Evaluating LLM Models for Model-to-Text Transformation: Singletons, Ensembles, and Statistical Analysis

Pedro Henrique Bernardes Caetano Milken[1], **Marcelo de Almeida Maia**[2], **Leandro Nogueira Couto**[2], **Nicole Maia Argondizzi**[1],**Marcelo Keese Albertini**[2]

[1]Graduate Program in Computer Science (PPGCO)
Federal University of Uberlândia (UFU) – Uberlândia, MG, Brazil

[2]Faculty of Computing (FACOM)
Federal University of Uberlândia (UFU) – Uberlândia, MG, Brazil

{pedro.milken,marcelo.maia,leandro.couto,nicole.argondizzi,albertini}@ufu.br

Abstract. *Business Process Model and Notation (BPMN) diagrams are essential for illustrating organizational procedures but can be inaccessible to non-technical stakeholders. Model-to-text (M2T) transformation helps bridge this gap. Research identifies GPT-4 as a leading model for M2T, but highlights hallucination, which introduces inaccuracies as a main issue. This study explores the potential reduction of hallucinations by employing contemporary LLMs in an iterative self-refinement cycle, testing eight models under various feedback and temperature conditions. Key performance indicators reveal that while self-refinement reduces task omission, it increases misrepresentation and spurious generation. The results indicate that top-performing configurations excel in one-shot generation, and modeling-register leakage is minimal due to effective jargon avoidance. However, automated feedback may inadvertently amplify hallucinations, challenging the assumptions of self-refinement strategies in M2T for requirements engineering. The study's reproducibility package is publicly available, detailing the methodology and results.*

1. Introduction

The creation of an information system (IS) is a complex process that requires requirements engineers to have a deep understanding of the technological, organisational, and operational characteristics of the application domain. This comprehension encourages engineers to engage with a variety of artefacts, including both traditional and newly developed textual artefacts such as user stories, use cases, and scenario descriptions, as well as graphical artefacts like conversation flows, UML diagrams, and visual business process descriptions. Business process models serve as vital tools for enhancing organisational performance by concentrating on business processes [Klievtsova et al. 2024].

The requirements engineering process is typically divided into two main phases: requirements development and requirements management. Requirements Development (RD) comprises four main phases: elicitation, analysis, specification, and validation. In contrast, Requirements Management (RM) encompasses essential services such as status monitoring, version control, and change control. Both RD and RM phases require a thorough understanding of artefacts to ensure their consistency and support effective management and (re-)validation [Klievtsova et al. 2024].

The recent development of natural language processing (NLP), particularly through large language models (LLMs) in chatbots like ChatGPT, has equipped requirements engineers with effective tools for managing, querying, and transforming extensive requirements databases into various artefacts. While different aspects of requirements elicitation, such as interviews with domain experts and chatbots, along with the transformation of textual process descriptions into graphical representations (text-to-model or T2M transformation), have been studied, further investigation in other domains is necessary [Klievtsova et al. 2024].

This research focuses on the elements of requirement elicitation and validation through the analysis of business process models and their conversion into textual descriptions. This initiative aims to improve the accessibility of information in graphical process models. Accessible artefacts for all participants in the requirements engineering process enhance the comprehension of complex process model characteristics, ensure their consistency validation, and transform textual data into improved graphical business process models. This study examines the effectiveness of large language models (LLMs) in generating process descriptions from process models, known as model-to-text (M2T) transformation, and contrasts this approach with conventional deterministic M2T methods [Klievtsova et al. 2024]. This study investigates the following research questions:

- **RQ1** What are the effects of feedback on the errors made in an initial transformation on the reprocessing by LLMs?
- **RQ2** Which LLM model has superior metrics compared to the others?
- **RQ3** What is the level of information loss that occurs during the M2T transformation, especially when using different LLMs?

This study addresses RQ1 through RQ3 as outlined below. Section 2 analyses contemporary M2T transformation methodologies and the relevant landscape of LLMs associated with this study. Section 3 examines the contexts for M2T transformation application within the Requirements Engineering (RE) process and delineates the M2T strategy proposed in this study, which includes the eight evaluated LLMs, the experimental framework, and the KPIs employed for assessment and comparison. Section 4 outlines the evaluation of the proposed methodology using a recognised "baseline" dataset [Bellan et al. 2023], incorporating contemporary large language models—namely Claude Opus 4.7, Claude Sonnet 4.6, Claude Haiku 4.5, GPT-5.5, GPT-5.4, GPT-4, DeepSeek v4 Flash, and DeepSeek v4 Pro. This study assesses the quality of outputs produced from different LLM prompts and feedback conditions ($\rightarrow$ RQ1 and RQ2), along with the information loss experienced during M2T transformation ($\rightarrow$ RQ3), the latter being quantified through similarity measures and the proposed KPIs. Section 5 concludes the paper and analyses the findings.

