University of Amsterdam at the CLEF 2025 Eloquent Track

Evaluating the Influence of Stylistic Prompt Variations on Semantic Interpretation

Bruno N. Sotic, Jaap Kamps

University of Amsterdam, Amsterdam, The Netherlands

Abstract

This paper reports on the University of Amsterdam's participation in the CLEF 2025 Eloquent Track's Robustness and Consistency Task. Our overall goal is to evaluate the influence of stylistic prompt variations on semantic interpretation. Our specific focus is to investigate how variations in prompt tone, structure, and persona affect the consistency and robustness of responses generated by large language models (LLMs).

We approach this through two complementary methods. First, we use a model-as-judge setup to quantify semantic consistency: each stylistic variant prompt is compared to its original base prompt using GPT-4.1 to rate the similarity of the generated responses on a 0–5 scale. Second, we conduct an inductive qualitative analysis on a selected prompt to closely examine how different stylistic framings influence content shifts in model outputs. Our results suggest that prompt reformulations can lead to variations in output, informational content, and tone.

Keywords

Stylistic Prompting, Generative Large Language Models, Robustness Semantic Consistency, Culturally Appropriate Responses

1. Introduction

The first CLEF 2024 Eloquent Track [1], featured the Quiz Task [2], the HaluciGen Task [3], and the Robust Task [4]. After a successful first year, the track continues as the CLEF 2025 Eloquent track [5]. There is a mix of ongoing and new tasks: the CLEF 2024 Eloquent Robust Task [4] continues as the CLEF 2025 Eloquent Robustness and Consistency Task [6], which is our main research focus in this paper. This paper contains an extended version of our main result, as the Task Overview [6] already includes a "Joint Report."

While the main design of the task focuses on robustness and consistency, we also examined the cultural or stylistic appropriateness of the response. We feel that both aspects are of key interest. On the one hand, the information value of the response must be invariant to what is assumed to be invariant prompts, for example, for a factual request phrased in a different language. On the other hand, responses must be culturally and stylistically appropriate. Here, we may expect the same informal content to be framed and phrased very differently depending on the context.

The goal of our participation was to explore how stylistic variation in prompts affects model behavior. The original English-language prompts provided by the track organizers as a base and were rewritten into nine distinct prompting styles. These styles were derived from a typology informed by academic literature.

The rewritten prompts preserved the original semantic and informational content but varied in phrasing, tone, structure, and stylistic framing. All variation was implemented exclusively through user-facing prompts. This submission includes only the English-language prompts, though versions in additional languages are currently in preparation.

The goal of this work is not to make broad claims about LLM behavior but to conduct an initial, exploratory analysis of how semantically equivalent prompts (differing only in style) may yield semantically divergent outputs - an increasingly important problem given the recent popularity of LLMs for informational search.

CLEF 2025 Working Notes, 9–12 September 2025, Madrid, Spain

nadalic.sotic@uva.nl (B. N. Sotic); kamps@uva.nl (J. Kamps)

0009-0003-2122-2235 (B. N. Sotic); 0000-0002-6614-0087 (J. Kamps)

Table 1
Prompt Styles and Their Definitions

Prompt Style/Name	Definition
Aggressive/Authoritative Tone	*Prompts characterized by commanding or forceful language, often lacking politeness or courtesy.*
Conversational Tone	*Prompts that mimic natural human dialogue, often informal and friendly in nature.*
Chain-of-Thought (CoT)	*A prompting technique where the model is guided to generate intermediate reasoning steps before arriving at a final answer.*
Formatting Differences	*Variations in the structural presentation of prompts, such as the use of lists, bullet points, or different punctuation.*
Persona-Based Prompts	*Prompts that assign a specific role or identity to the model, such as "You are a helpful assistant."*
Polite Tone	*Prompts that employ courteous language, including phrases like "please" and "thank you."*
Technical/Jargon-Heavy Prompts	*Prompts that utilize domain-specific terminology or complex language.*
System 1 Thinking Prompts	*Prompts that encourage fast, intuitive responses, aligning with the concept of System 1 thinking.*
System 2 Thinking Prompts	*Prompts that promote slow, deliberate reasoning, aligning with the concept of System 2 thinking.*

The remainder of this paper is structured as follows. Next, Section 2 presents our experimental setup. Section 3 presents our analytical approach. Section 4 discusses our detailed analysis and findings. Section 5 ends the paper with a discussion and conclusions.

