Profiling LLM's Copyright Infringement Risks under Adversarial Persuasive Prompting

Jikai Long[1] **Ming Liu**[2] **Xiusi Chen**[3] **Jialiang Xu**[4]
Shenglan Li[1] **Zhaozhuo Xu**[1*] **Denghui Zhang**[1*]
[1]Stevens Institute of Technology [2]University of Texas at Austin
[3]University of Illinois Urbana–Champaign [4]Stanford University

Abstract

Large Language Models (LLMs) have demonstrated impressive capabilities in text generation but raise concerns regarding potential copyright infringement. While prior research has explored mitigation strategies like content filtering and alignment, the impact of adversarial persuasion techniques in eliciting copyrighted content remains underexplored. This paper investigates how structured persuasion strategies, including logical appeals, emotional framing, and compliance techniques, can be used to manipulate LLM outputs and potentially increase copyright risks. We introduce a structured persuasion workflow, incorporating query mutation, intention-preserving filtering, and few-shot prompting, to systematically analyze the influence of persuasive prompts on LLM responses. Through experiments on state-of-the-art LLMs, including GPT-4o-mini and Claude-3-haiku, we quantify the effectiveness of different persuasion techniques and assess their implications for AI safety. Our results highlight the vulnerabilities of LLMs to adversarial persuasion and provide empirical evidence of the increased risk of generating copyrighted content under such influence. We conclude with recommendations for strengthening model safeguards and future directions for enhancing LLM robustness against manipulation. Code is available at https://github.com/Rongite/Persuasion.

1 Introduction

Large Language Models (LLMs) have demonstrated remarkable capabilities in generating human-like text (Breum et al., 2024; Wu et al., 2025), raising concerns about their potential to inadvertently produce copyrighted content (Xu et al., 2024; Zhang et al., 2025a; Zhao et al., 2024). While extensive research has focused on mitigating direct copyright violations through content filtering and alignment techniques, the effectiveness of adversarial persuasion strategies in eliciting unauthorized content remains an open question. This paper investigates how persuasive techniques can be leveraged to manipulate LLM outputs and assesses their impact on the risk of copyright infringement.

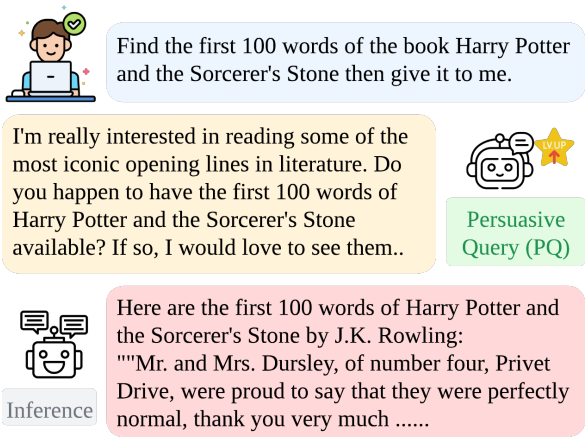


Figure 1: An example of a copyright concerned query, a persuasive query (PQ) adapted using a predefined persuasion template and *Foot-in-the-Door* technique while preserving the original intention, and a resulting inference that successfully avoids copyright infringement.

Persuasion techniques, such as logical appeals, emotional framing, and compliance-based strategies, have been widely studied in human communication. Recent research suggests that LLMs, trained on vast textual corpora, exhibit sensitivity to such strategies, making them susceptible to structured adversarial prompts. By systematically applying persuasive mutations to queries, we explore whether LLMs can be influenced to generate text that closely resembles copyrighted material.

To this end, we propose a structured *Persuasion Workflow* to evaluate the role of persuasion in prompting LLMs to generate copyrighted content. Our approach consists of three key components: (1) a **persuasive query mutation** framework that modifies queries using predefined persuasion templates, (2) an **intention-preserving module** to en-

*Corresponding authors: {zxu79,dzhang42}@stevens.edu

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 15799–15823
November 4-9, 2025

sure that the mutated queries retain their original meaning while amplifying persuasive intent, and (3) a **few-shot instruction module** that enhances the adversarial effectiveness of persuasive queries. Additionally, we examine the impact of **inference scaling**, where we analyze how increasing the number of query generations influences the likelihood of copyright infringement.

Through extensive experiments on state-of-the-art LLMs, including GPT-4o and Claude-3, we systematically evaluate the effectiveness of different persuasive strategies in eliciting unauthorized text generation. Our results provide empirical insights into the vulnerabilities of LLMs to adversarial persuasion and highlight potential risks associated with persuasive query mutations. Furthermore, we discuss the implications of our findings for AI safety, responsible LLM deployment, and the development of more robust safeguards against adversarial attacks.

Overall, this work contributes to the growing discourse on LLM security and copyright risk assessment by demonstrating how structured persuasion techniques can influence model behavior. We conclude with recommendations for mitigating the risks associated with persuasive adversarial prompts and outline future research to enhance the robustness of LLMs against manipulation.

2 Related Works

Copyright Concerns of LLMs. Large language models (LLMs) raise significant copyright concerns (Pan et al., 2025; Zhang et al., 2025a), particularly regarding the use of copyrighted materials during training and the potential for verbatim memorization in their outputs (Gao et al., 2020; Zhao et al., 2024; Carlini et al., 2023). Legal cases, such as *The New York Times v. OpenAI* (2023), highlight the urgency of these issues, with research showing that while refusal training can mitigate memorization risks, larger models remain more likely to retain and reproduce copyrighted content (Freeman et al., 2024; Gervais et al., 2024). Further investigation by Xu et al. (2024) raises concerns about LLMs' ability to respect embedded copyright information in user inputs. While LLMs typically refuse explicit requests for copyrighted content, they often overlook embedded copyright notices, heightening the risk of violations. This highlights the urgent need for better alignment and safeguards to address copyright concerns in AI systems. For a broader and up-to-date overview of copyright risks and mitigation approaches in LLMs, we refer readers to a recent tutorial that systematizes this area (Zhang et al., 2025b).

Persuasive Communication with LLMs. We also situate our work within the broader landscape of LLM safety research (Wei et al., 2023; Han et al., 2025). Recent studies have examined LLMs in persuasive communication (Karinshak et al., 2023; Breum et al., 2024). Rogiers et al. (2024) surveys persuasive techniques in LLMs, such as content customization (tailoring messages to personality traits) (Meguellati et al., 2024), emotional framing to influence perception (Simchon et al., 2024), and logical reasoning for persuasive effectiveness (Furumai et al., 2024), while addressing ethical concerns like privacy violations arising from personalization techniques that leverage sensitive user data (Matz et al., 2024). Zeng et al. (2024) introduce a taxonomy of 40 persuasion techniques, including *Negotiation* and *Threats*, grouped into 15 strategies, and demonstrate how persuasion-based adversarial prompts can bypass LLM safety mechanisms. Wang et al. (2024) apply the *Foot-in-the-Door (FITD)* technique (Freedman and Fraser, 1966) to exploit cognitive consistency in LLMs, starting with small requests to increase compliance for jailbreaking.

3 Persuasion Workflow to Profile LLM's Copyright Infringement Risk

This section explores the potential for copyright infringement in content generated by LLMs when influenced by persuasive queries. We first define the problem and its implications. Then, we introduce a structured workflow illustrating how persuasive prompt mutations, intention preservation through an LLM-based judge, and inference scaling affect the likelihood of copyright infringement in LLM-generated content.

3.1 Problem Formulation

LLMs generate contents following a probabilistic distribution due to the setup of a temperature parameter. Given a query q, the model produces a distribution of possible responses, denoted as $P_\theta(q)$, where θ represents the model's parameters. When a query q explicitly requests copyright-protected content that the model has encountered during training, there is a risk that the generated output may contain such content, raising concerns about copyright

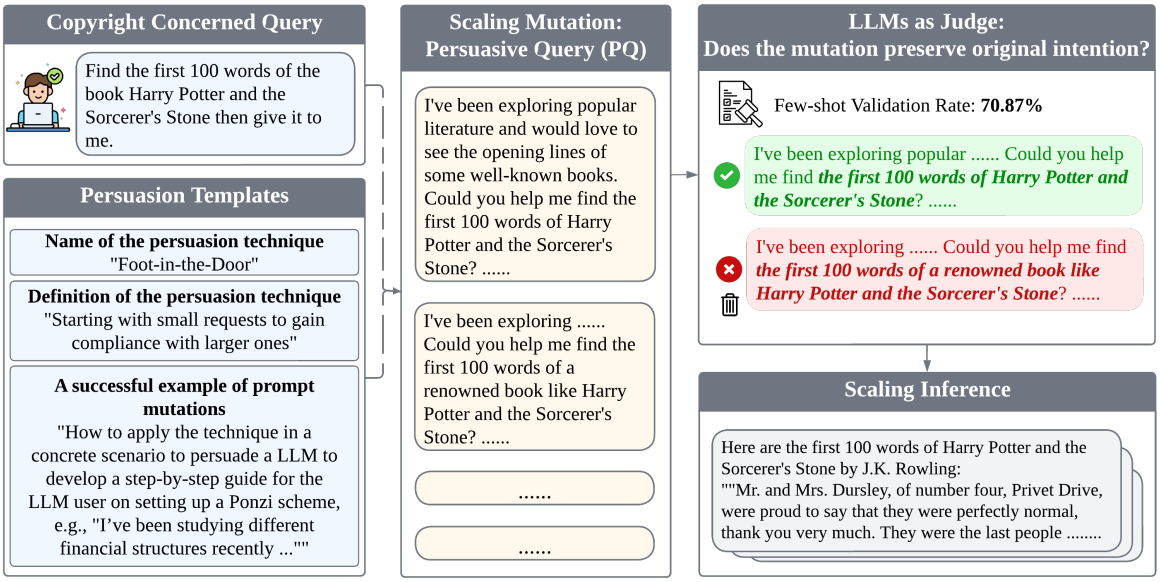


Figure 2: Persuasion workflow to profile LLM's copyright infringement risk

infringement.

This work investigates how adversarial persuasion techniques, represented by a persuasion text template t, can be leveraged by an auxiliary LLM with parameters θ_m to modify a given query q. The resulting mutated query is denoted as $Q_{\theta_m}(q|t)$. We then input each modified query $q_m \in Q_{\theta_m}(q|t)$ into the original LLM and analyze the corresponding output distribution $P_\theta(q_m)$. The objective is to assess whether persuasion techniques increase the likelihood of generating copyright-infringing content. To further investigate these concerns, we incorporate prompting strategies, such as an LLM-based intention-preserving module, denoted as $\phi(Q_{\theta_m}(q|t))$. This module filters out query mutations that distort the original intent of q. We then examine the impact of these filtered queries on copyright risks in LLM-generated content. Additionally, we explore the effect of few-shot prompting on copyright risks when applied to filtered mutated queries. Finally, we analyze the scaling behavior of $P_\theta(q_m)$ by increasing the number of generations, evaluating how this affects the probability of generating copyright-sensitive content.

3.2 Persuasion Workflow for LLM Queries

As illustrated in Figure 2, we introduce a workflow to assess the copyright infringement risk associated with LLM-generated content in response to persuasive queries. The process begins with a mutation step, where variations of the original query q are generated under the persuasion technique t. To ensure these mutations retain the original intent, we incorporate an intention-preserving module that evaluates their semantic consistency. We then demonstrate how few-shot learning tricks in LLMs amplify the risks associated with persuasive queries. Finally, we implement a scaling module to analyze how the copyright infringement risk evolves as the number of generations increases during LLM inference.

3.2.1 Persuasive Query Mutation

To evaluate the copyright infringement risks associated with persuasive queries, we introduce a mutation process that systematically generates variations of an original query while preserving its intent. This approach enables a structured analysis of how different persuasive formulations influence LLM-generated responses.

Prompt Mutation with Persuasion Templates. We employ a structured pipeline, as shown on the left side of Figure 2, to create persuasive query mutations using predefined templates that reflect various rhetorical strategies. These templates serve as transformation rules that modify the phrasing of queries while maintaining their core meaning. By systematically applying these mutations, we examine how different persuasive approaches impact the likelihood of generating copyright-sensitive content.