2. Related Work

This study is grounded in three interconnected literature areas: M2T transformation in business process management, comparative LLM benchmarking research, and statistical comparison methodologies across datasets.

2.1. LLMs

Large language models (LLMs) like ChatGPT and Claude, built on transformer architecture and trained on extensive internet datasets, have shown potential in scientific research across

various fields. Despite their advancements in reasoning and problem-solving, LLMs exhibit limitations such as flawed reasoning and reduced mathematical skills, making them less reliable for critical scientific applications [Jiang et al. 2025]. Ongoing research aims to evaluate their effectiveness in scientific tasks, emphasizing the need for domain-specific assessments to gauge their accuracy and reasoning capabilities [Jiang et al. 2025].

2.2. M2T transformation in business process management

The conversion of process models to natural language has a rich history in model-driven engineering. Wimmer and Burgueño [Wimmer and Burgueño 2013] highlight the challenges in M2T transformations, necessitating quality indicators over simple comparisons. Traditional BPMN methodologies often yield rigid outputs, while LLM-based approaches promise more natural language generation. Klievtsova et al. [Klievtsova et al. 2024] provide a relevant framework, assessing five LLMs using key performance indicators (KPIs) for readability and contextual similarity. This study builds on their work, utilizing the PETv11 corpus and Mermaid representations for comparability, and incorporates cross-lingual elements with multilingual encoders [Silva et al. 2025].

2.3. Statistical comparison and educational grounding

The evaluation methodology follows Demšar's [Demšar 2006] statistical comparison approach, employing the Friedman test and Nemenyi post-hoc analysis. This method, rooted in cognitive theories of multimedia learning, supports the integration of graphical and textual representations to enhance understanding, aligning with Klievtsova et al.'s scenarios for requirements engineering [Klievtsova et al. 2024].

2.4. Positioning of this work

This study contributes to the literature by adapting Demšar's statistical comparison protocol for M2T evaluation, replacing descriptive averages with inferential statistics and Critical Difference diagrams. It expands the LLM landscape from five models (2022-2023) to eight models (2024-2026), providing a quantitative assessment of LLM-based M2T transformation. Additionally, it introduces automated self-refine feedback as a new experimental factor and addresses the omission-versus-hallucination trade-off in LLM outputs, refining existing questions in the field [Klievtsova et al. 2024].

3. Experimental Procedures

3.1. Overview

This study enhances the Model-to-Text (M2T) evaluation framework by Klievtsova et al. [Klievtsova et al. 2024], expanding the analysis to eight contemporary large language models (LLMs) from Anthropic, OpenAI, and DeepSeek. It introduces iterative diagnostic feedback as an experimental variable, resulting in an 8×2 factorial design, and employs a multi-layer inferential statistical analysis, contrasting with the descriptive statistics of the original study. The experimental workflow is reproducible and publicly accessible [Caetano Milken 2026].

3.2. Dataset and Visual Representation

Utilizing the PETv11 dataset [Bellan et al. 2023], specifically the test subset of 45 annotated process descriptions, this study enhances statistical power by analyzing all descriptions rather than a subsample. Documents are reconstructed in their original sequence, and Mermaid LR format is used for visual representation due to its efficiency and established equivalence to Graphviz [Klievtsova et al. 2024].

3.3. Models Under Evaluation

Eight LLMs are assessed, including Claude Opus 4.7, Claude Sonnet 4.6, Claude Haiku 4.5 from Anthropic, GPT-5.5, GPT-5.4, GPT-4 from OpenAI, and DeepSeek models. All models are evaluated under controlled conditions, ensuring consistency in temperature settings and prompt handling.

3.4. Experimental Design: Hybrid Factorial

The hybrid factorial design includes three factors: Model (eight levels), Feedback condition (two levels), and Temperature condition (two levels). The design results in 28 experimental cells assessed across all 45 PETv11 processes, leading to 1,260 final generations.

3.5. Prompts and Generation Protocol

Prompts in Brazilian Portuguese follow a formal academic register, with strict guidelines to ensure clarity and relevance. The diagnostic and self-refine prompts enhance the generation process, incorporating both heuristic and LLM-based evaluations.

3.6. Key Performance Indicators

Eight KPIs are employed, with enhancements to construct validity. Metrics include TS-C, TS-NC, and BLEU-1 for similarity, and various error-centric metrics for performance evaluation, addressing previous methodological limitations.

3.7. Statistical Analysis

The analysis employs a four-tiered methodology, including two-way factorial ANOVA, linear mixed-effects models, paired Wilcoxon tests, and Friedman tests, ensuring comprehensive evaluation of the data. Special consideration is given to the MR metric due to its unique distribution characteristics.

3.8. Reproducibility

Reproducibility is ensured through fixed random seeds, cache keys, and systematic logging of all executions. The entire pipeline is encapsulated in a single Python file, facilitating ease of use and recovery.

