Explaining Financial LLMs: An Attribution-Based Interpretability Study in Multilingual Table QA in Dutch and English

Amalia Stuger
University of Amsterdam
Amsterdam, Netherlands
amalia.stuger@student.uva.nl

Lucas Lageweg
University of Amsterdam
Amsterdam, Netherlands
l.lageweg@uva.nl

Fina Polat
University of Amsterdam
Amsterdam, Netherlands
f.yilmazpolat@uva.nl

Abstract

The reliable deployment of Large Language Models (LLMs) in critical sectors, especially involving structured input like tabular data, necessitates mechanisms for transparency and accountability. This paper investigates the interpretability of domain-specific LLMs applied to Table Question Answering (TQA) tasks in the financial domain. We conduct a comparative attribution study between domain-adapted and general-purpose LLMs in both Dutch and English. The analysis employs parallel datasets sourced from ConvFinQA. Utilizing Input $\times$ Gradient attribution, we segment input tokens based on their semantic and structural roles. Domain adaptation yielded more balanced attributions in both languages, but manifested differently: Dutch models increased focus on tabular structures, whereas English models prioritized question context. However, attribution patterns were often diffuse and offered limited predictive value regarding model correctness. This underscores the fundamental limitations of current interpretability techniques, particularly under long-context conditions. Accordingly, there is a pressing need for more causally grounded and scalable methodologies to ensure transparency in critical domains such as finance.

1 Introduction

Table Question Answering (TQA) requires models to integrate structured tabular data with unstructured textual context in order to generate accurate and contextually grounded responses. This capability has become increasingly important in domains where analytical reasoning depends on the joint understanding of numerical and textual information, such as finance and healthcare [41]. In these settings, TQA enables the automation of complex tasks that traditionally rely on human expertise, including financial analysis, auditing, and compliance reporting. However, such applications also demand transparency and auditability, as decisions informed by model outputs must remain verifiable and explainable [1, 22].

Large Language Models (LLMs) have achieved strong performance in natural language understanding and reasoning tasks, but their internal mechanisms for processing structured information remain insufficiently understood [14, 35]. In the context of TQA, it is therefore essential to determine which components of the input, such as table regions, numerical values, or question spans, influence model predictions; attribution-based interpretability methods help quantify these token-level contributions and reveal underlying reasoning behavior [29, 31]. Such analyses are particularly relevant for assessing whether domain-specific fine-tuning leads to more semantically coherent and reliable reasoning behavior.

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: EurIPS'25 Workshop on AI for Tabular Data.

Financial applications present a demanding test case for interpretability research. Financial language is technically specialized and numerically precise, and regulatory requirements impose a strong need for model transparency [22]. English-language financial LLMs, including BloombergGPT [36] and FinGPT [18], have demonstrated that domain adaptation improves factual precision and robustness, but corresponding advances in low-resource languages remain limited. Recent developments in Dutch, such as GEITje-7B-ultra [33] and its domain-adapted counterpart FinGEITje-7B [23], highlight the emerging shift toward building and evaluating financial language models outside the English-speaking context. These models allow systematic comparison across both domain specialization and language, providing insight into how domain and language adaptation influence internal attribution behavior.

This study investigates the interpretability of domain-adapted and general-purpose LLMs in financial TQA across Dutch and English. Using datasets derived from ConvFinQA [8], we apply Input × Gradient (I×G) attribution to four open-source LLMs: FinGEITje-7B and GEITje-7B-ultra for Dutch, and FinMA-7B and Vicuna-7B for English. The analysis quantifies how saliency is distributed across semantic and structural input components, including tables, numeric values, and question spans. The contributions of this work are threefold: (1) a comparative interpretability framework for multilingual financial TQA, (2) empirical observations on attribution differences between domain-adapted and general-purpose models across input structures, and (3) a critical discussion of the scalability and fragility of attribution methods under long-context conditions [2, 14, 21]. To support reproducibility and enable further research in multilingual financial NLP, all experimental parameters, including prompt formats and model configurations, are made publicly available [1].