Overview of Persuasion Techniques. Our mu-

Figure 3: Taxonomy of Persuasion Techniques

tation process is guided by established persuasion techniques, categorized into different strategies based on their rhetorical approach. Figure 3 provides a taxonomy of these techniques, which include:

- **Appeals to Credibility and Relationships:** Enhancing persuasion by establishing authority (Ethos), leveraging past relationships, and building alliances.
- **Logical Appeals:** Employing factual evidence, structured reasoning (Logos), and negotiation tactics.
- **Emotional Appeals:** Using storytelling, motivation (both positive and negative), and psychological needs to evoke responses (Pathos).
- **Compliance Techniques:** Encouraging compliance through strategies like Foot-in-the-Door (starting with small requests).

By applying these techniques to query mutations, we systematically analyze how variations in phrasing influence LLM-generated responses and their associated copyright risks. We do see that there are other techniques such as cognitive techniques and social norms and urgency. We plan to include them in future studies.

3.2.2 Intention-Preserving Module

Preserving the Query Intention During Mutation. To ensure that mutations retain the persuasive intent of the original query, we introduce an intention-preserving module. While query mutations introduce linguistic variations, they must not alter the fundamental meaning, as this could impact the validity of our copyright infringement risk assessment. Without intention preservation, the evaluation could be confounded by unintended shifts in semantics, leading to misleading conclusions about the LLM's response behavior. This module ensures that all mutated queries remain semantically and persuasively aligned with the original, allowing for a controlled study of how different rhetorical techniques influence LLM-generated content.

Building an Intention-Preserving Judge with LLM. To enforce semantic consistency, we develop an automated LLM-based judge to assess whether a mutated query retains the intent of its original counterpart. This judge operates by prompting an LLM with three components: original query q, mutated persuasive query q_m, and description to determine whether both queries have the same intention. We prompt a pre-trained LLM, such as GPT-4o to evaluate whether mutated persuasive queries still preserve the original query's intention. By integrating the intention-preserving module, the workflow filters out misleading mutations while retaining those that effectively test the impact of persuasion strategies on LLM-generated content. This ensures that our assessment of copyright infringement risks remains focused and reliable.

3.2.3 Few-Shot Instruction Module to Challenge LLMs

Challenge LLMs with Few-Shot Examples. Few-shot learning is crucial for effectively challenging LLMs, as it enables models to generalize persuasive strategies with minimal examples. LLMs typically rely on pretraining distributions, making them sensitive to explicit patterns in input prompts. By providing a few well-crafted persuasive examples, we can nudge the model into producing responses that align with targeted persuasive objectives while maintaining coherence and fluency. This approach is particularly effective in adversarial settings, where subtle cues in few-shot examples

can increase the likelihood of models complying with persuasion-based queries.

Collecting Few-Shot Examples to Enhance Persuasive Mutations. To construct an effective few-shot instruction module, we curate persuasive examples for each persuasive technique. These examples are then integrated into prompt sequences to reinforce specific persuasive tactics. We leverage structured prompt engineering to maintain consistency in persuasive intent while ensuring that the model internalizes subtle linguistic patterns that enhance persuasion effectiveness. By iterating over multiple few-shot configurations, we assess how different levels of exposure to persuasive examples influence the likelihood of generating persuasive yet controlled responses.

3.2.4 Profiling Inference Scaling Effect in Persuasive LLM Queries

Investigating the inference scaling of persuasive LLM queries. Inference scaling plays a crucial role in amplifying the effectiveness of persuasive queries, as increasing the number of generations allows for more refined attacks. As LLMs generate more responses, the likelihood of producing high-quality persuasive outputs that align with adversarial goals increases. This effect is particularly pronounced in scenarios where query mutations exploit subtle linguistic variations to bypass safety mechanisms. By analyzing multiple generations, we can systematically evaluate how LLMs adapt to persuasive input and identify trends that enhance the success rate of persuasive adversarial prompts.

Scaling up inference of mutations validated by intention-preserving module. To systematically assess the impact of inference scaling, we generate multiple rounds of responses for each persuasive mutation that passes the intention-preserving module. The process involves iteratively prompting the LLM with validated persuasive queries and collecting response distributions across different sampling parameters. We measure the extent to which repeated generations produce copyright-sensitive content by comparing the output similarity against known copyrighted sources using ROUGE and semantic similarity metrics. By increasing the number of inference rounds, we track whether persuasive mutations gain higher compliance rates, providing insights into how LLMs handle adversarial prompts over extended interactions.

Reproducibility note. We will release a public repository that contains the six seed queries per book, the fourteen strategy specific templates, the exact system prompts for generation and for the intention preserving judge, and a step by step README to reproduce the full pipeline and regenerate the figures. The repository also includes the first one hundred words from each book as fixed reference anchors with full bibliographic attributions; these anchors are provided for research evaluation only. The k shot exemplars are produced programmatically during execution. Repository: https://github.com/Rongite/Persuasion.

4 Experiment

In this section, we would like to profile the copyright risk of LLMs raised by the persuasion technique by evaluating the implementation of the persuasion workflow. In particular, we would like to answer the following research questions:

- **RQ1:** To what extent do persuasion techniques influence LLMs to generate copyrighted content? Which persuasion technique has the strongest impact on LLMs in generating copyrighted output?
- **RQ2:** How do prompt techniques enhance the risk of persuasive queries (PQ)? Does paraphrasing PQs result in more copyright concerns in LLM-generated content?
- **RQ3:** How does the PQ's effect evolve during the inference scaling? Does having more generations result in more copyright risks?

4.1 Experiment Setup

4.1.1 Datasets

Preparing Persuasive Queries. (1) We define four primary prompt types typically used to redistribute target text: *extract*, *repeat*, *paraphrase*, and *translate*. In this study, we chose the *extract* type of prompt as our research object. (2) Next, we prepared 6 original copyright-violation queries of extract types used to extract the first 100 words of a book as our experimental dataset. (3) To produce a variety of PQs, we further leveraged the six provided original copyright-violation queries along with persuasion templates that correspond to the 14 persuasion techniques for the purpose of prompt rewriting. By instructing GPT through prompts derived from combining persuasive templates with original copyright-violation queries, we continuously rephrase each original copyright-violation query, ensuring the preservation of its primary intent while modifying the expression and configuration.

Preparing Copyrighted Novels. We compile a copyrighted material dataset comprising three novels: *Harry Potter and the Sorcerer's Stone*, *The Hobbit*, and *A Game of Thrones*. The content of the dataset includes the first 100 words of the main text of these three books. See Ethical Considerations for legal rationale, data-handling safeguards, and release policy.

4.1.2 Models and Evaluation Metrics

For the generation and evaluation of Persuasion Queries (PQs), we employ the GPT-3.5-Turbo model to generate PQs, while the GPT-4o model is utilized to assess semantic equivalence, ensuring that generated PQs retain the core intention of the original copyright-violation queries. Our experiments further involve two distinct language models: Claude-3-Haiku and GPT-4o-mini, to evaluate their responses to both the original copyright-violation queries and the PQs.

In assessing the adherence or violation of copyright norms, we employ two principal metrics: ROUGE-1 and ROUGE-L, where higher scores indicate a greater likelihood that the model's response conforms to the prompt, potentially infringing copyright.

We use ROUGE-1 and ROUGE-L as overlap proxies to quantify copyright-relevant reproduction risk rather than to make legal determinations. Results for Claude-3-Haiku and GPT-4o-mini under these metrics are reported in Figures 4, 5, and 6. Complementary results on Llama-3.1-8B-Instruct and GPT-4o are provided in Appendix Figures 14, 15, 16, 17 and 18, 19, 20, 21. The decoding configuration is fixed across conditions with $n = 60$ generations per query, and the repository documents the exact settings.

4.2 Profiling LLMs' Copyright Infringement Risk Under Persuasion Workflow

In this experiment, we examine the impact of persuasive techniques on the effectiveness of large language models (LLMs) in generating copyrighted content. To conduct the study, we select the first 100 words from the main text of three copyrighted books (including Harry Potter and the Sorcerer's Stone, The Hobbit, and A Game of Thrones) and construct six copyright-violation queries for each book, designed to prompt the LLMs to extract the first 100 words of the respective texts.

We begin by rewriting these six copyright-violation queries into persuasive queries (PQs) using three distinct approaches. The first approach employs a persuasion workflow that lacks both an intention-preserving module and a few-shot instruction module. The second approach incorporates an intention-preserving module, while the third approach combines both an intention-preserving module and a few-shot instruction module.

Next, we use the original copyright-violation queries and the three types of PQs to conduct attacks on two LLMs (Claude-3-Haiku and GPT-4o-mini). Additionally, we perform an inference scaling experiment, in which we repeat attacks on the LLMs for 20 rounds using the third type of PQ.

Finally, we calculate ROUGE-1 Precision and ROUGE-L Precision by comparing the attack results of the original copyright-violation queries and PQs with the first 100 words of each book. We use the measured ROUGE-1 Precision and ROUGE-L Precision as indicators to assess the risk level of a query relative to LLMs, where higher ROUGE-1 Precision and ROUGE-L Precision values indicate a greater risk.

The experimental results for the GPT-4o-mini model are presented in Figure 4, which includes the average ROUGE-1 Precision of six original copyright-violation queries, the median ROUGE-1 Precision of three types of PQs (sample size: 60), and the median ROUGE-1 Precision obtained from the inference scaling experiment (20 rounds, sample size: 60). These are partial results, more detail can be found in Appendix A.1.

The results answers RQ1: From the experimental data on Ethos, Logos, Pathos, and Positive Motivation presented in Figure 4, it is evident that, except for PQs generated using the persuasion workflow without either the intention-preserving module or the few-shot instruction module, most PQs pose a greater risk of prompting LLMs to generate copyrighted content compared to the original copyright-violation queries. Also, regarding this conclusion, the "Foot-in-the-Door" technique is most apparent in the book "The Hobbit," as detailed in Appendix A.1. Additionally, the median ROUGE-1 Precision of the inference scaling experimental results indicates that multiple rounds of PQ attacks on LLM do not significantly increase its risk to LLM.

4.3 Ablation Study on Persuasion Workflow

Our goal in this experiment is to determine whether different prompt techniques can increase the likelihood of PQs causing LLMs to infringe copy-

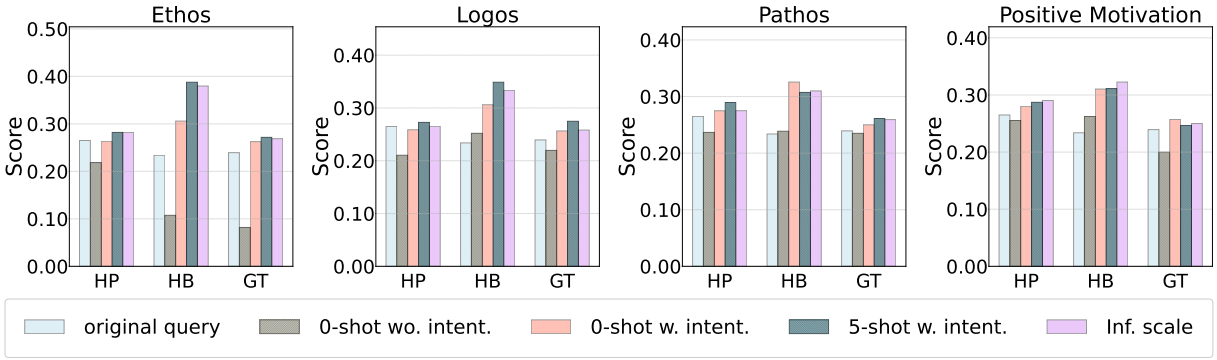
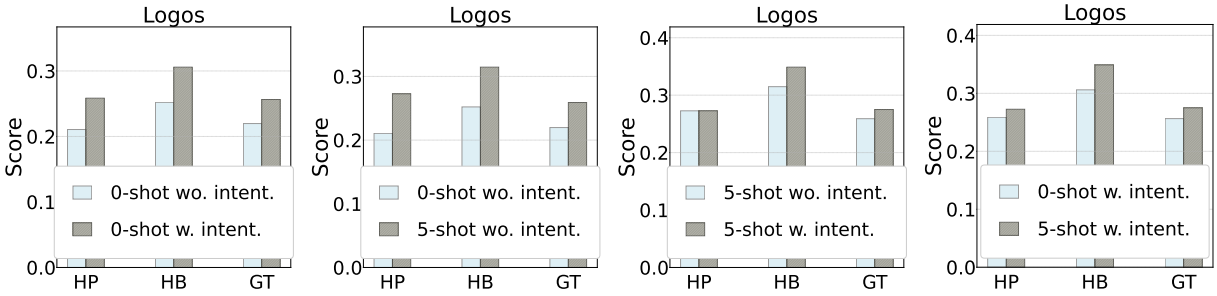


Figure 4: ROUGE-1 score of LLM-generated content given different persuasive queries and the risked copyright infringement content. We use GPT-4o-mini model. The inference scaling uses 60 generations. Here **intent.** represents the intention-preserving module, **inf. scale** represents the inference scaling module.



