

IntelliProof: An Argumentation Network-based Conversational Helper for Organized Reflection

Kaveh Eskandari Miandoab^{1*}, Katharine Kowalyshyn^{1*}, Kabir Pamnani^{2*}, Anesu Gavhera^{1*},
Vasanth Sarathy¹, Matthias Scheutz¹

¹Tufts University

²UST

kaveh.eskandari_miandoab, katharine.kowalyshyn, kabir.pamnani, anesu.gavhera, vasanth.sarathy,
matthias.scheutz{@tufts.edu}

Abstract

We present IntelliProof, an interactive system for analyzing argumentative essays through LLMs. IntelliProof structures an essay as an argumentation graph, where claims are represented as nodes, supporting evidence is attached as node properties, and edges encode supporting or attacking relations. Unlike existing automated essay scoring systems, IntelliProof emphasizes the user experience: each relation is initially classified and scored by an LLM, then visualized for enhanced understanding. The system provides justifications for classifications and produces quantitative measures for essay coherence. It enables rapid exploration of argumentative quality while retaining human oversight. In addition, IntelliProof provides a set of tools for a better understanding of an argumentative essay and its corresponding graph in natural language, bridging the gap between the structural semantics of argumentative essays and the user's understanding of a given text.

Code — github.com/collective-intelligence-lab/intelliproof
Demo — intelliproof.vercel.app

Introduction & Related Work

The rise of Large Language Models (LLMs) has drastically accelerated research in computational argumentation and automated writing support. Argumentative writing is uniquely challenging, requiring a balance of claims, supporting evidence, and counterarguments within a coherent, persuasive structure. Traditional analysis methods, from rule-based systems to neural encoders, frequently struggle to capture the nuanced interrelations between claims and evidence (Elaraby and Litman 2022).

We introduce *IntelliProof*, an LLM-powered tool that analyzes arguments by modeling them as graphs (Saveleva et al. 2021). In this model, claims are represented as nodes, with their strength quantified by evidence encoded as node properties. Weighted edges denote *support* or *attack* relations between claims. An LLM is used to score, classify, and justify

these relations, while allowing human overrides for transparency and control. The dynamic identification and visualization of these relationships are shown in Figure 1.

By transforming essays into structured argumentation graphs, IntelliProof aims to make argumentative reasoning more interpretable, providing writers and educators with insights into essay coherence and persuasiveness. This approach contributes to the discussions on how to integrate LLMs into workflows that demand interpretability, reliability, and pedagogical value simultaneously.

LLMs have shifted argument mining methods from encoder-based architectures to prompting and fine-tuning strategies (Cabessa, Hernault, and Mushtaq 2024; Favero et al. 2025). However, annotation bottlenecks and evaluation challenges remain (Schaefer 2025). Recent work also explores interactive systems that combine generative models with human input for constructing argument graphs (Lenz and Bergmann 2025). IntelliProof extends this work by integrating graph-based structuring directly into analysis while grounding scoring of arguments in quantifiable, mathematical metrics.

Educational applications increasingly use LLMs for essay scoring and feedback (Kim and Jo 2024; Chu et al. 2025). Although many approaches optimize predictive accuracy, few address the interpretability of argumentative quality. Surveys of persuasive applications highlight both the promise and ethical risks of LLM-driven reasoning systems (Rogiers et al. 2024). By grounding essay feedback in explicit argument graphs, IntelliProof contributes to more interpretable educational tools, which will lead to safer AI systems deployed in educational settings.

Intelliproof Overview

Intelliproof's functionality spans argument creation, scoring, classification, and generation techniques.

Graph Visualization IntelliProof is designed to structurally visualize argumentative essays while providing an LLM-powered (GPT-4o for the instance of the demo given its performance (Shahriar et al. 2024)) toolset for the analysis of the claims. As such, users can input claims, classify them (into Fact, Policy, or Value), and establish connections between the claims via the main GUI of the tool.

