Using Generative AI to Retrieve and Analyze CEO Compensation Consultant Information from Public Corpora

Kwan S. (Sean) Lee
Assistant Professor
College of Business
Texas A&M University Victoria
Victoria, TX 77901
leek3@tamuv.edu

Annie Kim
Assistant Professor
College of Business
University of Nevada Reno
Reno, NV 89557
anniekim@unr.edu

Abstract

We demonstrate how fine-tuning open-weight large language models can significantly improve the retrieval of information about compensation consultant engagement from company proxy statements, a task previously requiring costly manual collection in finance research and practice. We fine-tuned Gemma 3 12B-Instruct using parameter-efficient adapters (QLoRA) on 400 expert-labeled examples to identify consultant engagements and distinguish between retained advisers versus survey-only providers. Our fine-tuned model achieves 67.7% F1 score compared to 33.3% for the base model, a 103% improvement while reducing false positives by 68%. The approach addresses key challenges of processing non-standardized lengthy text data such as learning domain knowledge and requiring multi-hop reasoning. By trading throughput for fidelity through targeted chunking and conservative decoding, we provide a scalable, reproducible method for converting narrative-heavy filings into structured financial data, laying the foundation for a vertical agent capable of automating finance data-collection tasks.

1 Introduction

An increasing number of studies in economics and finance are investigating the impacts of generative artificial intelligence (GenAI) on finance research and practices [11, 14, 15, 23]. However, long-running debates have still persisted mainly because many questions hinge on information locked inside lengthy, legally dense filings that have historically required costly hand collection. To tackle this data collection challenge, this paper explores the potential benefits of GenAI that will aid in processing lengthy company documents at scale, particularly within compensation consulting and the analysis of CEO pay. Specifically, we fine-tuned Gemma 3 12B-Instruct, an open-weights large language model pretrained by Google, using parameter-efficient adapters (PEFT/LoRA) and 4-bit quantization (QLoRA), and show that domain adaptation with a set of quality training samples materially improves extraction quality relative to vanilla prompting [21].

Whereas early GenAI-in-finance work queried commercial LLM APIs or fine-tuned on readily available sentiment datasets to address relatively simple tasks (e.g., headline/news or earnings-text sentiment) [15, 18, 23], our contribution is to build expert-labeled training samples and then fine-tune an open model that extracts firm-level events from legal-financial documents at scale. This design achieves accuracy, reproducibility, and cost efficiency, and it directly addresses the practical challenges posed by long documents, heterogeneous formatting, and domain-specific distractors, thus laying the foundations for a domain-specific vertical agent system.

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

Processing long, unstructured filings from the U.S. Securities and Exchange Commission (SEC) such as proxy statements can push LLMs to their limits. Because these filings are verbose narratives with heterogeneous formatting, identifying how compensation consultants are engaged could require *multi-hop reasoning* over dispersed evidence rather than a simple needle-in-a-haystack (NIAH) lookup. Another major challenge is mitigating the context rot, which refers to a broader degradation in the performance of the LLM as the input increases [26, 4]. To prioritize accuracy under these constraints, our pipeline trades throughput for fidelity: we adjust the size of the text chunks that target consultant-relevant phrases and paragraphs to shorten text spans while preserving sufficient context needed for multi-hop aggregation [25, 12, 22, 7]. In addition, references to compensation consultants are sometimes either overly complicated or implicit in the narrative, requiring the LLM to reason carefully through the context to derive the correct information. Accordingly, we scaffold inference with lightweight reasoning prompts while training the LLM to stabilize multi-step extraction in long contexts [24, 13].

Our fine-tuned model achieves 67.7% F1 score compared to 33.3% for the base model, a 103% improvement, while reducing false positives by 68%. These promising results underscore how LLMs can reshape data collection in finance by converting previously intractable, narrative-heavy filings into reliable structured signals. Beyond research, the approach scales to practical users of SEC disclosures—analysts, investors, and data providers—by pairing expert-curated labels with parameter-efficient fine-tuning (LoRA) and 4-bit quantized backbones (QLoRA) to achieve accuracy on modest hardware. In turn, adapting open models this way provides a cost-effective, reproducible path to extract finance-relevant information at scale from Electronic Data Gathering, Analysis, and Retrieval (EDGAR)'s full-text filings, opening new avenues for empirical investigation.

2 Research background

Finance research on executive pay has been examining the influence of compensation consultants for the last two decades [2, 19]. Early studies argued consultants act as independent experts, helping boards benchmark pay against market practices. However, other research emphasized conflicts of interest, suggesting consultants may encourage higher CEO pay to maintain favorable relationships with management [1, 19]. Combined with public outcry over skyrocketing CEO compensation, this argument on the collusion between compensation consultants and CEOs has shaped policy concerns, including SEC disclosure rules requiring firms to identify retained consultants. Yet empirical progress has been limited by the difficulty of systematically extracting consultant information from lengthy proxy statements.