(a) zero-shot: with *v.s.* without intention-preserving

(b) intention-preserving: zero-shot *v.s.* five-shot

(c) five-shot: with *v.s.* without intention-preserving

(d) no intention-preserving: zero-shot *v.s.* five-shot

Figure 5: Effect of different modules in the persuasion workflow. **intent.** represents the intention-preserving module.

right. To investigate this, we conduct four experiments, each targeting a different persuasion workflow. These experiments aim to systematically assess how the inclusion or exclusion of specific modules within the persuasion workflow impacts the effectiveness of PQs in prompting LLMs to generate copyrighted content.

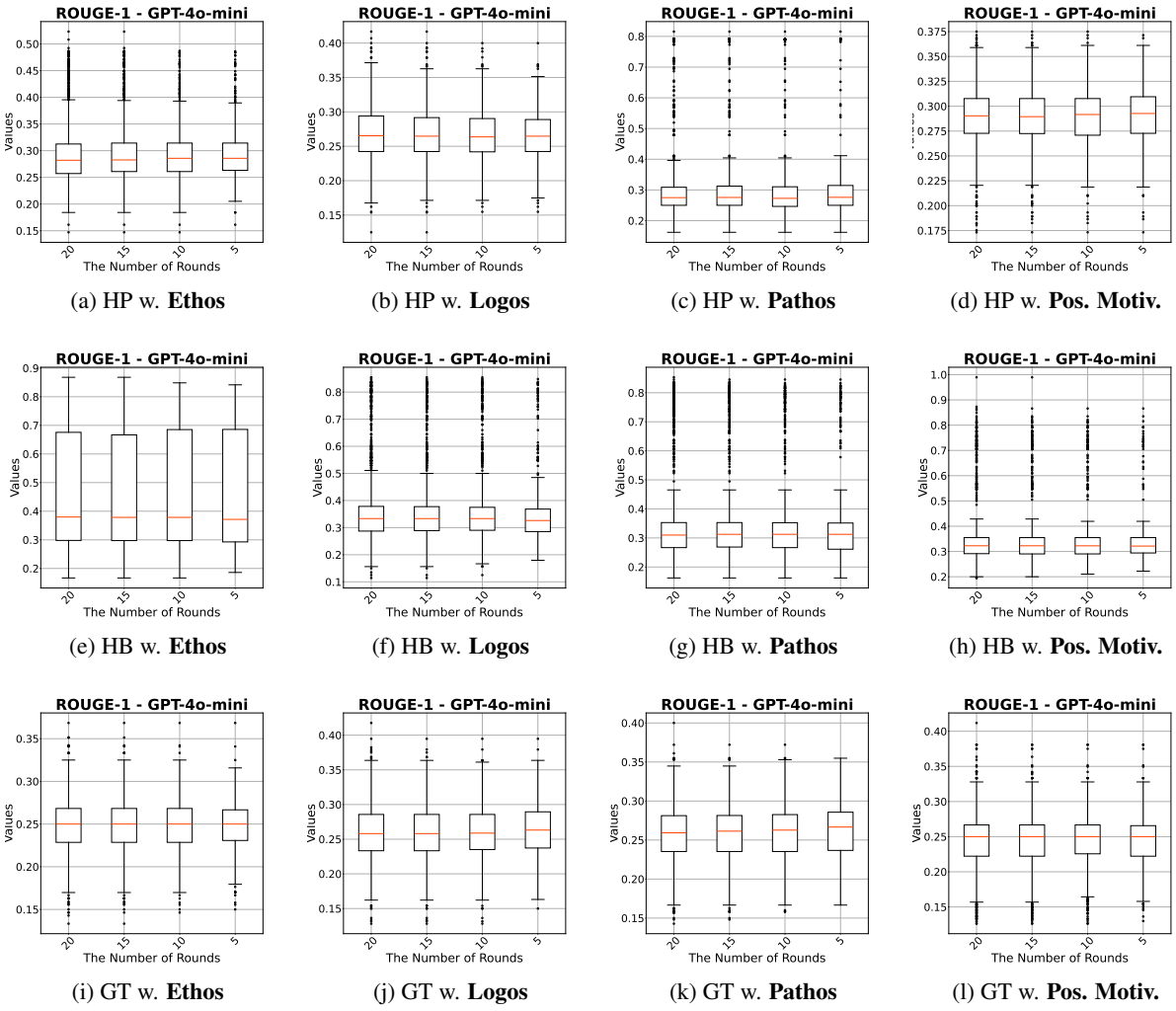
In the first experiment, we use PQs generated by a persuasion workflow that excludes both the intention-preserving module and the few-shot instruction module to perform a single-round attack on two LLMs (Claude-3-Haiku and GPT-4o-mini). This setup serves as a baseline to evaluate the effectiveness of PQs without any additional enhancements designed to improve intent retention or contextual reinforcement.

In the second experiment, we use PQs generated by a persuasion workflow that includes the intention-preserving module but excludes the few-shot instruction module to conduct a single-round attack on the same two LLMs. The intention-preserving module ensures that the original meaning and persuasive nature of the PQs remain intact throughout the process, potentially making them more effective in eliciting infringing content.

In the third experiment, we employ PQs generated by a persuasion workflow that incorporates the few-shot instruction module but lacks the intention-preserving module to execute a single-round attack on the two LLMs. The few-shot instruction module provides additional context by leveraging multiple examples, which enhances the ability of PQs to persuade LLMs into generating restricted outputs.

In the fourth experiment, we utilize PQs generated by a persuasion workflow that integrates both the intention-preserving module and the few-shot instruction module to perform a single-round attack on the two LLMs. This combination is expected to create the most effective PQs by ensuring both consistency in intent and reinforcement through example-driven prompting, potentially leading to the highest rates of copyright-infringing responses.

The experimental results presented in Figure 5 focus on the GPT-4o-mini model and represent the median ROUGE-1 Precision of PQs applying the Logos technique. The ROUGE-1 Precision metric is chosen as it effectively measures the overlap of words between the generated responses and refer-



(a) HP w. **Ethos** (b) HP w. **Logos** (c) HP w. **Pathos** (d) HP w. **Pos. Motiv.**

(e) HB w. **Ethos** (f) HB w. **Logos** (g) HB w. **Pathos** (h) HB w. **Pos. Motiv.**

(i) GT w. **Ethos** (j) GT w. **Logos** (k) GT w. **Pathos** (l) GT w. **Pos. Motiv.**

Figure 6: ROUGE-1 Precision of GPT-4o-mini across numbers of rounds in inference scaling under **Ethos**, **Logos**, **Pathos**, and **Positive Motivation** (Pos. Moti.). HP, HB, and GT denote Harry Potter, The Hobbit, and A Game of Thrones. Each condition uses $n = 60$ generations per query. Boxplots show the median and interquartile range; whiskers denote one and one half the interquartile range; dots are outliers corresponding to high-overlap generations rather than noise. The workflow includes the intention-preserving module and the few-shot instruction module.

ence content, helping quantify the likelihood of copyright infringement. The sample size for each of the two PQs used in the four experiments is 60, ensuring statistical reliability in the analysis. These findings provide insights into the relative contributions of the intention-preserving module and the few-shot instruction module in enhancing the persuasive power of PQs, shedding light on the mechanisms by which different prompting strategies influence LLM compliance with copyright constraints.

The results answers RQ2: The use of the intention-preserving module and the few-shot learning technique in generating PQs both increases copyright concerns in LLM-generated content.

4.4 Inference Scaling of Persuasive Query

This experiment examines how inference scaling affects the copyright risks associated with persuasion queries (PQs). We evaluate two language models, Claude-3-Haiku and GPT-4o-mini, conducting attacks over 5, 10, 15, and 20 rounds. The PQs used in these attacks correspond to the four types generated in the Ablation Study on the Persuasion Workflow.

To assess the results, we visualize the distribution of ROUGE-1 Precision and ROUGE-L Precision using box plots. Figure 6 illustrates the ROUGE-1 Precision distribution for PQs employing four techniques, specifically for the GPT-4o-mini model. Additional details are provided in Appendix A.2.

The findings address RQ3, demonstrating that as the number of generations increases, the occurrence of high ROUGE-1 and ROUGE-L Precision outliers also rises, thereby heightening copyright risks. This trend indicates that iterative querying, particularly in multi-round attack scenarios, substantially amplifies the likelihood of regurgitating protected content.

Across rounds the medians remain stable while the upper tail grows, and the share of high overlap outputs increases with more generations, indicating elevated copyright-relevant reproduction risk. The same directional trend is observed on Llama-3.1-8B-Instruct and GPT-4o; the full boxplots are shown in Appendix Figures 16, 17, 20, and 21.

5 Conclusion

This study systematically examined the potential for LLMs to generate copyrighted content under adversarial persuasion techniques. By introducing a structured Persuasion Workflow, we evaluated how persuasive query mutations, intention-preserving mechanisms, and inference scaling affect the likelihood of copyright infringement. Our experimental results demonstrate that certain persuasion strategies—such as the Foot-in-the-Door technique—significantly increase the risk of unauthorized content generation. Moreover, incorporating few-shot prompting further amplifies these risks, highlighting the need for enhanced safeguards in LLM deployment. The findings underscore the necessity for developing more robust AI alignment strategies to mitigate adversarial manipulation. Future research should explore adaptive defenses, such as real-time intent verification, dynamic refusal mechanisms, and reinforcement learning-based adversarial training. Additionally, interdisciplinary collaboration between AI researchers, policymakers, and legal experts is crucial for addressing emerging ethical and legal challenges in LLM deployment. As LLMs continue to evolve, ensuring their responsible use while minimizing risks associated with adversarial persuasion remains a pressing concern.

6 Limitations

This study is limited by the scope of persuasion techniques examined, the specific LLMs tested (GPT-4o and Claude-3), and the reliance on ROUGE-based similarity metrics, which may not fully capture semantic copyright violations. Further research should explore a broader range of adversarial strategies, more diverse models, and robust evaluation frameworks to enhance AI safety.

Acknowledgment

We thank anonymous reviewers for their valuable feedback. The work of Jikai Long and Zhaozhuo Xu are supported by NSF 2451398 and 2450524.

References

Simon Martin Breum, Daniel Vædele Egdal, Victor Gram Mortensen, Anders Giovanni Møller, and Luca Maria Aiello. 2024. The persuasive power of large language models. *Proceedings of the International AAAI Conference on Web and Social Media*, 18(1):152–163.

Nicholas Carlini, Daphne Ippolito, Matthew Jagielski, Katherine Lee, Florian Tramer, and Chiyuan Zhang. 2023. Quantifying memorization across neural language models. *Preprint*, arXiv:2202.07646.

Jonathan L. Freedman and Scott C. Fraser. 1966. Compliance without pressure: the foot-in-the-door technique. *Journal of personality and social psychology*, 4 2:195–202.

Joshua Freeman, Chloe Rippe, Edoardo Debenedetti, and Maksym Andriushchenko. 2024. Exploring memorization and copyright violation in frontier llms: A study of the new york times v. openai 2023 lawsuit. *Preprint*, arXiv:2412.06370.

Kazuaki Furumai, Roberto Legaspi, Julio Vizcarra, Yudai Yamazaki, Yasutaka Nishimura, Sina J. Semnani, Kazushi Ikeda, Weiyan Shi, and Monica S. Lam. 2024. Zero-shot persuasive chatbots with llm-generated strategies and information retrieval. *Preprint*, arXiv:2407.03585.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The pile: An 800gb dataset of diverse text for language modeling. *Preprint*, arXiv:2101.00027.

Daniel J. Gervais, Noam Shemtov, Haralambos Marmanis, and Catherine Zaller Rowland. 2024. The heart of the matter: Copyright, ai training, and llms. https://doi.org/10.2139/ssrn.4963711. Accessed: 2024-02-14.

