Same Meaning, Different Wording: Evaluating Paraphrase Robustness in Financial Natural Language Inference

Yike Li
Swarthmore College
yli11@swarthmore.edu

Adam Poliak
Bryn Mawr College
apoliak@brynmawr.edu

Abstract

We present FINNLI-PARA, a paraphrased version of financial natural language inference (FinNLI). We investigate how much predictions on FinNLI depend on the particular wording used to express a financial hypothesis. We generate three paraphrases for each FinNLI hypothesis and use them to test prediction consistency across eleven model–prompt configurations. Models change their predictions on 7.9–12.3% of paraphrased examples, despite 98.35% of paraphrases preserving the original NLI label, indicating that financial NLI systems remain sensitive to hypothesis wording.

1 Introduction

Natural language inference (NLI) is the task of determining whether one statement can likely be inferred from another (Cooper et al., 1996; Dagan et al., 2006; MacCartney and Manning, 2009; Bowman et al., 2015; Poliak, 2020). In financial settings, the same underlying proposition can often be expressed through multiple linguistic formulations. A robust financial NLI system should therefore reach the same inference when a hypothesis is reformulated without changing its meaning. FinNLI (Magomere et al., 2025) provides a benchmark for financial NLI, but each example contains only a single hypothesis formulation, so strong benchmark performance does not reveal whether predictions remain stable when hypotheses are paraphrased. Robustness to paraphrastic variability is particularly important in finance, where equivalent hypotheses may vary in wording across reports, filings, and earnings calls, while quantities, comparative direction, modality, and temporal relations must remain unchanged for the inference relation to hold.

We introduce FINNLI-PARA, a paraphrased version of FinNLI. For each hypothesis, we generate three sentence-level paraphrases — lexical, syntactic, and plain-language reformulations – while keeping the premise and gold NLI label fixed (Figure 1). Our evaluation complements standard benchmark performance by asking whether models' predictions remain stable when the same financial propositions are expressed differently.

Figure 1: Illustrative example of a FinNLI premise and original hypothesis together with its lexical, syntactic, and plain-language paraphrases. Because each reformulation preserves the same underlying financial claim, all four hypotheses retain the same NLI label, Entailment.

We validate the generated paraphrases through human annotation and find that 98.35% preserve the original NLI label. We then evaluate eleven model–prompt configurations spanning fine-tuned encoder models and instruction-tuned LLMs. Across all systems, paraphrasing reduces macro-F1 by 1.5–3.4 points and changes 7.9–12.3% of paired predictions. Syntactic reformulations and neutral hypotheses produce the highest prediction-change rates, and changes are substantially more likely to turn correct predictions into errors than to repair incorrect ones.

These results show that strong FinNLI performance does not imply robustness to alternative formulations of the same financial hypothesis. FINNLI-PARA complements conventional benchmark evaluation by exposing sensitivity to hypothesis wording that aggregate performance hides.

2 Related Work

Financial natural language inference FinNLI (Magomere et al., 2025) introduces a multi-genre financial NLI benchmark spanning financial reports, filings, and earnings calls. More broadly, recent work has developed increasingly comprehensive benchmarks for evaluating language models in finance. PIXIU combines a financial instruction dataset, a finance-specialized LLM, and a benchmark covering multiple financial NLP and prediction tasks (Xie et al., 2023), while FinBen expands this evaluation to 42 datasets spanning 24 financial tasks and evaluates a broad collection of contemporary LLMs (Xie et al., 2024). InvestorBench extends financial evaluation further to LLM-based agents performing financial decision-making tasks across products and market settings (Li et al., 2025). Related financial inference tasks address economic causal reasoning (Guo and Yang, 2024) and claim verification over long and multimodal financial documents (Zhao et al., 2024). Together, these resources demonstrate the growing breadth of financial language-model evaluation, but they do not test whether model predictions remain stable across alternative formulations of the same financial hypothesis. Building on recent work that positions NLI as a useful framework for evaluating modern language models (Madaan et al., 2025), recent financial NLP work has also used NLI to detect factual and causal errors in language-model reasoning (Wu et al., 2025). These applications further motivate evaluating whether financial NLI systems themselves remain stable under meaning-preserving reformulations.