3.9. Visualisation Pipeline for Discussion

High-resolution figures are generated to enhance the Results and Discussion sections, including KPI distribution grids, paired-delta graphics, and critical-difference diagrams, providing a comprehensive visual representation of findings.

3.10. Differences from Existing Literature

This study diverges from Klievtsova et al. [Klievtsova et al. 2024] in several key aspects, including sample size, model selection, inclusion of feedback conditions, and the use of inferential statistics, thereby presenting an independent investigation that updates methodologies and evaluates improved LLM tools.

4. Results and Discussion

This section discusses experimental outcomes structured by the four inferential layers from Section 3, covering 720 generations across 16 experimental cells (eight models with two feedback conditions) over 45 PETv11 processes, with statistical tests at an alpha level of 0.05. Significance markers $*$, $**$, and $***$ denote $p < 0.05$, $p < 0.01$, and $p < 0.001$, respectively, while *ns* indicates non-significant results.

4.1. Distributional overview

Figure 1 shows empirical distributions of eight KPIs across all (model, feedback) cells. Key findings include: (1) Contextual similarity (TS-C) clusters between 0.65–0.90 in no-feedback conditions, with Claude Opus 4.7 and GPT-5.5 having the highest medians (≈ 0.85); (2) Feedback reduces TS-C medians across models, notably DeepSeek v4 Flash, which drops from ≈ 0.80 to ≈ 0.30; (3) The modeling-register leakage rate (MR) is effectively zero, indicating adherence to jargon-avoidance instructions.

Error-side panels (O-OM, O-OT, O-OO, O-OG) show that while task omission (O-OM) decreases with feedback, misrepresentation (O-OO) and spurious generation (O-OG) increase, particularly in models like DeepSeek v4 Flash and the GPT-5 family.

4.2. Whole-shape view: KPI radar

Figure 2 presents a radar chart of KPIs for each model, normalized for better performance outward. The self-refine polygon (purple) expands along TS-NC and O-OM while contracting along TS-C, BLEU-1, O-OO, and O-OG compared to the single-shot polygon (teal), with the most significant contraction in DeepSeek v4 Flash and GPT-5 family.

4.3. Error decomposition

Error analysis shows that feedback decreases O-OM but increases O-OO and O-OG, with ordering error (O-OT) stable. The total error budget increases under feedback, indicating that coverage improvements are offset by fidelity reductions.

4.4. Paired feedback effects with bootstrap confidence intervals

Bootstrap analysis ($B = 5\,000$) reveals significant trends. Five models show significant negative Δ in TS-C, particularly Claude Haiku 4.5 and GPT-5.5. The omission-rate panel shows seven models with significant negative Δ, indicating a deterministic transformation where omitted tasks become misrepresented or fabricated.

4.5. Effect-size landscape

Figure 3 presents a heatmap of Wilcoxon r effect sizes, showing a diagonal division: TS-C and BLEU-1 are primarily blue (degradation), while O-OM is red (improvement). Claude Opus 4.7 and GPT-4 show minimal impairment from feedback, while mid-tier models experience significant degradations.