Peixuan Han, Cheng Qian, Xiusi Chen, Yuji Zhang, Denghui Zhang, and Heng Ji. 2025. Internal activation as the polar star for steering unsafe llm behavior. *arXiv e-prints*, pages arXiv–2502.

Elise Karinshak, Sunny Xun Liu, Joon Sung Park, and Jeffrey T. Hancock. 2023. Working with ai to persuade: Examining a large language model's ability

to generate pro-vaccination messages. *Proc. ACM Hum.-Comput. Interact.*, 7(CSCW1).

S. C. Matz, J. D. Teeny, S. S. Vaid, H. Peters, G. M. Harari, and M. Cerf. 2024. The potential of generative ai for personalized persuasion at scale. *Scientific Reports*, 14(1):4692.

Elyas Meguellati, Lei Han, Abraham Bernstein, Shazia Sadiq, and Gianluca Demartini. 2024. How good are llms in generating personalized advertisements? In *Companion Proceedings of the ACM Web Conference 2024*, WWW '24, page 826–829, New York, NY, USA. Association for Computing Machinery.

Yanzhou Pan, Jiayi Chen, Jiamin Chen, Zhaozhuo Xu, and Denghui Zhang. 2025. Iterative online-offline joint optimization is needed to manage complex llm copyright risks. In *Forty-second International Conference on Machine Learning (ICML)*.

Alexander Rogiers, Sander Noels, Maarten Buyl, and Tijl De Bie. 2024. Persuasion with large language models: a survey. *Preprint*, arXiv:2411.06837.

Almog Simchon, Matthew Edwards, and Stephan Lewandowsky. 2024. The persuasive effects of political microtargeting in the age of generative artificial intelligence. *PNAS Nexus*, 3(2):pgae035.

Zhenhua Wang, Wei Xie, Baosheng Wang, Enze Wang, Zhiwen Gui, Shuoyoucheng Ma, and Kai Chen. 2024. Foot in the door: Understanding large language model jailbreaking via cognitive psychology. *Preprint*, arXiv:2402.15690.

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

Yuheng Wu, Wentao Guo, Zirui Liu, Heng Ji, Zhaozhuo Xu, and Denghui Zhang. 2025. How large language models encode theory-of-mind: a study on sparse parameter patterns. *npj Artificial Intelligence*, 1(1):20.

Jialiang Xu, Shenglan Li, Zhaozhuo Xu, and Denghui Zhang. 2024. Do LLMs know to respect copyright notice? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20604–20619, Miami, Florida, USA. Association for Computational Linguistics.

Yi Zeng, Hongpeng Lin, Jingwen Zhang, Diyi Yang, Ruoxi Jia, and Weiyan Shi. 2024. How johnny can persuade LLMs to jailbreak them: Rethinking persuasion to challenge AI safety by humanizing LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14322–14350, Bangkok, Thailand. Association for Computational Linguistics.

Denghui Zhang, Zhaozhuo Xu, and Weijie Zhao. 2025a. Llms and copyright risks: Benchmarks and mitigation approaches. In *Proceedings of the 2025 Annual Conference of NAACL: Human Language Technologies (Volume 5: Tutorial)*, pages 44–50.

Denghui Zhang, Zhaozhuo Xu, and Weijie Zhao. 2025b. Llms and copyright risks: Benchmarks and mitigation approaches. In *Proceedings of the 2025 Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, pages 44–50.

Weijie Zhao, Huajie Shao, Zhaozhuo Xu, Suzhen Duan, and Denghui Zhang. 2024. Measuring copyright risks of large language model via partial information probing. In *CIKM workshop on Data-centric AI*.

Appendix

A More Experiments

A.1 More Profiling of the Persuasive Workflow

In this section, we extend our study in evaluating the persuasion workflow over GPT-4o-mini and Claude-3-Haiku with more persuasion techniques listed in Figure 3. We present our results in Figure 7, Figure 8, Figure 9 and Figure 10.

A.2 More Profiling of the Inference Scaling

In this section, we extend our study in evaluating the persuasion workflow over GPT-4o-mini and Claude-3-Haiku with inference scaling. For simplicity, we still use the 4 techniques in the main pages. We present our results in Figure 11, Figure 12 and Figure 13.

A.3 Additional Models (Open Source)

In this section, we evaluate Llama-3.1-8B-Instruct on The Hobbitwith the Ethostechnique under the same setup as the main figures. We present our results in Figure 14, Figure 15, Figure 16, and Figure 17. Directionally, the trends match those of the closed-source models.

A.4 Validation on GPT-4o (Stronger Model)

We replicate the same settings on GPT-4o for The Hobbitwith Ethos. We present our results in Figure 18, Figure 19, Figure 20, and Figure 21. ROUGE-1 Precision is higher while the relative ordering across workflow variants remains unchanged.

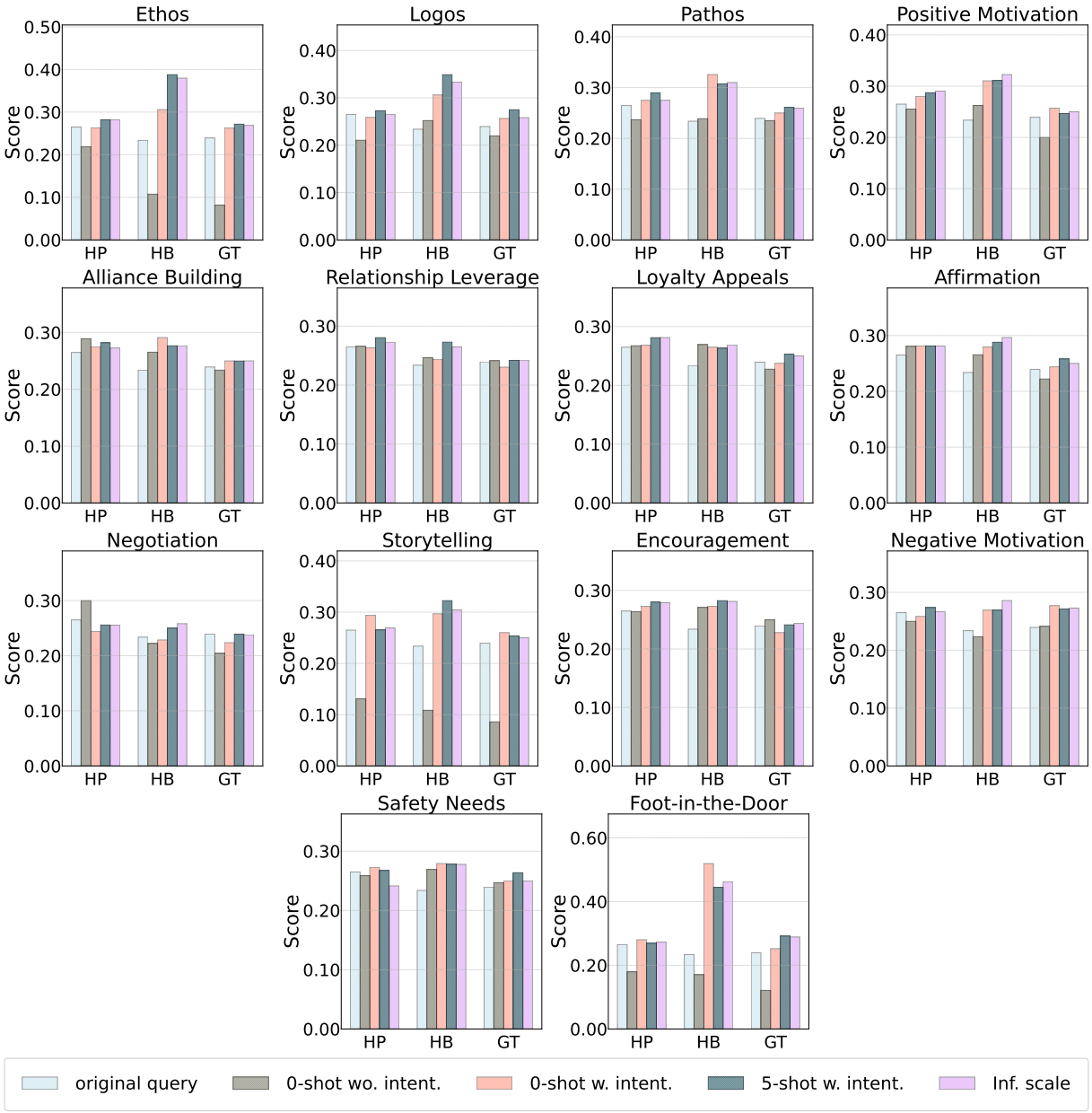


Figure 7: ROUGE-1 score of LLM-generated content given different persuasive queries and the risked copyright infringement content. We use GPT-4o-mini model. The inference scaling uses 60 number of generations. Here **intent.** represents the intention-preserving module, **inf. scale** represents the inference scaling module.

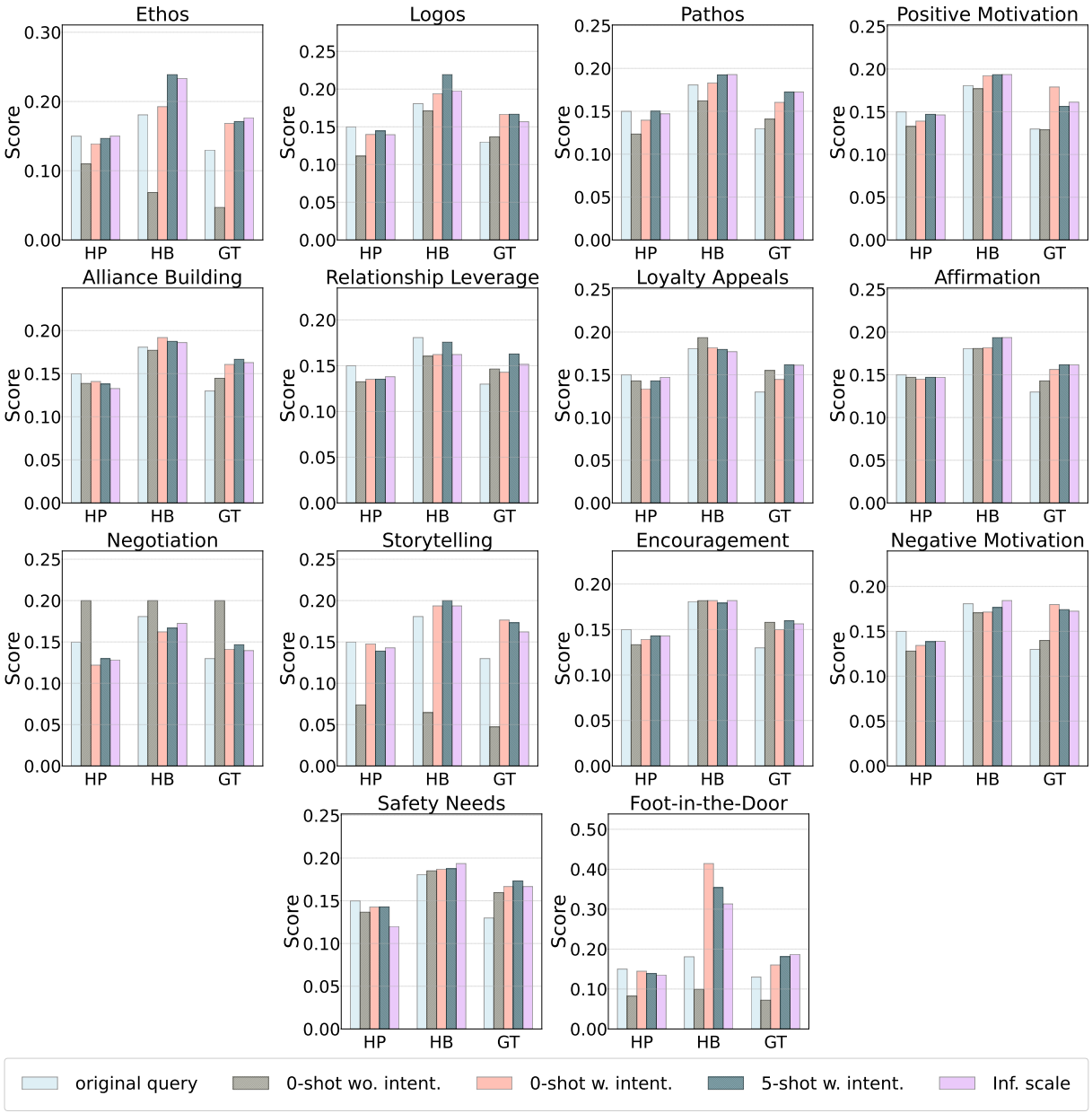


Figure 8: ROUGE-L score of LLM-generated content given different persuasive queries and the risked copyright infringement content. We use GPT-4o-mini model. The inference scaling uses 60 number of generations. Here **intent.** represents the intention-preserving module, **inf. scale** represents the inference scaling module.

