UCSC

QUD Empowered LLM Co-Writing Tool

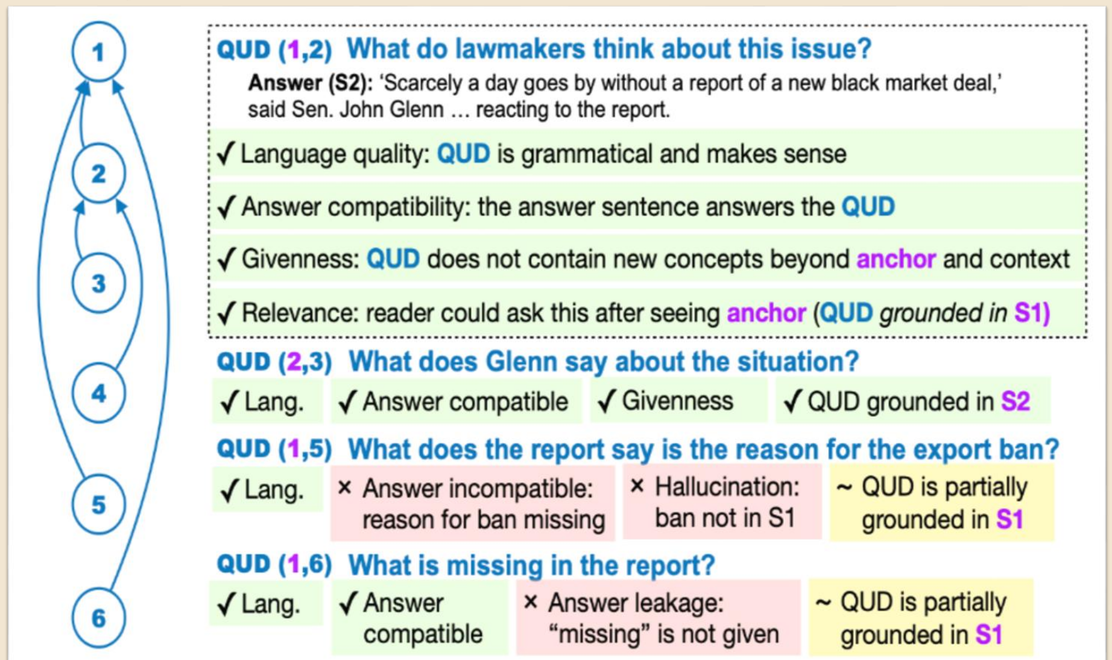
Ting-Yu Chou, Yousuf Golding, Kiara LaRocca, Neng Wan

INTRO

Many students use large language models (LLMs) as part of their writing process, but the impact on learning and comprehension remains unclear. **We aim to identify where LLMs can provide support without diminishing the learner's active reasoning and skill development.**

We leveraged **Question Under Discussion (QUD)** to provide targeted, question-based feedback that guides students to reason through their summaries to find missing content, without relying on the model to generate answers.

