Context-Masked Meta-Prompting for Privacy-Preserving LLM Adaptation in Finance

Sayash Raaj Hiraou
Fidelity Investments
Bengaluru, India
sayashraaj@alumni.iitm.ac.in

Abstract

The increasing reliance on Large Language Models (LLMs) in sensitive domains like finance necessitates robust methods for privacy preservation and regulatory compliance. This paper presents an iterative meta-prompting methodology designed to optimise hard prompts without exposing proprietary or confidential context to the LLM. Through a novel regeneration process involving feeder and propagation methods, we demonstrate significant improvements in prompt efficacy. Evaluated on public datasets serving as proxies for financial tasks such as SQuAD for extractive financial Q&A, CNN/DailyMail for news summarisation, and SAMSum for client interaction summarisation, our approach, utilising GPT-3.5 Turbo, achieved a 103.87% improvement in ROUGE-L F1 for question answering. This work highlights a practical, low-cost strategy for adapting LLMs to financial applications while upholding critical privacy and auditability standards, offering a compelling case for its relevance in the evolving landscape of generative AI in finance.

1 Introduction

The financial industry is exploring Large Language Models (LLMs) for tasks such as compliance Q&A, research summarisation, and automated risk assessment. However, strict regulations (e.g., GDPR, SEC guidelines) and internal governance prohibit exposing client data, proprietary models, or internal research to external systems. This rules out many common adaptation approaches that require sharing task context.

The challenge is therefore not just a natural language processing (NLP) problem but a financial integration problem: *how to tailor LLMs to domain needs without breaching confidentiality or auditability?*. We address this with a **context-masked meta-prompting framework** that refines human-readable "hard" prompts [1, 2] through an LLM-as-optimiser process [3, 4], while ensuring all sensitive data remains within a secure perimeter.

Evaluated on public datasets as proxies for financial NLP tasks, our approach delivers substantial performance gains using cost-efficient models, aligning LLM optimisation with the operational and regulatory realities of finance.

2 Related Work

Generative AI is increasingly applied in finance for tasks such as compliance checks, market news summarisation, and client-–advisor interaction analysis. These use cases involve sensitive data — proprietary strategies, client records, or internal research — that cannot leave secure systems due to

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Generative AI in Finance.

regulations like GDPR and strict internal governance. This makes direct fine-tuning or prompt-based adaptation of LLMs, even in few-shot settings [5], difficult to deploy.

Outside finance, lightweight adaptation techniques such as hard prompt optimisation [1], mixtures of soft prompts [2], automatic hint generation [6], and meta-prompting with LLMs as optimisers [3, 4] have shown strong task-specific gains. Other work explores self-referential prompt evolution [7, 8] and structured prompt pattern catalogues [9]. However, these approaches generally assume the model can access task context — an assumption incompatible with high-compliance financial environments.

Our work adapts these ideas into a context-masked meta-prompting framework that enables LLM optimisation for finance-relevant tasks without exposing any sensitive data.

3 Context-Masked Meta-Prompting Methodology

Our framework enables iterative prompt optimisation while respecting a strict context-masking principle. The optimisation process is driven by a meta-prompt that instructs an LLM to generate improved prompt templates [6] based on the performance of previous ones, without ever seeing the confidential data used for performance evaluation. The entire process occurs within a secure internal system [10] that only sends sanitised, context-free data to the external LLM API, as illustrated in Figure 1.