2. Experimental Set-up

2.1. Prompt Design

The 15 original prompts were manually rewritten into nine stylistic variations (with the tenth being the original baseline), resulting in a total of 135 prompts. Although the rewriting was done by hand, Gemini was used to validate grammar and fluency.

The aim was to keep the semantic content of each prompt consistent while varying stylistic/linguistic aspects (tone, framing, structure). The typology and definitions of each style is based on the literature [7, 8, 9, 10, 11, 12, 13, 14]. The prompts styles are presented in Table 1.

Only the English prompts were used in this analysis, as authors lacked native speaker understanding in the other languages, and machine translation might introduce too much ambiguity to reliably validate semantic similarity.

Figure 1 presents basic descriptive statistics illustrating how each stylistic prompt type differs in surface-level linguistic aspects. This only concerns the prompt itself (the request) and not the response of the model. These responses will be analyzed in detail in the rest of this paper.

Semantic Equivalence Validation Given the goal of inducing different answers from semantically equivalent prompts, an attempt was made to verify that the rewritten prompts are indeed semantically similar. This is a difficult task and remains an open challenge.

To approach this, a heuristic "AI-as-a-judge" method was used, inspired by [15]. GPT-4.1 was prompted to act as an oracle and evaluate whether the stylistic rewrites of each original prompt still asked the same thing in terms of meaning and intent.

The comparison was made between a rewritten variation and its original prompt, using the following system prompt:

```
SYSTEM_PROMPT = """
```