Paraphrase robustness Paraphrases and semantics-preserving transformations reveal behavior that aggregate held-out performance can obscure (Iyyer et al., 2018; Ribeiro et al., 2018, 2020). Prior work documented inconsistent predictions across semantically equivalent inputs (Jang et al., 2021; Elazar et al., 2021; Rabinovich et al., 2023; Cho and Watson, 2025; Fu and Barez, 2025) and showed that paraphrastic variability accounts for a meaningful fraction of errors in natural language reasoning systems (Srikanth et al., 2024). Arakelyan et al. (2024) study semantic sensitivity in NLI using adversarial surface-form variations filtered through symmetric entailment from the evaluated model. More recently, Zgreabăn et al. (2025) introduce MERGE, which constructs non-adversarial NLI variants through minimal expression replacement and finds consistency failures in both fine-tuned NLI models and LLMs.

Our work most closely follows Verma et al. (2023), who used a T5-based paraphraser fine-tuned on PAWS (Zhang et al., 2019) to rewrite examples from the Pascal RTE1–3 challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007). We extend this line of work to financial NLI by generating lexical, syntactic, and plain-language hypothesis reformulations. Unlike approaches that condition generation or validation on the evaluated NLI model, we generate paraphrases independently of model predictions and manually assess them for semantic and NLI-label preservation. This design tests sensitivity to broader reformulations without conditioning paraphrase construction on the evaluated model.

3 FinNLI-Para Construction

Source data We leave the FinNLI training split unchanged and use the 1,800-example development split to evaluate the paraphrasing procedure before applying it to the test set. Because the procedure achieved high semantic-equivalence and label-preservation rates on the development split, we apply the paraphrasing procedure to the 1,247 available examples in the test set.[1] For each test hypothesis, we generate three reformulations—lexical, syntactic, and plain-language—yielding 3,741 paraphrased hypotheses. To evaluate model robustness, we compare each model's prediction on each of the 1,247 original test examples with its predictions on the three corresponding paraphrased hypotheses.

Controlled paraphrase generation For each FinNLI hypothesis, we generate three variants: *lexical* paraphrases primarily change word choice while retaining approximately the same sentence structure; *syntactic* paraphrases substantially change sentence structure, voice, or clause organization; and *plain-language* paraphrases express the proposition more directly while retaining its financial meaning. Figure 1 illustrates the three variants, and the prompt is provided in Appendix A.

We generate paraphrases locally using Qwen3 8B through Ollama with temperature 0.8, top-p 0.9, and top-k 40. The prompt requires preservation of named entities, numbers, percentages, currencies,

[1]The publicly released FinNLI test set shared by the authors contains only the Llama-generated portion of the benchmark's full test set.

System	Prompt	Original F1	Para F1	Δ F1	Pred. change %	Harmful %	Helpful %
RoBERTa-Large	–	73.58	70.86	-2.71	12.16	6.84	4.17
BART-Large-MNLI	–	74.26	71.65	-2.61	11.20	6.44	3.90
Llama 3.3 70B	ZS	70.90	68.48	-2.42	7.91	4.70	2.35
Llama 3.3 70B	ZSAG	76.31	74.04	-2.27	9.76	5.72	3.45
Llama 3.3 70B	FSAG	78.41	76.31	-2.10	10.29	5.93	3.96
Mistral Medium 3.5	ZS	76.61	74.17	-2.44	9.80	5.89	3.35
Mistral Medium 3.5	ZSAG	78.23	76.72	**-1.50**	12.27	6.63	5.21
Mistral Medium 3.5	FSAG	**79.38**	**77.32**	-2.05	11.28	6.44	4.33
Phi-3 14B	ZS	75.31	72.28	-3.03	11.87	6.98	4.22
Phi-3 14B	ZSAG	76.44	73.03	-3.41	12.16	7.56	4.04
Phi-3 14B	FSAG	75.09	72.72	-2.36	10.26	6.12	3.72

Table 1: Model performance and paraphrase robustness on FinNLI. Original and Para F1 are macro-F1 scores; Pred. change is the percentage of original–paraphrase pairs with different predictions; Harmful and Helpful denote correct→incorrect and incorrect→correct changes

dates and units, negation, modality and uncertainty, comparative direction, quantifiers, and temporal relations. It explicitly prohibits modifying claims.

The paraphrasing model receives only the hypothesis. Providing the premise could encourage premise-conditioned rewriting – adding supporting evidence or correcting a contradictory hypothesis. Moreover, prior work shows that LLM-elicited NLI hypotheses can encode systematic label-associated artifacts (Proebsting and Poliak, 2025). Thus, we hide the premise and gold label during generation.