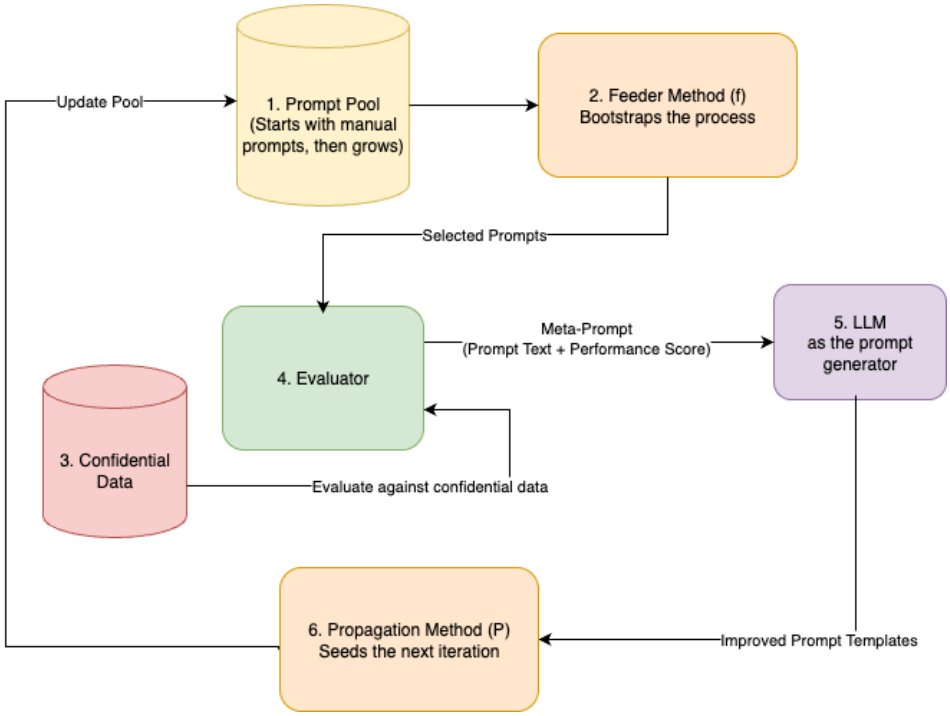


Figure 1: Conceptual overview of the context-masked meta-prompting loop

The core of our approach is a regeneration method, which consists of two components:

- **Feeder Methods (f):** These determine the initial sampling of prompts to bootstrap the process. We test two strategies: Feeder Method A (f_a), which samples only the top-n best-performing manually written prompts, and Feeder Method B (f_b), which samples both the top-n and bottom-n prompts to provide the LLM with both positive and negative examples.
- **Propagation Methods (P):** These govern how prompts from one iteration are used to seed the next. We investigate Propagation Method A (P_a), which cumulatively concatenates all previously generated prompts, and Propagation Method B (P_b), which uses the feeder logic to resample from the growing pool of generated prompts, thereby managing context window size.

By combining these components, we evaluate four distinct pipeline strategies: $f_a P_a$ (Method A), $f_b P_a$ (Method B), $f_a P_b$ (Method C), and $f_b P_b$ (Method D). This systematic exploration allows for a granular analysis of different optimisation strategies impacting performance and prompt diversity.

4 Experiments and Key Results

Experimental Setup All experiments were conducted using `GPT-3.5 Turbo` (2024 release) with a temperature of 1.0 to encourage diversity. This model was chosen deliberately to test our method on a cost-efficient, widely accessible LLM, simulating a realistic setting for financial institutions with budget, latency, and compliance constraints. We used established public datasets as finance-relevant proxies for common tasks:

- **SQuAD** [11]: extractive financial Q&A, analogous to retrieving figures from reports.
- **CNN/DailyMail** [12]: summarising news and market analysis.
- **SAMSum** [13]: summarising client–advisor or compliance-related interactions.

Performance was measured by the ROUGE-L F1 score over 10 iterations, comparing against the baseline of manually written prompts (S_m). The initial set of prompts selected by the feeder method before the first generative iteration is termed S_f.

Performance Improvement Our results show that iterative meta-prompting significantly improves performance over the baseline. As shown in Figure 2, methods using the resampling Propagation Method B (P_b) consistently outperformed those using the cumulative method (P_a), which suffered from context window limitations. The most striking result was on the Question-Answering task, where method $f_a P_b$ (Method C) achieved a mean ROUGE-L F1 score of 0.526 after 9 iterations, a 103.87% improvement over the manual baseline score of 0.258. This demonstrates that our privacy-preserving technique can more than double the effectiveness of prompts for precise information extraction tasks crucial in finance.

