Improving Verbatim Financial Causality Extraction with Supervised Fine-Tuning and Prompt Repetition

Sanae Attak, Mohammed Salah Chiadmi, Youssef Lamrani Alaoui
IFE-lab,LERMA, Mohammadia School of Engineers (EMI)
Mohammed V University in Rabat, Morocco
sanae.attak@research.emi.ac.ma, schiadmi@emi.ac.ma, lamrani@emi.ac.ma

Abstract

This paper investigates the application of generative Large Language Models (LLMs) for strict verbatim span extraction. We evaluate our methodology within the FinCausal 2026 shared task. Because generative LLMs optimize next-token probability rather than strict boundaries, they naturally suffer from over-generation and boundary drift in extraction tasks. To address this, we introduce a generalized structural training constraint, extending prompt repetition from a purely inference-time heuristic to a training-time supervision framework. By incorporating duplicated prompts directly into Supervised Fine-Tuning (SFT), we hypothesize that this encourages the model to internalize a form of unidirectional cross-reading behavior, leading to stronger alignment between generated spans and the source context for exact extraction. Evaluating on open-weights (`Qwen2.5-14B-Instruct-1M`) and proprietary (`GPT-4.1-Nano`) architectures, we find this soft attention constraint improves Exact Match scores for open models and helps balance cross-lingual performance disparities. Conversely, the proprietary model exhibited sensitivity to prompt duplication, achieving its highest score without repetition. Ultimately, our deterministic SFT approach secured 4th place in the Spanish subtask (4.73) and 6th place in the English subtask (4.70), indicating the viability of structurally simple, natively fine-tuned models compared to complex multi-stage pipelines.

Keywords: Causality Extraction, Generative LLMs, Prompt Repetition, Supervised Fine-Tuning (SFT)

1. Introduction

The increasing complexity of financial documents necessitates advanced methodologies to extract and analyze causality within such texts. The FinCausal 2026 shared task introduces a generative Question-Answering (QA) framework for detecting causal relationships in financial disclosures (Moreno-Sandoval et al., 2026). The task requires models to process abstractive questions regarding causes or effects and answer by extracting verbatim spans directly from the source text.

This generative formulation presents a structural conflict: generative models are inherently designed to synthesize and paraphrase information (Chrysostomou et al., 2024), yet the task evaluation strictly penalizes any generative deviation or hallucination from the original text. Because generative LLMs optimize next-token probability rather than strict boundaries, they naturally suffer from over-generation and boundary drift in extraction tasks. To address this, our study investigates the effectiveness of Supervised Fine-Tuning (SFT) coupled with targeted prompting techniques specifically Prompt Repetition to constrain generative models into performing exact span extraction.

In this study, we investigate the following Research Questions (RQs):

- **RQ1:** Can structurally simple SFT constrain generative LLMs to act as verbatim extractors without relying on multi-stage pipelines?
- **RQ2:** Does extending prompt repetition from an inference trick to a training paradigm improve extraction fidelity?
- **RQ3:** How do open-weight models compare to proprietary models under these extraction constraints across different languages?

Contributions. While prompt repetition has primarily been studied as an inference-time technique, we explore its integration directly into the supervised fine-tuning process for causal extraction tasks. Specifically, this paper makes three core contributions:

1. We propose a generalized structural training constraint for generative extraction. By extending prompt repetition from an inference-time heuristic to a training-time supervision methodology, we hypothesize that generative models internalize context anchoring for strict span extraction.
2. We provide a cross-lingual evaluation (English and Spanish) indicating that this method reduces language-specific extraction biases.
3. Our experiments indicate that mid-scale open-weight models (`Qwen2.5-14B-Instruct-1M`) can achieve competitive performance relative to proprietary models under strict extraction constraints.

Proceedings of the 7th Financial Narrative Processing Workshop @ LREC 2026, pages 152–159

2. Related Work

Causal relationship extraction remains a persistent challenge in financial NLP. Earlier FinCausal shared tasks (Mariko et al., 2022) predominantly framed the problem as a sequence-tagging task, utilizing BIO tagging schemes via token classifiers like BERT or BioBERT to identify causal spans (Saha et al., 2022; Lyu et al., 2022). While token classifiers and pointer networks perform well on small datasets, they often struggle when causal chains span multiple disconnected sentences. Consequently, the field recently shifted toward generative Q&A frameworks (Moreno-Sandoval et al., 2026). To combat the hallucinations inherent in generative architectures, prior approaches utilized complex lexically constrained decoding (Ghosh and Naskar, 2022), explicit pointer-generator copy mechanisms, or passed extractive outputs through LLMs for refinement (Trivedi et al., 2025). Our objective is to approach the strict verbatim fidelity of these classical copy mechanisms natively through generative SFT, without relying on modified decoding algorithms.