Automated approaches are attractive, but prior work has rarely applied them to unstructured filings. To date, most finance research still relies on manual coding or small samples, restricting both scope and replication. For example, existing studies on CEO pay and the retention of compensation consultants often note that data were "hand-collected" without disclosing extraction criteria or coder protocols, or were sourced from proprietary databases such as the ISS Incentive Lab, which is limited in scope [3, 19].

Extracting necessary information from SEC filings such as proxy statements often requires even expert annotators to read carefully across long, heterogeneous narratives. Consequently, retrieving information about how firms retain compensation consultants goes beyond classic NER or synthetic NIAH tasks [8]. In addition, processing extremely long texts could sharply degrade performance as context grows.

3 Applications of GenAI to finance

Early finance applications explored LLMs to facilitate research workflows [15, 23], but their implementations relied primarily on prompt engineering or rudimentary fine-tuning with existing datasets and thus underutilized domain adaptation. There remains immense potential to combine recent performance improvements in GenAI with domain-specific expertise to advance research in finance.

Although state-of-the-art base LLMs are trained on vast corpora and exhibit strong general language skills, a base model still lacks domain-specific knowledge. For example, while a base LLM may successfully detect whether a firm relies on outside advisers, it can still confuse compensation consultants with auditors or legal counsel when these entities appear in proximity, leading to role

misclassification. Moreover, a firm may retain multiple consultants and non-consultants as well (e.g., natural persons or business associations), yet an LLM may fail to differentiate among distinct entities. Proxy statements may also list former consultants to provide additional context, which can distort the data collection process by introducing duplicate entries. Moreover, some firms may purchase proprietary compensation survey data from consultants without formally retaining them as outside advisers. Finally, broad pretraining can inject strong priors about well-known firms (e.g., Aon, Frederic W. Cook, Mercer, Willis Towers Watson), potentially amplifying *false positives* when names coincidentally match tokens in a filing [10].

Compounding these challenges, proxy statements often extend to hundreds of thousands of tokens. Even after cleaning and extracting relevant text chunks, a single proxy statement may still contain tens of thousands of tokens. The latest research shows that LLMs' long-context performance typically degrades as input length grows—well before the nominal context window is reached—due to positional effects and reduced effective context usage [17, 8, 16, 4]. Consistent with these findings, we observe that state-of-the-art API models struggle to return fully correct extractions when fed very large, information-dense contexts. This motivates our choice to chunk inputs and target consultant-relevant spans rather than relying on monolithic prompts with full proxy statement texts. In addition to efficient text chunking, model performance improves materially when it is *fine-tuned* for the specific structure and language of SEC filings.

4 Fine-tuning strategy

To maximize the accuracy of fine-tuning, we adopted a *fidelity-over-throughput* approach: adapt an open 12B instruction-tuned backbone (Gemma 3) with parameter-efficient tuning, strict context budgeting, and deterministic decoding. This design targets long, heterogeneous proxy statements without requiring large GPU clusters.

We fine-tuned the base model on domain-specific, expert-labeled excerpts, so it learns to map text patterns to adviser *roles* (retained adviser vs. survey-only provider vs. unrelated party) rather than merely memorizing names. Practically, we adopt parameter-efficient fine-tuning (LoRA/PEFT) and 4-bit quantization (QLoRA) to adapt the model on modest hardware while preserving accuracy [9, 5, 6]. This domain adaptation improves precision on consultant extraction and reduces spurious matches.

We assembled approximately 400 expert-labeled examples spanning diverse issuers, years, and layouts. Each item includes: (i) the excerpt bundle produced by the keyword pipeline; (ii) a filing-year header for temporal disambiguation; (iii) role-aware detailed instructions; and (iv) a compact gold extraction string (See Appendix B for the instruction prompt).

5 Analysis

The fine-tuned model significantly outperforms the base model across all evaluation metrics as illustrated in Table 1 (see Appendix C for the sample test results). Most notably, the fine-tuned model achieves 67.7% overall F1 score compared to 33.3% for the base model, a 103% relative improvement. This dramatic performance gain demonstrates the value of domain-specific adaptation for complex financial document processing tasks. By contrast, the surprisingly low score for the base model reveals that modern LLMs may not be sufficient to perform complicated and ambiguous QA tasks without additional training.

6 Limitations

To our knowledge, this study is among the first to apply LLM fine-tuning with reproducible code to finance research, particularly for extracting compensation-consultant data. Despite our novel contribution, there are notable limitations.