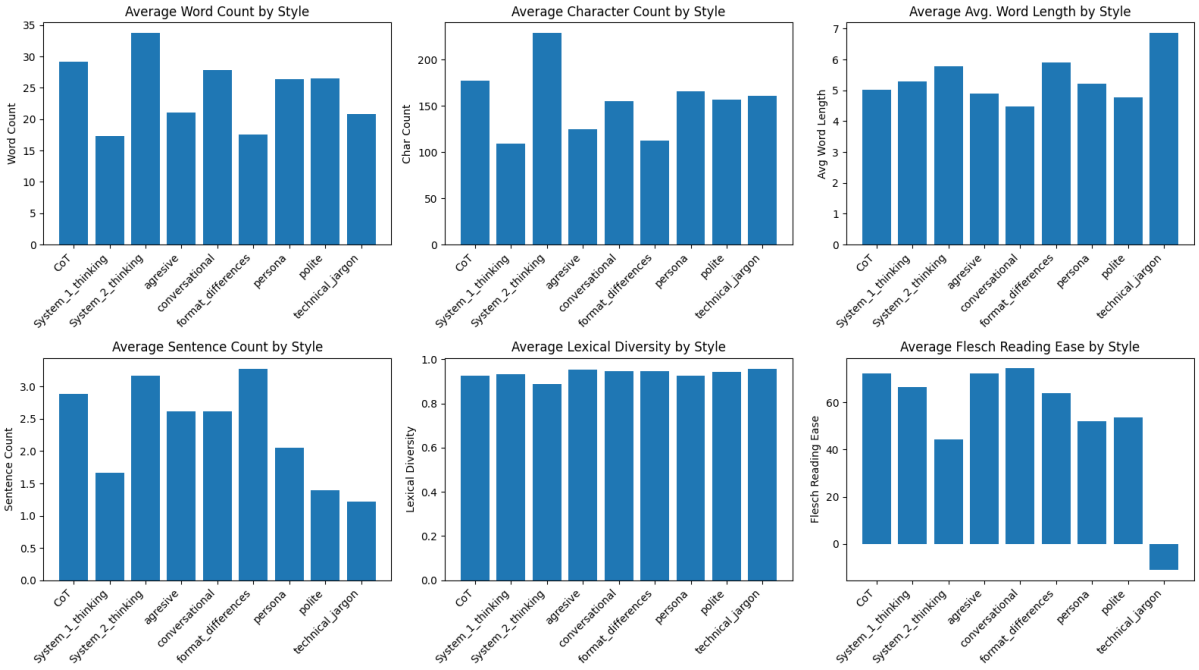

Figure 1: Descriptives per Prompt Style Variation (Questions)

Table 2
GPT-Judged Average Similarity Scores for Question Variations

Prompt Style	Mean Similarity Score
CoT	4.7778
System 1 Thinking	4.9444
System 2 Thinking	4.3333
Aggressive	5.0000
Conversational	4.8333
Formatting Differences	4.8889
Persona	4.7778
Polite	4.8889
Technical Jargon	4.2778

```
You will be given multiple questions. Your task is to assess whether each variation
conveys the same intent and meaning as the original. For each variation, assign a
similarity score from 0 to 5 - where 0 - completely different in meaning; and 5 -
expresses the exact same intent. Return exactly one JSON object with two keys:
  * "variation_id": the identifier for this variation (filename and item id).
  * "score": a number (integer) between 0 and 5, inclusive.
Do not output any extra text-only the JSON object.
Example output:
{
  "variation_id": "style1.json#item42",
  "score": x
}
"""
```

While this approach is not a perfect semantic measure, it serves as a pragmatic heuristic for approximating whether stylistic prompts are interpreted as conveying the same (and if they are understood as such by the system).

The results of this semantic validation are presented in Table 2.

