

Semi-Supervised Learning to Improve Detection of Factual and Causal Errors in Language Models' Explanations

Mingxuan Liu Han Yuan Yilin Wu Li Zhang Zheng Ma

Global Decision Science, American Express

{Mingxuan.Liu, Han.Yuan1, Yilin.Wu, Li.Zhang1, Zheng.Ma2}@amex.com

Abstract

Language models (LMs) are increasingly used to generate decisions and explanations for high-stakes financial tasks, but these explanations can contain factual inconsistencies with the underlying claim or invalid causal reasoning. We propose a two-step auditing framework that decomposes each explanation into reasoning points and verifies them using natural language inference (NLI): (i) claim $\Rightarrow$ point for factuality verification and (ii) point $\Rightarrow$ conclusion for causality verification. To reduce reliance on expensive expert annotations, we study effective fine-tuning strategies based on semi-supervised learning (SSL) for NLI verifiers, leveraging abundant unlabeled explanations via pseudo-labeling and FreeMatch with self-adaptive confidence thresholds. Experiments across small language models and both encoder-only and decoder-only verifiers under varying labeling budgets show that SSL consistently outperforms the pre-trained baseline and supervised fine-tuning. FreeMatch is generally more robust than simple pseudo-labeling, suggesting better utilization of unlabeled data for scaling error detection of LMs' generations.

1 Introduction

The integration of Generative Artificial Intelligence (GenAI) into financial services is accelerating, yet the "Black Box" nature of language models (LMs) presents significant operational risks (Loughran and McDonald, 2011). In financial analysis, an explanation that contains a factual error (e.g., stating "The applicant has no missed payments" when the recorded history shows delinquencies) or a causal fallacy (e.g., "Credit score improved due to recent hard inquiries") can lead to liability and regulatory non-compliance. Regulators, such as the U.S. Securities and Exchange Commission (SEC) and the European Securities and Markets Authority (ESMA), have increasingly emphasized explainability and robust controls to ensure the reliability and accu-

racy of AI-generated outputs in financial services (U.S. Securities and Exchange Commission, 2024; European Securities and Markets Authority, 2024). Consequently, there is a need for automated auditing tools that can validate the model explanations.

Explanation auditing requires fine-grained supervision (e.g., point-level decisions for factual and causal checks), but such labels are expensive in finance because they must be produced or verified by domain experts and often require cross-checking supporting evidence. As a result, labeled datasets are typically small relative to the volume of available unlabeled narratives (e.g., model-generated rationales), creating a bottleneck for training robust detectors. This motivates semi-supervised learning (SSL) that leverages abundant unlabeled data to reduce reliance on manual annotation, as demonstrated by data-efficient SSL methods (Lee, 2013; Berthelot et al., 2019; Wang et al., 2023b).

In this study, we enhance GenAI explanation auditing in financial services under limited supervision by combining reasoning-point verification with label-efficient SSL approaches. Specifically, we make the following contributions:

- Reasoning-point auditing for factual and causal errors.** We cast explanation auditing as binary NLI at the reasoning-point level, enabling fine-grained identification of which specific statements contradict the claim (factual) or fail to support the conclusion (causal).
- Practical SSL for scarce expert labels.** To reduce reliance on costly domain-expert annotation, we leverage unlabeled explanations using two complementary SSL strategies: (i) simple pseudo-labeling (Lee, 2013) and (ii) self-adaptive thresholding via FreeMatch (Wang et al., 2023b) to scale auditing models as unlabeled data grows.
- Realistic evaluation.** We evaluate across mul-