2.1. Attention Constraints and Prompt Repetition

Prior work has shown that repeating prompts during inference improves extraction accuracy by reinforcing attention alignment in causal language models (Leviathan et al., 2025). However, this technique has primarily been explored as an inference-time heuristic. In this study, we extend this idea by integrating prompt repetition directly into the supervised fine-tuning stage, enabling the model to internalize cross-reading be- havior during training.

Furthermore, evaluations from the FinCausal 2025 shared task demonstrated that standard prompt optimization and few-shot learning are fundamentally insufficient to prevent generative models from hallucinating during causal extraction (Niess et al., 2025). As noted by (Niess et al., 2025),fine-tuning generative architectures is absolutely essential for minimizing boundary drift and enforcing strict extraction constraints. Our work builds directly upon this premise: we not only adopt Supervised Fine-Tuning (SFT) as a baseline necessity,but we further constrain the generative process by introducing prompt repetition as a structural soft-attention mechanism during the SFT phase itself.

3. Methodology

3.1. Dataset and Preprocessing

This study utilizes the official FinCausal 2026 dataset (Moreno-Sandoval et al., 2026). The dataset comprises financial reports sourced from the UK and Spain, providing 2,000 annotated training samples per language (4,000 samples in total).

To validate our models and conduct internal ablation studies, we employed a two-stage data utilization strategy. First, we merged and randomly split the dataset into an internal Training Set (3,600 samples) and a held-out Development Set (400 samples: 200 English and 200 Spanish). This internal split was exclusively utilized to independently compute Exact Match (EM) and Semantic Answer Similarity (SAS) metrics for our comparative analyses. Second, for the final official blind test submissions, the models were retrained on the entirety of the provided dataset (all 4,000 samples) to maximize domain exposure. Prior to tokenization, all texts underwent a standard normalization pipeline (lowercasing and whitespace removal).

3.2. Internalizing Attention via Prompt Repetition

Unlike prior work which applies prompt repetition only at inference (Leviathan et al., 2025), we incorporate the repeated prompt structure directly during SFT. Generative LLMs, built upon decoder-only transformer architectures (Brown et al., 2020), utilize a lower-triangular causal attention mask to preserve the auto-regressive property; meaning a token t_i can only attend to previous tokens $t_{\leq i}$ (Vaswani et al., 2017). In a standard prompt (*Context* + *Question*), the question tokens can attend to the context, but the context tokens cannot attend to the question. By duplicating the input ($Context_1$ + $Question_1$ + $Context_2$ + $Question_2$), we fundamentally alter the attention graph: the tokens in $Context_2$ are now positioned after $Question_1$, allowing them to compute dense attention over both the source text and the task specification simultaneously. This restructuring of the input sequence may mitigate the limitations imposed by causal masking by allowing later context tokens to attend to both the source text and task instruction. We hypothesize that this structural duplication encourages stronger anchoring of the generated span to the original context.

3.3. Experimental Setup

Exact prompt templates utilized for the experiments are provided in Section A.

Open-Weights Configuration We utilized the `Qwen2.5-7B-Instruct-1M` and `Qwen2.5-14B-Instruct-1M` checkpoints. Models were loaded using 4-bit quantization via `unsloth`. We applied Low-Rank Adaptation (QLoRA) targeting all linear modules (`q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, `down_proj`) with a rank of $r = 64$ and $\alpha = 16$.

The models were trained for exactly 1 epoch using the 8-bit AdamW optimizer with a linear learning rate schedule peaking at 2×10^{-4}, a warmup ratio of 0.03, and an effective batch size of 8. Maximum sequence length was set to 2048, and sequence packing was explicitly disabled (`packing = False`) to maintain the structural integrity of the duplicated contexts.

Proprietary Configuration For proprietary comparisons, we applied SFT to the `gpt-4.1-nano` checkpoint via the official API for 3 epochs.

Inference Parameters To isolate the impact of our training methodology, decoding parameters were set to pure greedy decoding. For all architectures, generation was deterministic (`temperature` $= 0.0$ and `top_p` $= 1.0$). Inference for the open-weights models was executed on a single NVIDIA L40S 48GB GPU. While Prompt Repetition inherently increases the input sequence length, we observed only a moderate computational overhead during the prefill phase.