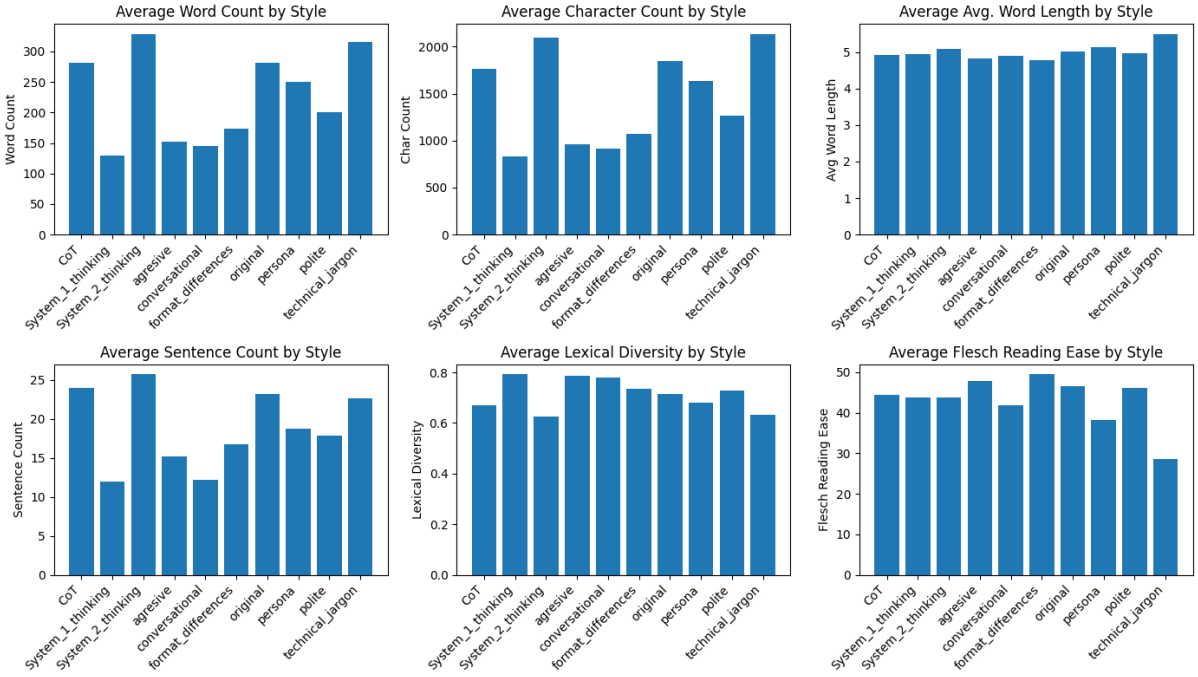

Figure 2: Descriptives per Output for Each Style Variation (LLM Responses)

3. Analytic Framework

The full set of stylistic prompts was used to query GPT-4.1 in isolated sessions to avoid memory effects. Each variation was treated as a new query, and the model's responses were recorded.

Descriptive metrics were averaged across prompt styles and are shown in Figure 2. As expected for advanced instruction tuned large language models, the style of the prompt indeed has a significant effect on the response. There are some interesting similarities between the statistics of the responses in Figure 2 and those of the prompts in Figure 1 earlier.

Evaluating the effect of prompt stylistics on LLM responses presents a methodological challenge. The prompts in our study pose culturally loaded, subjective questions (e.g., *"Is it more important to be honest or polite?"*) for which there is no correct answer in the conventional sense. These questions differ from typical factual QA tasks (e.g., *"Is X greater than Y?"*), where semantic similarity can be more easily computed using token overlap, embeddings, or entailment metrics.

In our case, LLM responses express stances, values, and culturally framed reasoning. Since the prompt variations also differ in structure, tone, and length (e.g., *"What is more important, honesty or politeness?"* vs. *"In your opinion, should one prioritize being polite over being honest?"*), the responses they elicit often vary in length, rhetorical form, and surface structure.

We use the same AI-as-judge method described in the previous section. However, in this phase, we provided the model with both the original question and its corresponding base response, allowing it to evaluate the semantic similarity of each stylistic variant's response relative to this reference. Ratings were again given on a 5-point scale.

In addition, we conducted an inductive qualitative content analysis on a selected prompt and its variants to gain deeper insight into how different styles influence the substance of responses. This involved descriptively labeling and comparing each response variant, identifying shifts in emphasis, reasoning patterns, tone, and framing. The goal was not to quantify but to trace how and where meaning drifted.

Table 3
GPT-Judged Semantic Similarity Scores by Prompt Style

Prompt Style	Mean	StdDev	Min	Max
CoT	3.50	0.99	1	5
System 1 Thinking	3.11	1.45	0	5
System 2 Thinking	3.56	1.10	0	5
agressive	3.39	1.29	0	5
conversational	3.39	1.04	0	5
format differences	3.44	0.86	1	4
persona	3.67	0.69	2	5
polite	3.50	1.10	0	5
technical jargon	3.11	1.08	1	5

4. Results

This section presents the results of our experiment. First, we report the outcomes of the AI-as-judge evaluation using GPT-4.1, followed by the findings from an inductive qualitative analysis of a smaller sample.

AI-as-Judge Table 3 shows the averaged similarity scores assigned by GPT-4.1 across the different prompting styles. Each score reflects how closely a candidate answer matched a reference answer in meaning and intent. The complete set of per-item scores is available in the appendix.

Across styles, average scores fell in the mid-range (3.1–3.6), suggesting that GPT-4.1 frequently judged responses as partially overlapping with the reference. A score near 3 indicates that some key points were shared, although details differed or were missing. The highest overall mean was by the *persona* style (3.67), which also had the lowest standard deviation (0.69) and no zero scores (indicating responses aligned with the reference).

System_1_thinking and *technical_jargon* styles had the lowest means (both 3.11), suggesting more frequent semantic drift. System 1 prompts, (which encourage fast, intuitive responses) often led to simplified or overgeneralized reasoning. Technical jargon, on the other hand, appeared to reframe or distort original meaning through abstract or (unsurprisingly) domain-specific language. Both styles also showed wider variability and occasional minimum scores of zero, meaning that in some cases the LLM's responses were entirely off-topic.