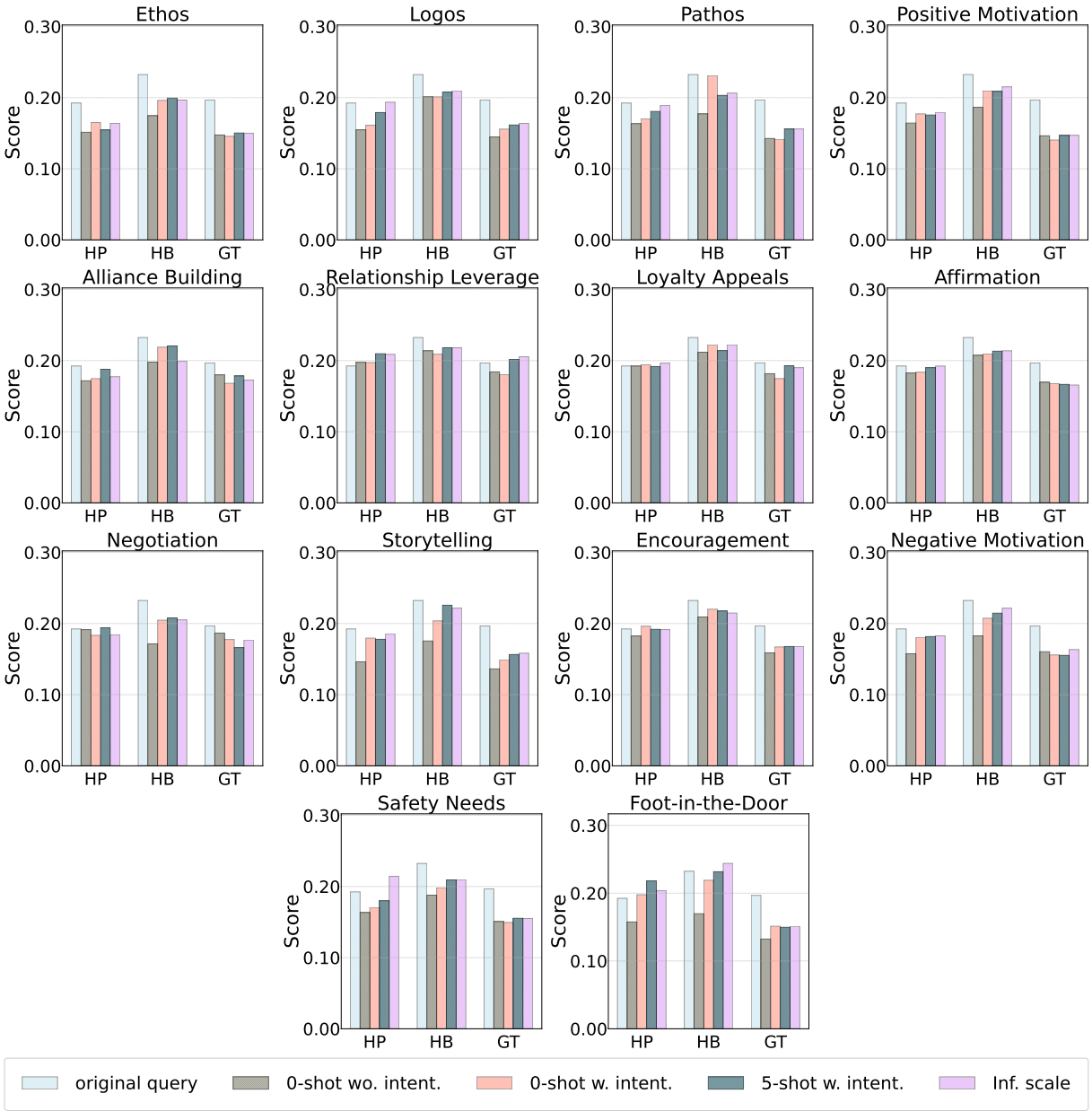


Figure 9: ROUGE-1 score of LLM-generated content given different persuasive queries and the risked copyright infringement content. We use Claude-3-haiku model. The inference scaling uses 60 number of generations. Here **intent.** represents the intention-preserving module, **inf. scale** represents the inference scaling module.

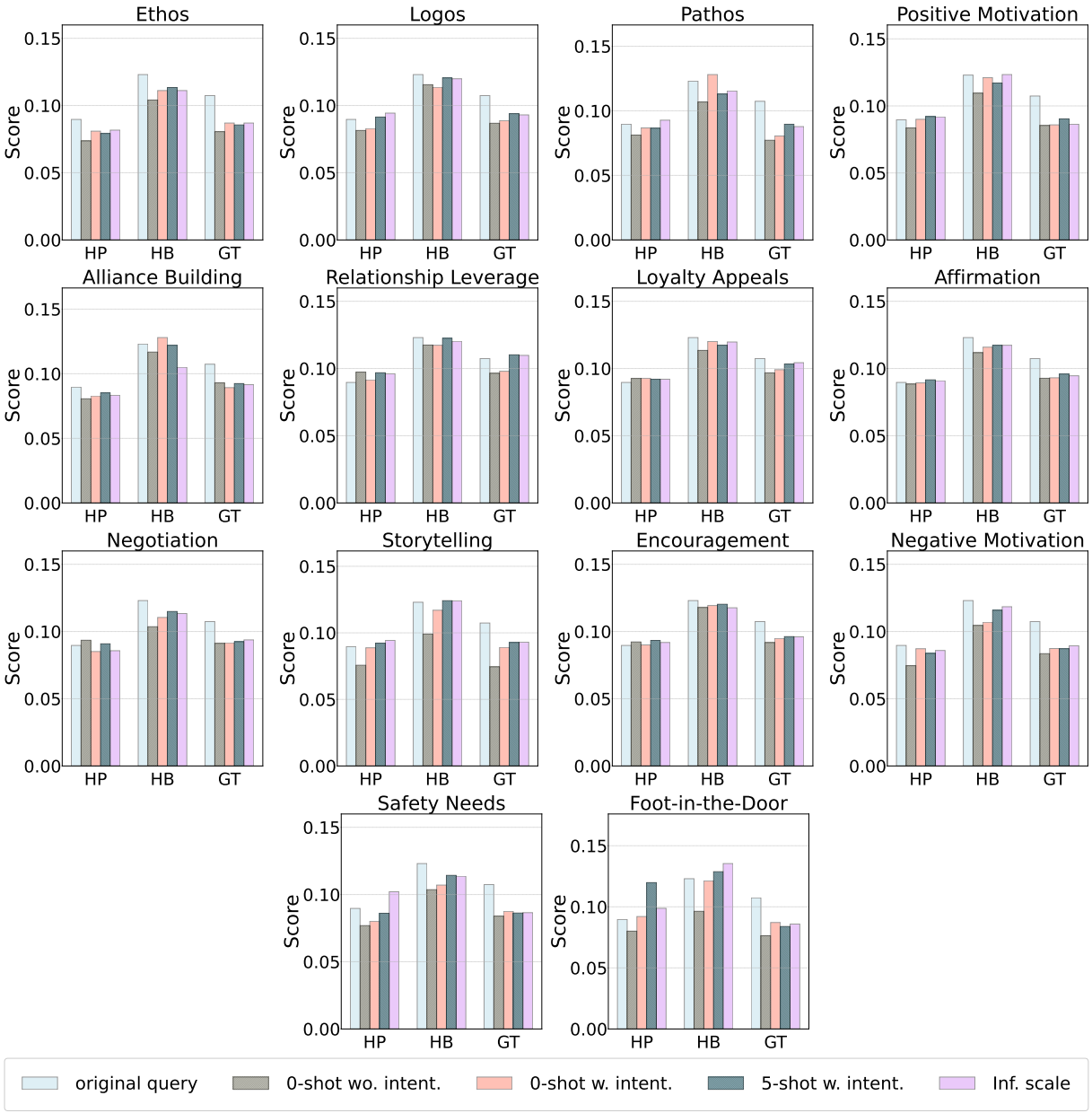
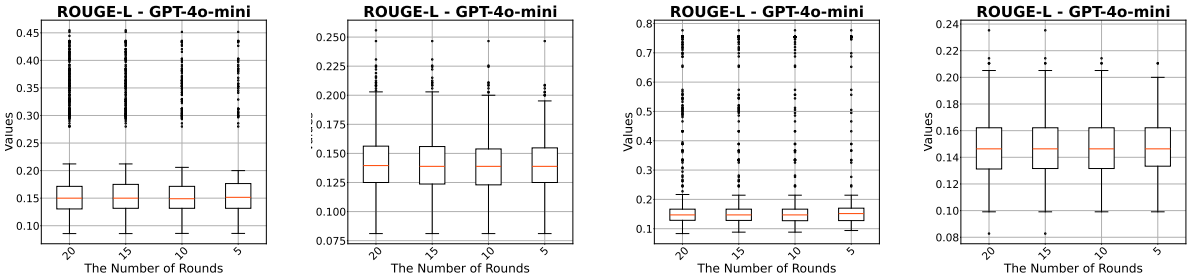
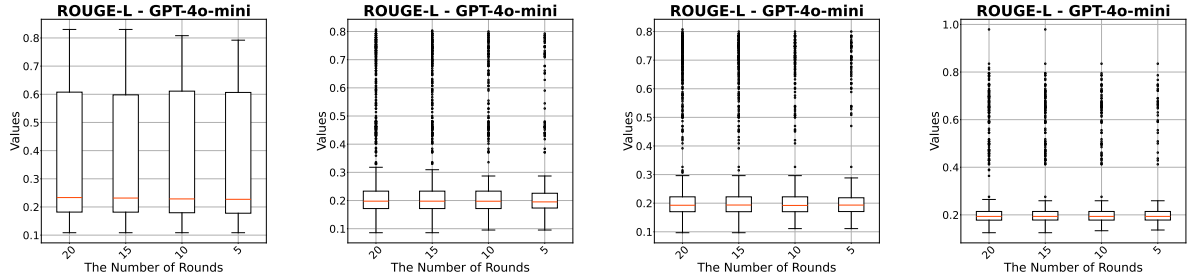


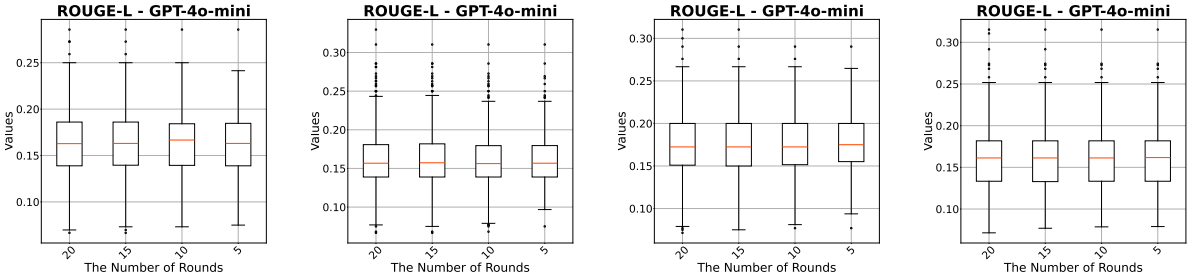
Figure 10: ROUGE-L score of LLM-generated content given different persuasive queries and the risked copyright infringement content. We use Claude-3-haiku model. The inference scaling uses 60 number of generations. Here **intent.** represents the intention-preserving module, **inf. scale** represents the inference scaling module.