2 Methodology

Dataset and TQA Context. Our work builds on the ConvFinQA dataset [8], a benchmark for financial Table Question Answering (TQA) that combines textual disclosures, associated tables, and conversational context to enable multi-step reasoning. We evaluate on two derivatives: FinGPT-ConvFinQA (English, 1,490 test examples) [18] and FinDutchBench (Dutch, 1,453 test examples) [23]. The Dutch set was created through machine translation and fluency-based filtering, achieving over 97% overlap in answer values with the English data. As input sequences often exceed 1,500 tokens, we focus on scalable attribution techniques based on gradients and attention rather than computationally intensive alternatives such as Shapley values.

Models and Interpretability Pipeline. We analyze four language models with approximately 7B parameters. The Dutch models, FinGEITje-7B and GEITje-7B ultra, are built upon the Mistral model family and represent finance domain-adapted and general-purpose variants of LLMs specialized in Dutch. The English models, FinMA-7B and Vicuna-7B, follow the LLaMA family design and provide a parallel pair for assessing the influence of domain specialization and language [42]. Detailed model descriptions and configurations are provided in Appendix E.

Method. Input × Gradient [29, 31] is one of the earlier gradient-based attribution techniques. It assigns importance scores to each input feature by multiplying its value with the gradient of the model's output. Formally, for a model output $f(x)$ and input vector x, the attribution for each feature x_i is given by:

$$\text{Attribution}_i = x_i \times \frac{\partial f(x)}{\partial x_i}$$

Input × Gradient offers a computationally tractable solution for multi-billion-parameter LLMs with long-context inputs. It is not a measure of causal influence, but works as a proxy for *local sensitivity*. Capturing local sensitivity, rather than causal pathways [43], this method quantifies feature contribution to predictions via gradient-based sensitivity (i.e. attribution). Our analysis focus on input segments: context, table, and question. Attribution scores are compared across these functional regions to detect over-reliance on context or under-utilization of tabular information. Additionally, its computational efficiency makes it suitable for long-context setting, scaling to 1,500-token inputs where combinatorial methods like Shapley values [5] become infeasible. More elaboration on limitations of Input × Gradien can be found in Appendix B.

[1] `https://github.com/amalia020/xai-financial-qa`

Segmentation of TQA Input. To interpret attributions in a linguistically and structurally meaningful way, each input was segmented into five predefined, occasionally overlapping categories as illustrated in Figure 1:
Numeric Tokens, identified via rule-based pattern matching to capture quantitative reasoning cues.
Financial Terms, extracted using a zero-shot GPT-4o prompt aligned with established financial taxonomy categories [25].
Table, denoting structured content enclosed within `<table>` tags.
Narrative, representing unstructured textual context outside tabular regions.
Question Span, corresponding to the current query and its immediate reasoning context.

Tokens could belong to multiple categories (e.g., a financial token inside the question). Attribution scores were averaged within each span and aggregated per category to quantify how saliency mass is distributed across semantic and structural components of the TQA inputs.

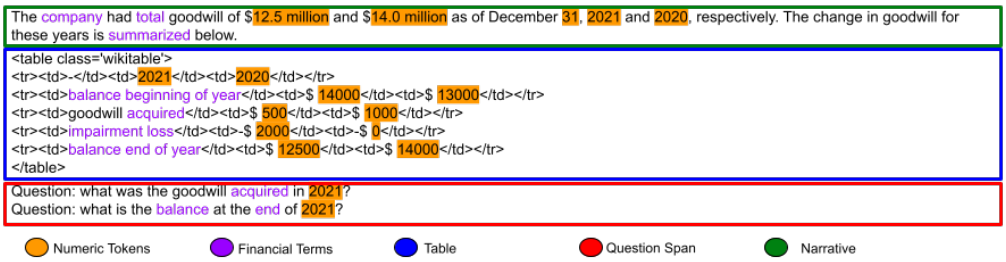

Figure 1: Illustrative TQA input instance with segmentation.

