Generative AI for Innovative Writing and Editing Support Systems

Nicoló Donati[a,*]

[a]University of Bologna, Zanichelli editore S.p.A.
ORCID (Nicoló Donati): https://orcid.org/0009-0000-5673-5274

Abstract. Generative AI models are revolutionizing writing and editing support systems, yet their integration into real-world workflows presents open challenges. This paper outlines my doctoral research on leveraging generative AI for innovative writing and editing support, summarizing related works, research questions, contributions to date, and future directions.

1 Problem Description and Challenges

The integration of generative AI into writing and editing workflows presents a complex and multifaceted problem. While large language models (LLMs) have demonstrated impressive capabilities in generating fluent and contextually appropriate text, their deployment in real-world writing support systems raises a number of unresolved questions, both technical and human-centred.

The core of the problem is the tension between the creative potential of generative models and the need for control, reliability, and user alignment. Writing is not merely a matter of producing grammatically correct sentences; it involves purpose, structure, audience awareness, and often domain-specific constraints. For AI systems to be genuinely useful in supporting writing, they must do more than generate plausible text; they must understand and respond to the intent of the writer, provide meaningful feedback and adapt to different genres, styles and levels of expertise.

One of the key challenges lies in evaluation. Unlike tasks such as translation or classification, where performance can be measured against a clear ground truth, writing and editing involve open-ended goals and subjective judgments. Traditional automatic metrics (such as BLEU or ROUGE) are poorly suited to capturing the quality of AI-generated suggestions in writing contexts. Human evaluation, while more nuanced, is expensive, subjective, and difficult to scale. This creates a bottleneck for both system development and user trust.

Another challenge is transparency. LLMs often produce outputs that are fluent but opaque in their reasoning. Users may find it difficult to understand why a particular suggestion was made or whether it can be trusted. This is particularly problematic in high-stakes domains such as education. The lack of interpretability also limits the ability of users to learn from the system or to engage in a productive dialogue with it.

Controllability is a further concern. While LLMs can generate a wide range of outputs, guiding them to produce text that aligns with specific goals or constraints remains difficult. Prompt engineering offers some leverage, but it is often brittle and unintuitive. Fine-tuning can improve alignment, but it requires computational resources, access to large, high-quality datasets and may reduce generalisation. Striking the right balance between flexibility and control is an ongoing research problem.

Personalisation adds another layer of complexity. Effective writing support often depends on understanding the user's background, preferences, and evolving needs. Current LLMs are largely stateless and impersonal, offering the same suggestions to all users regardless of context. Incorporating user modelling into generative systems is a significant open challenge.

Finally, there are broader social and ethical concerns. The use of AI in writing raises questions about authorship, creativity, and the role of human judgment. There is a risk that over-reliance on AI tools could deskill writers and editors or reinforce existing biases in training data. Ensuring that generative systems support rather than replace human creativity requires careful design.

Overall, while generative AI holds great promise for writing and editing support, its effective integration depends on addressing a range of challenges: from evaluation and transparency to control, personalization, and ethics. These challenges are technical but also conceptual and social, requiring interdisciplinary approaches that combine insights from natural language processing, human-computer interaction and education.

Research Questions

- How can generative AI be effectively integrated into writing and editing workflows to provide support?
- What are the most effective ways to evaluate the quality and usefulness of AI-generated suggestions in the context of writing and editing?
- How can user control and transparency be balanced with the creative potential of generative models?
- What are the main limitations and risks of current generative AI systems in writing support, and how can they be mitigated?

2 Related Work

Dale and Viethen [1] note that automated writing assistance has existed for decades, traditionally focusing on spelling, grammar, and style checks, but "recent advances in deep learning" have substantially expanded its capabilities. Modern LLM-powered writing assistants go beyond error correction to support idea generation and content structuring. For example, qualitative studies report that writers (including students) often use ChatGPT as a "brainstorming" partner

[*] Corresponding Author. Email: n.donati@unibo.it.

and for organizing arguments and structure, while also relying on it for local editing (syntax, diction, grammar) [10]. In educational settings, Evmenova et al. [4] found that ChatGPT-generated feedback on student essays partially aligned with teachers' identified needs, though it failed to account for student-specific factors (such as grade level or disability). Similarly, McGuire et al. [8] studied co-creative writing interfaces and showed that users achieve greater creativity when they co-create with AI (rather than merely post-edit AI drafts) – emphasizing the importance of keeping the user in a central, ideating role. At the system level, Lee et al. [7] propose the WritingPath framework, which explicitly uses user-provided outlines to guide LLM generation so that the output stays on topic and follows the intended structure. They demonstrate that this outline-driven approach significantly improves text quality, as judged by both automated and human evaluators [7]. These works illustrate a trend toward hybrid writing tools that combine LLM-driven content generation with explicit user control (e.g., planning, outlining, revision prompts) to support coherent, goal-directed writing.