Compute and model size. Our experiments were constrained by commodity hardware. To fit training and inference on a personal desktop, we employed parameter-efficient fine-tuning (LoRA/PEFT) and 4-bit quantization (QLoRA) on a smaller open-weights LLM rather than exploring larger base models. This choice usefully demonstrates that meaningful gains are attainable without expensive APIs or multi-GPU clusters, but it also limited our ability to test higher-parameter variants or extensive

Table 1: Model Performance Comparison on Compensation Consultant Extraction

Metric Category	Model	Precision	Recall	F1-Score	Correct/Total
Overall	Base Model	0.265	0.446	0.333	173/388
	Fine-tuned	0.679	0.675	0.677	262/388
	Improvement	*+156%*	*+51%*	*+103%*	*+51%*
RET (Retained)	Base Model	0.273	0.604	0.376	131/217
	Fine-tuned	0.698	0.724	0.710	157/217
	Improvement	*+156%*	*+20%*	*+89%*	*+20%*
SURV (Survey)	Base Model	0.243	0.246	0.244	42/171
	Fine-tuned	0.652	0.614	0.633	105/171
	Improvement	*+168%*	*+150%*	*+159%*	*+150%*
Pooled	Base Model	0.310	0.506	0.385	176/348
	Fine-tuned	0.767	0.756	0.761	263/348
	Improvement	*+147%*	*+49%*	*+98%*	*+49%*

Notes: The table presents performance metrics for compensation consultant extraction on 146 proxy statement excerpts. **Overall** metrics evaluate instance-based performance where each consultant mention must be correctly identified and categorized (RET vs. SURV); 388 total gold labels. **Pooled** metrics combine both categories and evaluate only consultant name identification, ignoring categorization; 348 unique consultant instances after removing duplicates appearing in both categories within the same document. Higher pooled performance indicates the model effectively identifies consultant names but occasionally misclassifies engagement types. Improvement percentages are calculated as (Fine-tuned $-$ Base) / Base $\times$ 100%.

ablations. With more robust compute, larger models and fuller sweeps could plausibly yield further improvements. Additional high-quality, expert-labeled training data would likewise improve accuracy.

Long-context management and context length mismatch between training and inference. Although *Gemma 3 12B-Instruct* supports a 128K-token window [21], we deliberately capped the model's effective context at roughly 2,100 tokens and processed each proxy statement in multiple chunks. Two considerations drove this decision. First, recent work documents degradation as input length grows—even within the nominal window—due to positional effects and reduced effective context usage ("context rot") [26, 4]. Second, our initial supervision consisted of paragraph-level examples; when supplied with much longer multi-paragraph inputs, the fine-tuned model exhibited higher error rates and occasional hallucinations, consistent with sensitivity to length discrepancies between training and evaluation. We gradually increased the mean length of training samples over time, but increasing the context window to the extent of several thousand tokens requires more dedicated GPU hardware. Consequently, we decided to prioritize fidelity over throughput at the cost of under-utilizing the model's maximum window. Future work should explore curriculum or scheduling strategies to bridge training–inference length gaps and improve long-context robustness [20].

Evaluation scope. Our evaluation emphasizes extraction fidelity for consultant identification and role labeling in DEF 14A filings; broader downstream analyses (e.g., heterogeneity across industries, dynamic effects with alternative identification strategies) are outside the scope. As the labeled corpus scales, we intend to expand evaluations and ablations to quantify trade-offs between context length, chunking strategy, and decoding settings on end-to-end research outcomes.

Finally, this work presents both positive and negative societal implications. On the positive side, our approach democratizes access to compensation consultant data that was previously available only through expensive proprietary databases or costly manual collection. However, automated extraction algorithms could produce misleading analyses or misguided policy implications if deployed without proper validation protocols.

References

[1] Lucian Arye Bebchuk and Jesse M. Fried. Executive compensation as an agency problem. *Journal of Economic Perspectives*, 17(3):71–92, 2003. doi: 10.1257/089533003769204362.

[2] Brian Cadman, Mary Ellen Carter, and Stephen Hillegeist. The incentives of compensation consultants and CEO pay. *Journal of Accounting and Economics*, 49(3):263–280, 2010.

[3] Jeh-Hyun Cho, Iny Hwang, Jeong-Hoon Hyun, and Jae Yong Shin. Compensation Consultant Fees and CEO Pay. *Journal of Management Accounting Research*, 32(1):51–78, 2020.

[4] Chroma Research Team. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma, July 2025. `https://research.trychroma.com/context-rot`.

[5] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. URL `https://arxiv.org/abs/2305.14314`.

[6] Zeyu Han, Chao Gao, Jinyang Liu, Jeff Zhang, and Sai Qian Zhang. Parameter-efficient fine-tuning for large models: A comprehensive survey, 2024. URL `https://arxiv.org/abs/2403.14608`.