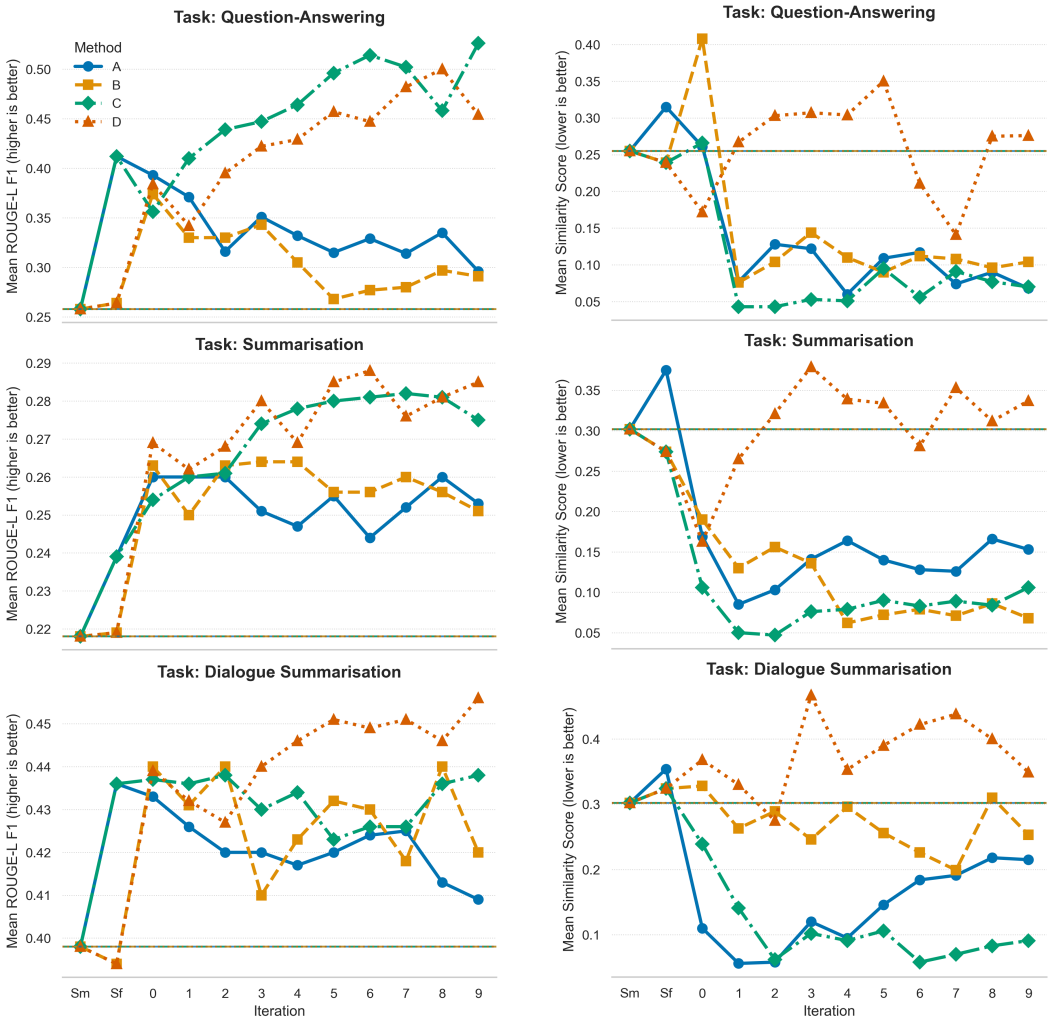
Analysis of prompt diversity revealed that method $f_a P_b$ also provides a strong balance between high performance and the generation of varied prompts, a key factor for robust deployment. Full per-iteration scores and similarity analysis are provided in Appendix A.

5 Financial Applications and Conclusion

Practical Deployment and Applications A financial firm can adopt this framework by keeping all proprietary data within its secure perimeter. An internal service would evaluate prompts against this data, then send only the context-free prompt text and its performance score to an external LLM via a secure API. The optimised prompt templates returned by the LLM are then integrated back into internal applications. This architecture strictly maintains data confidentiality. Key applications include:

- **Private Compliance Q&A:** Optimise prompts to answer questions against internal regulatory documents without exposing proprietary legal interpretations.
- **Proprietary Research Summarisation:** Create effective summarisation prompts for sensitive analyst reports without the reports ever leaving the firm's environment.
- **Auditable Risk Checkers:** Bootstrap and refine human-readable instruction templates for automated risk and fraud detection systems, ensuring transparency.

Limitations and Responsible Deployment While promising, this work has limitations. The methodology was validated on proxy datasets; future work should test it on financial data. Ethically, a key risk is that the meta-prompting process could amplify biases inherent in the LLM, whose training data is opaque. Responsible deployment therefore necessitates continuous bias auditing and robust human-in-the-loop governance for critical decisions. A comprehensive discussion of these points is available in Appendices B and C.



(a) Mean ROUGE-L F1 Scores (higher is better)

(b) Prompt Similarity Scores (lower is better)

Figure 2: Performance and diversity of prompts over 10 iterations. (a) ROUGE-L scores show sustained performance gains with Method C ($f_a P_b$) being most effective. (b) Similarity scores show the evolution of prompt diversity, with Method C maintaining a good balance.