QUD FRAMEWORK



STUDIES

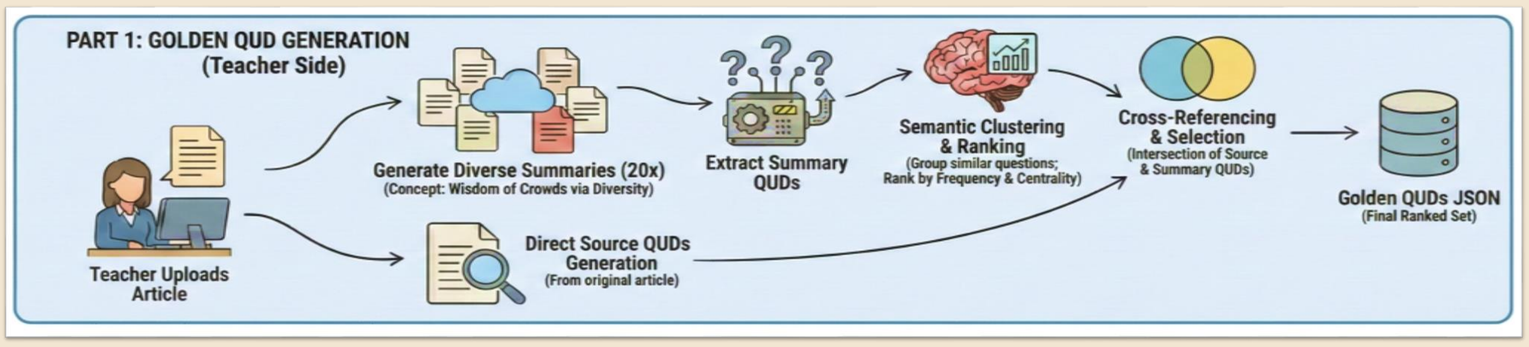
Study 1: Question Extraction Study

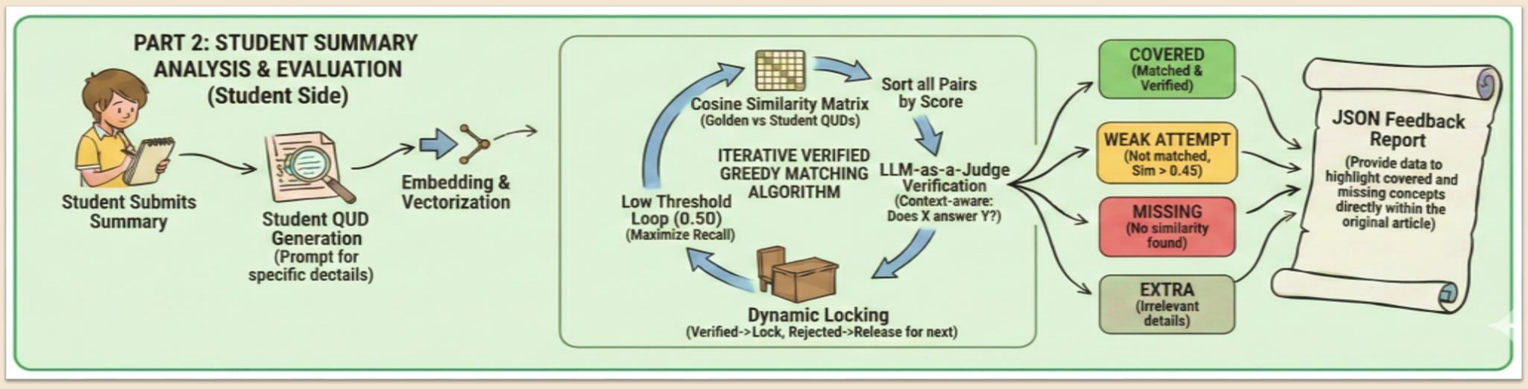
- Validates if AI frequency ranking correlates with human judgements.
- Participants rate ~15 questions per article on relevance.
- 10 raters per article across 10 articles.

Study 2: Human Preference Study

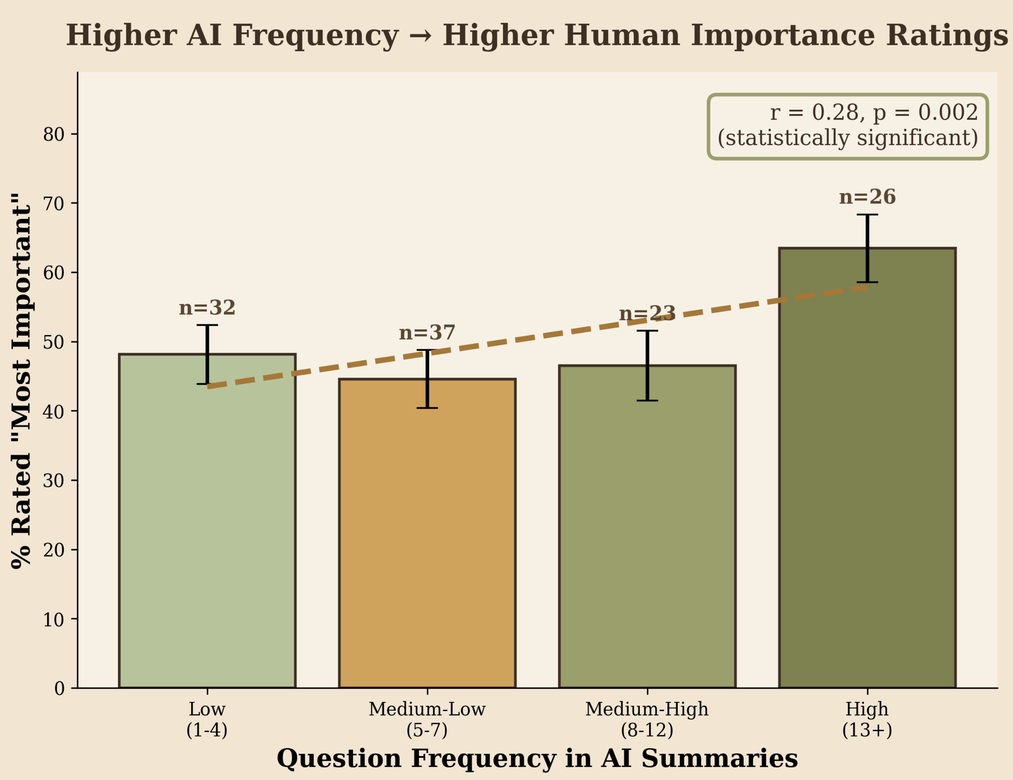
- Participants use both QUD system and Vanilla LLM feedback.
- Task: Read -> Summarize -> Receive Feedback -> Revise
- Metrics: Summary quality, user understanding, ownership and ease of use.