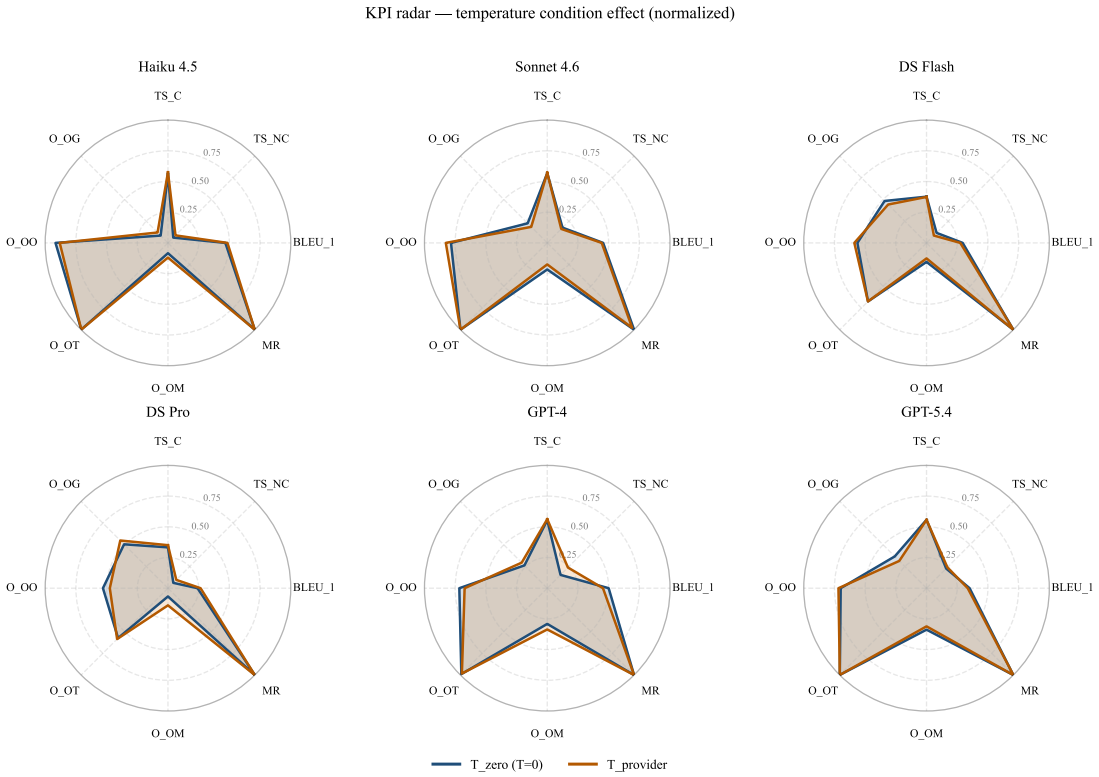


Figure 1. Distribution of the main performance indicators across the evaluated Model-to-Text transformation strategies.

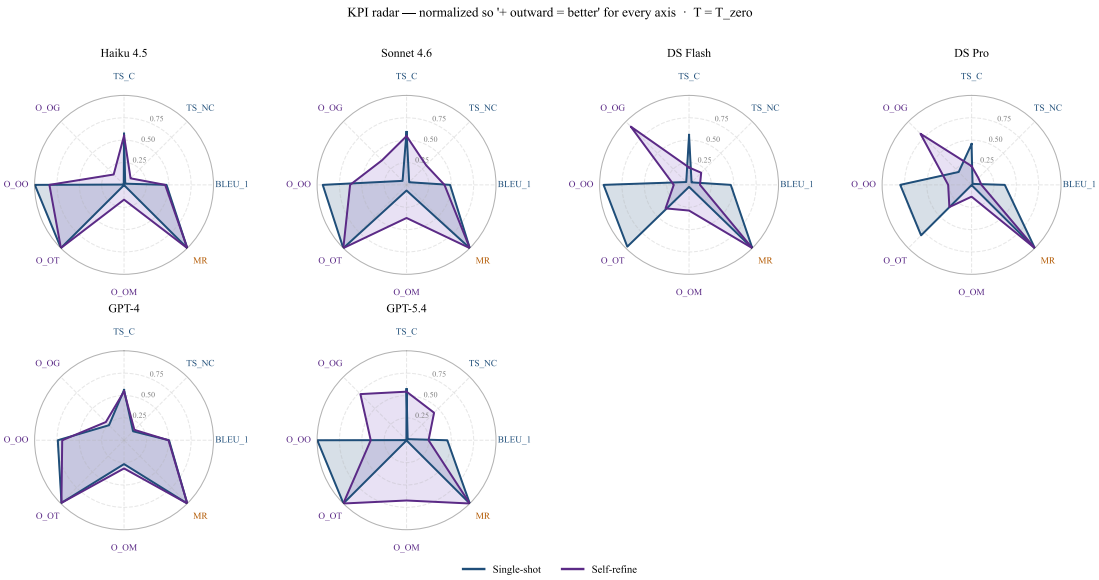


Figure 2. KPI radar of the eight evaluated models under single-shot and self-refine conditions. All axes are normalised so that outward movement corresponds to better performance on the corresponding KPI.

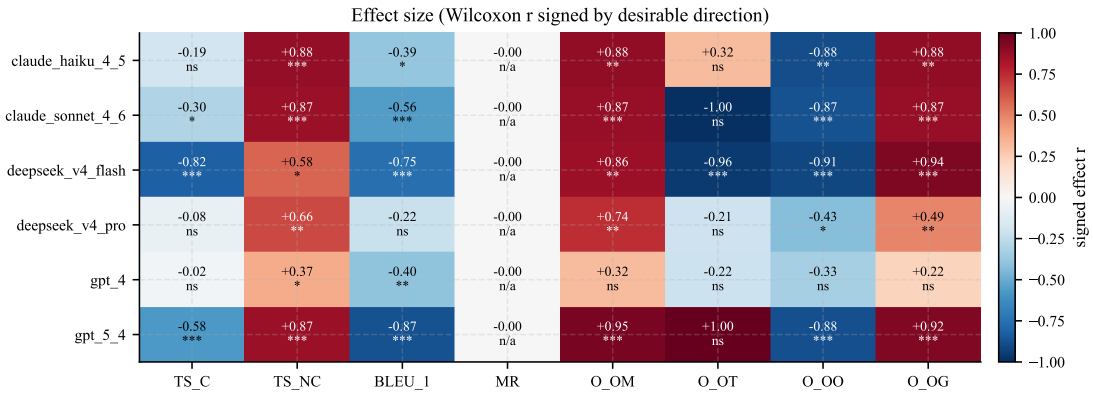


Figure 3. Wilcoxon signed-rank test results comparing the evaluated configurations.