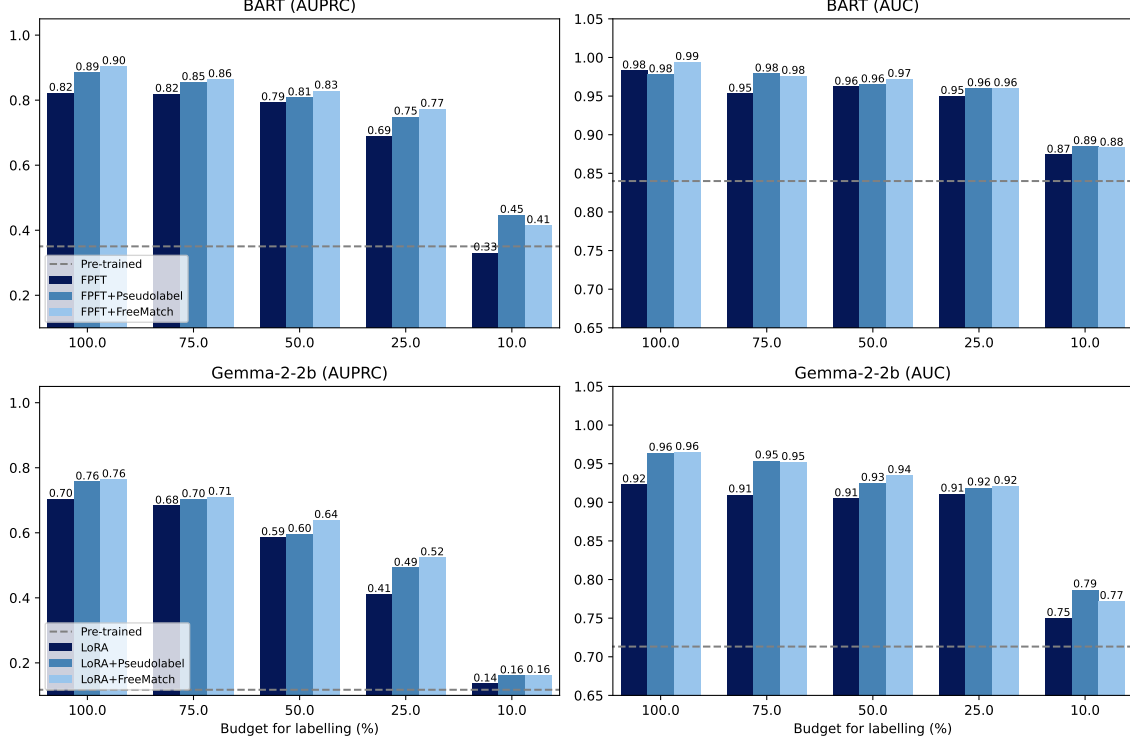


Figure 1: Factual error detection with labeling budget for encoder-only and decoder-only NLI verifiers under different fine-tuning strategies. FPFT: full-parameter fine-tuning. LoRA: Low-Rank Adaptation.

multiple small language models (SLMs) and multiple labeled-data ratios, demonstrating consistent benefits of SSL for explanation auditing in data-scarce settings.

2 Related work

2.1 LMs’ Reasoning and Process Supervision

Natural-language rationales can improve LMs’ reasoning and make model behavior more auditable in high-stakes domains (Xie et al., 2023). Chain-of-thought prompting and self-consistency use step-by-step reasoning to improve performance (Wei et al., 2022; Wang et al., 2023a), but such rationales are not always faithful and can appear plausible even when the final answer is wrong or biased (Turpin et al., 2023). This motivates process-level verification: instead of supervising only final outcomes, verifiers assess intermediate reasoning steps to identify unsupported or faulty claims (Cobbe et al., 2021; Uesato et al., 2022; Lightman et al., 2024; Wu et al., 2025). Related work studies self-checking, step-aware verifiers, factuality benchmarks such as TruthfulQA, and causal-reasoning evaluations such as CLadder (Miao et al., 2024; Li et al., 2023; Weng et al., 2023; Lin et al., 2021; Jin et al., 2023). Our work follows this direction by

verifying each reasoning point for factual consistency (claim $\Rightarrow$ step) and causal consistency (step $\Rightarrow$ conclusion) in a unified NLI formulation.

2.2 Semi-Supervised Learning

SSL improves label efficiency by leveraging unlabeled data alongside a small labeled set. One line of work uses *pseudo-labeling* (self-training), turning high-confidence model predictions into training targets (Lee, 2013); recent thresholding-based methods combine this idea with strong augmentation (FixMatch (Sohn et al., 2020); FlexMatch (Zhang et al., 2021); FreeMatch (Wang et al., 2023b)). A second line focuses on *consistency regularization*, enforcing stable predictions across perturbations or over time (Temporal Ensembling (Laine and Aila, 2017); Mean Teacher (Tarvainen and Valpola, 2017); UDA (Xie et al., 2020)), and can be paired with representation learning objectives such as contrastive losses (Chen et al., 2020; Khosla et al., 2020). Finally, *hybrid* frameworks such as MixMatch (Berthelot et al., 2019) and ReMixMatch (Berthelot et al., 2020) explicitly blend pseudo-labeling with consistency and distribution alignment. Despite broad success, SSL is still less explored for LMs’ explanation auditing, where un-

labeled rationales are plentiful but labels are step-level and pseudo-label noise can be amplified by fluent yet incorrect explanations.

3 Methodology

3.1 Problem Formulation

Given a financial claim C and a generated explanation E , we detect factual and causal errors by casting verification as NLI, as explored in former explanation reliability analysis (Yuan et al., 2025).

We decompose E into m reasoning points $E = (r_1, \dots, r_m)$ and denote the final conclusion as r_{concl} .

Factual NLI (Claim $\Rightarrow$ Point). For each r_i ,

$$P_i^{(f)} = C, \quad H_i^{(f)} = r_i.$$

Causal NLI (Point $\Rightarrow$ Conclusion). For each r_i ,

$$P_i^{(c)} = r_i, \quad H_i^{(c)} = r_{\text{concl}}.$$

We use a binary NLI label set $\mathcal{Z} = \{\text{Entail}, \text{Contradict}\}$. Neutral is ignored because, under our constructed pairs of premise P_i and hypothesis H_i , the verifier is expected to make a *decisive* judgment between Entail and Contradict; Neutral typically reflects missing/insufficient evidence rather than correctness, and can be handled as abstention outside the core detector.

An NLI model f_θ predicts

$$\hat{z}_i^{(t)} = \arg \max_{z \in \mathcal{Z}} f_\theta(P_i^{(t)}, H_i^{(t)})[z], \quad t \in \{f, c\},$$

where $\hat{z}_i^{(f)} = \text{Contradict}$ flags a factual inconsistency between C and r_i , and $\hat{z}_i^{(c)} = \text{Contradict}$ flags a causal inconsistency between r_i and r_{concl} .

3.2 NLI verifier

We use three SLMs as g_ϕ , e.g., Llama-3.2-3B, Phi-3.5-3.8B, and Gemma-2-2B, to generate an explanation $E = g_\phi(c)$ for a financial claim c (Yuan et al., 2026a). We use GPT-4.1 to split E into m reasoning points $E = (r_1, \dots, r_m)$; the prompts are detailed in the Appendix. After that, we use an NLI verifier f_θ to predict $\hat{z} \in \{\text{Entail}, \text{Contradict}\}$ for factual and causal pairs. The verifier can be instantiated as either an encoder-only model like BART or a decoder-only model such as Gemma-2-2B.

3.3 SSL framework

We consider two SSL approaches to leverage unlabeled explanations. We focus on pseudo-labeling and FreeMatch because they are simple, model-agnostic drop-in objectives for text NLI: they only require generating pseudo-labels and enforcing prediction consistency under light textual augmentations, without label-mixing/interpolation. In contrast, hybrid methods such as MixMatch (Berthelot et al., 2019) and ReMixMatch (Berthelot et al., 2020) rely on interpolation (e.g., MixUp) and distribution alignment, which are less robust for the current context where mixing or aggressive perturbations can twist the factual or causal meaning.

3.3.1 Pseudo-labeling

A teacher model f_{θ_T} generates hard pseudo-labels $\tilde{y} = \arg \max f_{\theta_T}(x)$ for unlabeled samples $x \in D_U$, and a student f_{θ_S} is trained on the union of labeled D_L and pseudo-labeled $\{(x, \tilde{y})\}$ samples.

3.3.2 FreeMatch

We adopt FreeMatch to perform confidence-adaptive pseudo-labeling with consistency regularization: a sample is assigned a pseudo-label only if $\max f_{\theta_T}(x)$ exceeds an adaptive threshold, and the student is trained to make consistent predictions under weak and strong augmentations. Concretely, we use the original input as the weak augmentation and synonym replacement as the strong augmentation. We also ablate augmentation type and strength and find that both designs affect performance.

3.4 Dataset

We use the German credit financial classification dataset (Hofmann, 1994), a widely recognized benchmark in financial natural language processing (Xie et al., 2023; Bhatia et al., 2024; Yuan et al., 2026b). In this dataset, label $L = 1$ indicates a good profile and $L = 0$ represents a bad profile. We first adopt LMs to generate profile decisions and corresponding explanations for the samples, and then focus on detecting factual and causal errors in the generated explanations to facilitate a more informed understanding of the outputs.

Specifically, we segment each LMs’ explanations into reasoning points and have multiple domain experts annotate each point with binary NLI labels (*Entail* vs. *Contradict*) for factual (claim $\Rightarrow$ point) and causal (point $\Rightarrow$ conclusion) verification. We vary the labeling budget by subsampling the labeled set while keeping the unlabeled set fixed: a

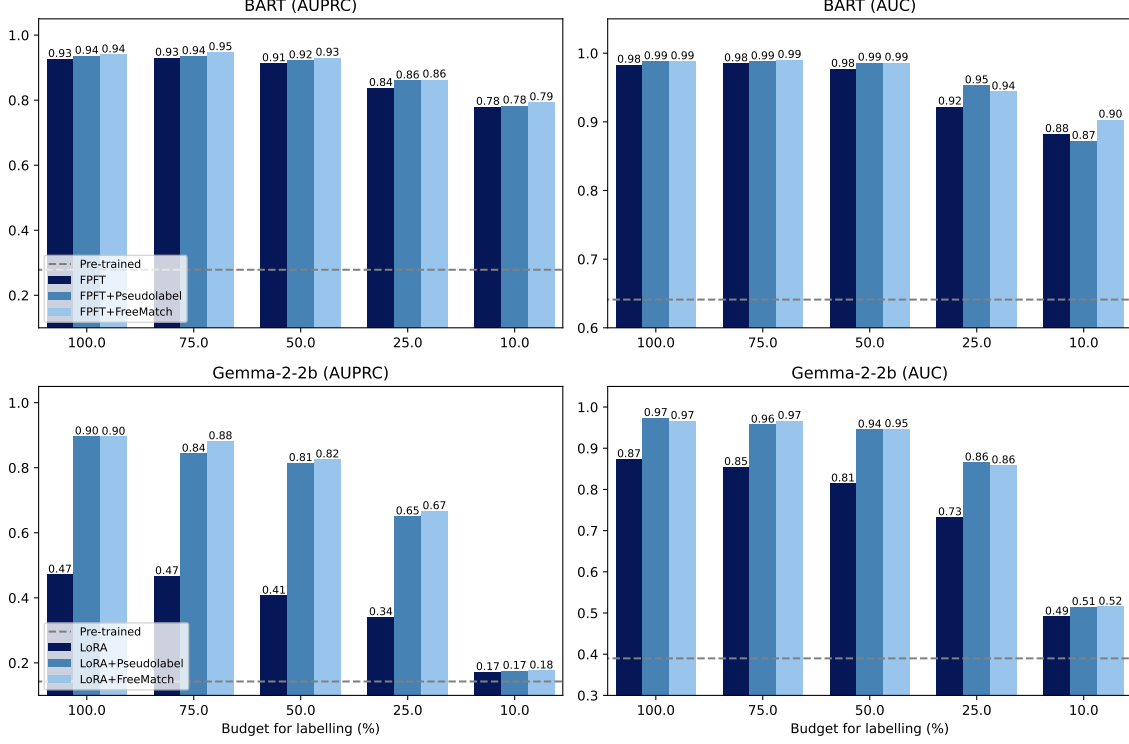


Figure 2: Causal error detection with labeling budget for encoder-only and decoder-only NLI verifiers under different fine-tuning strategies. FPFT: full-parameter fine-tuning. LoRA: Low-Rank Adaptation.

ratio of 1.0 uses all labeled data together with all unlabeled data, while a ratio of 0.5 uses 50% of the labeled data together with all unlabeled data.

4 Experiments

4.1 Factuality

We first evaluate factual error detection and study how SSL strategies affect performance under different labeling budgets.

Fig. 1 shows that SSL substantially improved factual error detection over the pre-trained model and supervised fine-tuning (FPFT / LoRA (Hu et al., 2022)) for both models. For each fine-tuning strategy, the benefit of additional supervision exhibited diminishing returns as the labeling budget increases (a concave trend). Overall, FreeMatch was consistently competitive and often stronger than pseudo-labeling at moderate-to-high budgets ($\geq 25\%$), indicating better control of pseudo-label noise; this is consistent with higher unlabeled-data utilization under FreeMatch in Fig. 3 for both verifiers.

4.2 Causality

Fig. 2 shows that SSL fine-tuning also improved causal error detection over both the pre-trained baseline and supervised fine-tuning. Consistent

with the factuality, the improvement of return diminishes as supervision increases, and FreeMatch is typically stronger than pseudo-labeling.

For the encoder-only verifier, the marginal benefit of adding SSL over supervised fine-tuning is relatively modest, whereas the SSL gains are more pronounced with the decoder-only verifier (lower panels in Fig. 2), where SSL yields a larger lift relative to its supervised fine-tuned model. For example, under the budget of 50% labeling, Gemma-2-2b achieved AUPRCs of 0.41 for LoRA, 0.81 for Pseudolabeling, and 0.82 for FreeMatch.

5 Conclusions

This study contributes to efficient explanation auditing for LMs in a high-stakes financial context by verifying factual and causal correctness at the reasoning-point level and by showing that semi-supervised fine-tuning can improve verifiers when expert labels are scarce. The resulting verifiers can support governance workflows by triaging generated rationales for another round of human review, flagging potentially misleading statements, and monitoring explanation reliability under model and data drift. Future work should test broader financial scenarios, more diverse LMs, lightweight

verifiers, stronger augmentations that preserve factual and causal meaning based on LMs, etc.

References

- David Berthelot, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Kihyuk Sohn, Han Zhang, and Colin Raffel. 2020. Remixmatch: Semi-supervised learning with distribution matching and augmentation anchoring. In *International Conference on Learning Representations*.
- David Berthelot, Nicholas Carlini, Ian Goodfellow, Nicolas Papernot, Avital Oliver, and Colin Raffel. 2019. Mixmatch: A holistic approach to semi-supervised learning. In *Advances in Neural Information Processing Systems*.
- Gagan Bhatia, El Moatez Billah Nagoudi, Hasan Cavusoglu, and Muhammad Abdul-Mageed. 2024. Fintral: A family of GPT-4 level multimodal financial large language models. In *Findings of the Association for Computational Linguistics: ACL*.
- Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning*.
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- European Securities and Markets Authority. 2024. ESMA provides guidance to firms using artificial intelligence in investment services. Statement.
- Hans Hofmann. 1994. Statlog (German credit data). UCI Machine Learning Repository.
- Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.
- Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kai Blin, Fernando Gonzalez Adauro, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: Assessing causal reasoning in language models. In *Advances in Neural Information Processing Systems*.
- Prannay Khosla, Piotr Teterwak, Chen Wang, Aaron Sarna, Yonglong Tian, Phillip Isola, Aaron Maschinot, Ce Liu, and Dilip Krishnan. 2020. Supervised contrastive learning. In *Advances in Neural Information Processing Systems*.
- Samuli Laine and Timo Aila. 2017. Temporal ensembling for semi-supervised learning. In *International Conference on Learning Representations*.
- Dong-Hyun Lee. 2013. Pseudo-label: The simple and efficient semi-supervised learning method for deep neural networks. In *ICML Workshop on Challenges in Representation Learning*.
- Yifei Li, Zeqi Lin, Shizhuo Zhang, Qiang Fu, Bei Chen, Jian-Guang Lou, and Weizhu Chen. 2023. Making language models better reasoners with step-aware verifier. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*.
- Hunter Lightman, Vineet Kosaraju, Yura Burda, Harri Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2024. Let’s verify step by step. In *International Conference on Learning Representations*.
- Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. TruthfulQA: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.
- Tim Loughran and Bill McDonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-Ks. *The Journal of Finance*, 66(1):35–65.
- Ning Miao, Yee Whye Teh, and Tom Rainforth. 2024. Selfcheck: Using LLMs to zero-shot check their own step-by-step reasoning. In *International Conference on Learning Representations*.
- Swaroop Mishra, Matthew Finlayson, Pan Lu, Laria Tang, Sean Welleck, Chitta Baral, Tanmay Rajpurohit, Oyvind Tafjord, Ashish Sabharwal, Peter Clark, and Ashwin Kalyan. 2022. LILA: A unified benchmark for mathematical reasoning. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*.
- Kihyuk Sohn, David Berthelot, Chun-Liang Li, Zizhao Zhang, Nicholas Carlini, Ekin D. Cubuk, Alex Kurakin, Han Zhang, and Colin Raffel. 2020. Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In *Advances in Neural Information Processing Systems*.
- Antti Tarvainen and Harri Valpola. 2017. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Advances in Neural Information Processing Systems*.
- Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don’t always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems*.
- Jonathan Uesato, Nate Kushman, Ramana Kumar, Francis Song, Noah Siegel, Lisa Wang, Antonia Creswell, Geoffrey Irving, and Irina Higgins. 2022. Solving math word problems with process- and outcome-based feedback. *arXiv preprint arXiv:2211.14275*.
- U.S. Securities and Exchange Commission. 2024. SEC division of examinations announces 2025 priorities. Press Release 2024-172.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023a. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*.

Yidong Wang, Hao Chen, Qiang Heng, Wenxin Hou, Yue Fan, Zhen Wu, Jindong Wang, Marios Savvides, Takahiro Shinozaki, Bhiksha Raj, Bernt Schiele, and Xing Xie. 2023b. Freematch: Self-adaptive thresholding for semi-supervised learning. In *International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yixuan Weng, Min Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, Bin Sun, Jun Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *Findings of the Association for Computational Linguistics: EMNLP*.

Yilin Wu, Han Yuan, Li Zhang, and Zheng Ma. 2025. Natural language inference as a judge: Detecting factuality and causality issues in language model self-reasoning for financial analysis. In *Proceedings of The 10th Workshop on Financial Technology and Natural Language Processing*, pages 210–220.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. PIXIU: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Minh-Thang Luong, and Quoc V. Le. 2020. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*.

Han Yuan, Yilin Wu, Li Zhang, and Zheng Ma. 2026a. Empowering small language models with factual hallucination-aware reasoning for financial classification. In *AAAI 2026 Workshop on Trust and Control in Agentic AI (TrustAgent)*.

Han Yuan, Li Zhang, and Zheng Ma. 2025. Exploring the reliability of self-explanation and its relationship with classification in language model-driven financial analysis. In *ICLR 2025 Workshop on Advances in Financial AI: Opportunities, Innovations and Responsible AI*.

Han Yuan, Yue Zhao, Li Zhang, Wuqiong Luo, and Zheng Ma. 2026b. Quantifying the impact of structured output format on large language models through causal inference. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 1771–1795.

Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. FA*IR: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM Conference on Information and Knowledge Management*.

Bowen Zhang, Yidong Wang, Wenxin Hou, Houwen Wu, Jindong Wang, Manabu Okumura, and Takahiro Shinozaki. 2021. Flexmatch: Boosting semi-supervised learning with curriculum pseudo labeling. In *Advances in Neural Information Processing Systems*.

A Data Preprocessing

The original dataset includes outdated information and pre-existing bias (Zehlike et al., 2017) that pose challenges for SLMs. For instance, certain features are denominated in Deutsche Marks, a currency that has been obsolete for over two decades; moreover, SLMs often exhibit limited sensitivity to numeric reasoning (Mishra et al., 2022). To address these issues, we exclude features misaligned with contemporary SLM development contexts and convert numeric features into percentile representations. All processing steps are performed *ad hoc* and do not involve any operations related to the labels, ensuring that performance is not affected by information leakage. We also conduct all experiments using all minority cases paired with an equal number of majority cases to avoid the adverse impact of data imbalance on the analyses.

B Prompt for Point Segmentation

Segmentation Prompt

Role. You are an expert in credit risk assessment with strong logical reasoning skills.

Task. Given another expert’s assessment, split the reasoning into individual sentences, where each sentence represents a distinct reasoning point.

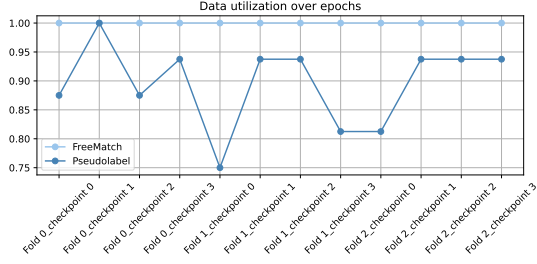
Requirements. Omit summary sentences that aggregate reasoning points; do not alter the original wording—only split the text into separate items.

Transition rules. Split addition, comparison, sequence, and clarification links; keep the closest contrast, cause, and emphasis links together; remove the point that directly states the final good/bad credit conclusion; explicitly resolve references such as “these factors.”

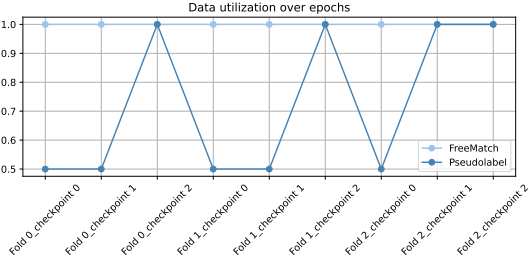
Example. Original reasoning: {Original reasoning}. Structured reasoning: {Structured reasoning}.

C Experimental Details

We obtain 8989 reasoning points from explanations generated by the SLMs and segmented using GPT-4.1. Among these, 1782 points are manually annotated by two authors of this paper, each with more than two years of experience in financial services. Disagreements are adjudicated by



(a) BART



(b) Gemma-2-2b

Figure 3: Utilization of unlabeled data under different SSL strategies.

a senior author with more than five years of experience in financial services. The annotated set contains 4.04% factual errors and 18.46% causal errors. We split the labeled points into three folds for cross-validation, ensuring that all reported evaluation results are computed on manually labeled samples.

For BART, we perform FPFT using AdamW with a learning rate of $1e-5$, weight decay of 0.01, batch size of 32, and train for 5 epochs. For Gemma-2-2B, given its larger model size relative to the encoder-only verifier, we use LoRA with rank 8 and scaling factor 8, and optimize the trainable parameters using AdamW with a learning rate of $5e-5$, weight decay of 0.01, batch size of 4, for 1 epoch. Our code is publicly available on GitHub.¹

D Utilization of Unlabeled Data

For BART (Fig. 3a), FreeMatch maintains near-full utilization throughout training, whereas simple pseudo-labeling fluctuates noticeably, typically staying around 0.85–1.0 but dropping to roughly 0.75 in its lowest-utilization epoch. For Gemma-2-2b (Fig. 3b), FreeMatch again stays close to 1.0 across epochs, while pseudo-labeling exhibits a bimodal pattern, alternating between very low utilization (about 0.2) and full utilization (about 1.0)

¹<https://github.com/Han-Yuan-Med/semi-supervised-learning>

and indicating unstable or overly conservative acceptance of unlabeled samples.

Disclaimer

This paper is provided solely for informational purposes as an academic contribution by the authors to the research community and does not represent, reflect, or constitute the views, policies, positions, or practices of American Express or its affiliates. Nothing in this paper should be cited or relied upon as evidence of, or support for, the business views, policies, positions, or practices of American Express or its affiliates.