Conclusion This paper presented a context-masked meta-prompting framework that enables significant LLM performance gains while adhering to the stringent privacy and auditability requirements of the financial industry. By demonstrating a 103.87% improvement in a key proxy task representative of financial NLP applications using a resource-efficient model, we have shown a practical, low-cost method for adapting LLMs to sensitive domains. This work provides a viable and responsible pathway for deploying effective and interpretable generative AI in finance.

Disclaimer The views expressed herein are those of the author and do not necessarily reflect the views of Fidelity Investments.

References

[1] Yuxin Wen, Neel Jain, John Kirchenbauer, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Hard prompts made easy: Gradient-based discrete optimization for prompt tuning and discovery, 2023.

[2] Guanghui Qin and Jason Eisner. Learning how to ask: Querying lms with mixtures of soft prompts, 2021.

[3] Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, and Xinyun Chen. Large language models as optimizers, 2024.

[4] Ruotian Ma, Xiaolei Wang, Xin Zhou, Jian Li, Nan Du, Tao Gui, Qi Zhang, and Xuanjing Huang. Are large language models good prompt optimizers?, 2024.

[5] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[6] Hong Sun, Xue Li, Yinchuan Xu, Youkow Homma, Qi Cao, Min Wu, Jian Jiao, and Denis Charles. Autohint: Automatic prompt optimization with hint generation, 2023.

[7] Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. Promptbreeder: Self-referential self-improvement via prompt evolution, 2023.

[8] Priyanshu Gupta, Shashank Kirtania, Ananya Singha, Sumit Gulwani, Arjun Radhakrishna, Gustavo Soares, and Sherry Shi. MetaReflection: Learning instructions for language agents using past reflections. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8369–8385, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[9] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. A prompt pattern catalog to enhance prompt engineering with chatgpt, 2023.

[10] Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. Prompt consistency for zero-shot task generalization, 2022.

[11] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text, 2016.

[12] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[13] Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. Samsum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. Association for Computational Linguistics, 2019.

A Supplementary Results and Details

A.1 Detailed Performance and Similarity Scores

The following tables provide the exact numerical results for the mean ROUGE-L F1 scores (Table 1) and similarity scores (Table 2) for all methods across all 10 iterations of the experiment. These tables form the basis for the analysis presented in the main paper.

Table 1: Detailed Mean ROUGE-L F1 Scores for All Tasks and Methods Across Iterations. This table preserves the exact numerical results from the original study.

	Question-Answering				Summarisation				Dialogue Summarisation			
Method	A	B	C	D	A	B	C	D	A	B	C	D
Sm	0.258	0.258	0.258	0.258	0.218	0.218	0.218	0.218	0.398	0.398	0.398	0.398
Sf	0.412	0.264	0.412	0.264	0.239	0.219	0.239	0.219	0.436	0.394	0.436	0.394
0	0.393	0.374	0.356	0.384	0.260	0.263	0.254	0.269	0.433	0.440	0.437	0.439
1	0.371	0.330	0.410	0.342	0.260	0.250	0.260	0.262	0.426	0.431	0.436	0.432
2	0.316	0.330	0.439	0.395	0.260	0.263	0.261	0.268	0.420	0.440	0.438	0.427
3	0.351	0.343	0.447	0.422	0.251	0.264	0.274	0.280	0.420	0.410	0.430	0.440
4	0.332	0.305	0.464	0.429	0.247	0.264	0.278	0.269	0.417	0.423	0.434	0.446
5	0.315	0.268	0.496	0.457	0.255	0.256	0.280	0.285	0.420	0.432	0.423	0.451
6	0.329	0.277	0.514	0.447	0.244	0.256	0.281	0.288	0.424	0.430	0.426	0.449
7	0.314	0.280	0.502	0.482	0.252	0.260	0.282	0.276	0.425	0.418	0.426	0.451
8	0.335	0.297	0.458	0.500	0.260	0.256	0.281	0.281	0.413	0.440	0.436	0.446
9	0.296	0.291	0.526	0.454	0.253	0.251	0.275	0.285	0.409	0.420	0.438	0.456

Table 2: Detailed Similarity Scores for All Tasks and Methods Across Iterations. This table preserves the exact numerical results from the original study.