*These authors contributed equally.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

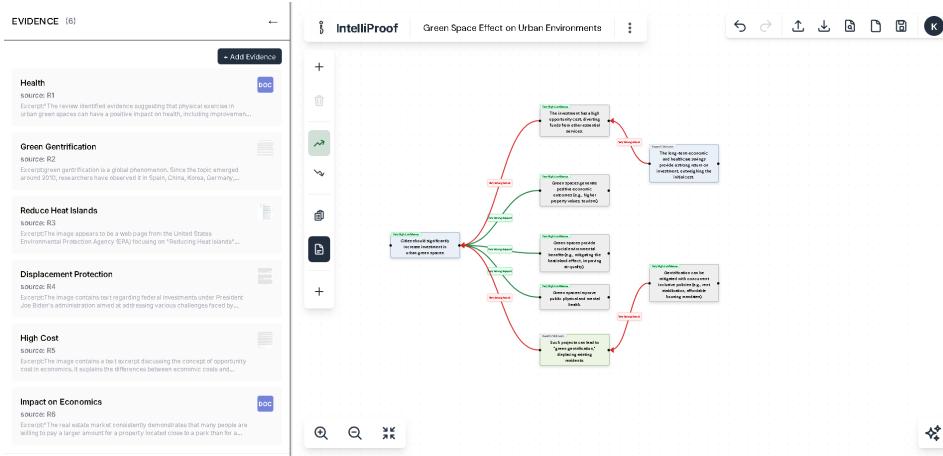


Figure 1: IntelliProof user interface overview using an example graph on the effect of green space on urban environments.

LLM Document Analysis To establish claims, users upload supporting documents as evidence in PDF or image format. A dedicated LLM instance then processes these files, suggesting relevant text or image extracts for a specific claim. The user attaches the suggested evidence to the claim via a drag-and-drop interface, which in turn prompts the LLM to assess the claim’s strength by analyzing all attached evidence. Any number of supporting or negating evidence pieces can be associated with a single claim.

Claim Credibility Score To assess overall claim strength, we combine evidence and edge scores to obtain the claim credibility score S_t where $S_t = \tanh(\Delta \frac{1}{n} \sum_{i=0}^n f_E(e_i) + \sum_{j=0}^m f_{ED}(k_j) * S_{t-1})$. f_E and f_{ED} are calculated based on the LLMs assessment of claim support based on an evidence, and based on an incoming edge, respectively, and Δ is a tunable hyperparameter. Note that given the weakness of LLMs in directly generating scores (Schroeder and Wood-Doughty 2025; Cui 2025), we first generate a qualitative classification as the LLM’s assessment, and then utilize the Evans coefficient interpretation (Evans 1996) to convert the qualitative assessment to numerical scores.

Report Generation Another feature of Intelliproof is automatic report generation from the graph implementation of an argument. These reports combine evidence evaluation, edge validation, assumptions analysis, and graph critique into a singular unified report. Our system processes graph structure, evidence quality, relationship strengths, and logical patterns simultaneously and creates an eight section, comprehensive report of the argumentative essay.

AI Copilot Chat Interface Using our integrated chatbot, natural language queries are parsed, and one may ask questions about the graph. The responding AI is context aware, and users can get insights on arguments’ strengths, weaknesses, and gaps to fill. As arguments are built, the LLM context window is also updated in real-time to contain the new information.

Assumption Generation Intelliproof analyzes claim relationships to identify three implicit assumptions that would

strengthen support between claims. It also finds hidden premises and bridges assumptions needed to make arguments more robust. Each assumption includes an importance rating and a justification for why it strengthens the relationship generated based on a few-shot learning approach (Brown et al. 2020) prepared by an argumentation field expert.

Critique Graph To identify essay weaknesses, we deploy a state-of-the-art LLM (GPT-4o) to match the overall argument against our comprehensive Argument Patterns Bank. This Bank is a built-in YAML database, developed by an argumentation expert, containing patterns for logical fallacies, good arguments, and absurd reasoning. This process allows us to specifically identify issues like circular reasoning, straw man arguments, and false causes.

System Implementation

Intelliproof’s architecture consists of three core components. The **frontend** is built with *Vite* and *React.js* to create a dynamic user interface that handles all back-end API and database calls. The **backend** uses a *Python* server with *FastAPI* for handling requests and *SupaBase* (PostgreSQL) for managing user data such as profiles, evidence files, and graphs. For the large language model, we utilize GPT-4o via OpenAI’s Python library, chosen for its balance of performance, cost, and availability. The design is modular, allowing GPT-4o to be easily substituted with other locally or remotely deployed LLMs.

Conclusion

Intelliproof is an interactive LLM platform that creates argument graphs based on provided evidence. This system is designed to be helpful in devising strong arguments, filling gaps in arguments, and utilizing an LLM to provide a detail-oriented look at an argumentative essay. While this system may be expanded further in the future, at present, we provide a robust, functional system that demonstrates the feasibility of Intelliproof as a powerful tool for structured, LLM-driven argumentation.