Human quality audit We manually evaluated all three paraphrases for a stratified sample of 585 FinNLI examples, yielding 1,755 new NLI pairs. The first author judged whether each paraphrase preserved the original hypothesis meaning and NLI label when paired with the premise; the second author spot-checked annotations and adjudicated ambiguous cases. Overall, 95.90% were judged semantically equivalent and 98.35% preserved the original NLI label. As Table 2 shows, label preservation remained high across all three conditions, ranging from 97.78% for plain-language to 99.15% for lexical paraphrases. These results support using the generated reformulations as controlled perturbations for evaluating paraphrase robustness.

Condition	Semantic Equivalence.	Label Pres.
Lexical	97.27	99.15
Syntactic	96.24	98.12
Plain-language	94.19	97.78

Table 2: Human evaluation of paraphrase quality of 1,755 examples. Values represent percentages.

4 Experimental Setup

We evaluate two encoder-based NLI models, RoBERTa-large and BART-large-MNLI, fine-tuned on the original FinNLI training split using the hyperparameters of Magomere et al. (2025). We also evaluate three instruction-tuned LLMs—Phi-3 14B, Llama 3.3 70B, and Mistral Medium 3.5—chosen to span different model families and scales. All LLMs are run locally through Ollama.

For each LLM, we evaluate three prompting configurations: zero-shot (ZS), zero-shot with answer guidance (ZSAG), and few-shot with answer guidance (FSAG). In all settings, the model receives the premise and hypothesis and predicts entailment, contradiction, or neutral. We use the same prompting configuration for each original hypothesis and its three paraphrases, and no model is trained or adapted on paraphrased examples.

We report F1 on the original and paraphrased sets and measure robustness using prediction-change rate, the percentage of original–paraphrase pairs for which the predicted label changes. We distinguish *harmful* changes, where a correct prediction becomes incorrect, from *helpful* changes, where an incorrect prediction becomes correct.

5 Results

Table 1 shows that paraphrasing consistently reduces performance across all evaluated systems. Macro-F1 decreases for every model–prompt configuration, with drops ranging from 1.5 points for Mistral Medium 3.5 under zero-shot answer guidance to 3.4 points for Phi-3 14B under the same condition. Prediction-change rates range from 7.9% to 12.3%, showing substantial instance-level instability despite relatively modest aggregate performance

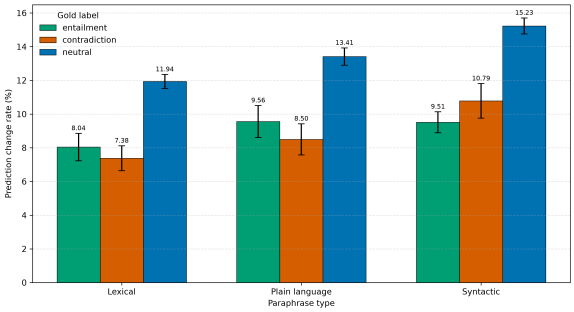

Figure 2: Prediction-change rate by paraphrase type and gold NLI label, averaged across systems. Error bars show standard error of the mean.

degradation. Higher original FinNLI performance does not guarantee greater paraphrase robustness, as some of the strongest systems also exhibit relatively high prediction-change rates.

Our model ranking also differs from that reported by Magomere et al. (2025). In their experiments, fine-tuned RoBERTa-large outperformed all evaluated LLMs except Llama 3.1 70B. On the publicly released test set, several LLM configurations outperform RoBERTa-large on both the original and paraphrased examples. For example, Mistral Medium 3.5 with few-shot answer guidance reaches 79.4 macro-F1 on the original examples and 77.3 on the paraphrases, compared with 73.6 and 70.9 for RoBERTa-large.

Figure 2 shows that prediction changes vary across paraphrase types and NLI labels. Syntactic reformulations produce the highest overall change rates, particularly for contradiction and neutral examples, while lexical paraphrases generally change the least. Neutral hypotheses are the most unstable across all three paraphrase conditions.

Prediction changes are also more harmful than beneficial. Averaged across systems, 6.3% of original–paraphrase pairs change from correct to incorrect, compared with 3.9% that change from incorrect to correct. Among pairs whose prediction changes, 58.3% are harmful, 35.7% are helpful, and 6.0% move between two incorrect labels. Syntactic paraphrases produce the greatest net harm, at 3.2 percentage points, compared with approximately 2.0 points for lexical and plain-language reformulations. Although neutral examples have the highest harmful and helpful change rates, contradiction exhibits the largest net harm at 3.7 points. Thus, prediction changes under paraphrasing are more likely to create errors than to correct them.