Evaluating Open-Ended Generation Evaluating freeform AI-generated text remains challenging. Traditional overlap metrics like BLEU or ROUGE were designed for tasks with clear references and tend to correlate poorly with human judgments of open-ended writing [9]. Recent surveys note that such task-agnostic scores (even semantic variants like BERTScore) often have "weak correlation with human" assessment for fluency, coherence, or creativity [9]. As a result, researchers have explored more human-aligned and multi-dimensional evaluation methods. One approach is rubric-based evaluation, where models or annotators score output on several quality dimensions. For instance, Hashemi et al. [5] introduce LLM-RUBRIC, which prompts an LLM with each rubric question and combines multiple responses to predict human ratings. In human–AI dialogue tasks, LLM-RUBRIC achieved an RMS error <0.5 in predicting user satisfaction, roughly doubling the accuracy of an uncalibrated baseline. Likewise, Ni'mah et al. [9] propose a metric "preference checklist" to assess how well automatic metrics reflect human preferences across several writing tasks.

Despite these advances, human evaluation remains the gold standard. As Hashemi et al. [5] point out, human judges typically consider multiple criteria holistically, but such manual evaluation is "expensive, time-consuming, and not without its own quality and reliability issues". More recently, some studies have tried using LLMs themselves as evaluators to reduce cost, but aligning LLM-based ratings with expert human judgments is an open problem. In practice, researchers often perform human–AI comparisons or side-by-side assessments: for example, expert writers consistently outperform LLMs on creative writing measures, and human judges report that AI-generated text tends to be repetitive or formulaic [8]. Such findings underscore that while AI text is improving, direct comparisons against human benchmarks are still essential for robust evaluation.

Alignment, Interpretability, and Feedback Challenges A key challenge is ensuring AI output aligns with user goals and remains transparent. LLMs can produce fluent text, but their suggestions may not match the writer's intent or constraints. Evmenova et al. [4] observed that ChatGPT's writing feedback did not consider student-specific needs (e.g., developmental level, learning disabilities). This highlights the difficulty of alignment: models trained on large generic corpora may overlook individual context or objectives. Efforts like WritingPath address this by encoding user intentions (desired topics, structure, keywords) into the prompt, but perfect alignment is elusive [7]. In human–AI co-writing, the issue manifests as a choice between passive editing versus active co-creation. McGuire et al. [8] show that treating the AI as a passive editor (simply revising its output) can degrade human creativity, whereas framing the user as a collaborator restores creative efficacy.

Interpretability is also limited: LLMs are effectively black boxes, making it hard to justify or control their edits. Some work adds intermediate reasoning steps or explicit planning to improve transparency. For example, "chain-of-thought" prompts or outline generation can expose reasoning paths in the text generation process [7]. In educational feedback, systems sometimes generate explanations or identify argumentation elements to make AI suggestions more inspectable (see below). Finally, in terms of feedback, studies of learners using AI point out concerns about losing one's own voice or learning opportunities. Wang [10] reports that students using ChatGPT grappled with "balancing AI to enhance writing and maintaining their authentic voice," as well as fearing that reliance on AI might short-circuit their learning. These observations reinforce the need for AI tools that not only generate text but also provide interpretable feedback and preserve user agency.

Structured Reasoning and Argumentation Frameworks Given the difficulty of evaluating free text, researchers have borrowed from structured reasoning and argumentation theory. One line of work explicitly integrates planning structures into generation. WritingPath [7] uses detailed outlines to anchor the LLM, and their results show that planning-based generation significantly improves text consistency and quality. Beyond outlines, formal argumentation models (e.g. Toulmin's scheme) are used to analyze and assess generated or student-written text. For instance, several recent systems break down an essay into argument components (claim, data, warrant, backing, qualifier, rebuttal) and use AI to identify and score each part. Jho and Ha [6] developed a web-based system that interfaces with ChatGPT to extract these six Toulmin elements from student essays, score them, and compare the AI's extraction against human annotations using NLP metrics (ROUGE, cosine similarity). Similarly, other studies employ cognitive diagnostic or rubric-based models to evaluate argumentation: Zhai et al. [12] trained classifiers on claim/evidence/warrant dimensions, and Wilson et al. [11] optimized a rubric-guided ensemble to score argumentative essays. These structured approaches allow more fine-grained judgment of reasoning quality and coherence than simple text overlap. In summary, incorporating explicit reasoning frameworks has proven a promising direction for guiding generation and for assessing complex, open-ended text.