3.4. Evaluation Metrics and Statistical Testing

To comprehensively evaluate performance, we utilize three distinct metrics. **Exact Match (EM)** measures whether the predicted answer exactly matches the gold reference span. Both the generated predictions and the withheld gold references undergo the exact same normalization pipeline (lowercasing, diacritic stripping, and whitespace trimming) prior to EM calculation. **Semantic Answer Similarity (SAS)** evaluates the semantic equivalence of the answers (Risch et al., 2021). To align precisely with the official evaluation framework established by the FinCausal 2025 organizers (Moreno-Sandoval et al., 2026), we extract 768-dimensional text embeddings using the cross-lingual `paraphrase-multilingual-mpnet-base-v2` Sentence Transformer model and compute their pairwise cosine similarity. Finally, the **LLM-as-a-Judge** assesses answer adequacy on a scale of 1 to 5. We conducted paired bootstrap resampling to verify that improvements in Exact Match are statistically significant ($p < 0.05$).

4. Results

4.1. Validation Set Performance

Because our internal development split differs from the official 2025 test set, directly comparing these scores against historical extractive baselines would be methodologically unsound. Instead, we use this held-out set strictly as an internal ablation to quantify the absolute impact of Prompt Repetition on textual fidelity.

As shown in Table 1, incorporating Prompt Repetition during SFT yields a consistent improvement in Exact Match for the open-weight models. Internalizing cross-reading behavior increases the `Qwen2.5-14B-Instruct-1M` model's EM score from 0.8200 to 0.8550 in English. This confirms that generative models, when constrained via SFT and prompt duplication, effectively reduce boundary errors compared to standard zero-shot prompting. Across both model scales and languages, prompt repetition consistently improves Exact Match performance for open-weight architectures, suggesting that simple structural constraints during SFT can improve verbatim span fidelity.

4.2. Official Blind Test Results

For the final evaluation on the official blind test set (where gold references are withheld), our models were retrained on the full 4,000-sample dataset. Because independent calculation of EM and SAS is impossible for these final submissions, Table 2 reports the official standardized LLM-as-a-judge scores provided by the organizers.

5. Discussion and Analysis

1. Language Balance via Prompt Repetition: An observation from our internal evaluation is the impact of training-time Prompt Repetition on cross-lingual disparities. In the baseline "Simple" setting, the `Qwen2.5-14B-Instruct-1M` model exhibits a slight bias toward Spanish (EM of 0.8350 in ES vs. 0.8200 in EN). Applying the Repeated technique appears to balance this performance (0.8550 EN and 0.8450 ES). The consistent improvements observed for the open-weight models suggest that structural prompt duplication may help stabilize span boundaries during generation, particularly in tasks requiring strict verbatim extraction. This observation reinforces the hypothesis that simple structural constraints applied during supervised fine-tuning can improve the reliability of generative models in high-precision extraction tasks.

2. The Conflict Between RLHF and Structural Constraints: While open-weights models showed benefits from Prompt Repetition, the proprietary `GPT-4.1-Nano` exhibited divergent behavior. On the internal dev set (Table 1), repetition maintained raw Exact Match extraction boundaries. Yet, under the official blind test evaluated by the LLM-as-a-judge metric (Table 2), applying Prompt Repetition resulted in a noticeable degradation (from $\sim$4.73 down to $\sim$3.98). We posit this reveals a

Generative Configuration	English (EN)		Spanish (ES)	
	SAS	EM	SAS	EM
Qwen2.5-7B-Instruct-1M (Simple Prompt)	0.9667	0.8000	0.9699	0.7600
Qwen2.5-7B-Instruct-1M (Repeated Prompt)	0.9729	0.8300	0.9723	0.7950
Qwen2.5-14B-Instruct-1M (Simple Prompt)	0.9626	0.8200	0.9749	0.8350
Qwen2.5-14B-Instruct-1M (Repeated Prompt)	**0.9730**	**0.8550***	0.9746	0.8450
GPT-4.1-Nano (Simple Prompt)	0.9578	0.8050	0.9759	0.8450
GPT-4.1-Nano (Repeated Prompt)	0.9683	0.8000	**0.9787**	**0.8600***

Table 1: Semantic Answer Similarity (SAS) and Exact Match (EM) Results evaluated strictly on our 400-sample Internal Development Set. Because this data split differs from historical blind test sets, these scores serve specifically as an internal ablation to isolate the impact of Prompt Repetition. (*) denotes a statistically significant improvement over the Simple baseline for the respective model architecture ($p < 0.05$).