Orig. \ Para.	Entailment	Contradiction	Neutral
Entailment	–	4.18%	29.42%
Contradiction	5.15%	–	18.79%
Neutral	27.69%	14.79%	–

Table 3: Prediction transitions across all 11 model–prompt configurations. Cells show the percentage of paraphrase-induced prediction changes and add to 100.02 due to rounding.

Sensitivity to paraphrase quality To ensure that prediction instability is not driven by imperfect paraphrases, we repeat the analysis on the paraphrases manually judged to preserve both the meaning of the original hypothesis and its NLI label. Prediction-change rates remain substantial, ranging from 8.7% to 13.5% across the model–prompt configurations, closely matching and slightly exceeding the 7.9–12.3% range observed in the full evaluation. This result indicates that the observed sensitivity to hypothesis wording persists even when evaluation is restricted to human-verified meaning- and label-preserving reformulations.

Prediction transitions Prediction-change rates do not capture how the inferred NLI relation changes across specific labels. We analyze the direction of every paraphrase-induced prediction change, aggregating across all 11 model–prompt configurations. For each original–paraphrase pair whose predicted label differs, we record the transition from the prediction on the original to the prediction on the paraphrased hypothesis.

Table 3 shows that these changes overwhelmingly involve the neutral class: 90.69% of all prediction changes cross the neutral boundary, while only 9.33% directly reverse entailment and contradiction. The most common transitions are between entailment and neutral. This concentration around the neutral class may reflect the inherently less determinate nature of neutral examples: unlike entailment or contradiction, neutrality often corresponds to insufficient evidence for either relation (Williams et al., 2018; Jiang and Marneffe, 2022; Nighojkar et al., 2023). Models may therefore be particularly sensitive to surface-form changes when distinguishing neutral hypotheses from those expressing an entailed or contradicted claim. This pattern is consistent across paraphrase types: 89.3% of lexical, 91.3% of plain-language, and 91.3% of syntactic prediction changes cross this boundary. Although syntactic paraphrases produce the highest overall

change rates, they do not disproportionately induce entailment–contradiction reversals; most changes still involve the neutral class. Transitions are also asymmetric, with contradiction→neutral occurring more often than neutral→contradiction (18.79% vs. 14.79%).

6 Conclusion

We introduced FINNLI-PARA, a human-validated evaluation of paraphrase robustness in financial natural language inference. Across eleven model–prompt configurations, paraphrasing reduced macro-F1 by 1.5–3.4 points and changed 7.9–12.3% of paired predictions. Instability was greatest for syntactic reformulations and neutral hypotheses, and prediction changes were more likely to turn correct predictions into errors than to repair incorrect ones. These findings show that strong FinNLI performance did not guarantee robustness to alternative formulations of the same financial hypothesis. FINNLI-PARA therefore provides a complementary evaluation of model behavior that is not captured by aggregate benchmark performance alone.

7 Limitations

Our study has several limitations. We use a single LLM to generate paraphrases, and other models may produce different reformulations. Semantic and label preservation cannot be guaranteed automatically; however, we mitigate this risk through human validation. Our lexical, syntactic, and plain-language categories are operational rather than perfectly isolated linguistic transformations. Additionally, human validation was conducted primarily by a single author rather than through independent annotation by multiple annotators. Finally, FinNLI's license restricts redistribution of derivative text; where necessary, we will instead release generation code, prompts, model configuration, validation procedures, and permitted metadata.

References

Erik Arakelyan, Zhaoqi Liu, and Isabelle Augenstein. 2024. Semantic sensitivities and inconsistent predictions: Measuring the fragility of NLI models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 432–444, St. Julian's, Malta. Association for Computational Linguistics.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, and Bernardo Magnini. 2006. The second pascal recognising textual entailment challenge.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.

Nicole Cho and William Watson. 2025. Multiq&a: An analysis in measuring robustness via automated crowdsourcing of question perturbations and answers. In *AAAI 2025 Workshop on Preventing and Detecting LLM Misinformation (PDLM)*.