3 Experiments and Results

Performance Benchmarking.

Table 1 compares the benchmarking accuracy of four 7B-parameter language models across finance and general domains without instruction tuning (ChatML format). The finance domain–adapted models (FinGEITje-7B and FINMA-7B-NLP) clearly outperform their general-purpose counterparts in both Dutch and English.

Table 1: Exact answer accuracy of domain-adapted and general-purpose models (ChatML, no instructions) on the FinDutchBench (NL) and FinGPT-ConvFinQA (EN) benchmarks.

Model	Domain	Accuracy (%)
FinGEITje-7B (NL)	Finance	27.60
GEITje-7B-ultra (NL)	General	6.06
FINMA-7B-NLP (EN)	Finance	44.36
Vicuna-7B-v1.3 (EN)	General	16.44

This initial benchmarking as shown in Table 1, confirms that domain adaptation in finance produces substantial performance improvements in both Dutch and English, establishing a reliable baseline for the interpretability study [18, 36].

Input $\times$ Gradient Attribution Results.

Aggregated I$\times$G scores identify the dominant attribution segment for each datapoint in both the Dutch and English test sets. Table 2 reports the number of attribution tokens across input segments for Dutch (FinGEITje-7B, GEITje-7B) and English (FINMA-7B, Vicuna-7B v1.3/v1.5) models. Additional distribution plots are provided in Appendix A for reference. Across languages, numeric and question segments attract the most attribution tokens, indicating higher model attention to quantitative and interrogative content.

Dutch Results. The domain-adapted FinGEITje-7B distributes its highest attribution more frequently across semantically relevant segments (Question, Financial Terms, and Table regions) compared to its general-purpose counterpart, GEITje-7B-ultra. In contrast, GEITje-7B-ultra exhibited a pronounced focus on Numeric Tokens. The relative increase in focus on the Table segment for FinGEITje suggests that domain adaptation encourages the model to recognize and utilize the structural relevance of tabular data for TQA.

Table 2: Attribution token counts by input segment for Dutch (FinGEITje-7B, GEITje-7B) and English (FINMA-7B, Vicuna-7B v1.3/v1.5) models.

Segment	**NL**		**EN**		
	FinGEITje	**GEITje**	**FINMA**	**Vicuna-7B-v1.3**	**Vicuna-7B-v1.5**
Numeric Tokens	432	525	359	608	562
Narrative	235	245	237	208	171
Financial Terms	216	209	148	177	220
Table	256	200	217	187	131
Question	314	274	529	311	406

English Results. The English models show a parallel, if distinct, shift. FINMA-7B-NLP attributes heavily to the Question segment while deemphasizing Numeric Tokens compared to Vicuna. Unlike the Dutch adaptation, which increased focus on the Table, the English model's balance is achieved through stronger question and narrative alignment.

Table 2 counts tokens with highest absolute I×G score per example, showing which segment most influenced each prediction. Across both languages, domain-adapted models distribute attribution more evenly than their general-purpose counterparts. This suggests robust financial QA relies on integrating multiple input types, not just raw numbers.

4 Discussion and Conclusion

Fragility and Scaling Limits of Explainable AI: Attribution Methods.

A key finding of this work, relevant to the AI for Tabular Data community, is the limited robustness of current attribution methods when applied to complex Table QA tasks.

1. **Scaling Limits,** arise from input sequences exceeding 1,500 tokens, which dilute attribution signals and render combinatorial methods such as Shapley values computationally infeasible [2, 5, 19]. These challenges are amplified in financial TQA, where tabular evidence and textual context interact across heterogeneous structures.
2. **Non-robustness,** stems from the sensitivity of gradient-based methods to tokenization, prompt phrasing, and numerical formatting. In practice, small variations in table layout or linguistic structure yield substantial shifts in attribution, undermining reproducibility. These effects highlight that current attribution techniques capture local sensitivity rather than the true causal pathways that connect tabular content to model predictions [2, 14, 27, 35].