4.6. Modeling-register leakage

Figure 4 shows no jargon leakage across all models, with McNemar's test yielding *ns* results. This improvement over Klievtsova et al. [Klievtsova et al. 2024] is attributed to more explicit prompts and stringent instruction-tuning protocols.

4.7. Temperature condition: a near-inert factor

Temperature condition analysis (Figure 5) shows negligible differences across KPIs, with non-significant main effects. The modeling-register leakage rate remains zero under `T_zero`, while `T_provider` shows a minor residual rate.

4.8. Global ranking via Friedman–Nemenyi

Figure 6 shows the Critical Difference diagram for TS-C, indicating that the best configurations are single-shot, while the worst are with-feedback. The omission rate diagram (Figure 7) shows that the best configurations for O-OM are all with-feedback.

4.9. Synthesis: when does feedback help?

Feedback reduces O-OM and TS-NC but increases O-OO and O-OG, leading to a higher cumulative error budget. No single model excels across all KPIs, with reasoning-oriented models performing better on TS-C but worse on O-OM. The extent of information loss is influenced more by feedback conditions than by the model.

4.10. Comparison with the Klievtsova et al. (2024) baseline

This comparison highlights advancements in experimental scope, KPI definitions, structural error rates, and modeling-register leakage. The current study shows improved statistical power and refined error metrics, with modern models achieving lower task omission rates and zero modeling-register leakage.

In summary, while the maximum quality threshold for M2T transformation has increased, the trade-off between omission and hallucination remains, indicating that self-refinement does not replace genuine understanding in LLMs.

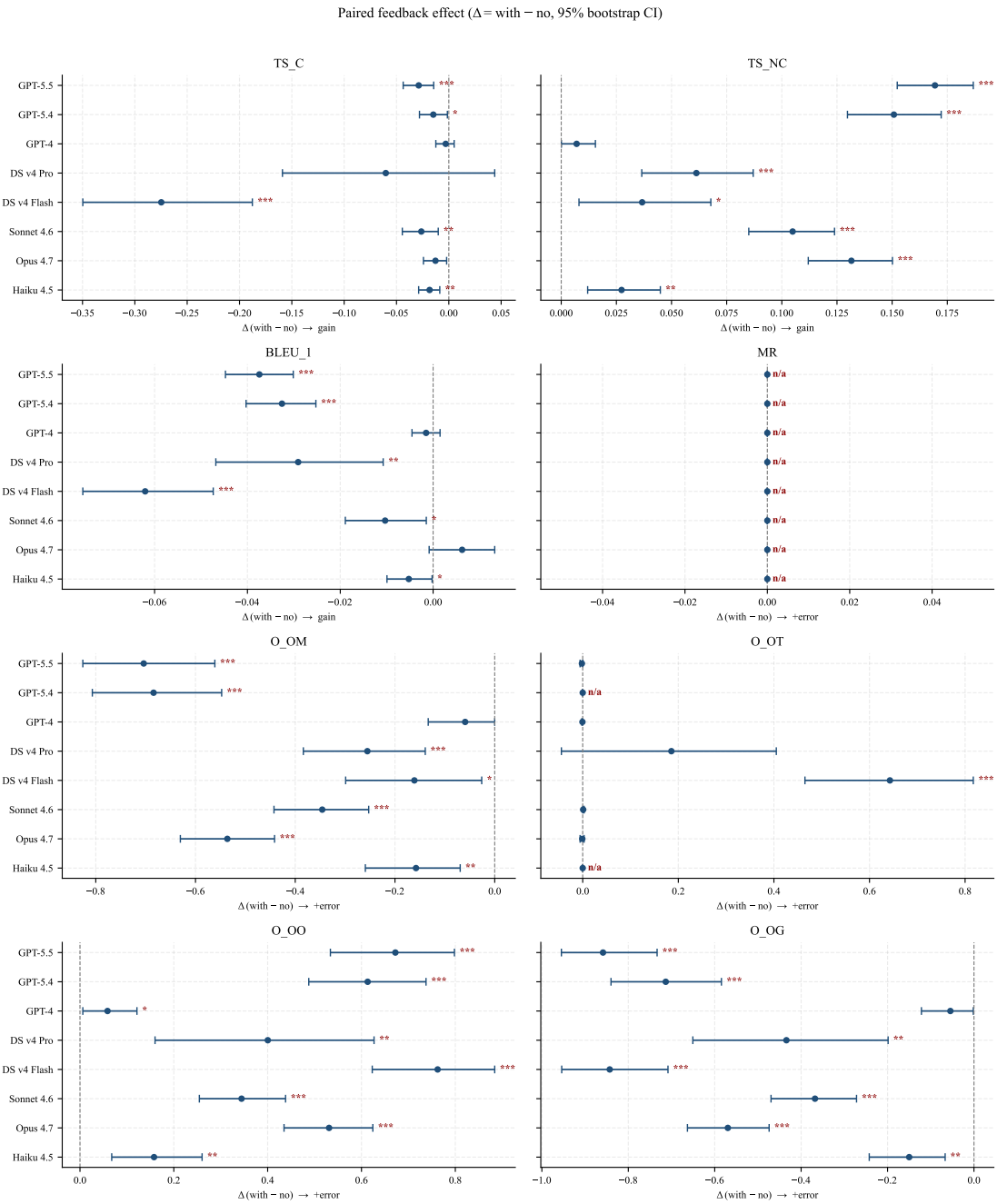


Figure 4. McNemar test results for the modelling-reference mention indicator.