Robin Cooper, Dick Crouch, Jan Van Eijck, Chris Fox, Jan Jaspars, Hans Kamp, David Milward, Manfred Pinkal, Massimo Poesio, et al. 1996. Using the framework.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2006. The pascal recognising textual entailment challenge. In *Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment*, pages 177–190. Springer.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Tingchen Fu and Fazl Barez. 2025. Same question, different words: A latent adversarial framework for prompt robustness. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31305–31319, Suzhou, China. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. 2007. The third PASCAL recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, pages 1–9, Prague. Association for Computational Linguistics.

Yue Guo and Yi Yang. 2024. EconNLI: Evaluating large language models on economics reasoning. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 982–994, Bangkok, Thailand. Association for Computational Linguistics.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1875–1885, New Orleans, Louisiana. Association for Computational Linguistics.

Myeongjun Jang, Deuk Sin Kwon, and Thomas Lukasiewicz. 2021. Accurate, yet inconsistent? consistency analysis on language understanding models. *arXiv preprint arXiv:2108.06665*.

Nan-Jiang Jiang and Marie-Catherine de Marneffe. 2022. Investigating reasons for disagreement in natural language inference. *Transactions of the Association for Computational Linguistics*, 10:1357–1374.

Haohang Li, Yupeng Cao, Yangyang Yu, Shashidhar Reddy Javaji, Zhiyang Deng, Yueru He, Yuechen Jiang, Zining Zhu, K.P. Subbalakshmi, Jimin Huang, Lingfei Qian, Xueqing Peng, Jordan W. Suchow, and Qianqian Xie. 2025. INVESTORBENCH: A benchmark for financial decision-making tasks with LLM-based agent. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2509–2525, Vienna, Austria. Association for Computational Linguistics.

Bill MacCartney and Christopher D. Manning. 2009. An extended model of natural logic. In *Proceedings of the Eight International Conference on Computational Semantics*, pages 140–156, Tilburg, The Netherlands. Association for Computational Linguistics.

Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2025. Lost in inference: Rediscovering the role of natural language inference for large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 9229–9242, Albuquerque, New Mexico. Association for Computational Linguistics.

Jabez Magomere, Elena Kochkina, Samuel Mensah, Simerjot Kaur, and Charese Smiley. 2025. FinNLI: Novel dataset for multi-genre financial natural language inference benchmarking. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4545–4568, Albuquerque, New Mexico. Association for Computational Linguistics.

Animesh Nighojkar, Antonio Laverghetta Jr., and John Licato. 2023. No strong feelings one way or another: Re-operationalizing neutrality in natural language inference. In *Proceedings of the 17th Linguistic Annotation Workshop (LAW-XVII)*, pages 199–210, Toronto, Canada. Association for Computational Linguistics.

Adam Poliak. 2020. A survey on recognizing textual entailment as an NLP evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 92–109, Online. Association for Computational Linguistics.

Grace Proebsting and Adam Poliak. 2025. Biases in large language model-elicited text: A case study in natural language inference. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 5836–5851, Abu Dhabi, UAE. Association for Computational Linguistics.

Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby Tavor. 2023. Predicting question-answering performance of large language models through semantic consistency. In *Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 138–154, Singapore. Association for Computational Linguistics.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging NLP models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 856–865, Melbourne, Australia. Association for Computational Linguistics.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Neha Srikanth, Marine Carpuat, and Rachel Rudinger. 2024. How often are errors in natural language reasoning due to paraphrastic variability? *Transactions of the Association for Computational Linguistics*, 12:1143–1162.

Dhruv Verma, Yash Kumar Lal, Shreyashee Sinha, Benjamin Van Durme, and Adam Poliak. 2023. Evaluating paraphrastic robustness in textual entailment models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 880–892, Toronto, Canada. Association for Computational Linguistics.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. 2018. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, New Orleans, Louisiana. Association for Computational Linguistics.

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, Suzhou, China. Association for Computational Linguistics.

Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, Yijing Xu, Haoqiang Kang, Ziyan Kuang, Chenhan Yuan, Kailai Yang, Zheheng Luo, Tianlin Zhang, Zhiwei Liu, Guojun Xiong, and 15 others. 2024. Finben: A holistic financial benchmark for large language models. In *Advances in Neural Information Processing Systems*,

volume 37, pages 95716–95743. Curran Associates, Inc.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A comprehensive benchmark, instruction dataset and large language model for finance. In *Advances in Neural Information Processing Systems*, volume 36, pages 33469–33484. Curran Associates, Inc.