Conclusion and Future Work.

This study examines the internal behavior of finance domain-adapted and general-purpose LLMs in multilingual TQA using Input $\times$ Gradient attribution. Domain adaptation yields more balanced attributions in both Dutch and English, but these improvements in accuracy do not translate into more interpretable or stable explanations.

Future work may extend beyond post-hoc saliency toward methods capable of isolating how models combine table structure and text when making decision about numerical relationships. Promising directions include probing classifiers to test layer-wise encoding of tabular information and mechanistic interpretability to identify circuits responsible for table–text alignment [12, 28]. Progress in this area is essential for developing transparent, auditable systems that meet the accountability requirements of financial decision-making.

References

[1] Alejandro Barredo Arrieta, Natalia Díaz-Rodríguez, Javier Del Ser, Adrien Bennetot, Siham Tabik, Alberto Barbado, Salvador García, Sergio Gil-López, Daniel Molina, Richard Benjamins, et al. Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion* **58**:82–115, 2020.

[2] Giuseppe Attanasio, Debora Nozza, Eliana Pastor, Dirk Hovy, et al. Benchmarking posthoc interpretability approaches for transformer-based misogyny detection. In *Proceedings of NLP Power! The First Workshop on Efficient Benchmarking in NLP*. Association for Computational Linguistics, 2022.

[3] Jasmijn Bastings & Katja Filippova. The elephant in the interpretability room: Why use attention as explanation when we have saliency methods? *arXiv preprint* arXiv:2010.05607, 2020.

[4] Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu & Muhammad Abdul-Mageed. FinTral: A family of GPT-4 level multimodal financial large language models. *arXiv preprint* arXiv:2402.10986, 2024.

[5] Javier Castro, Daniel Gómez & Juan Tejada. Polynomial calculation of the Shapley value based on sampling. *Computers & Operations Research* **36**(5):1726–1730, 2009.

[6] Hugh Chen, Ian C. Covert, Scott M. Lundberg & Su-In Lee. Algorithms to estimate Shapley value feature attributions. *Nature Machine Intelligence* **5**(6):590–601, 2023.

[7] Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, et al. FinQA: A dataset of numerical reasoning over financial data. *arXiv preprint* arXiv:2109.00122, 2021.

[8] Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah & William Yang Wang. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. *arXiv preprint* arXiv:2210.03849, 2022.

[9] Paul F. Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg & Dario Amodei. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems 30*, 2017.

[10] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman & Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems 36*, pp. 10088–10115, 2023.

[11] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang & Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey. *arXiv preprint* arXiv:2403.14608, 2024.

[12] John Hewitt & Christopher D. Manning. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*, pp. 4129–4138, 2019.

[13] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. LoRA: Low-rank adaptation of large language models. *ICLR* **1**(2):3, 2022.

[14] Sarthak Jain & Byron C. Wallace. Attention is not explanation. *arXiv preprint* arXiv:1902.10186, 2019.

[15] Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel & Desmond Elliott. MultiFin: A dataset for multilingual financial NLP. In *Findings of the Association for Computational Linguistics: EACL 2023*, pp. 894–909, 2023.

[16] Jean Lee, Nicholas Stevens & Soyeon Caren Han. Large language models in finance (FinLLMs). *Neural Computing and Applications*, pp. 1–15, 2025.

[17] Hanxi Liu, Xiaokai Mao, Haocheng Xia, Jian Lou & Jinfei Liu. Prompt valuation based on Shapley values. *arXiv preprint* arXiv:2312.15395, 2023.

[18] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang & Daochen Zha. FinGPT: Democratizing internet-scale data for financial large language models. *arXiv preprint* arXiv:2307.10485, 2023.