Among all styles, *format_differences* (where only the layout of the output changed) produced the narrowest score range (Min=1, Max=4, StdDev=0.86). Changes to surface formatting seem to preserve meaning more reliably than changes in tone or reasoning structure, although no style consistently produced perfect matches.

Qualitative Analysis An inductive qualitative analysis was conducted on a single prompt across all styles. The selected prompt (Question ID 12) asked the model to rank six societal values:

> *"Which of the following societal values is most important? Rank them with the most important first: [individual freedom, promotion of industry, support for creativity, safety, equality, law and order]"*

This task was chosen because ranking prompts enable straightforward observation of content shifts, priority changes, and semantic variation.

Table 4 summarizes the comparative findings. In terms of order preservation, the *persona* and *chain-of-thought* (CoT) styles remained closest to the base ranking, making only minor adjustments. In contrast, *aggressive*, *conversational*, *format_difference*, *system_1_thinking*, *system_2_thinking*, and *technical_jargon* frequently reordered top-ranked values, indicating that the tone or reasoning style affected how the model prioritized the list.

Table 4
Inductive Qualitative Codes for Prompt Variations (including Original)

Style	Inductive Codes	Observation
Original	Rationale Present, Single Perspective	Provides a detailed ranking with "why" explanations for each value, acknowledges subjectivity but maintaining one perspective / ranking.
Aggressive	Rank Change, Missing Explanation	The list order shifts and all "why" details vanish.
Conversational	Extra Perspectives, AI Framing	Offers several ranking examples and meta-text ("I don't have opinions").
CoT	Rationale Present, Rank Change	Keeps "why" logic but swaps a couple of top values.
Format Difference	Single-List Only, Rank Change	Delivers a bare list (no paragraphs) and reorders the top item.
Persona	Structured Reasoning, Rationale Present	Uses expert voice and bullet explanations; order stays mostly aligned.
Polite	AI Framing, Scope Shift	Heavy prefacing and describes multiple value priorities—no single list.
System 1 Thinking	AI Framing, Extra Perspectives	Shows two societal-type rankings instead of one coherent answer.
System 2 Thinking	Structured Reasoning, Rank Change	Gives step-by-step "why" but places a different value at #1.
Technical Jargon	Formal Tone, Rank Change	Uses high-register language and changes the top priority.

Rationale also played a role. Styles that embedded explicit justifications, i.e., *CoT*, *persona*, *system_2*, and *technical_jargon*, tended to maintain closer alignment with the logic of the base ranking, even when order shifted slightly. In contrast, outputs that omitted reasoning or presented flat, unexplained lists, i.e., *aggressive* and *format_difference*, showed greater divergence from the original rationale.

Some styles also expanded the scope of the task. The *conversational*, *polite*, and *system_1_thinking* prompts often introduced multiple perspectives or emphasized the subjectivity of ranking values. Rather than providing a single prioritized list, these responses framed the task as open-ended or contingent, fundamentally shifting the prompt's intention from a single viewpoint to a multi-perspective discussion.

Framing effects were also evident, particularly in *conversational*, *polite*, and *persona*, which included epistemic markers such as "As an AI…" or referred to expert communities (e.g., "political scientists might say…"). These framings shifted the tone and sometimes led the model away from direct rankings and toward speculative responses.

Finally, the use of formal or technical language influenced interpretability. The *technical_jargon* style frequently translated everyday values into academic terminology. While the core logic was often intact, this reframing affected accessibility and occasionally altered perceived intent.