Mădălina Zgreabăn, Tejaswini Deoskar, and Lasha Abzianidze. 2025. Merge: Minimal expression-replacement generalization test for natural language inference. *arXiv preprint arXiv:2510.24295*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. PAWS: Paraphrase adversaries from word scrambling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1298–1308, Minneapolis, Minnesota. Association for Computational Linguistics.

Yilun Zhao, Yitao Long, Tintin Jiang, Chengye Wang, Weiyuan Chen, Hongjun Liu, Xiangru Tang, Yiming Zhang, Chen Zhao, and Arman Cohan. 2024. FinD-Ver: Explainable claim verification over long and hybrid-content financial documents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14739–14752, Miami, Florida, USA. Association for Computational Linguistics.

A Prompts

We provide the prompt templates used to generate the paraphrases and evaluate LLMs. The prompts to evaluate the LLMs are direct copies of the corresponding prompts used in Magomere et al. (2025).

Paraphrase Generation Prompt

You are paraphrasing a hypothesis from a financial natural language inference dataset.

Produce three sentences that express exactly the same proposition as the original sentence:

1. **lexical**: primarily change word choice;
2. **syntactic**: substantially change sentence structure;
3. **plain_language**: express the same proposition more directly.

Strict requirements:

- Do not add or remove any claim or implication.
- Preserve all named entities.
- Preserve all numbers, dates, percentages, currencies, and units.
- Preserve negation.
- Preserve modality and uncertainty, including words such as may, must, likely, could, and will.
- Preserve comparative direction, including increase/decrease, greater/less, before/after, and higher/lower.
- Preserve quantifiers such as all, some, most, only, and at least.
- Preserve tense and temporal relationships.
- Do not explain your response.
- Return valid JSON.

Original hypothesis:
{HYPOTHESIS}

Return:

```
{
  "lexical": "...",
  "syntactic": "...",
  "plain_language": "..."
}
```

Zero-Shot (ZS)

```
You will be given pairs of sentences: a premise and a hypothesis. Assign one of
the following labels to each pair: ‘entailment’, ‘neutral’, or ‘contradiction’.
Return only the label.
Premise: {premise}
Hypothesis: {hypothesis}
Label:
```

Zero-Shot w/ Annotation Guidelines (ZSAG)

You will be given pairs of sentences: a premise and a hypothesis. The premise sentence comes from one of three types of financial texts:

- A company's SEC filings
- A company's annual report
- A company's earnings call transcript

Your task is to assign one of the following labels to each pair based on these definitions:

- **entailment**: The truth of the premise guarantees the truth of the hypothesis.
- **neutral**: The truth of the premise neither guarantees nor contradicts the truth of the hypothesis.
- **contradiction**: The truth of the premise guarantees that the hypothesis is false.

Given the premise and hypothesis below, determine the relationship between the sentences and return 'entailment', 'neutral', or 'contradiction':
Premise: {premise}
Hypothesis: {hypothesis}
Label:

Few-Shot w/ Annotation Guidelines (FSAG)

You will be given pairs of sentences: a premise and a hypothesis. The premise sentence comes from one of three types of financial texts:

- A company's SEC filings
- A company's annual report
- A company's earnings call transcript

Your task is to assign one of the following labels to each pair based on these definitions:

- **entailment**: The truth of the premise guarantees the truth of the hypothesis.
- **neutral**: The truth of the premise neither guarantees nor contradicts the truth of the hypothesis.
- **contradiction**: The truth of the premise guarantees that the hypothesis is false.

Here are a few examples:

1. **Premise**: And then the actual funded balances that we get from various other lines **Hypothesis**: The funded balances obtained from various lines are expected to increase in the next fiscal quarter. **Label**: neutral
2. **Premise**: We remember, I think about 126 stores that we've already remodeled on that new store prototype. **Hypothesis**: The company has completed the remodeling of approximately 126 stores using the new store prototype. **Label**: entailment
3. **Premise**: We maintain allowances for doubtful accounts for estimated losses resulting from the inability of customers to make required payments. **Hypothesis**: We do not anticipate any losses from customers being unable to make their required payments, so no allowances for doubtful accounts are maintained. **Label**:contradiction

Given the premise and hypothesis below, determine the relationship between the sentences and return 'entailment', 'neutral', or 'contradiction':
Premise: {premise}
Hypothesis: {hypothesis}
Label: