be measured through embedding-based pairwise distances or by comparing structural and lexical variation.

E1.2 Question Diversity: How diverse are the 100-200 extracted questions? Clustering metrics could assess whether questions cover distinct themes rather than redundant variations.

E1.3 Human Alignment: How similar are the generated summaries to those humans produce, and how similar are the extracted questions to those humans would ask? Following Nenkova and Passonneau (2004), human-written summaries could serve as references for comparison.

5.1.2 Phase 2: Clustering and Golden Question Selection

E2.1 Theme Capture (Study 1): Do questions ranked higher by AI frequency correlate with human importance judgments? Human raters rank questions by importance; we check correlation with system rankings. *This study was completed.*

E2.2 Clustering Quality: How well do the semantic clusters group genuinely similar questions? Humans could be given questions and asked to cluster them, then compared against system clusters.

E2.3 Source Mapping: How accurate is the mapping of questions to relevant sentences in the source article? Humans could identify which sentence(s) each question responds to, compared against system mappings.

5.1.3 Phase 3: Student Summary Analysis

E3.1 Missing Question Detection: How accurately does the system identify questions the student's summary fails to address? Humans rate "to what extent does this summary answer the following question?" and we correlate with similarity scores.

E3.2 Question Placement: How well does the system link missing questions to relevant parts of the student's essay for feedback insertion? Humans identify where questions should most naturally be placed, similar to teacher comment insertion tasks.

5.1.4 Phase 4: Feedback Effectiveness

E4.1 Human Preference Study (Study 2): Does QUD-based feedback improve summary quality compared to vanilla LLM critique? Participants revise summaries using either approach, measuring revision quality, sense of ownership, understanding of material, and ease of use. *This study was designed but not completed.*

E4.2 Longitudinal Impact: Does sustained use of the system improve writing skills over time? This would require semester-long deployment with pre/post assessment.

Of these evaluations, only E2.1 (Study 1) was completed during the project period.

5.2 Study 1: Question Extraction Validation

5.2.1 Study Design

The objective of Study 1 was to validate whether our AI frequency ranking (the question extraction phase) correlates with human judgments of question importance. If questions that appear frequently across diverse LLM-generated summaries are also rated as more important by humans, this validates our core approach.

Participants: We recruited 100 participants through Prolific, assigning 10 raters per article across 10 articles.

Materials: We selected 10 articles covering diverse topics including AI and robotics, climate science, astronomy, sustainability, and AI ethics. For each article, we generated approximately 12-15 questions using our pipeline. While we did consider other types of articles (such as short stories, fairytales, and other genres), we decided to stick with academic and scientific articles as to not introduce too many variables in our pending experiments. We would be happy to experiment with different genres in later experiments, especially as academic writing, even in post-graduate settings, covers all types of content.

Procedure: Participants read the assigned article, then rated each of approximately 15 questions on a four-point scale:

- Most Important: This question addresses a central theme of the article
- Somewhat Important: This question addresses a relevant but secondary theme
- Low Importance: This question is only tangentially related
- Not Answered: The article does not address this question

We included decoy/attention check questions to ensure data quality. We also debated different scoring procedures, but as experiments usually go, we needed to start somewhere, and we are happy to experiment with different scales in future work.

5.2.2 Metrics

Our primary metric was Pearson correlation between AI-predicted frequency rank (based on how many of the 20 summaries addressed each question's theme) and human importance ratings (percentage of raters selecting Most Important).

5.3 Study 1 Results

5.3.1 Overall Correlation

Analysis of approximately 118 questions across 10 articles revealed a statistically significant positive correlation between AI-predicted frequency and human importance ratings:

$$r = 0.28, \; p = 0.002$$

This indicates that questions appearing more frequently across diverse LLM summaries are indeed rated as more important by human judges.