[19] Scott M. Lundberg & Su-In Lee. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30*, 2017.

[20] Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk & Alexandra Balahur. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the Web Conference 2018*, pp. 1941–1942, 2018.

[21] Vivek Miglani, Aobo Yang, Aram H. Markosyan, Diego Garcia-Olano & Narine Kokhlikyan. Using Captum to explain generative language models. *arXiv preprint* arXiv:2312.05491, 2023.

[22] Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen & Stefan Zohren. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint* arXiv:2406.11903, 2024.

[23] Sander Noels, Jorne De Blaere & Tijl De Bie. A Dutch financial large language model. In *Proceedings of the 5th ACM International Conference on AI in Finance, ICAIF '24*, pp. 283–291. Association for Computing Machinery, 2024.

[24] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems 35*, pp. 27730–27744, 2022.

[25] Lingfei Qian, Weipeng Zhou, Yan Wang, Xueqing Peng, Han Yi, Jimin Huang, Qianqian Xie & Jianyun Nie. FinO1: On the transferability of reasoning enhanced LLMs to finance. *arXiv preprint* arXiv:2502.08127, 2025.

[26] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon & Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 36*, pp. 53728–53741, 2023.

[27] Sofia Serrano & Noah A. Smith. Is attention interpretable? *arXiv preprint* arXiv:1906.03731, 2019.

[28] Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, et al. Open problems in mechanistic interpretability. *arXiv preprint* arXiv:2501.16496, 2025.

[29] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina & Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *arXiv preprint* arXiv:1605.01713, 2016.

[30] Mingjie Sun, Zhuang Liu, Anna Bair & J. Zico Kolter. A simple and effective pruning approach for large language models. *arXiv preprint* arXiv:2306.11695, 2023.

[31] Mukund Sundararajan, Ankur Taly & Qiqi Yan. Axiomatic attribution for deep networks. In *International Conference on Machine Learning*, pp. 3319–3328. PMLR, 2017.

[32] Bram Vanroy. Fietje: An open, efficient LLM for Dutch. *arXiv preprint* arXiv:2412.15450, 2024.

[33] Bram Vanroy. Geitje 7B Ultra: A conversational model for Dutch. 2024.

[34] Jesse Vig, Ali Madani, Lav R. Varshney, Caiming Xiong, Richard Socher & Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint* arXiv:2006.15222, 2020.

[35] Sarah Wiegreffe & Yuval Pinter. Attention is not not explanation. *arXiv preprint* arXiv:1908.04626, 2019.

[36] Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg & Gideon Mann. BloombergGPT: A large language model for finance. *arXiv preprint* arXiv:2303.17564, 2023.

[37] Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira & Jimin Huang. Pixiu: A large language model, instruction data and evaluation benchmark for finance. 2023.

[38] Siqiao Xue, Tingting Chen, Fan Zhou, Qingyang Dai, Zhixuan Chu & Hongyuan Mei. FAMMA: A benchmark for financial domain multilingual multimodal question answering. *arXiv preprint* arXiv:2410.04526, 2024.

[39] Xianlong Zeng. Enhancing the interpretability of SHAP values using large language models. *arXiv preprint* arXiv:2409.00079, 2024.

[40] Xiao Zhang, Ruoyu Xiang, Chenhan Yuan, Duanyu Feng, Weiguang Han, Alejandro Lopez-Lira, Xiao-Yang Liu, Meikang Qiu, Sophia Ananiadou, Min Peng, et al. Dólares or dollars? Unraveling the bilingual prowess of financial LLMs between Spanish and English. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 6236–6246, 2024.

[41] Lucas Lageweg & Fieke Smit. On the use of large language models for question answering in official statistics. In *UNECE Expert Meeting on the Use of Generative AI in Official Statistics*, Geneva, Switzerland, 2025. Available at: `https://unece.org/sites/default/files/2025-05/GenAI2025_S1_Netherlands_Lageweg_D.pdf`

[42] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar et al. Llama: Open and efficient foundation language models.*arXiv preprint* arXiv:2302.13971, 2023.