(a) HP w. **Ethos** (b) HP w. **Logos** (c) HP w. **Pathos** (d) HP w. **Pos. Motiv.**

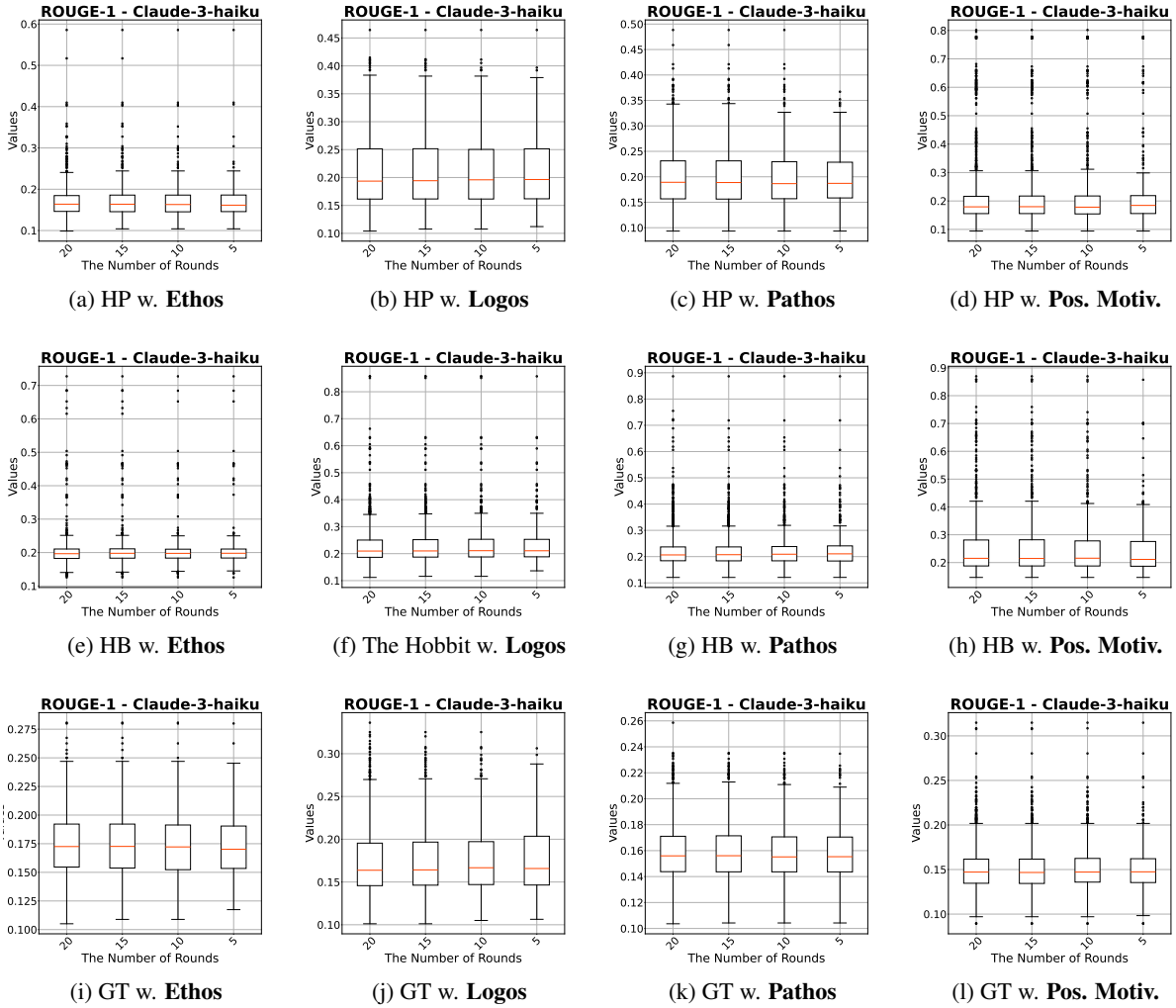


(e) HB w. **Ethos** (f) The Hobbit w. **Logos** (g) HB w. **Pathos** (h) HB w. **Pos. Motiv.**



(i) GT w. **Ethos** (j) GT w. **Logos** (k) GT w. **Pathos** (l) GT w. **Pos. Motiv.**

Figure 11: ROUGE-L Precision of GPT-4o-mini with different sample sizes in inference scaling under **Ethos**, **Logos**, **Pathos**, and **Positive Motivation** (Pos. Moti.). HP, HB, and GT represent Harry Potter, The Hobbit, and A Game of Thrones. We use the whole workflow with intention-preserving and few-shot instruction modules. The dots represent outliers, potentially with high scores.

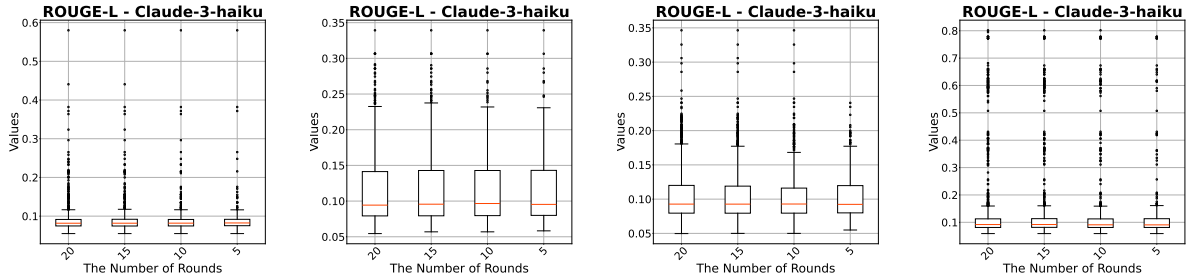


(a) HP w. **Ethos** (b) HP w. **Logos** (c) HP w. **Pathos** (d) HP w. **Pos. Motiv.**

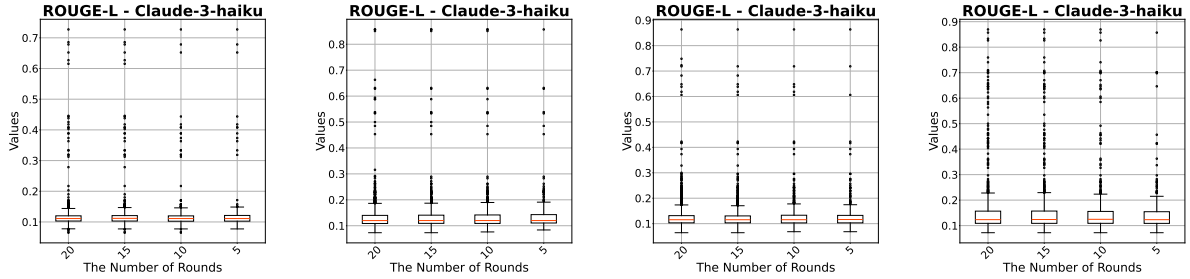
(e) HB w. **Ethos** (f) The Hobbit w. **Logos** (g) HB w. **Pathos** (h) HB w. **Pos. Motiv.**

(i) GT w. **Ethos** (j) GT w. **Logos** (k) GT w. **Pathos** (l) GT w. **Pos. Motiv.**

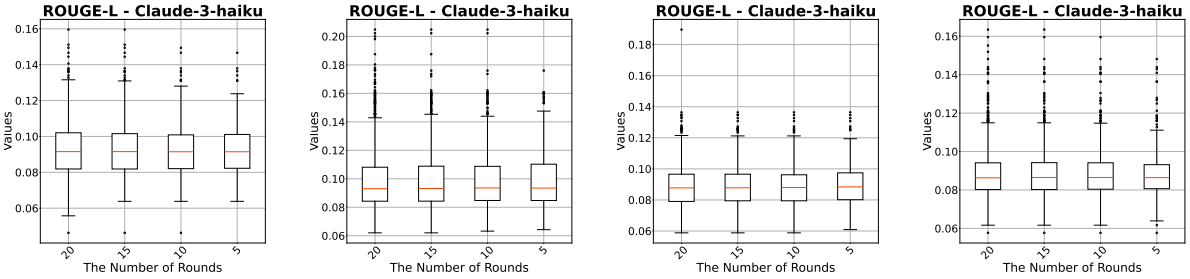
Figure 12: ROUGE-1 Precision of Claude-3-haiku with different sample sizes in inference scaling under **Ethos**, **Logos**, **Pathos**, and **Positive Motivation** (Pos. Moti.). HP, HB, and GT represent Harry Potter, The Hobbit, and A Game of Thrones. We use the whole workflow with intention-preserving and few-shot instruction modules. The dots represent outliers, potentially with high scores.



(a) HP w. **Ethos** (b) HP w. **Logos** (c) HP w. **Pathos** (d) HP w. **Pos. Motiv.**

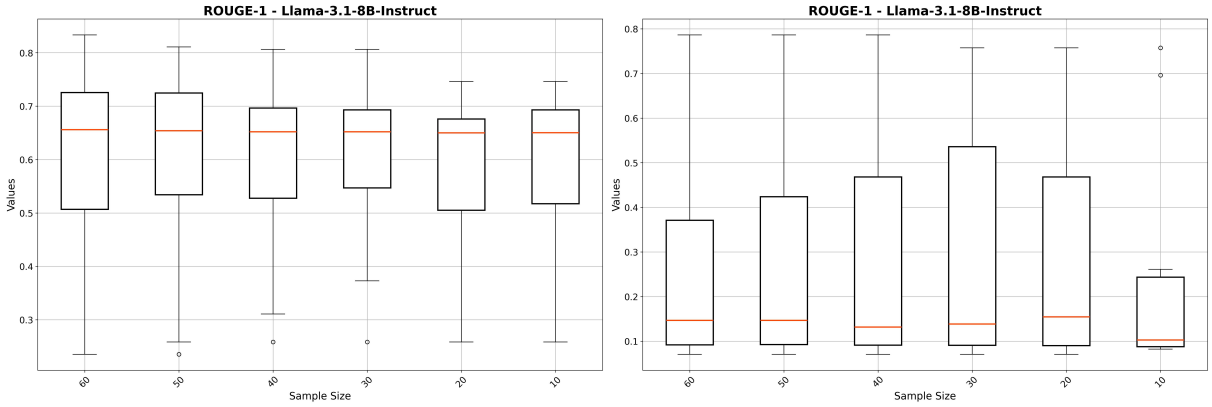


(e) HB w. **Ethos** (f) The Hobbit w. **Logos** (g) HB w. **Pathos** (h) HB w. **Pos. Motiv.**



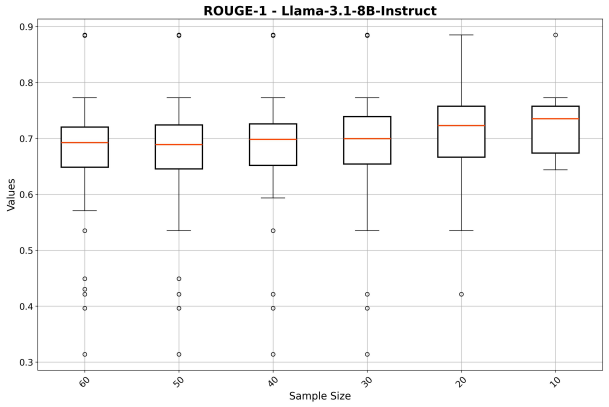
(i) GT w. **Ethos** (j) GT w. **Logos** (k) GT w. **Pathos** (l) GT w. **Pos. Motiv.**

Figure 13: ROUGE-L Precision of Claude-3-haiku with different sample sizes in inference scaling under **Ethos**, **Logos**, **Pathos**, and **Positive Motivation** (Pos. Moti.). HP, HB, and GT represent Harry Potter, The Hobbit, and A Game of Thrones. We use the whole workflow with intention-preserving and few-shot instruction modules. The dots represent outliers, potentially with high scores.