3 Contributions to Date

My doctoral research has focused on the role of generative AI in writing and editing support systems, with particular attention to how large language models (LLMs) can be used not only to generate text but also to evaluate and structure it. This work has developed along three interconnected strands: the generation of educational content, the evaluation of open-ended text, and the exploration of argumentation as a framework for understanding and enhancing LLM reasoning.

One of the first areas I investigated was the use of LLMs to generate English grammar exercises, specifically multiple-choice cloze (MCC) items. These exercises are widely used in language learning but are traditionally time-consuming to create (requiring both linguistic expertise and pedagogical sensitivity). In collaboration with colleagues, I explored whether LLMs could be used to automate this process in a way that preserved both grammatical correctness and contextual relevance. The resulting system was based on a fine-tuned version of LLaMA 3, trained on a carefully curated dataset covering 19 grammar topics. Particular attention was paid to the generation of distractors (incorrect but plausible alternatives), which are essential for maintaining the pedagogical value of MCC items. To evaluate the quality of the generated exercises, we developed a set of rule-based metrics that assessed structural compliance and lexical diversity, and we complemented these with a human evaluation conducted by a linguist with experience in language teaching. The results, presented at CLiC-it 2024 [2], showed that the model was able to produce exercises that were not only grammatically sound but also varied and appropriate for classroom use. Approximately 79% of the generated items were judged suitable for learners, and the automatic structural compliance metric showed strong agreement with human judgments. This work suggested that LLMs can be adapted to support educational content creation in a way that balances automation with pedagogical quality.

A second line of research has examined the use of LLMs as evaluators of open-ended text, particularly in tasks such as summarization. Traditional evaluation metrics like ROUGE and BLEU are limited in their ability to capture the semantic and structural qualities of a good summary. Human evaluation, while more nuanced, is costly and difficult to scale. To address this, I contributed to a study that proposed a rubric-based evaluation framework designed to guide LLMs in assessing summaries along five editorial dimensions (coherence, consistency, fluency, relevance, and ordering). The rubric was implemented in a few-shot prompting setup and tested on a purpose-built dataset of Italian news summaries (each designed to isolate specific evaluation criteria). The results showed that LLMs could approximate human judgments to a moderate degree, particularly for more concrete criteria like relevance. However, the models also exhibited systematic biases (such as a tendency to overestimate fluency and coherence). These findings suggest that while LLMs can serve as scalable evaluators, their outputs need to be interpreted with care, and further work is needed to improve their alignment with human editorial standards. The study also highlighted the importance of rubric design and the challenges of interpreting abstract or hierarchical criteria (which are often difficult for models to apply consistently). This work was accepted at the LUHME workshop hosted at ECAI 2025.

The third field of study I worked on was argument mining as a way to study and potentially enhance the reasoning capabilities of LLMs. Argument mining involves identifying argumentative structures in text (such as claims, premises, and their relationships) and has traditionally been approached through rule-based or supervised learning methods. With the advent of LLMs, new possibilities have emerged for tackling these tasks. In this context, I have examined how LLMs perform on a range of argument mining tasks, including argument detection, classification, relation identification, and evaluation. This work has involved both a survey of existing approaches and empirical experiments using models such as BART, LLaMA 2, GPT-4, and many more. One focus has been on the comparison between different post training techniques used to improve model performance on AM tasks. Another has been on the integration of discourse structure and symbolic reasoning into LLM-based pipelines. For example, some of the systems I studied used graph-based representations of discourse to guide the generation of argumentative structures, while others incorporated formal argumentation frameworks to evaluate the strength of competing claims. These experiments suggest that LLMs can benefit from structured guidance when performing complex reasoning tasks, and that argument mining provides a useful lens for evaluating their capabilities in this area. This work will be published as a chapter in a Humane AI book.

Taken together, these three lines of work contribute to a broader understanding of how generative AI can support writing and editing. They show that LLMs can be used not only to generate text but also to evaluate and structure it in ways that are sensitive to both linguistic form and communicative function. They also point to the importance of combining generative capabilities with structured evaluation frameworks (whether in the form of rubrics, argumentation schemes, or pedagogical constraints). While many challenges remain (particularly in areas such as personalization, transparency, and domain adaptation), these initial results provide a foundation for further exploration of hybrid human-AI workflows in writing support systems.

