Research Statement

Mohamed Elaraby

Large language models (LLMs) are increasingly shaping our world. Between January and July 2025, Reuters reported a 125% increase in automated bot access to the internet[1]. As LLM-driven automation expands, these systems will not only retrieve massive amounts of online information but also generate new content at unprecedented scale. Recent studies indicate a desire to automate nearly half of computer-based tasks in a typical workspace Shao et al. (2025). However, LLMs remain prone to producing hallucinated Huang et al. (2025), incomplete, or unsafe content A. Wei, Haghtalab, and Steinhardt (2023). These risks highlight the urgent need for robust evaluation methods and improvement strategies to systematically measure the reliability, consistency, and safety of LLM-generated outputs. Establishing dependable evaluation frameworks is critical to ensuring a trustworthy, responsible future for LLM-powered technologies. My research directly addresses this challenge by focusing on evaluating and improving the **reliability** of LLM outputs in high-stakes natural language generation tasks. My research efforts are organized into three projects that combine summarization, computational argumentation, and explainable AI — three fields that have traditionally been studied in isolation. Through these projects, I aim to bridge these areas to advance more robust, transparent, and dependable generation outputs.

Argument-Aware Summarization #1: My first research project was among the earliest to demonstrate the benefits of integrating *Argument Mining* with *Abstractive Summarization*. We showed that incorporating argument roles — text units with specific argumentative functions — into the fine-tuning process of pretrained language models can improve summarization quality, particularly for long legal opinions Elaraby and Litman (2022). Building on this, we further explored the impact of cascading argument-aware fine-tuning with an argument-aware reranking module designed to select summaries that best overlap with the input's argumentative structure Elaraby, Zhong, and Litman (2023). This demonstrated the complementary benefits of leveraging argument signals in high-stakes domains like the legal field, both during training and inference, by enabling models to generate and rank more argument-consistent summaries. To validate our approach, we conducted a human evaluation study Elaraby, Xu, Gray, Ashley, and Litman (2024) with legal experts, who assessed the reliability and completeness of the generated summaries. This work was among the first to explicitly highlight the challenge of *summary completeness* by evaluating how well generated summaries covered salient argumentative content. Our findings confirmed the value of these techniques from an expert perspective. However, the recent rise of instruction-following LLMs has suggested that summarization might be close to a solved problem, with zero-shot summaries in some domains even preferred to human-written ones Zhang et al. (2024). In our recent preprint Elaraby and Litman (2025a)[2], we introduce a new metric, `ARC`, to analyze whether LLMs have indeed solved this problem by measuring argument coverage across three instruction-following models in both the legal opinions we previously included and expanded our analysis to the scientific documents. Our findings indicate that summarization is far from obsolete, and there remains substantial room to improve LLM-generated summaries to better match expert-selected argumentative content.

Summarizing Student Reflections #2: My initial contribution to this project explored the use of multitask learning to improve reflection summaries produced by pretrained language models Magooda, Litman, and Elaraby (2021). The success of this approach motivated us to expand the resources available for studying this problem. In particular, our subsequent work introduced a large-scale benchmark, `ReflectSumm` Zhong, Elaraby, Litman, Butt, and Menekse (2024), designed for collective summarization tasks spanning abstractive, extractive, and phrase-level summarization. This dataset is publicly available for use by the community[3] and serves as a valuable resource for advancing summarization research on student reflection data.

Evaluating Generated Self-explanations in Argumentation and Educational Settings #3: The rise of LLMs has been accompanied by remarkable emergent capabilities J. Wei et al. (n.d.). Among these is the ability to generate text-based explanations, allowing models to augment their outputs with more interpretable and user-friendly justifications. We introduced two research

[1] `https://www.reuters.com/technology/ai-intelligencer-what-matters-ai-this-week-2025-07-03`
[2] currently under submission for EMNLP 2025
[3] `https://huggingface.co/datasets/mse30/ReflectSumm`

studies that analyze and evaluate such explanations beyond traditional prediction tasks. First, we designed a novel human evaluation study Elaraby, Litman, Li, and Magooda (2024) to examine the *persuasiveness* of LLM-generated explanations in recommending support for or opposition to a given topic. This question is critical for advancing instruction-tuned LLMs in subjective decision-making scenarios, where responses cannot be strictly grounded in factual correctness. Our findings showed that an LLM's ability to generate persuasive explanations is not necessarily correlated with its ability to correctly solve the task, underscoring the need to better understand what these models reveal beyond standard evaluation metrics.

A subsequent line of work extended this investigation, building on our previously introduced `ReflectSumm` benchmark. In this study, we analyzed explanations in specificity scoring tasks Elaraby and Litman (2025b) to assess whether LLMs can produce meaningful rationales that support scoring decisions. Our analysis suggests that LLMs are capable of generating explanations that could be further explored as personalized scaffolding messages, helping students understand why they receive a particular specificity score for their reflections.

Future Research Directions: My future research aims to expand argument-aware summarization across two promising directions that connect argumentation with new summarization challenges.

(1) **Argumentative agents for multi-document perspective summarization:** This direction addresses a core question in multi-document summarization: *How can we search and plan summaries across multiple documents?* While single-document summarization has made significant progress, multi-document settings demand reasoning about how to aggregate and fuse complementary information. I argue that arguments can serve as guiding structures for summarization agents, helping pinpoint what document i reveals that document j does not, and linking these insights through argument mining. I plan to explore this framework within perspective summarization Deas and McKeown (2025), especially in domains where contrasting viewpoints, such as political discourse, need to be synthesized.

(2) **Deep argument-aware research agents for complex narrative understanding:** This direction envisions extending argument-aware summarization to interpret narratives as implicit forms of argumentation from different narrators' perspectives. Many events can be narrated from divergent viewpoints, raising the question of how to build objective summaries from these accounts. A compelling application is summarizing historical events, where historians rely on partial and subjective sources. Argument-based reasoning could help identify what can be learned across multiple, sometimes conflicting, narrations, enabling more objective and explainable interpretations of our collective past.

References

Deas, N., & McKeown, K. (2025, January). Summarization of opinionated political documents with varied perspectives. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. D. Eugenio, & S. Schockaert (Eds.), *Proceedings of the 31st international conference on computational linguistics* (pp. 8088–8108). Abu Dhabi, UAE: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2025.coling-main.539/`

Elaraby, M., & Litman, D. (2022, October). ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In N. Calzolari et al. (Eds.), *Proceedings of the 29th international conference on computational linguistics* (pp. 6187–6194). Gyeongju, Republic of Korea: International Committee on Computational Linguistics. Retrieved from `https://aclanthology.org/2022.coling-1.540/`

Elaraby, M., & Litman, D. (2025a). Arc: Argument representation and coverage analysis for zero-shot long document summarization with instruction following llms. *arXiv preprint arXiv:2505.23654*.

Elaraby, M., & Litman, D. (2025b, July–August). Lessons learned in assessing student reflections with llms. In *Proceedings of the 20th workshop on innovative use of nlp for building educational applications (bea)*. Vienna, Austria.

Elaraby, M., Litman, D., Li, X. L., & Magooda, A. (2024, November). Persuasiveness of generated free-text rationales in subjective decisions: A case study on pairwise argument ranking. In Y. Al-Onaizan, M. Bansal, & Y.-N. Chen (Eds.), *Findings of the association for computational linguistics: Emnlp 2024* (pp. 14311–14329). Miami, Florida, USA: Association for Compu-

tational Linguistics. Retrieved from https://aclanthology.org/2024.findings-emnlp.836/ doi: 10.18653/v1/2024.findings-emnlp.836

Elaraby, M., Xu, H., Gray, M., Ashley, K., & Litman, D. (2024, May). Adding argumentation into human evaluation of long document abstractive summarization: A case study on legal opinions. In S. Balloccu, A. Belz, R. Huidrom, E. Reiter, J. Sedoc, & C. Thomson (Eds.), *Proceedings of the fourth workshop on human evaluation of nlp systems (humeval) @ lrec-coling 2024* (pp. 28–35). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.humeval-1.3/

Elaraby, M., Zhong, Y., & Litman, D. (2023, July). Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In A. Rogers, J. Boyd-Graber, & N. Okazaki (Eds.), *Findings of the association for computational linguistics: Acl 2023* (pp. 7601–7612). Toronto, Canada: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2023.findings-acl.481/ doi: 10.18653/v1/2023.findings-acl.481

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., ... others (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55.

Magooda, A., Litman, D., & Elaraby, M. (2021, November). Exploring multitask learning for low-resource abstractive summarization. In M.-F. Moens, X. Huang, L. Specia, & S. W.-t. Yih (Eds.), *Findings of the association for computational linguistics: Emnlp 2021* (pp. 1652–1661). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-emnlp.142/ doi: 10.18653/v1/2021.findings-emnlp.142

Shao, Y., Zope, H., Jiang, Y., Pei, J., Nguyen, D., Brynjolfsson, E., & Yang, D. (2025). Future of work with ai agents: Auditing automation and augmentation potential across the us workforce. *arXiv preprint arXiv:2506.06576*.

Wei, A., Haghtalab, N., & Steinhardt, J. (2023). Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, *36*, 80079–80110.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., ... others (n.d.). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

Zhang, T., Ladhak, F., Durmus, E., Liang, P., McKeown, K., & Hashimoto, T. B. (2024). Benchmarking large language models for news summarization. *Transactions of the Association for Computational Linguistics*, *12*, 39–57.

Zhong, Y., Elaraby, M., Litman, D., Butt, A. A., & Menekse, M. (2024, May). ReflectSumm: A benchmark for course reflection summarization. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti, & N. Xue (Eds.), *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (lrec-coling 2024)* (pp. 13819–13846). Torino, Italia: ELRA and ICCL. Retrieved from https://aclanthology.org/2024.lrec-main.1207/