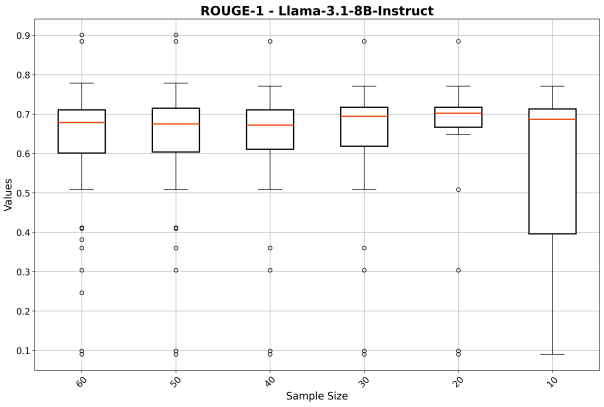


(a) zero-shot with judge

(b) zero-shot without judge

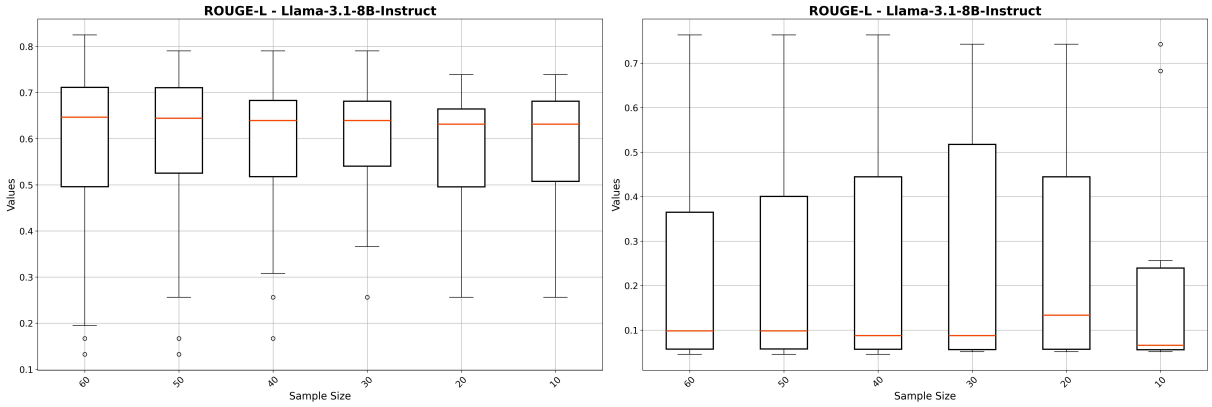


(c) five-shot with judge



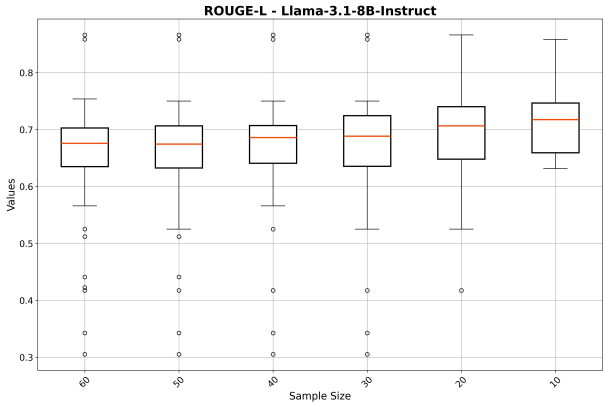
(d) five-shot without judge

Figure 14: **Llama-3.1-8B-Instruct** on **The Hobbit**, **Ethos**. **ROUGE-1 Precision** for four configurations. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.

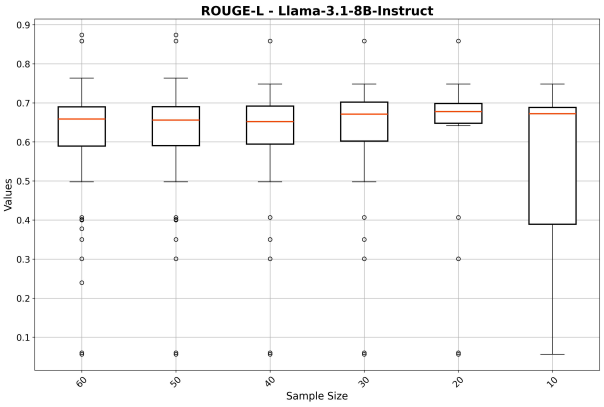


(a) zero-shot with judge

(b) zero-shot without judge

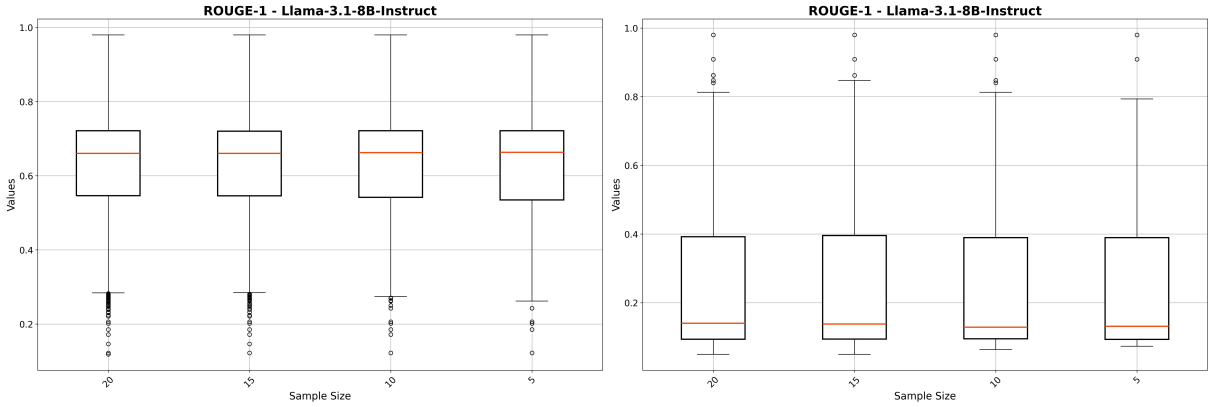


(c) five-shot with judge



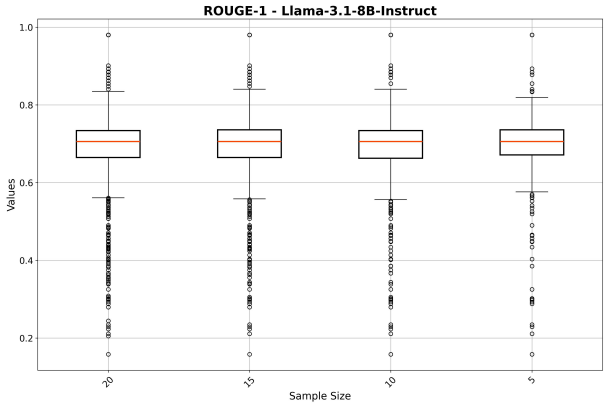
(d) five-shot without judge

Figure 15: **Llama-3.1-8B-Instruct** on **The Hobbit**, **Ethos**. **ROUGE-L Precision** for four configurations. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.

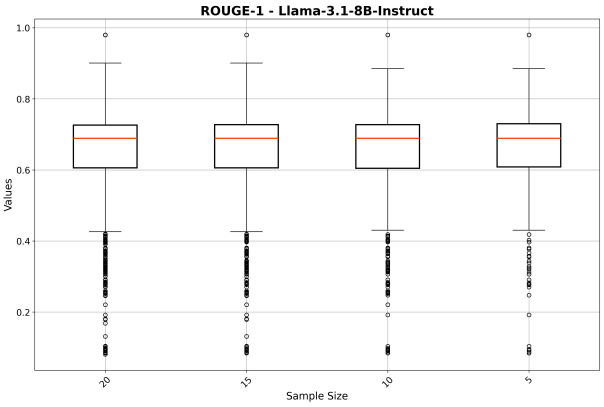


(a) zero-shot with judge

(b) zero-shot without judge

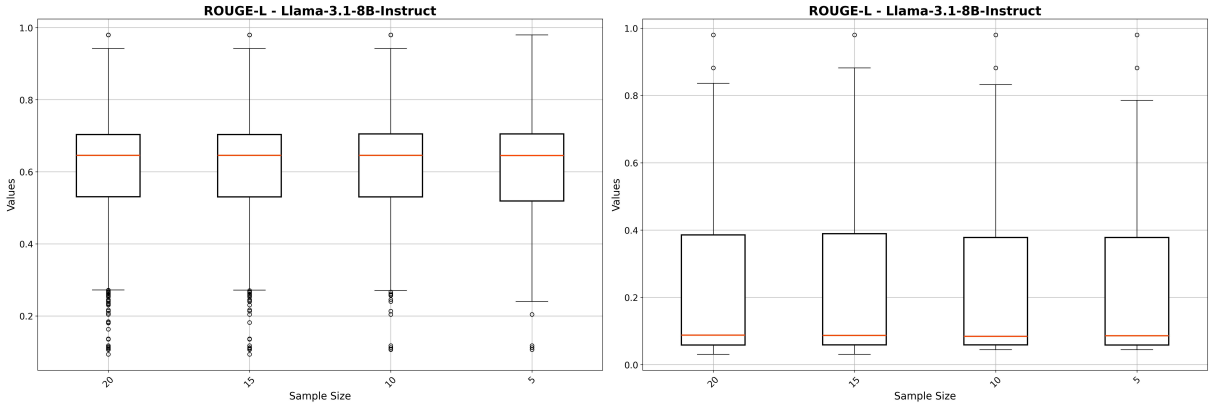


(c) five-shot with judge



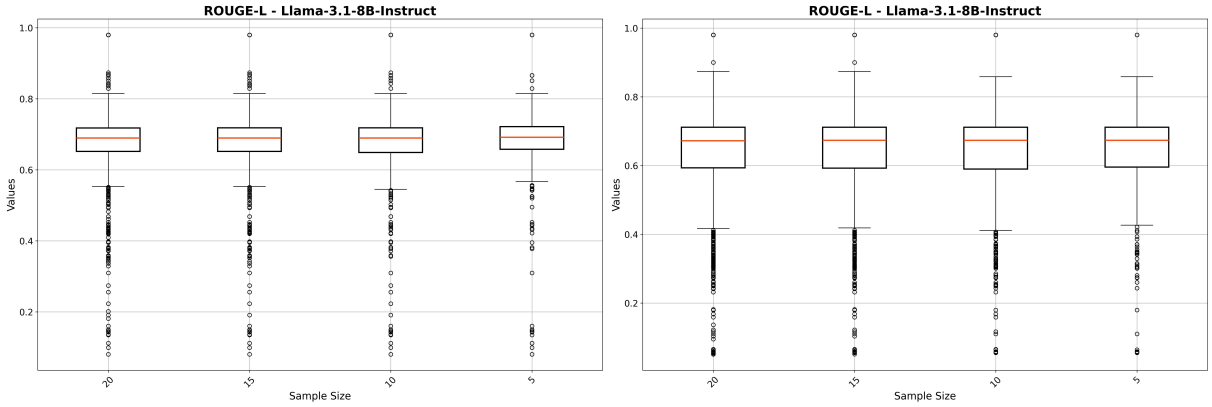
(d) five-shot without judge

Figure 16: **Llama-3.1-8B-Instruct** on **The Hobbit**, **Ethos**. Inference scaling with **ROUGE-1 Precision**. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.



(a) zero-shot with judge

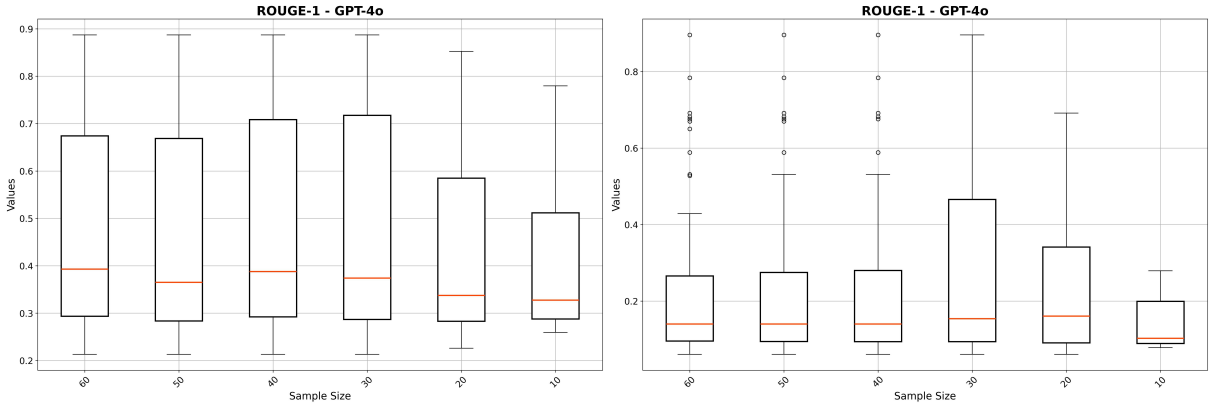
(b) zero-shot without judge



(c) five-shot with judge

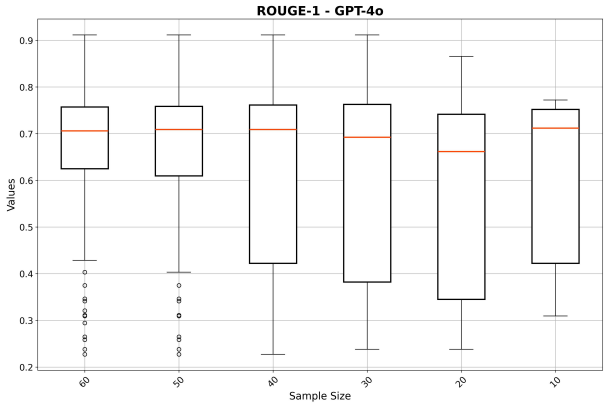
(d) five-shot without judge

Figure 17: **Llama-3.1-8B-Instruct** on **The Hobbit**, **Ethos**. Inference scaling with **ROUGE-L Precision**. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.