[43] Emanuele Ancona, Edoardo Ceolini, Cengiz Öztireli, Markus Gross. Gradient-based attribution methods. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, pp. 169–191. Springer International Publishing, 2019.

[44] Alexander Binder, Wojciech Samek, Gregoire Montavon, Sebastian Lapuschkin & Klaus-Robert Müller. Layer-wise relevance propagation for neural networks with local renormalization layers. In *International Conference on Artificial Neural Networks*, pp. 63–71. Springer International Publishing, 2016.

A Supplementary Bar Plots for I×G Attribution

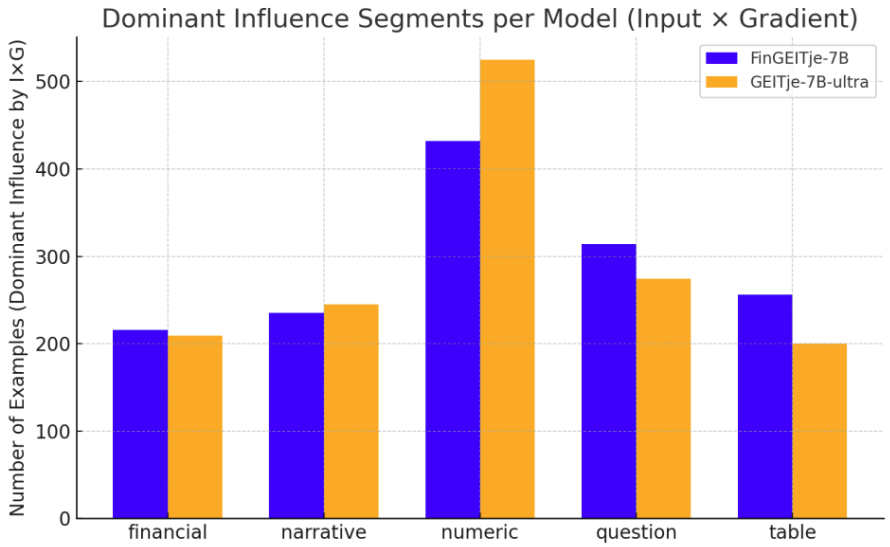

Figure 2: Distribution of dominant attribution categories for FinGEITje-7B and GEITje-7Bultra. Each bar shows the number of examples in which the majority of I×G saliency was concentrated on one of five input categories.

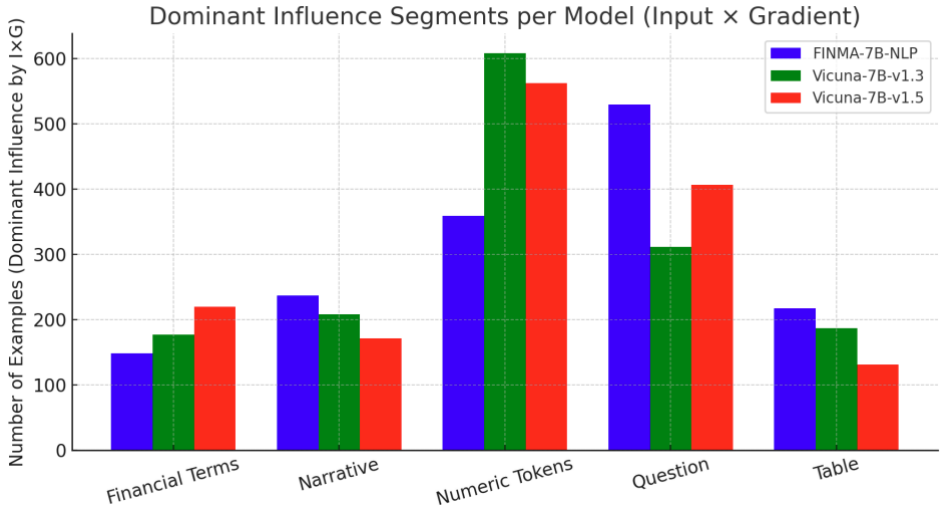

Figure 3: Distribution of dominant attribution categories for FinMA-7B and both Vicuna-7B versions