Model	Configuration	EN	ES
Open-Weights SFT (Single Model)			
Qwen2.5-7B-Instruct-1M	Simple Prompt	4.3120	4.0915
Qwen2.5-7B-Instruct-1M	Repeated Prompt	4.3500	4.3877
Qwen2.5-14B-Instruct-1M	Repeated Prompt	**4.6720**	**4.6740**
Proprietary SFT (Single Model)			
GPT-4.1-Nano	Zero-Shot	3.9540	3.9841
GPT-4.1-Nano	Repeated Prompt	3.9800	3.9940
GPT-4.1-Nano*	Simple Prompt	**4.7040**	**4.7396**
Ablation: Inference-Only Repetition			
GPT-4.1-Nano	Train Simple + Infer Rep.	4.6880	4.6143
Ablation: Complex Pipelines			
Qwen2.5-14B-Instruct-1M	Repeated + Ensemble	4.5800	4.5785
GPT-4.1-Nano	Simple + Ensemble	4.6660	4.6501
GPT-4.1-Nano	Simple + RAG (3-Shot)	4.6600	4.7117
GPT-4.1-Nano	Simple + GPT-4o Judge	4.2560	4.2445

Table 2: Official LLM-as-a-Judge performance on the FinCausal 2026 Blind Test Set (Scored 1 to 5). The ablations indicate that inference-only repetition and multi-stage pipelines (Ensembles, RAG, Correctors) generally resulted in lower scores compared to single-stage, natively fine-tuned models. (*) indicates the official submission.

potential conflict between structural training constraints and Reinforcement Learning from Human Feedback (RLHF). Heavily aligned models optimized for conversational naturalness may penalize or misinterpret highly unnatural, duplicated input structures during generative decoding. This indicates that structural prompting techniques effective on base-aligned open models may conflict with the safety alignment layers of proprietary models.

3. Training-Time vs. Inference-Time Repetition: To examine whether the performance gains stem from the training paradigm or merely from inference-time context duplication, we conducted a targeted ablation on the blind test set. When the proprietary model was fine-tuned on the *Simple* configuration but evaluated using the *Repeated* prompt during inference, its score decreased (from 4.7396 down to 4.6143 in Spanish). This supports the idea that forcing prompt repetition solely at inference introduces out-of-distribution formatting noise, and that the model benefits from internalizing the attention mechanism during the SFT phase.

4. The Limitations of Multi-Stage Pipelines: Recent NLP extraction tasks frequently deploy multi-stage pipelines. However, our ablations (Table 2) suggest these architectures can be less effective for strict verbatim extraction. Injecting dynamic context via Retrieval-Augmented Generation (RAG, utilizing 3-shot semantic retrieval for in-context examples) introduced external noise, slightly lowering the score. Similarly, utilizing a powerful meta-model (`GPT-4o`) as a post-generation corrector via strict formatting prompts caused a decrease in performance (from 4.7040 to 4.2560). Qualitative analysis indicated that the corrector model prioritized grammatical completeness over verbatim copying, disrupting the required span boundaries. Furthermore, Ensemble Voting (majority consensus across

5 high-temperature generations) introduced token-level variations that diluted the Exact Match consistency.

6. Error Analysis and Qualitative Comparison

To understand the mechanism by which Prompt Repetition improves Exact Match (EM) scores, we conducted a qualitative analysis of the residual errors in the baseline models. As demonstrated in Table 4 (Appendix C), the generative formulation introduces specific hallucination patterns.

For instance, consider a boundary error hallucination where the baseline *Simple Prompt* misses the exact starting boundary by appending an introductory connector (e.g., extracting "***As** external threats become more sophisticated*" instead of "*external threats become more sophisticated*"). Furthermore, when faced with causal chains, the baseline model occasionally truncates the extraction, or exhibits generative stutters (Example 1) and mid-generation stops (Example 4).

As shown in Table 4 (Appendix C), internalizing the *Repeated Prompt* during Supervised Fine-Tuning acts as an attention constraint, encouraging the model to respect verbatim spans and reducing these generative tendencies across the tested open-weights architectures. The high verbatim copy-rate achieved by the Repeated configuration supports our hypothesis that the SFT process enforces a unidirectional "cross-reading" mechanism.