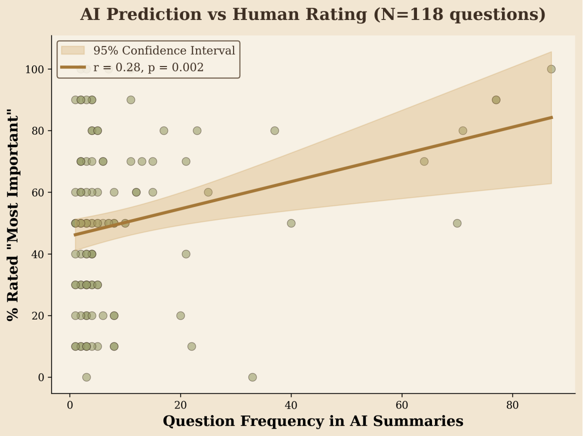

Figure 2: AI prediction vs. human rating across 118 questions showing statistically significant positive correlation ($r = 0.28$, $p = 0.002$).

5.3.2 Response Distribution

Table **??** shows the distribution of human responses across all questions.

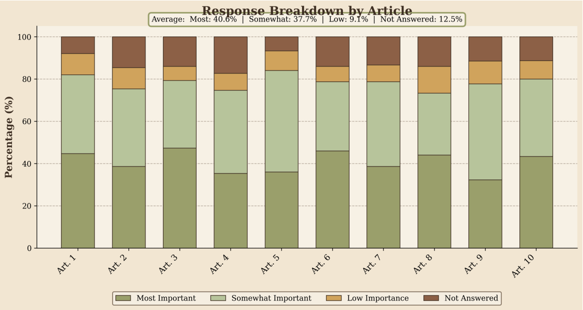

Figure 3: Response distribution by article. On average, 78.3% of questions were rated as Most Important or Somewhat Important.

Notably, 84.7% of questions were rated as either Most Important or Somewhat Important, suggesting that our question generation pipeline produces relevant questions overall. The low percentage of Not Answered responses (4.9%) indicates that generated questions are generally grounded in article content.

5.3.3 Per-Article Analysis

Per-article correlation varied considerably, ranging from -0.47 to 0.72. Importantly, 80% of articles (8 out of 10) showed positive correlations between AI frequency and human importance ratings.

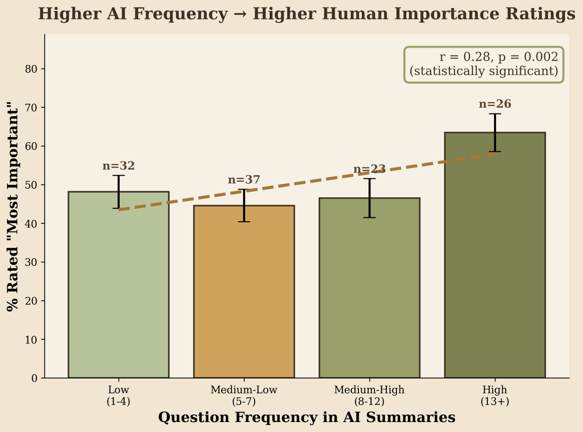

Figure 4: Questions with higher AI frequency receive higher human importance ratings on average.

The variation suggests that the method performs differently across article types, potentially due to differences in article structure, topic complexity, or the distribution of themes.

6 Discussion

6.1 Interpretation of Results

The statistically significant correlation ($r = 0.28$, $p = 0.002$) validates our core hypothesis that frequency across diverse summaries serves as a reasonable proxy for question importance. The effect size is modest but meaningful for an educational application where even small improvements in feedback quality can compound over many student interactions.

The high percentage of questions rated as important (84.7% rated Most Important or Somewhat Important) suggests that our QUD extraction and clustering pipeline produces educationally relevant questions. The low Not Answered rate (4.9%) indicates good grounding in source material.