	Question-Answering				Summarisation				Dialogue Summarisation			
Method	A	B	C	D	A	B	C	D	A	B	C	D
Sm	0.255	0.255	0.255	0.255	0.302	0.302	0.302	0.302	0.302	0.302	0.302	0.302
Sf	0.315	0.239	0.315	0.239	0.375	0.274	0.375	0.274	0.354	0.324	0.354	0.324
0	0.261	0.408	0.266	0.172	0.169	0.190	0.106	0.163	0.110	0.328	0.239	0.368
1	0.077	0.076	0.043	0.267	0.085	0.130	0.050	0.265	0.056	0.263	0.141	0.330
2	0.128	0.104	0.043	0.303	0.103	0.156	0.047	0.321	0.058	0.289	0.062	0.275
3	0.122	0.144	0.053	0.307	0.141	0.136	0.076	0.379	0.120	0.246	0.102	0.467
4	0.060	0.110	0.051	0.304	0.164	0.062	0.079	0.339	0.095	0.296	0.091	0.353
5	0.109	0.090	0.095	0.350	0.140	0.072	0.090	0.334	0.146	0.256	0.106	0.390
6	0.117	0.112	0.056	0.211	0.128	0.079	0.083	0.281	0.184	0.226	0.058	0.422
7	0.074	0.108	0.091	0.141	0.126	0.071	0.089	0.353	0.191	0.199	0.070	0.438
8	0.090	0.096	0.077	0.275	0.166	0.086	0.084	0.312	0.218	0.310	0.083	0.400
9	0.068	0.104	0.070	0.276	0.153	0.068	0.106	0.337	0.215	0.253	0.091	0.349

B Limitations and Future Work

Our framework demonstrates significant promise, but we acknowledge several limitations that present avenues for future work.

- **Proxy Datasets:** Experiments were conducted on public NLP datasets as proxies. Future work should prioritise validation on anonymised or synthetic financial datasets to confirm efficacy in a direct financial context.
- **Single LLM:** All experiments used GPT-3.5 Turbo. Future research should include LLM ablations with other models (including open-source alternatives) to test the generalisability of the optimisation process.
- **Decision-Quality Metrics:** Evaluation relied on ROUGE-L. Future work could benefit from using downstream, task-specific financial metrics (e.g., accuracy of extracted financial data, portfolio signal quality) to measure practical value.

C Broader Impact & Ethics

The primary positive impact of this work is enabling privacy-preserving AI in finance, reducing data leakage risks and fostering trust. However, any effective optimisation technique carries risks. A key ethical consideration is bias amplification. The meta-prompting process could inadvertently reinforce biases present in either the initial prompts or the LLM itself, leading to skewed outputs in sensitive applications like credit assessment. To mitigate this, we strongly recommend that any deployment of this method be accompanied by:

- **Rigorous Bias Auditing:** Continuous monitoring of both input and output prompts for demographic or other biases.
- **Human-in-the-Loop Governance:** Ensuring human oversight for all critical financial decisions derived from LLM outputs.
- **Full Auditability:** Maintaining transparent logs of the prompt evolution process to ensure that the logic driving the LLM remains interpretable and compliant.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper explicitly discusses its limitations in the "Limitations and Responsible Deployment" subsection of Section 5. It notes that the methodology was validated only on public proxy datasets rather than real financial data, and therefore future work should extend evaluation to proprietary datasets. It also acknowledges that the meta-prompting process could amplify biases inherent in the underlying LLM and stresses the need for continuous bias auditing and human-in-the-loop governance. These statements clearly outline the scope, assumptions, and constraints of the work.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: the paper discusses its limitations in the "Financial Applications and Conclusion" section, specifically under the "Limitations and Responsible Deployment" paragraph, where it notes the reliance on proxy datasets, the need for future testing on real financial data, and potential bias amplification risks.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [No]

   Justification: The paper does not present any formal theoretical results, assumptions, or proofs. Its contributions are methodological and experimental, focusing on the design, implementation, and evaluation of the context-masked meta-prompting framework rather than on formal derivations or theoretical guarantees.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides sufficient methodological detail for reproducing the main experimental results, including descriptions of the context-masked meta-prompting framework, the feeder and propagation strategies, and the iterative optimisation process. It specifies the model used (`GPT-3.5 Turbo`), temperature settings, number of iterations, and the public datasets (SQuAD, CNN/DailyMail, SAMSum) serving as proxies for financial tasks. The evaluation metrics (ROUGE-L F1) and baseline comparisons are clearly stated. While proprietary financial data is not used, the choice of openly available datasets ensures that the experimental setup can be replicated without requiring access to restricted resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example