7. Conclusion

This study investigated the effectiveness of a structural training constraint combining Supervised Fine-Tuning with prompt repetition for strict span extraction. We demonstrated that extending prompt repetition from an inference-time heuristic to a training-time supervision approach acts as a context anchoring mechanism, allowing generative causal LLMs to internalize cross-reading behaviors. The results suggest that structurally simple, natively fine-tuned generative models can perform reliable verbatim extraction, neutralizing language-specific biases to achieve balanced multilingual performance without relying on complex multi-stage pipelines. Although evaluated in the financial domain, the proposed structural constraint may extend to other high-precision extraction tasks where strict verbatim copying is required.

8. Limitations and Future Work

Our approach demonstrates practical empirical performance but also highlights several avenues for future research. The models were fine-tuned on a relatively small, domain-specific dataset (3,600 bilingual samples), which was effective in this context; however, extending structural training constraints to larger, multi-domain corpora would help assess broader generalization. While our study focused on the financial sector, adapting prompt repetition for verbatim extraction in other domains, such as biomedical relation extraction or legal evidence analysis, remains an open challenge.

The LLM-as-a-judge metric provides a useful assessment of answer adequacy but may introduce implicit alignment biases. Future work could explore hybrid evaluation frameworks to better relate these subjective judgments to deterministic metrics like Exact Match.

Computationally, prompt repetition imposes only minor overhead. As shown in Table 3 (Appendix B), it increased training time by 0.02 hours, added a negligible $+0.0009$ kg of CO_2 emissions, and caused only a slight rise in peak VRAM (38.67 GB vs. 39.31 GB), indicating feasibility for mid-scale models.

Finally, while our results show a statistically significant improvement in Exact Match scores (+3.5%), achieving absolute boundary determinism remains a challenge for generative architectures, even under strict SFT constraints. Our interpretation that prompt repetition may promote a unidirectional “cross-reading” mechanism is supported by behavioral evidence, such as reduced generative stutters and boundary offsets. Nonetheless, a formal extraction and visualization of multi-head attention maps could provide further validation and represents a promising direction for future interpretability research.

9. Acknowledgements

We thank the organizers of the FinNLP and FNP workshops for their dedication to the financial NLP community. We acknowledge the creators of the FinCausal 2026 dataset (Moreno-Sandoval et al., 2026), whose bilingual corpus enabled the cross-lingual analyses presented in this research.

10. Bibliographical References

Tom B Brown, Benjamin Mann, Nick Ryder, et al. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

George Chrysostomou, Zhixue Zhao, Miles Williams, and Nikolaos Aletras. 2024. Investigating hallucinations in pruned large language models for abstractive summarization. *Transactions of the Association for Computational Linguistics*, 12:1163–1181.

Sohom Ghosh and Sudip Kumar Naskar. 2022. Lipi at fincausal 2022: Mining causes and effects from financial texts. In *Proceedings of the 4th Financial Narrative Processing Workshop*, pages 121–123.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2025. Prompt repetition improves large language models. *arXiv preprint arXiv:2512.14982*.

Zhiheng Lyu et al. 2022. Dcu-lorcan at fincausal 2022. In *Proceedings of the 4th Financial Narrative Processing Workshop*.

Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. The financial causality extraction shared task (fincausal 2022). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107.

Antonio Moreno-Sandoval, Jordi Porta, Yanco Torterolo, Alexia Stanescu, Melina Chatzi, and Sofía Roseti. 2026. The financial document causality detection shared task (fincausal 2026). In *Proceedings of the 7th Financial Narrative Processing Workshop (FNP 2026) at LREC 2026*, Palma de Mallorca, Spain. ELRA. To appear.

Georg Niess, Houssam Razouk, Stasa Mandic, and Roman Kern. 2025. Addressing hallucination in causal q&a: The efficacy of fine-tuning over prompting in llms. In *Proceedings of the Joint Workshop of the 9th FinNLP, 6th FNP, and 1st LLMFinLegal*, pages 253–258.

Julian Risch, Timo Möller, Julian Gutsch, and Malte Pietsch. 2021. Semantic answer similarity for evaluating question answering models. In *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, pages 149–157.

Anik Saha, Jian Ni, Oktie Hassanzadeh, et al. 2022. Spock at fincausal 2022: Causal information extraction using span-based and sequence tagging models. In *Proceedings of the 4th Financial Narrative Processing Workshop*, pages 108–111.