[7] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing a multi-hop qa dataset for comprehensive evaluation of reasoning steps, 2020. URL `https://arxiv.org/abs/2011.01060`.

[8] Cheng-Ping Hsieh, Simeng Sun, Samuel Kriman, Shantanu Acharya, Dima Rekesh, Fei Jia, Yang Zhang, and Boris Ginsburg. Ruler: What's the real context size of your long-context language models?, 2024. URL `https://arxiv.org/abs/2404.06654`.

[9] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL `https://arxiv.org/abs/2106.09685`.

[10] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL `http://dx.doi.org/10.1145/3703155`.

[11] Sanjay Joshi. A literature review of gen ai agents in financial applications: Models and implementations, 01 2025. ISSN 2319-7064. URL `https://doi.org/10.21275/sr25125102816`.

[12] Tushar Khot, Peter Clark, Michal Guerquin, Peter Jansen, and Ashish Sabharwal. Qasc: A dataset for question answering via sentence composition, 2020. URL `https://arxiv.org/abs/1910.11473`.

[13] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners, 2023. URL `https://arxiv.org/abs/2205.11916`.

[14] David Kuo Chuen Lee, Chong Guan, Yinghui Yu, and Qinxu Ding. A comprehensive review of generative ai in finance, 09 2024. ISSN 2674-1032. URL `https://doi.org/10.3390/fintech3030025`.

[15] Huaxia Li, Haoyun Gao, Chengzhang Wu, and Miklos A. Vasarhelyi. Extracting financial data from unstructured sources: Leveraging large language models. *SSRN Electronic Journal*, 01 2023. ISSN 1556-5068. doi: 10.2139/ssrn.4567607. URL `https://doi.org/10.2139/ssrn.4567607`.

[16] Mo Li, Songyang Zhang, Taolin Zhang, Haodong Duan, Yunxin Liu, and Kai Chen. Needlebench: Can llms do retrieval and reasoning in information-dense context?, 2025. URL `https://arxiv.org/abs/2407.11963`.

[17] Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts, 2023. URL `https://arxiv.org/abs/2307.03172`.

[18] Kangtong Mo, Wenyan Liu, Xuanzhen Xu, Chang Yu, Yuelin Zou, and Fangqing Xia. Fine-tuning gemma-7b for enhanced sentiment analysis of financial news headlines, 2024. URL https://arxiv.org/abs/2406.13626.

[19] Kevin J. Murphy and Tatiana Sandino. Compensation consultants and the level, composition, and complexity of ceo pay. *The Accounting Review*, 95:311–341, 05 2019. ISSN 0001-4826, 1558-7967. doi: 10.2308/accr-52439. URL https://doi.org/10.2308/accr-52439.

[20] Saurav Pawar, S. M Towhidul Islam Tonmoy, S M Mehedi Zaman, Vinija Jain, Aman Chadha, and Amitava Das. The what, why, and how of context length extension techniques in large language models – a detailed survey, 2024. URL https://arxiv.org/abs/2401.07872.