6.2 Strengths

Our approach offers several advantages over existing methods:

1. **Preserves Student Agency:** Question-based feedback prompts students to find and articulate answers themselves, maintaining the cognitive engagement that produces learning.
2. **Diversity-Driven Coverage:** By generating 20 diverse summaries from different perspectives, we capture a broader range of important themes than any single summary would.
3. **Semantic Rather Than Surface:** Embedding-based similarity with LLM verification catches semantic coverage even when students use different words than the source.
4. **Grounded in Linguistic Theory:** The QUD framework provides a principled basis for what constitutes coverage of a theme.
5. **Practical for Deployment:** The separation of expensive article setup from cheap student analysis makes classroom use feasible.

6.3 Limitations

Several limitations constrain the current work:

1. **Moderate Correlation Strength:** While statistically significant, $r = 0.28$ indicates substantial variance not explained by our frequency ranking. Human judgments of importance are influenced by factors beyond frequency.
2. **Per-Article Variation:** Correlation ranging from -0.47 to 0.72 suggests inconsistent performance across article types. Two articles showed negative correlations, indicating the method may fail for certain content.
3. **API Dependence:** Reliance on OpenAI's API introduces cost constraints and availability concerns for large-scale deployment.
4. **Incomplete Evaluation:** Study 2 (Human Preference Study) comparing QUD feedback to vanilla LLM critique was not completed, leaving RQ3 unanswered.
5. **Task Specificity:** Our evaluation focused on summary writing; generalizability to other writing tasks (argumentative essays, research papers) is untested.

6.4 Project Evolution

It is worth noting that the project evolved significantly from its initial proposal. The original vision focused broadly on LLM co-writing support with uncertain methodology. Through iterative development and feedback, we converged on the specific QUD-based approach with the two-phase architecture described here.

This evolution reflects the exploratory nature of the broader research project of which our capstone is a part—a multi-year initiative investigating how LLMs impact learning through writing. Our work provides a foundation and validated methodology for continued research.

7 Future Work

Several directions emerge from this work:

1. **Complete Study 2:** The Human Preference Study comparing QUD-based feedback to vanilla LLM critique would directly address whether question-based feedback improves user experience, understanding, and sense of ownership.

2. **Custom Summarization Model:** Training a model specifically for diverse educational summarization could improve consistency and reduce API costs.

3. **Threshold Optimization:** Systematic study of similarity threshold settings across different article types could improve per-article performance.

4. **Longitudinal Evaluation:** Tracking students over a semester with vs. without the system would measure sustained impact on writing skills and conceptual understanding.

5. **Task Generalization:** Extending the approach to argumentative essays, research papers, and other writing tasks would test broader applicability.

6. **Question Libraries:** Investigating whether questions transfer across similar articles could enable reusable question banks for common topics.

7. **API Exploration:** Along with a custom model, we should test other models. We used OpenAI's models as we received free credit through our capstone. Exploring other LLMs for any part of our experimentation could lead us to other results.

8. **Psychological Studies:** We are already waiting for results from ongoing research projects to guide our design and evaluations, but if other studies in the same field are published, they could be significant evidence in all parts of our work that are fully experimental.

9. **Further Ablations:** We were not able to test many variables to see how significant they may have been. In a perfect world, our research would be backed up by many ablations to find the ideal value of each variable. We acknowledge here that we have a lot of work to do in general for this to be an evidence-backed project that could be a legitimate foundation for further research.

Additionally, our evaluation framework document identifies at least seven more potential evaluations beyond the two studies described, including missing question identification accuracy, question-to-text linking evaluation, and comparison with ROUGE baselines.

8 Conclusion

We presented the QUD Writing Advisor, a system that applies the Questions Under Discussion framework to provide question-based feedback on student summaries. Our approach generates diverse summaries to capture varied perspectives on article themes, extracts and clusters implicit questions, and identifies missing themes in student work through semantic similarity matching.