(a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: The paper does not provide open access to the code or scripts used for running the experiments, nor does it include supplemental material with reproduction instructions. While the datasets used (SQuAD, CNN/DailyMail, SAMSum) are publicly available and cited, the specific implementation of the context-masked meta-prompting framework and the exact prompts, configurations, and pipeline logic are not released. This omission limits the ability of others to reproduce the exact experimental results, even though the methodological description is detailed enough to guide an independent reimplementation.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimiser, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the key experimental details needed to understand the results, including the choice of model (GPT-3.5 Turbo, 2024 release), temperature setting (1.0), datasets used (SQuAD, CNN/DailyMail, SAMSum), the nature of these datasets as proxies for financial tasks, the number of optimisation iterations (10), baseline comparison

method (manually written prompts), and evaluation metric (ROUGE-L F1). While hyperparameters in the traditional ML sense (e.g., learning rates, optimisers) are not applicable due to the use of an API-based LLM, the paper provides sufficient description of the feeder and propagation methods, their combinations, and how they were applied, enabling readers to fully interpret the reported results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The paper presents mean ROUGE-L F1 scores over 10 iterations for different feeder–propagation method combinations, but it does not report error bars, confidence intervals, or statistical significance tests. Variability in performance across iterations is shown qualitatively in the figures, but no formal statistical measures are provided to quantify uncertainty or assess the robustness of observed differences. As such, while trends are clearly visualized, the statistical significance of the reported improvements is not explicitly established.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [No]

Justification: While the paper specifies that all experiments were conducted using the GPT-3.5 Turbo API with a fixed temperature of 1.0, it does not detail the computational resources required beyond the model choice. Information such as API usage costs, execution time per iteration, total number of requests, memory requirements, or any local preprocessing environment specifications is not provided. Without these details, reproducing the experiments with equivalent resource allocation would be challenging.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

Answer: [Yes]

Justification: The research complies with the NeurIPS Code of Ethics. The methodology is explicitly designed for privacy preservation, ensuring that no sensitive financial or client data is exposed outside secure institutional boundaries. Ethical considerations, including potential bias amplification and the need for human-in-the-loop oversight, are discussed in the "Limitations and Responsible Deployment" subsection. The work avoids harmful applications, maintains transparency in its claims, and aligns with both regulatory and ethical standards relevant to AI deployment in finance.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The paper addresses both positive and negative societal impacts in the "Limitations and Responsible Deployment" subsection. On the positive side, the framework enables financial institutions to harness LLM capabilities for compliance, risk management, and research summarisation while maintaining strict privacy and auditability standards, potentially improving efficiency, transparency, and regulatory alignment. On the negative side, the paper acknowledges risks such as amplification of biases inherent in the underlying LLMs, which could lead to unfair or misleading outputs in financial decision-making. It emphasizes the necessity of bias auditing, human-in-the-loop governance, and responsible oversight to mitigate these risks.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimising neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper does not release any new datasets or pretrained models that would present a high risk of misuse. The methodology is described conceptually, and all experiments are performed on established, publicly available datasets. No proprietary financial data, sensitive information, or deployable models are shared, eliminating the need for additional release safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: The paper uses established public datasets — SQuAD [11], CNN/DailyMail [12], and SAMSum [13] — all of which are properly cited in the text. The original creators are credited through these citations, and no modifications or redistribution of the datasets are performed. While explicit license names are not stated in the main text, the datasets are publicly available under permissive research-use terms, and their usage in this work complies with those licenses and terms of use.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: The paper does not introduce any new datasets, models, or code assets. All experiments are conducted using existing, publicly available datasets and a commercially available LLM API, with no novel asset release requiring documentation.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The manual prompts used as baselines were written by the authors themselves. No crowdsourcing platforms or external participants were involved, and thus no formal participant instructions or compensation apply. All human input was part of the authors' own contribution to the work rather than a structured human-subjects study.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The work does not involve research with human subjects as defined by IRB guidelines. All manual prompts were authored by the paper's authors themselves, without participation from external individuals or study participants, and therefore no ethical review or IRB approval was required. No participants were placed at risk.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.

- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: This work employs a large language model (GPT-3.5-Turbo) as an essential and original component of the proposed context-masked meta-prompting framework. The LLM functions as an optimiser, iteratively refining task-specific prompts without accessing any sensitive financial context, thereby adhering to stringent privacy, auditability, and compliance requirements. Its role and configuration are described in the experimental setup to maintain methodological transparency in accordance with the NeurIPS 2025 LLM Policy. All inputs to the LLM in this process were constructed to exclude confidential or proprietary information, and outputs were verified for correctness before integration into experiments.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (https://neurips.cc/Conferences/2025/LLM) for what should or should not be described.