Avinash Trivedi, Gauri Toshniwal, Sivanesan Sangeetha, and S.R. Balasundaram. 2025. Sarang at fincausal 2025: Contextual qa for financial causality detection combining extractive and generative models. In *Proceedings of the Joint Workshop of the 9th FinNLP, 6th FNP, and 1st LLMFinLegal*, pages 242–247.

Ashish Vaswani, Noam Shazeer, Niki Parmar, et al. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

11. Language Resource References

Language Resources

Antonio Moreno-Sandoval, Yanco Amor Torterolo Orta, Maria Alexia Stanescu, and Melina Chatzi. 2026. The Financial Document Causality Detection Shared Task (FinCausal 2026): Dataset.

A. Prompt Templates

We detail the exact prompt structures utilized during both Supervised Fine-Tuning and inference.

Simple Prompt (Zero-Shot SFT Baseline):

```
System:
You are a financial expert participating in FinCausal 2026.
Task: Extract the exact cause or effect from the provided financial text.
Rules:
1. The answer must be a VERBATIM extraction.
2. If the text contains a complex causal chain,extract the full relevant se-
quence.
3. Do not add introductory words.

User:
Context: {context}
Question: {question}
```

Repeated Prompt (Cross-Reading Configuration):

```
System:
You are a financial expert participating in FinCausal 2026. Task: Extract
the exact cause or effect from the provided financial text.
Rules:
1. The answer must be a VERBATIM extraction.
2. If the text contains a complex causal chain,extract the full relevant se-
quence.
3. Do not add introductory words.

User:
Context: {context}
Question: {question}
Context: {context}
Question: {question}
```

B. CodeCarbon Environmental Tracking

Model	Setting	Time (h)	CO_2 (kg)	Peak VRAM (GB)
Qwen2.5-14B-Instruct-1M	Simple	0.20	0.0051	38.67
Qwen2.5-14B-Instruct-1M	Repeated	0.22	0.0060	39.31

Table 3: Empirical computational cost and environmental impact for fine-tuning on a single NVIDIA L40S (48GB) GPU. Total training time, peak VRAM utilization, and CO_2 emissions were directly tracked using the CodeCarbon library. Measurements were taken during a dedicated reproducibility run using identical hyperparameters, data, and hardware.

C. Qualitative Examples

Target Span	Simple Prompt Baseline	Repeated Prompt
Type 1: Boundary Offset & Over-generation		
[Qwen2.5-14B-Instruct-1M - EN] the £9.4m improvement in underlying operating cash flows	the £9.4m improvement in underlying operating cash flows **offset by a £2.0m increase in outflows...**	**the £9.4m improvement in underlying operating cash flows**
[GPT-4.1-Nano - EN] external threats become more sophisticated, and the potential impact of service disruption increases	**As** external threats become more sophisticated, and the potential impact of service disruption increases	**external threats become more sophisticated, and the potential impact of service disruption increases**
Type 2: Generative Stutter & Typographical Drops		
[Qwen2.5-7B-Instruct-1M EN] Our business serving the grocery sector benefited from several new accounts although the additional business won, combined with a competitive marketplace	Our business serving the grocery sector benefited from several **new new** accounts	**Our business serving the grocery sector benefited from several new accounts although the additional business won, combined with a competitive marketplace**
[Qwen2.5-14B-Instruct-1M - ES] no incluyen intereses, dividendos, ganancias o pérdidas procedentes de venta de inversiones o de operaciones de rescate o extinción de deuda	no incluyen intereses, dividendos, ganancias o pérdidas procedentes de venta de inversiones o de operaciones de rescate o extinción de **deud**	**no incluyen intereses, dividendos, ganancias o pérdidas procedentes de venta de inversiones o de operaciones de rescate o extinción de deuda**
Type 3: Truncation (Premature Stops)		
[GPT-4.1-Nano - ES] parámetros como el suministro de materias primas, la utilización de los quemadores, los sensores instalados o el balance entre energía fósil y eléctrica pueden ser gestionados de una manera más rápida, moderna y eficiente	**parámetros como el s**	**parámetros como el suministro de materias primas, la utilización de los quemadores, los sensores instalados o el balance entre energía fósil y eléctrica pueden ser gestionados de una manera más rápida, moderna y eficiente**
[Qwen2.5-14B-Instruct-1M - EN] The key partnerships established with leading European manufacturers	**The key partnerships**	**The key partnerships established with leading European manufacturers**

Table 4: Qualitative comparison of extraction errors. A focused evaluation of six samples illustrates how Prompt Repetition reduces generative stutters, boundary misalignments, and premature truncations across multiple architectures without requiring an external corrector model.