Study 1 validated a core assumption of our approach: questions appearing more frequently across diverse summaries are rated as more important by human judges ($r = 0.28$, $p = 0.002$). While the correlation is modest, it demonstrates that our frequency-based ranking captures meaningful signal about question importance. The high proportion of questions rated as important (84.7%) and low proportion rated as unanswered (4.9%) suggest the pipeline produces relevant, grounded questions.

Our work contributes a novel operationalization of QUD theory for educational NLP, demonstrating that linguistic frameworks can inform practical tool design. The diversity-driven generation approach and semantic clustering methodology may apply beyond summary evaluation to other tasks requiring comprehensive coverage assessment.

More broadly, this project addresses the challenge of leveraging LLMs for education without sacrificing the cognitive engagement that produces learning. By using questions rather than answers as the medium of feedback, we aim to preserve student agency while still providing meaningful guidance. The foundation established here enables continued research into how AI can support rather than supplant the learning benefits of writing. If given the time, resources, and opportunity, we hope to turn this simple tool into a thoroughly evaluated product that can be used in a variety of settings by anyone who might benefit from it.

Acknowledgments

We thank our capstone advisor Professor Pranav Anand and the UCSC Psycholinguistics Research Team for guidance throughout this project. We also thank the Prolific participants who provided the human judgments for Study 1. This work is part of a broader research initiative investigating LLMs and learning through writing.

References

Kirsten Berthold, Matthias Nückles, and Alexander Renkl. 2007. Do learning protocols support learning strategies and outcomes? The role of cognitive and metacognitive prompts. *Learning and Instruction*, 17(5):564–577.

Janet Emig. 1977. Writing as a mode of learning. *College Composition and Communication*, 28(2):122–128.

David Galbraith. 1992. Conditions for discovery through writing. *Instructional Science*, 21:45–71.

Yanjun Gao, Chen Sun, and Rebecca J. Passonneau. 2019. Automated pyramid summarization evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 404–418, Hong Kong, China. Association for Computational Linguistics.

Perry D. Klein and Aartje van Dijk. 2019. Writing as a learning activity. In John Dunlosky and Katherine A. Rawson, editors, *The Cambridge Handbook of Cognition and Education*, pages 296–320. Cambridge University Press.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ramya Namuduri, Yating Wu, Anshun Asher Zheng, Manya Wadhwa, Greg Durrett, and Junyi Jessy Li. 2025. QUDsim: Quantifying discourse similarities in LLM-generated text. In *Second Conference on Language Modeling*.

Ani Nenkova and Rebecca Passonneau. 2004. Evaluating content selection in summarization: The pyramid method. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, pages 145–152, Boston, Massachusetts, USA. Association for Computational Linguistics.

Matthias Nückles, Julian Roelle, Inga Glogger-Frey, Julia Waldeyer, and Alexander Renkl. 2020. The self-regulation-view in writing-to-learn: Using journal writing to optimize cognitive load in self-regulated learning. *Educational Psychology Review*, 32:1089–1126.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and Pragmatics*, 5(6):1–69.

Marlene Scardamalia and Carl Bereiter. 1991. Literate expertise. In K. Anders Ericsson and Jacqui Smith, editors, *Toward a general theory of expertise: Prospects and limits*, pages 172–194. Cambridge University Press.

Anje van Dijk, Amos van Gelderen, and Folkert Kuiken. 2022. Which types of instruction in writing-to-learn lead to insight and topic knowledge in different disciplines? a review of empirical studies. *Review of Education*, 10(1):e3359.

Yating Wu, Ritika Mangla, Greg Durrett, and Junyi Jessy Li. 2023. QUDeval: The evaluation of questions under discussion discourse parsing. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5344–5363, Singapore. Association for Computational Linguistics.

Shiyue Zhang, David Wan, Arie Cattan, Ayal Klein, Ido Dagan, and Mohit Bansal. 2025. QAPyramid: Fine-grained evaluation of content selection for text summarization. In *Second Conference on Language Modeling*.