5. Discussion and Conclusions

The goal of this experiment was to explore whether semantically equivalent variations of subjective questions could nonetheless produce semantically different outputs when posed to an LLM. Using a set of manually written base prompts, each question was reformulated into multiple stylistic variants inspired by a literature review. To ensure that the rewrites remained semantically close to the original intent, we employed GPT-4.1 as a semantic judge in a separate evaluation stage. While this method is not flawless, it allowed us to establish that the prompts were, in theory, perceived as similar by the same model architecture used to generate the answers.

After generating LLM responses to each prompt variant, we evaluated the semantic similarity of the outputs using a 0–5 scale, where 5 indicates near-complete overlap in meaning with the base answer, and 0 denotes a fundamentally different or contradictory response.

Across most styles, the average similarity scores fell between 3.1 and 3.6, suggesting that while the responses often shared some common ground with the originals, they frequently diverged in specifics or focus. The "Persona" style stood out with the highest average score of 3.67 and the least variation, implying it reliably produced answers closest to the originals. By contrast, styles like "System 1" and "Technical Jargon" averaged the lowest score of 3.11, with responses that sometimes strayed far from the intended meaning, including cases where they were completely off the mark.

However, it is important to note that relying on GPT-4.1 to both generate and judge the responses raises the possibility of bias, as sometimes those middle-of-the-road scores might reflect the model's own uncertainty rather than real differences.

Further qualitative inspection of a single prompt response and its variant found that certain styles (*"Persona" and "Chain-of-Thought,"*) tended to mirror the original ranking order more closely. *"Aggressive," "Format," "System 1," and "Technical Jargon"* often shuffled the order, hinting that the style itself influenced how the model weighed the values. Styles that prompted the model to explain its reasoning (*"Chain-of-Thought" or "System 2"*) generally stayed truer to the base prompt by offering justifications that echoed the original prompt response. However, when the style favored brevity or simply listed the values without elaboration, some of the original meaning was lost. In addition, *"Conversational" and "Polite"* sometimes broadened the scope by encouraging multiple perspectives or highlighting subjectivity, resulting in more open-ended responses. Assigning the LLM a role (*"Persona"*) led the model to adopt a more cautious or expansive tone. When technical language was used, the model occasionally recast everyday values into more abstract terms, which could pull the response away from what was initially intended.

Acknowledgments

We thank the track and task organizers for their outstanding service and effort in making realistic benchmarks available for evaluating generative language model quality.

Bruno Sotic is partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Declaration on Generative AI

During the preparation of this work, the authors used *NotebookLM* and *Grammarly* in order to: **Grammar and spelling check** and **Paraphrase and reword**. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Karlgren, L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, M. Sahlgren, A. Talman, S. Zahra, Overview of ELOQUENT 2024 - shared tasks for evaluating generative language model quality, in: L. Goeuriot, P. Mulhem, G. Quénot, D. Schwab, G. M. D. Nunzio, L. Soulier, P. Galuscáková, A. G. S. de Herrera, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 15th International Conference of the CLEF Association, CLEF 2024, Grenoble, France, September 9-12, 2024, Proceedings, Part II, volume 14959 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 53–72. URL: https://doi.org/10.1007/978-3-031-71908-0_3. doi:10.1007/978-3-031-71908-0_3.

[2] J. Karlgren, A. Talman, ELOQUENT 2024 - topical quiz task, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 687–690. URL: https://ceur-ws.org/Vol-3740/paper-65.pdf.

[3] L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, S. Zahra, Overview of the CLEF-2024 eloquent lab: Task 2 on hallucigen, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the

Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 691–702. URL: https://ceur-ws.org/Vol-3740/paper-66.pdf.

[4] M. Sahlgren, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, S. Zahra, ELOQUENT 2024 - robustness task, in: G. Faggioli, N. Ferro, P. Galuscáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), Grenoble, France, 9-12 September, 2024, volume 3740 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 703–707. URL: https://ceur-ws.org/Vol-3740/paper-67.pdf.