Acknowledgments

This research was supported in part by Other Transaction award HR00112490378 from the U.S. Defense Advanced Research Projects Agency (DARPA) Friction for Accountability in Conversational Transactions (FACT) program.

References

Brown, T. B.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; Agarwal, S.; Herbert-Voss, A.; Krueger, G.; Henighan, T.; Child, R.; Ramesh, A.; Ziegler, D. M.; Wu, J.; Winter, C.; Hesse, C.; Chen, M.; Sigler, E.; Litwin, M.; Gray, S.; Chess, B.; Clark, J.; Berner, C.; McCandlish, S.; Radford, A.; Sutskever, I.; and Amodei, D. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165.

Cabessa, J.; Hernault, H.; and Mushtaq, U. 2024. Argument Mining in BioMedicine: Zero-Shot, In-Context Learning and Fine-tuning with LLMs. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, 122–131.

Chu, S.; Kim, J. W.; Wong, B.; and Yi, M. Y. 2025. Rationale Behind Essay Scores: Enhancing S-LLM’s Multi-Trait Essay Scoring with Rationale Generated by LLMs. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Findings of the Association for Computational Linguistics: NAACL 2025*, 5796–5814. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-195-7.

Cui, H. 2025. LLMs Are Not Scorers: Rethinking MT Evaluation with Generation-Based Methods. arXiv:2505.16129.

Elaraby, M.; and Litman, D. 2022. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. In Calzolari, N.; Huang, C.-R.; Kim, H.; Pustejovsky, J.; Wanner, L.; Choi, K.-S.; Ryu, P.-M.; Chen, H.-H.; Donatelli, L.; Ji, H.; Kurohashi, S.; Paggio, P.; Xue, N.; Kim, S.; Hahm, Y.; He, Z.; Lee, T. K.; Santus, E.; Bond, F.; and Na, S.-H., eds., *Proceedings of the 29th International Conference on Computational Linguistics*, 6187–6194. Gyeongju, Republic of Korea: International Committee on Computational Linguistics.

Evans, J. D. 1996. *Straightforward statistics for the behavioral sciences*. Thomson Brooks/Cole Publishing Co.

Favero, L.; Pérez-Ortiz, J.; Käser, T.; and Oliver, N. 2025. *Leveraging Small LLMs for Argument Mining in Education: Argument Component Identification, Classification, and Assessment*.

Kim, S.; and Jo, M. 2024. Is GPT-4 Alone Sufficient for Automated Essay Scoring?: A Comparative Judgment Approach Based on Rater Cognition. In *Proceedings of the Eleventh ACM Conference on Learning @ Scale*, 315–319. ArXiv:2407.05733 [cs].

Lenz, M.; and Bergmann, R. 2025. *ArgueMapper Assistant: Interactive Argument Mining Using Generative Language Models*, volume 15446 of *Lecture Notes in Computer Science*, 189–203. Cham: Springer Nature Switzerland. ISBN 978-3-031-77914-5.

Rogiers, A.; Noels, S.; Buyl, M.; and Bie, T. D. 2024. Persuasion with Large Language Models: a Survey. (arXiv:2411.06837). ArXiv:2411.06837 [cs].

Saveleva, E.; Petukhova, V.; Mosbach, M.; and Klakow, D. 2021. Graph-based Argument Quality Assessment. In Mitkov, R.; and Angelova, G., eds., *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 1268–1280. Held Online: INCOMA Ltd.

Schaefer, R. 2025. On Integrating LLMs Into an Argument Annotation Workflow. In Chistova, E.; Cimiano, P.; Hadadian, S.; Lapesa, G.; and Ruiz-Dolz, R., eds., *Proceedings of the 12th Argument mining Workshop*, 87–99. Vienna, Austria: Association for Computational Linguistics. ISBN 979-8-89176-258-9.

Schroeder, K.; and Wood-Doughty, Z. 2025. Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge. arXiv:2412.12509.

Shahriar, S.; Lund, B.; Mannuru, N. R.; Arshad, M. A.; Hayawi, K.; Bevara, R. V. K.; Mannuru, A.; and Batool, L. 2024. Putting GPT-4o to the Sword: A Comprehensive Evaluation of Language, Vision, Speech, and Multimodal Proficiency. arXiv:2407.09519.