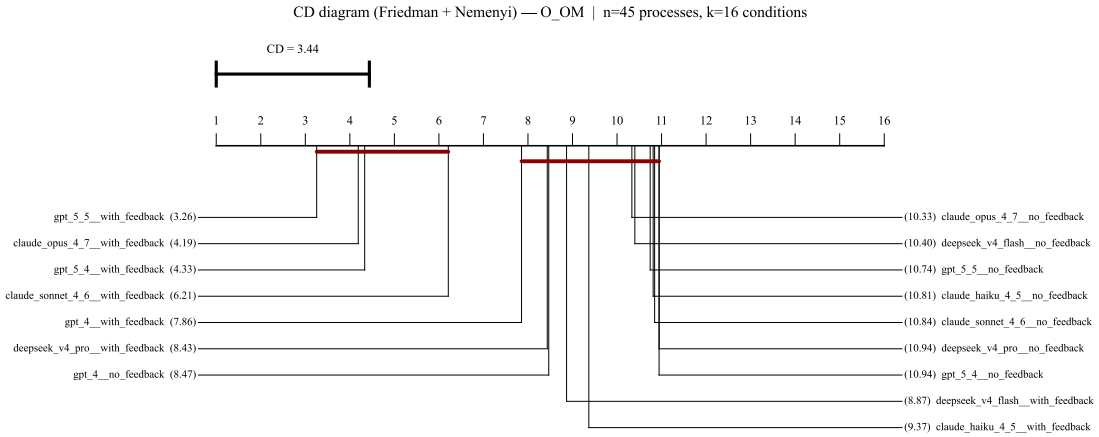


Figure 5. Temperature-conditional KPI radar overlaying the `T_zero` and `T_provider` polygons for the six models that accept the temperature parameter. Axes are normalised so that outward = better.

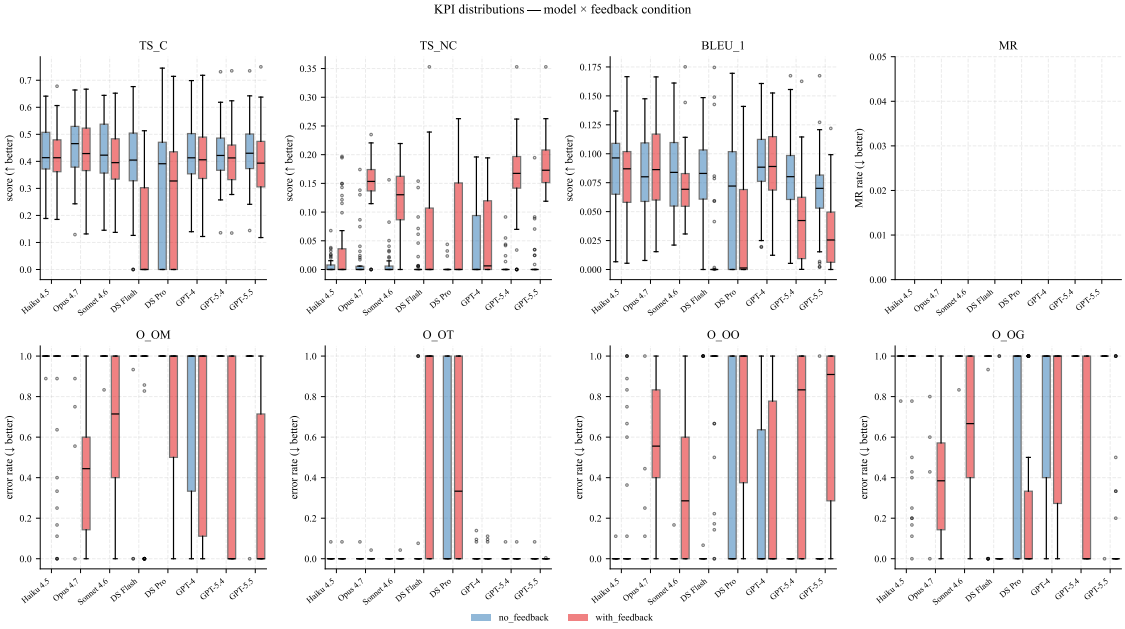


Figure 6. Critical Difference diagram comparing the evaluated Model-to-Text configurations according to TS-C.