[21] Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, Robert Busa-Fekete, Alex Feng, Noveen Sachdeva, Benjamin Coleman, Yi Gao, Basil Mustafa, Iain Barr, Emilio Parisotto, David Tian, Matan Eyal, Colin Cherry, Jan-Thorsten Peter, Danila Sinopalnikov, Surya Bhupatiraju, Rishabh Agarwal, Mehran Kazemi, Dan Malkin, Ravin Kumar, David Vilar, Idan Brusilovsky, Jiaming Luo, Andreas Steiner, Abe Friesen, Abhanshu Sharma, Abheesht Sharma, Adi Mayrav Gilady, Adrian Goedeckemeyer, Alaa Saade, Alex Feng, Alexander Kolesnikov, Alexei Bendebury, Alvin Abdagic, Amit Vadi, András György, André Susano Pinto, Anil Das, Ankur Bapna, Antoine Miech, Antoine Yang, Antonia Paterson, Ashish Shenoy, Ayan Chakrabarti, Bilal Piot, Bo Wu, Bobak Shahriari, Bryce Petrini, Charlie Chen, Charline Le Lan, Christopher A. Choquette-Choo, CJ Carey, Cormac Brick, Daniel Deutsch, Danielle Eisenbud, Dee Cattle, Derek Cheng, Dimitris Paparas, Divyashree Shivakumar Sreepathihalli, Doug Reid, Dustin Tran, Dustin Zelle, Eric Noland, Erwin Huizenga, Eugene Kharitonov, Frederick Liu, Gagik Amirkhanyan, Glenn Cameron, Hadi Hashemi, Hanna Klimczak-Plucińska, Harman Singh, Harsh Mehta, Harshal Tushar Lehri, Hussein Hazimeh, Ian Ballantyne, Idan Szpektor, Ivan Nardini, Jean Pouget-Abadie, Jetha Chan, Joe Stanton, John Wieting, Jonathan Lai, Jordi Orbay, Joseph Fernandez, Josh Newlan, Ju yeong Ji, Jyotinder Singh, Kat Black, Kathy Yu, Kevin Hui, Kiran Vodrahalli, Klaus Greff, Linhai Qiu, Marcella Valentine, Marina Coelho, Marvin Ritter, Matt Hoffman, Matthew Watson, Mayank Chaturvedi, Michael Moynihan, Min Ma, Nabila Babar, Natasha Noy, Nathan Byrd, Nick Roy, Nikola Momchev, Nilay Chauhan, Noveen Sachdeva, Oskar Bunyan, Pankil Botarda, Paul Caron, Paul Kishan Rubenstein, Phil Culliton, Philipp Schmid, Pier Giuseppe Sessa, Pingmei Xu, Piotr Stanczyk, Pouya Tafti, Rakesh Shivanna, Renjie Wu, Renke Pan, Reza Rokni, Rob Willoughby, Rohith Vallu, Ryan Mullins, Sammy Jerome, Sara Smoot, Sertan Girgin, Shariq Iqbal, Shashir Reddy, Shruti Sheth, Siim Põder, Sijal Bhatnagar, Sindhu Raghuram Panyam, Sivan Eiger, Susan Zhang, Tianqi Liu, Trevor Yacovone, Tyler Liechty, Uday Kalra, Utku Evci, Vedant Misra, Vincent Roseberry, Vlad Feinberg, Vlad Kolesnikov, Woohyun Han, Woosuk Kwon, Xi Chen, Yinlam Chow, Yuvein Zhu, Zichuan Wei, Zoltan Egyed, Victor Cotruta, Minh Giang, Phoebe Kirk, Anand Rao, Kat Black, Nabila Babar, Jessica Lo, Erica Moreira, Luiz Gustavo Martins, Omar Sanseviero, Lucas Gonzalez, Zach Gleicher, Tris Warkentin, Vahab Mirrokni, Evan Senter, Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, Yossi Matias, D. Sculley, Slav Petrov, Noah Fiedel, Noam Shazeer, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Jean-Baptiste Alayrac, Rohan Anil, Dmitry, Lepikhin, Sebastian Borgeaud, Olivier Bachem, Armand Joulin, Alek Andreev, Cassidy Hardin, Robert Dadashi, and Léonard Hussenot. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[22] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique: Multihop questions via single-hop question composition, 2022. URL https://arxiv.org/abs/2108.00573.

[23] Jiahao Wang and Victor Xiaoqi Wang. Leveraging large language models to democratize access to costly financial datasets for academic research. *arXiv (Cornell University)*, 12 2024. doi: 10.48550/arxiv.2412.02065. URL http://arxiv.org/abs/2412.02065.

[24] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models, 2023. URL https://arxiv.org/abs/2201.11903.

[25] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. Hotpotqa: A dataset for diverse, explainable multi-hop question answering, 2018. URL https://arxiv.org/abs/1809.09600.

[26] Tongyao Zhu, Qian Liu, Haonan Wang, Shiqi Chen, Xiangming Gu, Tianyu Pang, and Min-Yen Kan. Skyladder: Better and faster pretraining via context window scheduling, 2025. URL https://arxiv.org/abs/2503.15450.

A Screenshot of ChatGPT log where CHATGPT failed to follow the given instruction

ChatGPT 5, the latest LLM by OpenAI, was first given the same instruction as shown in Appendix B. We queried a series of text chunks. It continued to give the correct answers until it finally misclassified a consultant name after a succession of lengthy queries. When reminded of the rule, ChatGPT5 immediately acknowledged that it had failed to follow the instruction although it was aware of the rule; it says, it will "re"-apply the rule.

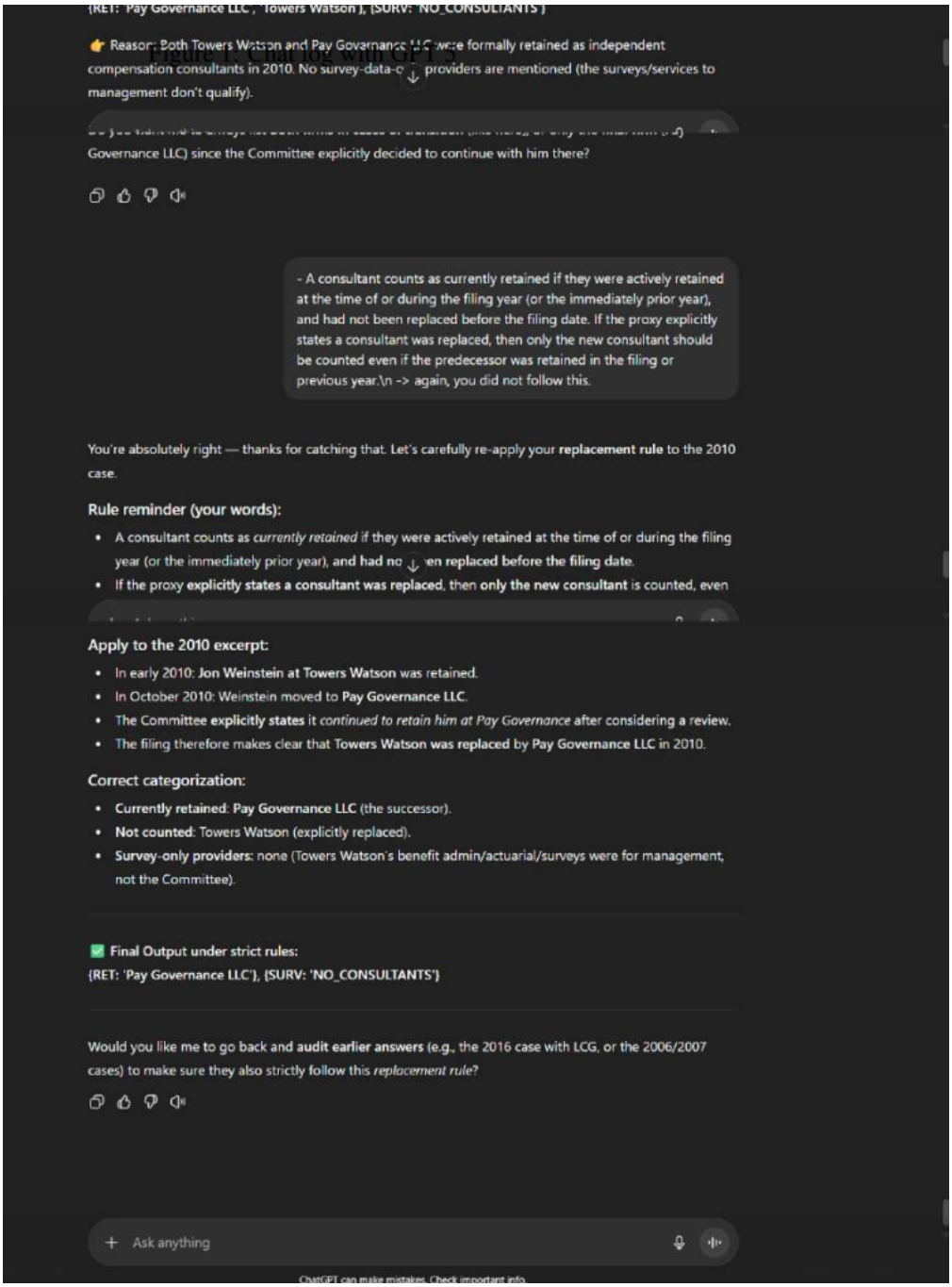

Figure 1: Excerpt from ChatGPT log illustrating acknowledgment of its mistake after a series of long-context queries.

B Prompt texts for training and retrieving information on compensation consultants

---

You are a meticulous reader of proxy-statement excerpts.task is to identify two categories of compensation consultants mentioned in the text:
1. RET: Consultants retained/engaged/selected/hired by the committee or management to advise on executive compensation (e.g., conduct peer group analysis, write up a report).
2. SURV: Consultants that provided compensation surveys and/or benchmarking/market data but were NOT explicitly retained/engaged. If there is a reference that a retained entity explicitly provides those, list it also under SURV.

INPUT FORMAT
You will receive a collection of excerpts separated by the token < < <EXCERPT_BREAK> > >. Treat them as one document.

Strict Output Rules (CUSTOM FORMAT; NOT JSON):
1. Return two dictionaries, exactly in this order and with these keys: RET: 'Firm A', 'Firm B', SURV: 'Firm C'
2. When a firm appears under multiple names, output *only the LONGEST name* found in the whole document/excerpts.
3. List the firms in *alphabetical order* in each dictionary, with every firm enclosed in single quotes.
4. If no firms are found for a category, use 'NO_CONSULTANTS'. If a firm's name is not disclosed, use 'UNNAMED'.

IMPORTANT TIMING RULES:
- You will see a filing year at the start of each document. If a following excerpt/context contains no specific year, assume the information is current.
- Fiscal year: Proxy statement may refer to a fiscal year. A fiscal year is 12 months but starts later than January 1, depending on business needs or regulatory requirements. Thus, Fiscal Year 2022 could refer to a period in 2023.
- A consultant counts as currently retained if they were actively retained at the time of or during the filing year (or the immediately prior year), and had not been replaced before the filing date. If the proxy explicitly states a consultant was replaced, then only the last consultant should be counted even if the predecessor was retained in the filing or previous year.

TEMPORAL CONTEXT:
This document was filed in `{filing_year}`. Consultants retained in `{filing_year}` or `{filing_year-1}` are considered 'currently retained'.

Text to Analyze: `{text}`

C Sample text for sanity check

Sample input text used in the sanity check and comparison of answers from the base and fine-tuned Gemma3 12B-Instruct (4-bit quantized).
The correct consultant names are written in green. The red herrings that managed to confuse the base model are marked in red.

INPUT:
Currently, Ms. Young is a senior consultant for TPH Asset Management, LLC, an affiliate of **Tudor, Pickering, Holt & Co.**
«<EXCERPT_BREAK»>

Committee meetings are attended by, for all or a portion of each meeting, not only the Committee members but also our Chief Executive Officer, our Vice President of Human Resources, an independent compensation consultant from **Compensia, Inc.**, an independent auditor from KPMG, and legal counsel from Fenwick West LLP.
«<EXCERPT_BREAK»>

In fulfilling its oversight responsibilities, the Audit Committee reviewed and discussed the Company's audited financial statements for the fiscal year ended December 31, 2006, with AMD management and Ernst & Young LLP, our independent registered public accounting firm. The Audit Committee also discussed with Ernst & Young LLP the matters required to be discussed by Statement on Auditing Standards No. 61 (Communications with Audit Committees, as amended by Statement on Auditing Standards No. 90 (Audit Committee Communications)). This included a discussion of the independent registered public accounting firm's judgments as to the quality, not just the acceptability, of the Company's accounting principles and such other matters that generally accepted auditing standards require to be discussed with the Audit Committee. The Audit Committee also received the written disclosures and the letter from Ernst & Young LLP required by Independence Standards Board Standard No. 1 (Independence Discussion with Audit Committees) and the Audit Committee discussed the independence of Ernst & Young LLP with that firm.
«<EXCERPT_BREAK»>

Ms. **Towers Watson** served on the Compensation and Human Resources Committee until her resignation from the Board on July 4, 2014, and Mr. Frederic Cook was appointed to the Compensation Committee effective January 1, 2020.
«<EXCERPT_BREAK»>

To assist the Committee in its review of named executive compensation, the Company's Human Resources Department, the Company's primary external compensation consultant, **Compensia, Inc.**, (retained both by management and by the Committee) as well as **Pearl, Meyer & Partners** (retained solely by the Committee for assistance relative to CEO compensation), provide compensation data compiled both from executive compensation surveys (including proprietary surveys conducted by **Radford Surveys + Consulting, Inc., Buck Consultants, LLC and Compensia, Inc.**), and annual reports and proxy statements from companies that the Committee selects as a "peer group" of technology companies for executive compensation analysis purposes.

ANSWER (BASE): RET: [Compensia, Inc., Pearl, Meyer & Partners, Tudor, Pickering, Holt & Co.], SURV: [Buck Consultants, LLC, Compensia, Inc., Radford Surveys + Consulting, Inc., Towers Watson]
ANSWER (FINE-TUNED): RET: [Compensia, Inc., Pearl, Meyer & Partners], SURV: [Buck Consultants, LLC, Compensia, Inc., Radford Surveys + Consulting, Inc.]

D Excerpts from proxy statements

Text chunks from proxy statements describing past consultants

Prior to selecting Veritas as its independent advisor, the committee would from time to time retain Towers Watson (formerly Towers Perrin), an outside compensation and benefits consulting firm, regarding management pay, compensation design and other related matters. In 2009, Towers Watson was engaged by the committee to attend certain meetings of the committee in order to provide insights on current executive compensation trends and practices. For 2009, the committee incurred fees of approximately $10,000 for Towers Watson's meeting attendance. The use of Towers Watson by the committee did not preclude management from hiring the same firm. In 2009, Towers Watson was engaged by management to provide executive compensation advice and recommendations on the following: (Note: Only Veritas should be retrieved as the current consultant.)
The compensation and governance committee first engaged Mercer in 1999 and then again in 2004 and 2007 to review our existing compensation practices and suggest changes based on trends and developments impacting executive compensation and its best practices knowledge. In 2012, the committee engaged FTI Consulting, a nationally recognized compensation consultant specializing in the real estate industry, to review our executive and non-employee director compensation programs. It was favorably noted that our stockholders overwhelmingly approved our executive compensation program at the 2014 annual meeting. Holders of 77.7 million shares of our common stock, or over 97% of the total votes cast, voted FOR the advisory vote on executive compensation. (HIW, 2015) (Note: There is no information about its current compensation consultant since the filing year is 2015.)

Text chunks from proxy statements describing data-providing consultants

The compensation surveys that we reviewed and summarized in the aggregate for the Committee in connection with establishing compensation for 2009 were published by: Carlson Dettmann Associates, Compensation Resources, Conference Board, Hewitt Associates, MACA, Mercer, Stanton Group (a division of Gallagher Benefits Services), Watson Wyatt Data Services and WorldAtWork.', 'In October 2006, FWC provided the Committee with a report on the primary components of our executive compensation program (base salary, annual bonus and longer-term incentives) (MGIC INVESTMENT, 2010) (Note: there are multiple consultants that only provided data, but there is no retained consultant given that the filing year is 2010. FWC is not the current consultant.)
After we developed this new retailer-only peer group with Watson Wyatt, Watson Wyatt provided the Committee with comparative executive compensation data it obtained from the proxy statements and other reports made public by these retailers.', 'The Committee charged Watson Wyatt with assisting it to meet the following primary objectives:', 'Additionally, the Committee reviewed executive compensation data from a broader base of companies that was aggregated in one or more of the compensation surveys obtained from Mercer Human Resource Consulting, Towers Perrin, Hewitt Associates and Watson Wyatt.', 'For fiscal 2008, the Committee worked with Watson Wyatt to develop a group of retailers that we believe is similarly situated to us and with whom we compete for talent.', 'In establishing executive compensation for fiscal 2008, the Committee subscribed to multiple compensation surveys and engaged Watson Wyatt Worldwide ("Watson Wyatt") to provide research, comparative compensation data and executive compensation program design expertise. (BIG, 2009)

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: Our paper clearly states its contribution as demonstrating performance gains by fine-tuning an open-weights LLM for extracting compensation consultant data from SEC filings. The abstract and introduction accurately reflect the 103% F1-score improvement achieved and the specific challenges addressed.
   Our paper is the first research in the intersection between artificial intelligence (AI) and finance that demonstrates the performance gains by fine-tuning a open-weights large language model (LLM) with carefully curated training set. It shows that the latest LLMs can help collect data that were not previously available with much higher accuracy than hand-collection. It also approaches the existing empirical problem in economics/finance through a new lens of machine learning (ML). This study is also the first project that employs double machine learning (DML) that aims to measure true treatment effects on the outcome, addressing endogeneity issues, prevalent in social phenomena.

   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 includes a dedicated Limitations section discussing compute constraints, long-context management challenges, training-inference mismatch, and evaluation scope limitations.
   The paper explicitly discusses a few limitations of the study, including: (1) restricted computational resources that necessitated using a smaller open-weights LLM with PEFT, LoRA, and 4-bit quantization, thereby preventing the use of higher-parameter models; (2) limited quantity of high-quality training data; (3) constraints on context length to prevent context rot, rooted in length discrepancies between training data and text chunks of proxy statements that contain relevant information; and the resultant need to cap the model's context length at 3,000 tokens, leading to inefficiencies in processing proxy statement texts. It also highlights the challenge LLMs face in maintaining optimal performance when there is a length discrepancy between training and evaluation data. Addressing this will require continued research into context extension techniques that allow models to process full documents without performance degradation.

   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: [NA]

   Justification: This paper does not include theoretical results; it is an empirical study of fine-tuning LLMs for financial document processing.

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 detailed fine-tuning settings including model architecture (Gemma 3 12B-Instruct), LoRA configuration (rank=8, alpha=16, dropout=0.05), quantization method (4-bit NF4), context window (2,048 tokens), and decoding parameters. Code is available on GitHub.

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: [Yes]

Justification: The code is available on GitHub. The expert-annotated training set is proprietary and still under development with evolving label logics.

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 optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4 provides complete fine-tuning details including LoRA parameters, quantization settings, context window, optimizer settings, and evaluation protocol on 146 proxy statement excerpts.

   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 reports single evaluation results without error bars or confidence intervals. This is a limitation we acknowledge, though the 103% performance improvement is substantial.

   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: [Yes]

   Justification: The paper specifies using a single 24GB consumer GPU with 4-bit quantization for both training and inference.

   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 paper follows all aspects of the NeurIPS Code of Ethics.

   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 Limitations section discusses both positive impacts (democratizing access to compensation data, improving transparency) and negative impacts (potential misuse for misleading financial analyses).

    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 optimizing 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 poses no such risks as it focuses on extracting publicly available information from SEC filings.

    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 properly cites Gemma 3 model and all open-source libraries used (LoRA, QLoRA, etc.) with appropriate references.

    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 current version of this paper does not release new assets beyond the code repository.

    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: [NA]

    Justification: The paper does not involve crowdsourcing nor research with human subjects.

    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 paper does not involve crowdsourcing nor research with516 human subjects.

    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: [NA]

    Justification: The main arguments of this study along with the research design are the original ideas of the authors.

    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.