Building Reliable NLP Systems through Argumentation-Grounded Language Models

Mohamed Salem Elaraby
Department of Computer Science
University of Pittsburgh
Pittsburgh, PA 15232
`mse30@pitt.edu`

Language models have reshaped modern AI and are often celebrated as general-purpose problem solvers capable of rivaling—and, with scale, even surpassing—human experts across numerous benchmarks [2]. Yet in practice, they continue to hallucinate [10], produce unsafe content [12], and generate incomplete or biased responses [9]. My research tackles some of these challenges by **integrating computational argumentation** techniques into both the training and evaluation of language-model-based NLP systems, with the goal of improving the **faithfulness**, **accountability**, and **alignment** of model outputs with domain experts. I focus on high-stakes domains such as **law** and **education**, where nuanced reasoning and interpretability are essential, and where conventional benchmarks often fail to capture real-world complexity.

Since beginning my PhD in 2021, my research has focused on the intersection of computational argumentation and natural language generation. While traditional argument mining has emphasized NLP tasks such as stance detection, persuasion, and debate analysis, I view argumentation as a broader framework for understanding, improving, and evaluating language generation itself. I argue that argumentative elements—such as argument roles (e.g., claims, counterclaims, evidence) and their relations (e.g., support, rebuttal, attack)—provide interpretable intermediate signals that make model behavior both **auditable** and **improvable**. They also enable principled and reproducible evaluation, moving beyond aggregate scores at a time when the field is still debating how to standardize evaluation practices. **The core contribution of my research lies in showing how integrating argument roles can improve long-context generation tasks such as abstractive summarization in high-stakes domains such as law, and how these roles can serve as a foundation for evaluating LLM-generated responses.**

Toward this vision, my work has demonstrated the value of argumentation across multiple fronts. In `ArgLegalSumm` [4], I showed that incorporating argument roles improves summarization with pretrained language models in the legal domain. I also developed inference-time strategies that guide models toward more complete and meaningful summaries through argument-aware selection [6]. On the evaluation side, I introduced a human evaluation framework [8] for measuring the coverage of long-form legal summaries using argument roles as structured atomic units, revealing that common automatic metrics correlate poorly with expert judgments of structured coverage. Building on this, I proposed `ARC` [5], a general framework that uses argument roles to provide interpretable evaluations of LLM-generated summaries in high-stakes domains such as law and science.

Beyond summarization tasks, I extended argument-based evaluation to core argumentative tasks. In *argument quality ranking*, I introduced one of the first studies evaluating generated explanations in subjective decision making [7], framing rationales as arguments and assessing their *persuasiveness*. This work lays the foundation for future debate assistance tools, where LLMs must not only decide but also justify their reasoning in subjective and contested domains.

1 Core Contributions

1.1 Argumentation for More Faithful and Aligned Summaries

Much of the existing work in abstractive summarization has been benchmarked on **news articles**, with datasets such as CNN/DailyMail and XSum serving as de facto evaluation standards for over a decade [11]. In contrast, summarization in domains such as law and science remains relatively underexplored, largely due to the challenges of obtaining high-quality expert summaries and the significant cost of annotation. This gap highlights the need for **standardized techniques** that both improve summary generation and establish **reliable evaluation frameworks** for these domains. More importantly, such techniques must reflect **domain experts' reasoning**—capturing their judgments about what constitutes a good, accurate, and useful summary rather than relying solely on surface-level lexical overlap or heuristic measures.

Legal documents, particularly *court opinions*, exemplify one of the most challenging forms of high-stakes summarization. These documents are typically lengthy and complex, often extending across thousands of words, yet they contain only a small fraction of content that is truly salient (important to include in summaries). Notably, salient information—typically aligned with the document's underlying argumentative structure [13]—constitutes less than 10% of the input text while representing over 60% of expert-written summaries.

In **ArgLegalSumm** [4], I introduced a methodology that explicitly integrates argument roles into the summarization process by injecting special tokens during training to mark argumentative spans. This approach provides pretrained language models with an additional supervisory signal, helping them identify and prioritize sparse but crucial argumentative content. The work represents one of the earliest attempts to both **generate abstractive summaries for long legal opinions** and to **demonstrate the value of argument roles in improving summarization quality within high-stakes domains**.

Building on this, I explored inference-time methods that do not require retraining. In my subsequent work [6], I showed that argument-based reranking of candidate summaries generated by pretrained LMs significantly improves quality over both vanilla fine-tuning and argument-token training alone. Together, these contributions established argumentation as an effective auxiliary signal for abstractive summarization, helping bridge research in legal summarization and argument mining in high-stakes domains.

1.2 Evaluation Frameworks Grounded in Expert Judgment

A recurring challenge in trustworthy AI is the gap between automatic evaluation metrics and the judgments of domain experts. This gap is especially acute in high-stakes summarization, where expert notions of saliency are structured and domain-specific — qualities that aggregate metrics like ROUGE systematically fail to capture.

To address this, I introduced **argument roles** as a structured foundation for evaluating salient content coverage in summaries [8] for long legal summaries that reframes evaluation as measuring argument-role-specific coverage. Rather than asking annotators to judge holistic quality — a cognitively demanding and inconsistent task — this framework focuses their attention on whether summaries preserve the key argumentative components of the source document. The result is a more reproducible, expert-aligned evaluation methodology that reduces annotator burden while increasing validity.

I generalized this into **ARC (Argument Representation and Coverage)** [5], a framework for zero-shot argument-based evaluation of LLM-generated summaries in law and science. ARC quantifies how well model outputs align with expert saliency judgments by measuring role-specific coverage of key arguments. Beyond improving correlation with human evaluations, ARC exposes systematic biases in state-of-the-art instruction-tuned models' behavior — for instance, whether models disproportionately represent certain argument types over others — providing concrete, actionable insights for improving alignment with community standards of quality and fairness.

1.3 Improving Core Argumentative Tasks

Beyond summarization, I also investigate how LLMs can enhance **core argumentative tasks** that underpin reasoning and decision-making.

Improving Evidence Detection with Semi-supervised Techniques In [3], I explored how pre-trained language models can benefit from semi-supervised learning—specifically, *self-training*—to improve key argument-mining tasks such as **evidence detection**. The findings demonstrated the effectiveness of leveraging **in-domain unlabeled data** to boost detection performance, highlighting the potential of self-training as a scalable and low-cost approach for improving argument understanding in data-scarce domains.

Evaluating LLM Reasoning Transparency in Subjective Decisions A growing concern in accountable AI is whether LLMs can not only produce decisions but also generate faithful, transparent justifications for those decisions — a capability essential for human oversight. In [7], I conducted one of the first systematic evaluations of LLM-generated rationales in a subjective decision-making context: pairwise argument quality ranking. Framing model-generated rationales as arguments, I introduced both a large-scale human evaluation study and an automatic persuasiveness metric to assess how convincingly LLMs justify their judgments. Our findings reveal concrete factors that make model-generated rationales more or less persuasive, offering insights relevant to the design of AI systems that produce faithful, transparent justifications in consequential domains such as legal reasoning, policy evaluation, and content moderation.

2 Future Vision

My long-term goal is to develop **argumentation-grounded language models** that reason reliably over complex information and whose outputs can be evaluated using interpretable argumentative structure. I view argumentation not merely as a linguistic phenomenon but as a **computational substrate for reasoning**, where claims, evidence, and counterarguments provide structured signals that support both faithful generation and reliable evaluation. To advance this vision, my future research will pursue two complementary directions: (1) developing **evaluation frameworks for argumentative reasoning in language models**, and (2) building **argumentation-grounded systems for high-stakes applications**.

2.1 Direction 1: Evaluation Frameworks for Argumentative Reasoning in LLMs

A central limitation of current evaluation practices is that they primarily measure surface-level correctness or fluency rather than assessing whether models reason coherently over competing arguments. As language models increasingly operate as agents in complex information environments, we need evaluation frameworks that measure *how models construct, maintain, and update arguments across long interactions*. My research will develop benchmarks and evaluation methodologies targeting three key dimensions of argumentative reasoning.

Belief consistency across long contexts. Language models often contradict earlier statements or shift positions across extended interactions. I will construct benchmarks that measure whether models maintain coherent argumentative stances across long documents and multi-turn dialogues. These benchmarks will combine long-context inputs with structured argument annotations to identify stance drift and reasoning inconsistencies.

Opponent modeling and theory-of-mind reasoning. Effective argumentation requires anticipating and responding to the beliefs of other participants. I will design evaluation tasks that measure whether models can infer an opponent's beliefs and generate arguments that appropriately respond to them. These tasks will leverage multi-party debate corpora and controlled experimental settings in which belief states evolve over time.