[5] J. Karlgren, E. Artemova, O. Bojar, M. I. Engels, V. Mikhailov, P. Šindelář, E. Velldal, L. Øvrelid, Overview of ELOQUENT 2025: shared tasks for evaluating generative language model quality, in: J. Carrillo de Albornoz, J. Gonzalo, L. Plaza, A. García Seco de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, 2025.

[6] J. Karlgren, M. I. Engels, M. Barrett, R. R. Gunti, M. Hoveyda, B. N. Sotic, J. Kamps, M. Koistinen, E. Zosa, Overview and Joint Report for Robustness and Consistency Task of the ELOQUENT 2025 Lab for Evaluating Generative Language Model Quality, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025: Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[7] Z. Yin, H. Wang, K. Horio, D. Kawahara, S. Sekine, Should we respect LLMs? a cross-lingual study on the influence of prompt politeness on LLM performance, in: J. Hale, K. Chawla, M. Garg (Eds.), Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024), Association for Computational Linguistics, ????, pp. 9–35. URL: https://aclanthology.org/2024.sicon-1.2/. doi:10.18653/v1/2024.sicon-1.2.

[8] Y. Li, A practical survey on zero-shot prompt design for in-context learning, in: Proceedings of the Conference Recent Advances in Natural Language Processing - Large Language Models for Natural Language Processings, ????, pp. 641–647. URL: http://arxiv.org/abs/2309.13205. doi:10.26615/978-954-452-092-2_069. arXiv:2309.13205 [cs].

[9] Y. Tang, J. Ou, C. Liu, F. Zhang, D. Zhang, K. Gai, Enhancing role-playing systems through aggressive queries: Evaluation and improvement, ???? URL: http://arxiv.org/abs/2402.10618. doi:10.48550/arXiv.2402.10618. arXiv:2402.10618 [cs], version: 1.

[10] B. Chen, Z. Zhang, N. Langrené, S. Zhu, Unleashing the potential of prompt engineering for large language models (????) 101260. URL: http://arxiv.org/abs/2310.14735. doi:10.1016/j.patter.2025.101260. arXiv:2310.14735 [cs].

[11] L. Ein-Dor, O. Toledo-Ronen, A. Spector, S. Gretz, L. Dankin, A. Halfon, Y. Katz, N. Slonim, Conversational prompt engineering, ???? URL: http://arxiv.org/abs/2408.04560. doi:10.48550/arXiv.2408.04560. arXiv:2408.04560 [cs].

[12] J. Zamfirescu-Pereira, R. Y. Wong, B. Hartmann, Q. Yang, Why johnny can't prompt: How non-AI experts try (and fail) to design LLM prompts, in: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, Association for Computing Machinery, ????, pp. 1–21. URL: https://dl.acm.org/doi/10.1145/3544548.3581388. doi:10.1145/3544548.3581388.

[13] J. He, M. Rungta, D. Koleczek, A. Sekhon, F. X. Wang, S. Hasan, Does prompt formatting have any impact on LLM performance?, ???? URL: http://arxiv.org/abs/2411.10541. doi:10.48550/arXiv.2411.10541. arXiv:2411.10541 [cs].

[14] L. Pawlik, How the choice of LLM and prompt engineering affects chatbot effectiveness 14 (????) 888. URL: https://www.mdpi.com/2079-9292/14/5/888. doi:10.3390/electronics14050888, number: 5 Publisher: Multidisciplinary Digital Publishing Institute.

[15] M. Sahlgren, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, S. Zahra, Eloquent 2024 — robustness task, in: G. Faggioli, N. Ferro, P. Galuščáková, A. García Seco de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), CEUR Workshop Proceedings, CEUR-WS.org, Germany, 2024, pp. 703–707. Publisher Copyright: © 2024 Copyright for this paper by its authors.; Conference and Labs of the Evaluation Forum, CLEF 2024 ; Conference date: 09-09-2024 Through 12-09-2024.

A. Appendix

All data, code, and material necessary to reproduce this study are publicly available at: https://anonymous.4open.science/r/Eloquent-2025-Robustness-and-Consistency-6C71/README.md