4 Future Work

Looking ahead, I plan to explore how the reasoning capabilities of large language models can be improved through the integration of argumentative data and structured reasoning frameworks. While current LLMs are capable of producing fluent and contextually appropriate text, their reasoning processes often remain opaque and shallow. One promising direction is to train models using Direct Reasoning Optimization (DRO) [3], where each learning step is guided not only by token-level likelihoods but also by explicit argumentative structures. This would involve exposing models to rich argumentative corpora, where reasoning is made explicit through premises, inference rules, and rebuttals. By interleaving DRO with lightweight argumentation models, it may be possible to generate and vet candidate reasoning traces. This could help models learn to propose, evaluate, and refine their own chains of thought, producing explanations that are not only fluent but also auditable. While challenges such as sandbagging (where models game the evaluation criteria) remain, incorporating an external argumentative model to verify reasoning traces may help mitigate these risks.

Another area of future work concerns the evaluation of open-ended text. Current approaches often rely on pointwise scoring, which can be noisy and difficult to calibrate. I plan to investigate two complementary strategies to address this. The first is a two-stage explanation-plus-scoring pipeline. Rather than asking a single model to both identify flaws and assign a numerical score, these functions would be decoupled. In the first stage, a powerful LLM would generate a detailed natural language explanation of the flaws of the generated text based on a structured rubric (identifying which criteria are met or violated and why). In the second stage, a lightweight ML model (regression or classification network, or gradient-boosted trees) would map the explanation, along with the source and candidate text, to a final score. This approach could improve both interpretability and consistency (especially if trained on a diverse set of expert-annotated examples). The second strategy involves reframing evaluation as a series of pairwise comparisons. Instead of assigning absolute scores, the model would compare a candidate summary to a set of generated reference summaries (each representing a different quality level with respect to a given criterion) and determine which it most closely resembles. This relative judgment approach mirrors how human evaluators often reason and may reduce calibration issues. By aggregating multiple such comparisons across criteria and references, it would be possible to derive a more robust overall assessment. Of course, this approach introduces trade-offs in terms of computational cost and latency.

In the domain of grammar exercise generation, I am currently exploring the integration of linguistic level calibration into the generation process. In collaboration with Claudia Roberta Combei, we are investigating how to annotate and model learner proficiency levels so that the generator can produce exercises tailored to specific linguistic competencies. This would allow for more targeted and pedagogically informed content generation. We are also considering alternative evaluation strategies that rely on LLMs to assess the appropriateness of generated items.

Across these directions, a common theme is the need for systems that are not only generative but also reflective, able to reason about their own outputs, explain their decisions, and adapt to user needs. Achieving this will require advances in model post-training, evaluation design, and human-AI interaction. I am particularly interested in hybrid workflows that combine the strengths of LLMs with structured reasoning tools and human oversight. These workflows could support more transparent, controllable, and trustworthy writing assistance.

5 Open Questions

Despite the progress made so far, several open questions remain, some of which are fundamental and may not have straightforward solutions. One of the most pressing challenges concerns the integration of reasoning into generative models in a way that is both effective and verifiable. While approaches such as DRO offer a promising direction, it is still unclear whether models trained in this way will generalize beyond narrow benchmarks or whether they will simply learn to mimic reasoning patterns without genuine understanding. The idea of combining DRO with the argumentation frameworks is appealing, but the practicalities of aligning these symbolic structures with the internal representations of LLMs remain largely unexplored. Even if such integration is technically feasible, it is not obvious that it will lead to more trustworthy or controllable outputs in real-world writing tasks.

A particularly difficult open question concerns how to evaluate whether improvements in a model's argumentative reasoning capabilities actually translate into better writing assistance. While it is possible to design benchmarks that test a model's ability to identify claims, generate counterarguments, or follow logical structures, it is much harder to determine whether these skills lead to more helpful, trustworthy, or pedagogically effective support for writers. There is currently no standard methodology for measuring the downstream impact of argumentative reasoning on writing quality, user satisfaction, or learning outcomes. One possible approach is to conduct controlled user studies comparing writing sessions with and without access to argument-aware models (but such studies are time-consuming and difficult to generalize). Another option is to develop proxy tasks (such as explanation generation or revision suggestion) where the influence of reasoning can be more directly observed. However, even in these cases, it remains unclear how to disentangle the effects of reasoning from other factors (such as fluency or style). This is an area where I have some hypotheses and experimental ideas, but no definitive answers yet. Progress will likely require a combination of qualitative and quantitative methods, as well as collaboration with editors, writers and domain experts.