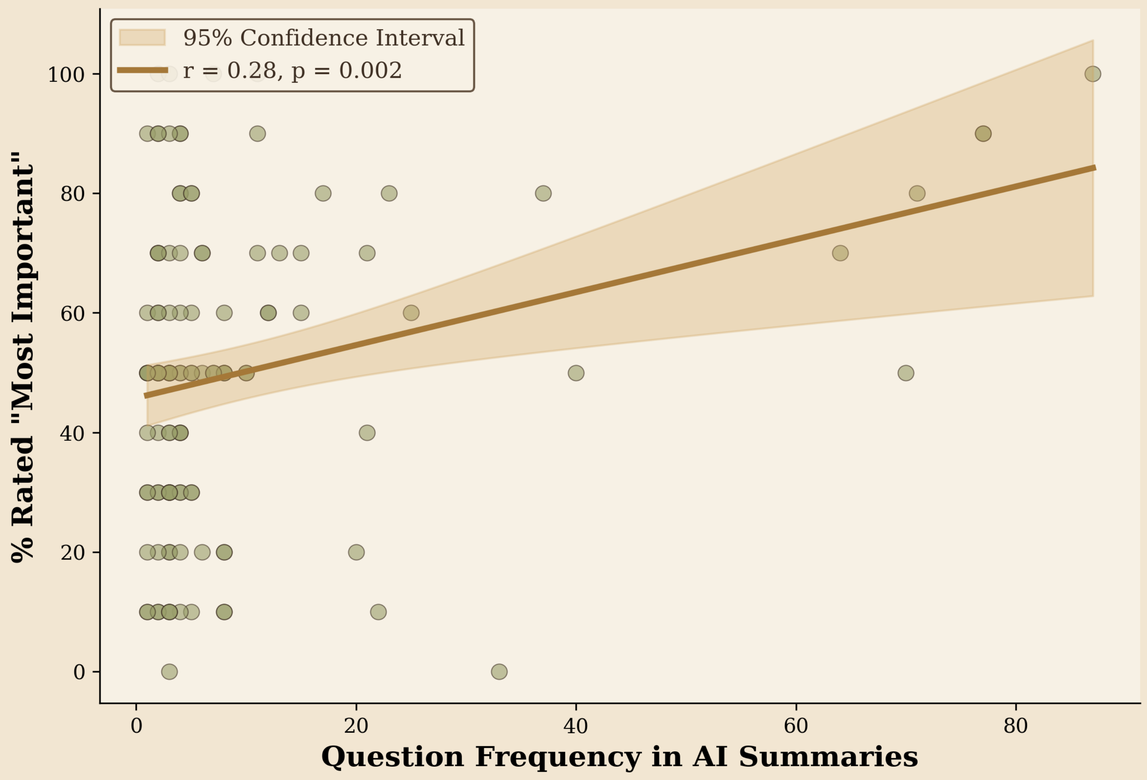
METHODS





RESULTS





CONCLUSION & ONGOING WORK

- **Analysis of ~118 questions over 10 articles showcased a statistically significant correlation between AI-predicted significance and human ratings.**
- **Per-article correlation varied (-0.47 to 0.72) but showcased positive relationships in 80% of the articles.**
- **The variation suggests the proposed method is feasible but requires refinement to better capture meaningful aspects of human judgement.**
- There are at least 7 more potential tool evaluations we can do
- We would like to train our own summarization model
- Testing the confidence threshold of the system might give us insights into comparing the summaries more accurately and efficiently

SELECTED REFERENCES
Gao, Y., Sun, C., & Passonneau, R. J. (2019). *Automated Pyramid Summarization Evaluation.* In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)* (pp. 404–418). https://aclanthology.org/K19-1038/
Heilman, M., & Smith, N. A. (2010). *Good Question! Statistical Ranking for Question Generation.* In *Proceedings of the 2010 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 609–617). https://aclanthology.org/N10-1086/
Zhang, S., Wan, D., Cattan, A., Klein, A., Dagan, I., & Bansal, M. (2024). *QAPyramid: Fine-grained Evaluation of Content Selection for Text Summarization.* arXiv:2412.07096. https://arxiv.org/abs/2412.07096