B Limitations of Input × Gradient Attribution

Input × Gradient is employed in this work with the explicit understanding that transformer-based language models are highly non-linear, and any attribution technique grounded in local linearity must be interpreted with caution. We therefore use Input × Gradient not as a measure of causal influence, but as a proxy for *local sensitivity*: it identifies which input regions the model is most responsive to under small perturbations. To generate explanations in our work, each datapoint is segmented into its main components: context, table, and question. Then, salience is then compared across these functional regions to detect over-reliance on context or under-utilization of tabular information.

Although Input × Gradient is one of the earlier gradient-based attribution techniques, it offers a computationally tractable solution for multi-billion-parameter LLMs with long-context inputs. More computationally intensive methods (e.g., Integrated Gradients [31], DeepSHAP [19], LRP [44]) become infeasible even on large-scale compute systems. Input × Gradient therefore provides a consistent and scalable approach across all experiments. Developing interpretability methods that combine stronger theoretical foundations with practical scalability remains an important direction for future work.

C Use of AI Assistance

During the development of this paper, language models such as ChatGPT were occasionally used to support the writing and debugging process. Specifically, these tools assisted in identifying grammatical issues, rephrasing technical sentences for clarity, and resolving minor implementation errors.

All conceptual contributions, research design, analyses, and interpretations presented in this work are the author's original intellectual work. The use of AI tools served only to improve efficiency and expression, without contributing to the core ideas or findings.

D Illustrative Prompt Structures: ChatML

```
<|system|>
System message goes here

<|user|>
User message goes here

<|assistant|>
Assistant response goes here
```

Figure 4: Example ChatML prompt format with system, user, and assistant roles.

E Model Cards

Abbreviations Used

This subsection provides brief explanations of the abbreviations used in the model comparison cards and related methodology sections.

- **SFT** – Supervised Fine-Tuning on instruction-style data. [24]
- **DPO** – Direct Preference Optimization.[26]
- **RLHF** – Reinforcement Learning from Human Feedback.[9]
- **PEFT** – Parameter-Efficient Fine-Tuning (e.g., LoRA, QLoRA).[1]
- **LoRA** – Low-Rank Adaptation.[13]
- **QLoRA** – Quantized LoRA, enabling efficient training via quantization[10]

FinGEITje-7B[a]

Language: Dutch
Domain: Finance
Architecture: Mistral 7B
Tokenizer: Mistral BPE
Tuning: SFT (QLoRA)
PEFT: QLoRA + LoRA
Open Weights: Yes

[a] https://huggingface.co/snoels/FinGEITje-7B-sft

GEITje-7B-ultra[a]

Language: Dutch
Domain: General-purpose
Architecture: Mistral 7B
Tokenizer: Mistral BPE
Tuning: SFT + DPO
PEFT: None (full DPO)
Open Weights: Yes

[a] https://huggingface.co/BramVanroy/GEITje-7B-ultra

FinMA-7B [a]

Language: English
Domain: Finance
Architecture: LLaMA 7B
Tokenizer: LLaMA Tokenizer
Tuning: SFT (full fine-tune)
PEFT: None
Open Weights: Yes

[a] https://huggingface.co/ChanceFocus/finma-7b-nlp

Vicuna-7B [a]

Language: English
Domain: General-purpose
Architecture: LLaMA 7B
Tokenizer: LLaMA Tokenizer
Tuning: SFT (full fine-tune)
PEFT: None
Open Weights: Partially

[a] https://huggingface.co/lmsys/vicuna-7b-v1.5

Figure 5: Core model cards for Dutch and English LLMs used in this study.