In the context of grammar exercise generation, the idea of calibrating outputs to linguistic proficiency levels is intuitively valuable, but operationalizing this remains difficult. Defining and annotating linguistic levels is itself a subjective and resource-intensive task. Moreover, it is not yet clear how well LLMs can internalize such distinctions or whether they will default to generating exercises that are fluent but pedagogically misaligned. We are currently exploring ways to evaluate this using LLM-based assessment pipelines, but these too, raise questions about reliability and bias.

More broadly, there is an open question about how to balance automation with human oversight in writing support systems. While hybrid workflows that combine LLMs with structured reasoning tools and human feedback seem promising, they also introduce complexity and potential friction. It is not yet clear how much structure is helpful before it becomes burdensome, or how to design interfaces that make reasoning processes transparent without overwhelming the user with a wall of text.

In all of these areas, I have proposed directions that I believe are worth exploring, but I do not yet know whether they will succeed. Some may turn out to be dead ends, while others may require rethinking core assumptions about how generative models should be post-trained, evaluated, and deployed. What is clear is that the path forward will require not only technical innovation but also careful empirical study and sustained interdisciplinary dialogue.

References

[1] R. Dale and J. Viethen. The automated writing assistance landscape in 2021. *Nat. Lang. Eng.*, 27(4):511–518, 2021. doi: 10.1017/S1351324921000164. URL https://doi.org/10.1017/S1351324921000164.

[2] N. Donati, M. Periani, P. Di Natale, G. Savino, and P. Torroni. Generation and evaluation of English grammar multiple-choice cloze exercises. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 325–334, Pisa, Italy, Dec. 2024. CEUR Workshop Proceedings. ISBN 979-12-210-7060-6. URL https://aclanthology.org/2024.clicit-1.39/.

[3] Y. X. et al. Direct reasoning optimization: Llms can reward and refine their own reasoning for open-ended tasks, 2025. URL https://arxiv.org/abs/2506.13351.

[4] A. Evmenova, K. Regan, R. Mergen, and R. Hrisseh. Improving writing feedback for struggling writers: Generative ai to the rescue? *TechTrends*, 68, 05 2024. doi: 10.1007/s11528-024-00965-y.

[5] H. Hashemi, J. Eisner, C. Rosset, B. Van Durme, and C. Kedzie. LLM-rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In *Proceedings of the 62nd ACL (Volume 1: Long Papers)*, pages 13806–13834, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.745. URL https://aclanthology.org/2024.acl-long.745/.

[6] H. Jho and M. Ha. Towards effective argumentation: Design and implementation of a generative ai-based evaluation and feedback system. *Journal of Baltic Science Education*, 23:280–291, 04 2024. doi: 10.33225/jbse/24.23.280.

[7] Y. Lee, S. Ka, B. Son, P. Kang, and J. Kang. Navigating the path of writing: Outline-guided text generation with large language models. In *Proceedings of the 2025 NAACL Conference (Volume 3: Industry Track)*, pages 233–250, Albuquerque, New Mexico, Apr. 2025. Association for Computational Linguistics. ISBN 979-8-89176-194-0. doi: 10.18653/v1/2025.naacl-industry.20. URL https://aclanthology.org/2025.naacl-industry.20/.

[8] J. McGuire, D. Cremer, and T. Cruys. Establishing the importance of co-creation and self-efficacy in creative collaboration with artificial intelligence. *Scientific Reports*, 14, 08 2024. doi: 10.1038/s41598-024-69423-2.

[9] I. Nimah, M. Fang, V. Menkovski, and M. Pechenizkiy. NLG evaluation metrics beyond correlation analysis: An empirical metric preference checklist. In *Proceedings of the 61st ACL (Volume 1: Long Papers)*, pages 1240–1266, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.69. URL https://aclanthology.org/2023.acl-long.69/.

[10] C. Wang. Exploring students' generative ai-assisted writing processes: Perceptions and experiences from native and nonnative english speakers. *Technology, Knowledge and Learning*, 30:1825–1846, 05 2024. doi: 10.1007/s10758-024-09744-3.

[11] C. e. a. Wilson. Using automated analysis to assess middle school students' competence with scientific argumentation. *Journal of Research in Science Teaching*, 61:38–69, 05 2023. doi: 10.1002/tea.21864.

[12] X. Zhai, K. Haudek, and W. ma. Assessing argumentation using machine learning and cognitive diagnostic modeling. *Research in Science Education*, 53, 07 2022. doi: 10.1007/s11165-022-10062-w.