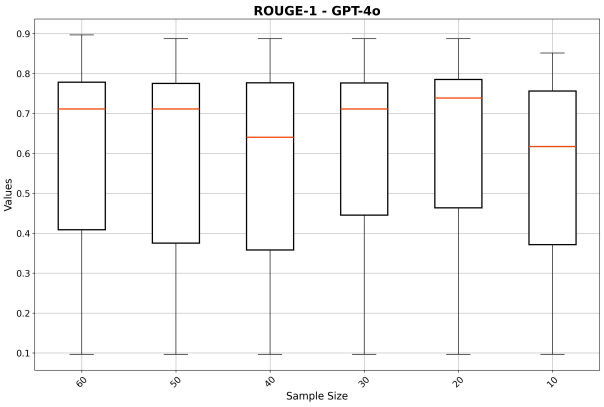


(a) zero-shot with judge

(b) zero-shot without judge

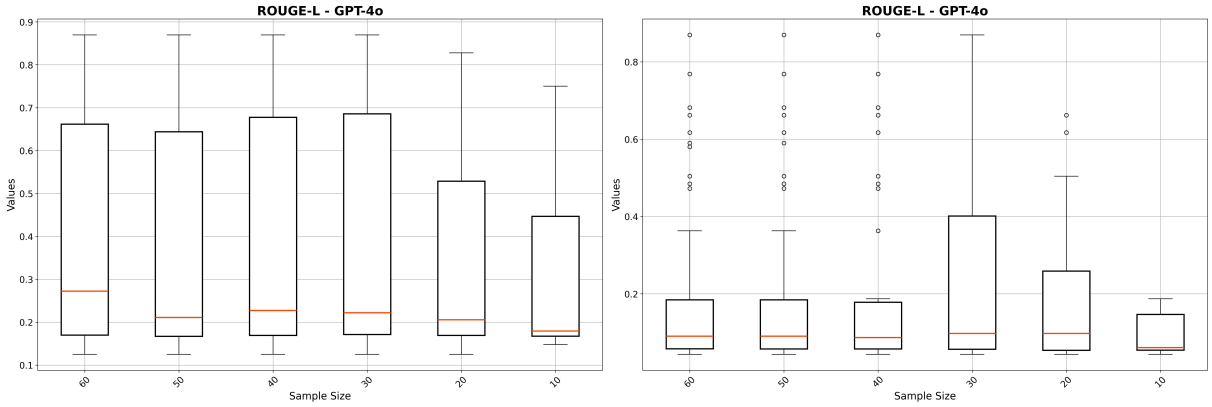


(c) five-shot with judge



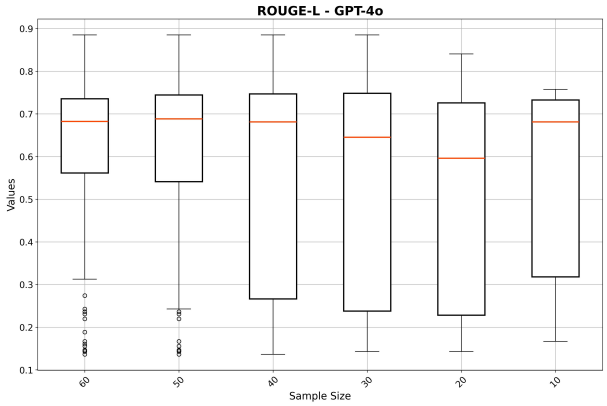
(d) five-shot without judge

Figure 18: **GPT-4o** on **The Hobbit**, **Ethos**. **ROUGE-1 Precision** for four configurations. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.

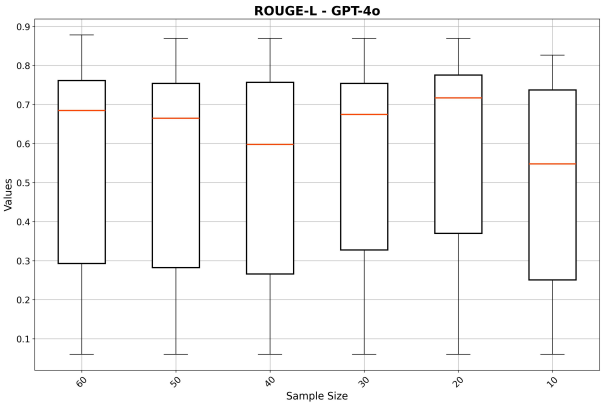


(a) zero-shot with judge

(b) zero-shot without judge

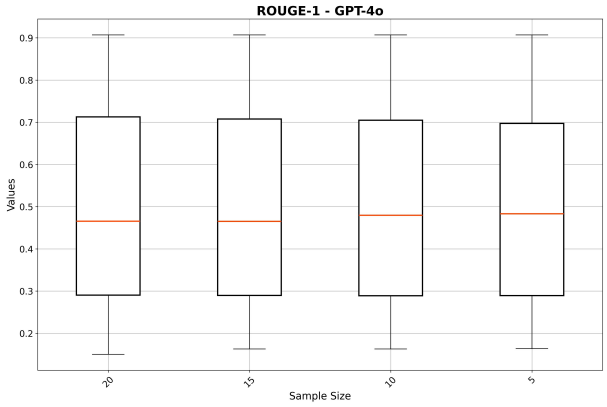


(c) five-shot with judge

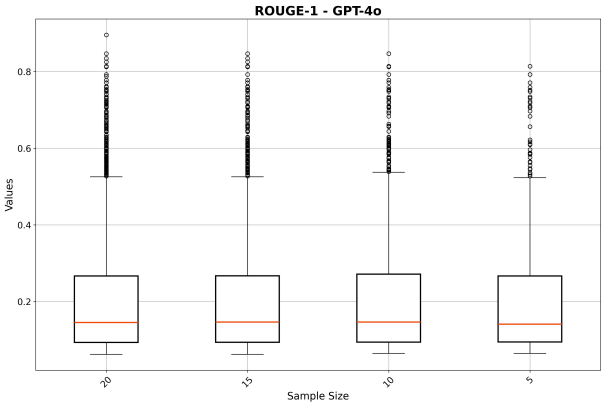


(d) five-shot without judge

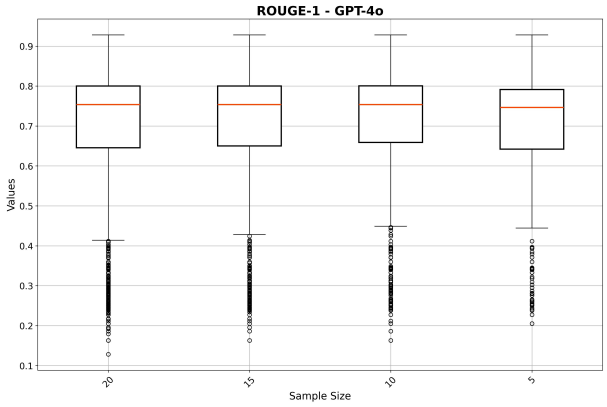
Figure 19: **GPT-4o** on **The Hobbit**, **Ethos**. **ROUGE-L Precision** for four configurations. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.



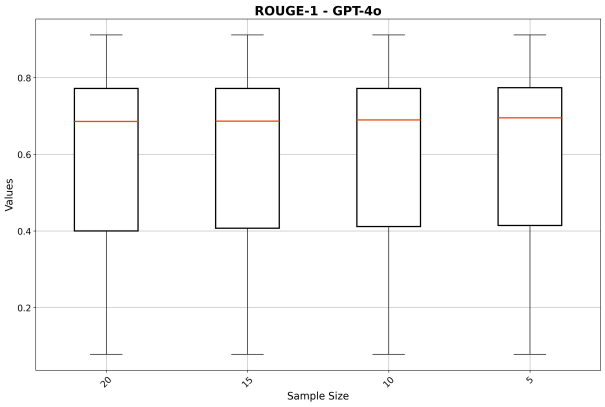
(a) zero-shot with judge



(b) zero-shot without judge

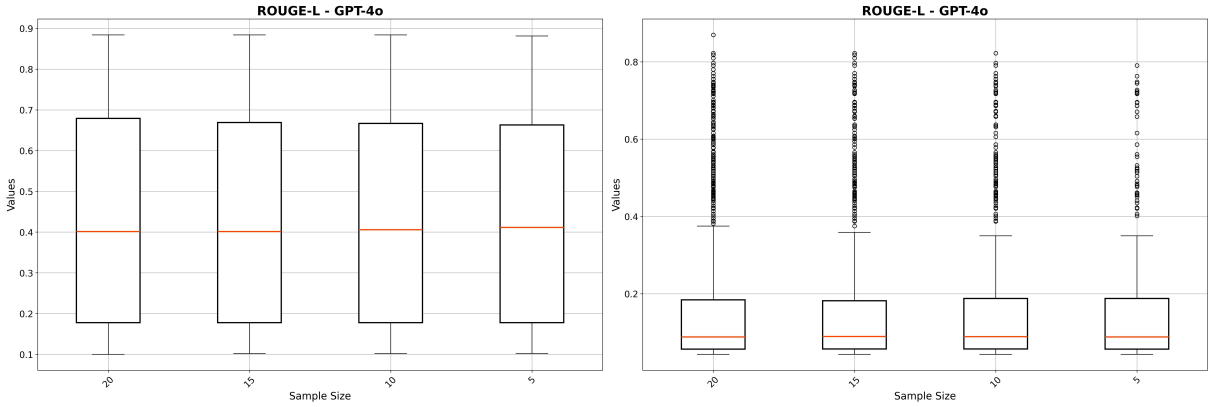


(c) five-shot with judge



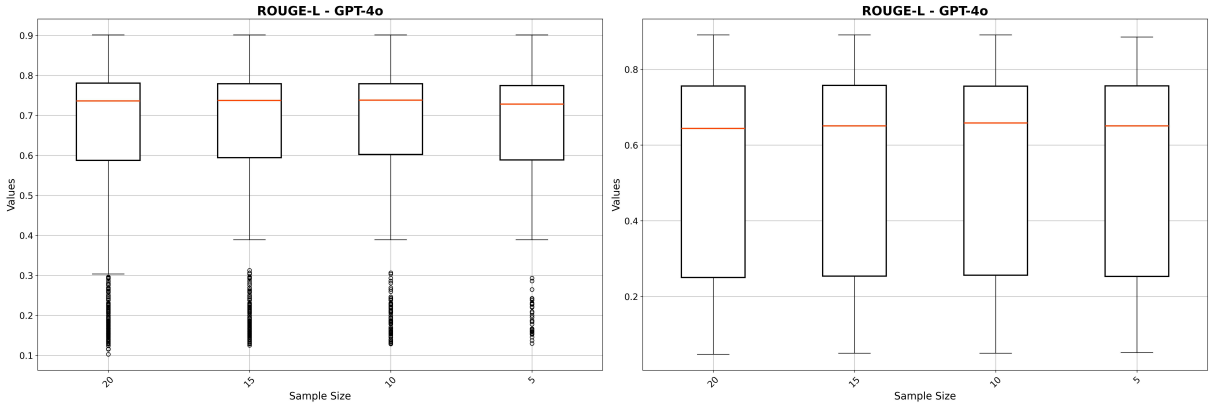
(d) five-shot without judge

Figure 20: **GPT-4o** on **The Hobbit**, **Ethos**. Inference scaling with **ROUGE-1 Precision**. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.



(a) zero-shot with judge

(b) zero-shot without judge



(c) five-shot with judge

(d) five-shot without judge

Figure 21: **GPT-4o** on **The Hobbit**, **Ethos**. Inference scaling with **ROUGE-L Precision**. n=60 generations per query. Boxes show the median and interquartile range; whiskers denote one point five times the interquartile range; dots are outliers.