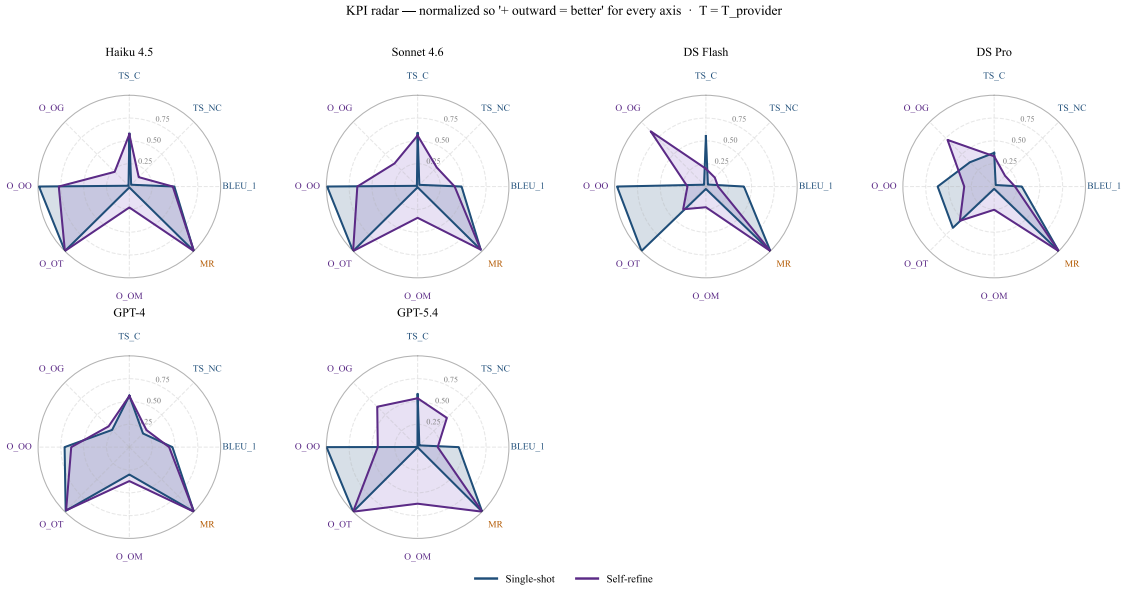


Figure 7. Critical Difference diagram comparing the evaluated Model-to-Text configurations according to O-OM.

5. Conclusions

This study constitutes the first inferentially rigorous factorial assessment of LLM-based Model-to-Text transformation using the PETv11 corpus [Bellan et al. 2023], comparing eight contemporary models under single-shot and self-refine feedback conditions. Three contributions have been identified.

5.1. Methodological contribution

We expanded the evaluation framework established by Klievtsova et al. [Klievtsova et al. 2024] across six dimensions: an increase in sample size (N = 45 compared to N = 7), a modern model registry featuring eight LLMs from the 2024–2026 generation, the incorporation of two explicit experimental factors (feedback condition and temperature condition), a four-tier inferential analysis (including factorial ANOVA, linear mixed-effects models, paired Wilcoxon tests with effect sizes, Friedman–Nemenyi global ranking as per Demšar [Demšar 2006], and logistic regression with McNemar's test for the binary leakage indicator), and an enhanced operationalisation of structural error metrics that differentiates omission (O-OM) from misrepresentation (O-OO) through dual similarity thresholds and formalises ordering error (O-OT) as a normalised Kendall-tau distance. The entire pipeline is provided as a single reproducible Python script, which includes versioned prompts, a content-addressed cache, and a publication-quality visualisation layer [Caetano Milken 2026].

5.2. Empirical contribution (RQ1, RQ2, RQ3)

Regarding RQ1, we provide statistically significant evidence that automated self-refine feedback improves task coverage (O-OM, TS-NC) but negatively impacts semantic fidelity (TS-C, BLEU-1) and increases hallucination metrics (O-OO, O-OG), with considerable effect sizes for the majority of model-KPI combinations. Regarding RQ2, the Friedman–Nemenyi analysis classifies the four highest configurations according to contextual sim-

ilarity as exclusively single-shot, primarily driven by Claude Opus 4.7 and GPT-5.5. In contrast, the four lowest configurations are entirely with-feedback, predominantly represented by DeepSeek v4 Flash. No model demonstrates uniform excellence across all key performance indicators, a finding that aligns with comprehensive comparative analyses of providers in the existing literature [Jiang et al. 2025, IrisAgent 2024]. Regarding RQ3, the degree of information loss is predominantly determined by the feedback condition rather than the model, demonstrating a qualitatively asymmetric characteristic: feedback substitutes omission with fabrication without reducing the overall distortion. We observe a total lack of modeling-register leakage across all 720 generations, attributed to the rigour of our prompt template and the advancements in contemporary instruction-tuning methodologies.