Cross-lingual and cross-cultural argumentative reasoning. Existing reasoning benchmarks are largely English-centric and reflect limited argumentative norms. Building on my prior work in multilingual datasets [1], I will develop cross-lingual evaluation resources that capture diverse reasoning styles and cultural perspectives in argumentative discourse.

Together, these benchmarks will provide **diagnostic tools for identifying reasoning failure modes in language models**, enabling researchers to systematically evaluate argumentative consistency, belief updating, and reasoning transparency.

2.2 Direction 2: Argumentation-grounded AI for High-stakes Applications

Alongside evaluation research, I aim to develop **argumentation-grounded AI systems** that support reasoning-intensive tasks in high-stakes domains where decisions depend on interpreting complex and potentially conflicting information.

AI assistants for scientific reasoning and writing. Scientific research increasingly relies on automated tools for literature exploration and manuscript preparation. I will develop AI systems that critique and improve scientific manuscripts by decomposing feedback into interpretable argumentative functions: validating the evidential support of claims, assessing the novelty of contributions relative to prior work, and identifying logical inconsistencies across sections of a paper. By grounding feedback in explicit argumentative components, these systems will provide **transparent and actionable guidance** rather than vague quality scores.

Argument-aware information synthesis. Many information-seeking tasks require synthesizing competing explanations across multiple sources. I will build retrieval and summarization pipelines that explicitly model argumentative relationships between sources, enabling systems to surface supporting and opposing evidence when generating responses. Such systems will produce outputs that reflect **diverse perspectives and areas of uncertainty**, rather than collapsing toward a single dominant narrative.

Across these directions, my guiding principle remains consistent: language models should generate outputs that reflect the **argumentative structure of knowledge sources** and that can be evaluated against expert reasoning standards. By integrating structured argument representations into both evaluation and generation pipelines, this work aims to produce AI systems that are not only fluent, but also **faithful, interpretable, auditable, and improvable**.

References

[1] Muhammad Abdul-Mageed, Hassan Alhuzali, and Mohamed Elaraby. You tweet what you speak: A city-level dataset of Arabic dialects. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Koiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan, May 2018. European Language Resources Association (ELRA). URL `https://aclanthology.org/L18-1577/`.

[2] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[3] Mohamed Elaraby and Diane Litman. Self-trained pretrained language models for evidence detection. In *Proceedings of the 8th Workshop on Argument Mining*, pages 142–147, 2021.

[4] Mohamed Elaraby and Diane Litman. ArgLegalSumm: Improving abstractive summarization of legal documents with argument mining. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*, pages 6187–6194, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL `https://aclanthology.org/2022.coling-1.540/`.

[5] Mohamed Elaraby and Diane Litman. Arc: Argument representation and coverage analysis for zero-shot long document summarization with instruction-following llms. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, Rabat, Morocco, 2026.

[6] Mohamed Elaraby, Yang Zhong, and Diane Litman. Towards argument-aware abstractive summarization of long legal opinions with summary reranking. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 7601–7612, Toronto, Canada, July 2023. Association for Computational

Linguistics. doi: 10.18653/v1/2023.findings-acl.481. URL `https://aclanthology.org/2023.findings-acl.481/`.

[7] Mohamed Elaraby, Diane Litman, Xiang Lorraine Li, and Ahmed Magooda. Persuasiveness of generated free-text rationales in subjective decisions: A case study on pairwise argument ranking. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14311–14329, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.836. URL `https://aclanthology.org/2024.findings-emnlp.836/`.

[8] Mohamed Elaraby, Huihui Xu, Morgan Gray, Kevin Ashley, and Diane Litman. Adding argumentation into human evaluation of long document abstractive summarization: A case study on legal opinions. In Simone Balloccu, Anya Belz, Rudali Huidrom, Ehud Reiter, Joao Sedoc, and Craig Thomson, editors, *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @ LREC-COLING 2024*, pages 28–35, Torino, Italia, May 2024. ELRA and ICCL. URL `https://aclanthology.org/2024.humeval-1.3/`.

[9] Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179, 2024.

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025.

[11] Xiao Pu, Mingqi Gao, and Xiaojun Wan. Summarization is (almost) dead. *arXiv preprint arXiv:2309.09558*, 2023.

[12] Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. Jailbroken: How does llm safety training fail? *Advances in Neural Information Processing Systems*, 36:80079–80110, 2023.

[13] Huihui Xu, Jaromir Savelka, and Kevin D Ashley. Toward summarizing case decisions via extracting argument issues, reasons, and conclusions. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 250–254, 2021.