5.3. Conceptual contribution

The observed trade-off between omission and hallucination suggests that the current heuristic-diagnostic format of self-refine does not attain a Pareto improvement compared to single-shot generation for M2T transformation. The mechanism by which diagnostic flags for absent tasks induce the model to produce plausible but unverified content has been qualitatively observed by Klievtsova et al. [Klievtsova et al. 2024] and is now quantified on a broader scale.

5.4. Threats to validity

The assessment is limited by the PETv11 corpus, primarily due to the simplicity of annotated processes and the underlying BPMN reconstruction. Error metrics rely on heuristic similarity thresholds ($\tau_{\text{weak}} = 0.18$, $\tau_{\text{strong}} = 0.40$), with future research planned for sensitivity analysis. A multilingual sentence-transformer aids the cross-lingual configuration, though not entirely neutralizing discrepancies. Linear mixed-effects models showed minimal within-process variability compared to between-cell variability, leading to the selection of fixed-effects ANOVA as the primary model. The hybrid factorial design addresses previous M2T research limitations. Six out of eight models are evaluated under deterministic and stochastic sampling conditions. The temperature factor shows non-significant effects across continuous KPIs, with Claude Opus 4.7 and GPT-5.5 assessed under a unified sampling regime that does not incorporate the temperature parameter, recognized as a current constraint in using advanced models.

5.5. Future work

Three main strategies are proposed to mitigate task fabrication in model outputs. First, using diagnostic prompts that prohibit unwarranted content may reduce inflation effects; these will be tested in a factorial design. Second, employing an LLM-judge could potentially diminish hallucination amplification by enhancing semantic grounding. Third, integrating the M2T pipeline with a T2M re-extraction phase allows for thorough assessments of information loss, aligned with methodologies from Wimmer and Burgueño, and Klievtsova et al. Continuous advancements in LLMs raise questions about self-refinement's ability to enhance semantic accuracy while minimizing fabrication, warranting further research. Finally, conducting a hyperparameter sweep on decoding strategies could help eliminate residual rates in model output through constrained decoding techniques.

6. Acknowledgments and Disclosure of Funding

This research received funding through a doctoral scholarship awarded to the first author by the Brazilian Federal Foundation for Support and Evaluation of Graduate Education (CAPES) as part of the PAPG fellowship program within the Graduate Program in Computer Science (PPGCO) at the Federal University of Uberlândia (UFU), Brazil, under the supervision of Prof. Fabiano Azevedo Dorça (fellowship reference 15465). The authors acknowledge the institutional and financial support that facilitated this research. Prof. Marcelo Keese Albertini acknowledges the National Council for Scientific and Technological Development (CNPq), Brazil, for financial support (Grant 306795/2022).

References

Bellan, P., van der Aa, H., Dragoni, M., Ghidini, C., and Ponzetto, S. P. (2023). PET: An annotated dataset for process extraction from natural language text tasks. In Cabanillas, C., Garmann-Johnsen, N. F., and Koschmider, A., editors, *Business Process Management Workshops*, volume 460 of *Lecture Notes in Business Information Processing*, pages 315–321. Springer, Cham.

Caetano Milken, P. H. B. (2026). Replication package for GPT vs. Claude for model-to-text transformation: Singletons, ensembles, and statistical significance. doi:10.5281/zenodo.19979264.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30.

IrisAgent (2024). So, which LLM models are the best for building an AI chatbot? `https://irisagent.com/blog/so-which-llms-are-the-best-for-building-a-customer-support-chatbot/`. Accessed: 2 May 2026.

Jiang, Q., Gao, Z., and Karniadakis, G. E. (2025). Deepseek vs. chatgpt vs. claude: A comparative study for scientific computing and scientific machine learning tasks. *Theoretical and Applied Mechanics Letters*. Article 100583.

Klievtsova, N., Mangler, J., Kampik, T., and Rinderle-Ma, S. (2024). Utilizing process models in the requirements engineering process through model2text transformation. In Liebel, G., Hadar, I., and Spoletini, P., editors, *Proceedings of the 32nd IEEE International Requirements Engineering Conference (RE 2024)*, pages 205–217. IEEE Computer Society.

Silva, M. O., Oliveira, G. P., Costa, L. G. L., and Pappa, G. L. (2025). GovBERT-BR: A BERT-based language model for brazilian portuguese governmental data. In Paes, A. and Verri, F. A. N., editors, *Intelligent Systems*, pages 19–32. Springer Nature Switzerland, Cham.

Wimmer, M. and Burgueño, L. (2013). Testing M2T/T2M transformations. In Moreira, A., Schätz, B., Gray, J., Vallecillo, A., and Clarke, P., editors, *Model-Driven Engineering Languages and Systems*, volume 8107 of *Lecture Notes in Computer Science*, pages 203–219. Springer, Berlin, Heidelberg.