

# FinNLI: Novel Dataset for Multi-Genre Financial Natural Language Inference Benchmarking

Jabez Magomere<sup>1,2,\*†</sup> Elena Kochkina<sup>2\*</sup> Samuel Mensah<sup>2</sup>  
Simerjot Kaur<sup>2</sup> Charese H. Smiley<sup>2</sup>

<sup>1</sup>University of Oxford <sup>2</sup>JPMorgan AI Research

jabez.magomere@keble.ox.ac.uk {name}.{surname}@jpmorgan.com

## Abstract

We introduce FinNLI, a benchmark dataset for Financial Natural Language Inference (FinNLI) across diverse financial texts like SEC Filings, Annual Reports, and Earnings Call transcripts. Our dataset framework ensures diverse premise-hypothesis pairs while minimizing spurious correlations. FinNLI comprises 21,304 pairs, including a high-quality test set of 3,304 instances annotated by finance experts. Evaluations show that domain shift significantly degrades general-domain NLI performance. The highest Macro F1 scores for pre-trained (PLMs) and large language models (LLMs) baselines are 74.57% and 78.62%, respectively, highlighting the dataset’s difficulty. Surprisingly, instruction-tuned financial LLMs perform poorly, suggesting limited generalizability. FinNLI exposes weaknesses in current LLMs for financial reasoning, indicating room for improvement.

## 1 Introduction

Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023b,a; Abdin et al., 2024) excel on a wide variety of tasks. However, their generalizability to specialized domains remains an open research question (Li et al., 2023; Zhang et al., 2024). Financial domain presents unique challenges such as financial terminology (Araci, 2019) and mathematical reasoning (Chen et al., 2021b). Thus specialized benchmarks are necessary to ensure models effectively handle the complexities of this field. Financial Natural Language Processing (NLP) is a growing research area with resources (Nie et al., 2024; Li et al., 2023) and models (Yang et al., 2023a,b; Xie et al., 2023) developed for a wide range of tasks. However, Natural Language Inference (NLI), one of the core tasks in NLP, has been overlooked in this domain.

\*These authors contributed equally to this work.

†Work done during an internship at JPMorgan AI Research.

**Premise:** The company's cash flow has remained stable, but it has experienced substantial decline in market share and rising operational costs over the past two years.  
**Hypothesis:** The company is likely facing increasing financial pressure.  
Entailment

Figure 1: An example of NLI in financial risk assessment, which might seem Neutral as there is no explicit mention of “financial pressure” in the premise.

NLI is a task of determining whether a given hypothesis is true (entailment), false (contradiction), or undetermined (neutral) based on a given premise. NLI is essential for downstream tasks such as question answering, fact-checking, stance classification and relation extraction (Samarinas et al., 2020; Chen et al., 2021a; Sainz et al., 2021). These are widely applicable in the financial domain as well, underlying potential applications in risk assessment, stock market prediction, and automated financial reporting. Figure 1 shows an example of an NLI task in the financial domain. While the hypothesis might initially seem neutral due to the lack of explicit mention of “financial pressure”, a deeper understanding of financial risk assessment shows that stable cash flow, rising operational costs, and declining market share typically signal increasing financial pressure; thus a robust NLI model should classify the pair as entailment. Without a resource allowing to evaluate the state of the art on financial NLI, the effectiveness of any improvements, or adaptation methods can not be assessed either.

To address this gap, we introduce the Multi-Genre **Financial Natural Language Inference (FinNLI)** dataset that enables us to assess and improve the capabilities of language models in understanding and inferring financial information. Previous studies have identified artifacts in NLI datasets that can diminish model effectiveness (Gururangan

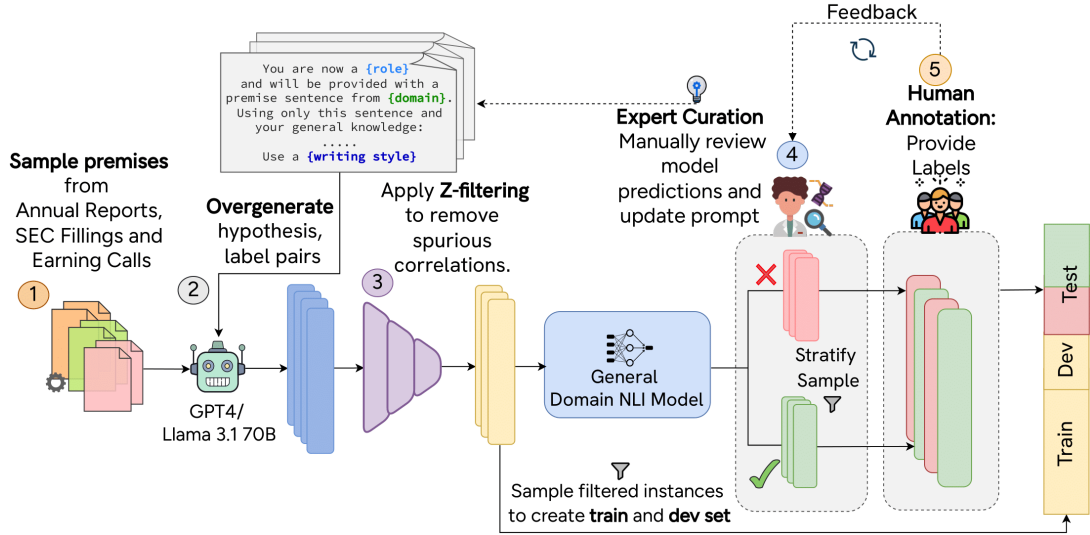


Figure 2: **Overview of the FinNLI data generation pipeline.** (1) We sample premises from real-world financial documents across multiple genres. (2) Hypothesis-label pairs are generated using multiple LLMs. (3) *Z-filtering* (Wu et al., 2022) removes spurious correlations. (4) The prompt is refined based on feedback from a general-domain NLI model and expert curation. (5) Finally, instances correctly predicted and misclassified by the NLI model are reviewed by finance experts for gold label annotation.

et al., 2018; Zhang et al., 2019; Ye et al., 2024). In this paper, we meticulously design a generation pipeline to address these prevalent challenges. Our framework for dataset generation shown in Figure 2 focuses on diversity, mitigation of spurious correlations, including challenging examples and having high quality annotations for the testing set. We then present a comprehensive evaluation of various state-of-the-art models on the FinNLI benchmark. These include general domain and financial LLMs and smaller pretrained language models (PLMs). We make the following contributions:

- We introduce a novel dataset for evaluating NLI in the financial domain.
- We present a dataset generation framework and discuss the contribution of its parameters.
- We provide a thorough evaluation and its analysis of multiple NLP models on FinNLI.
- We demonstrate significant room for improvement for Financial NLI across all models.

## 2 Related Work

**Financial NLP** Financial NLP is a rapidly evolving research area, with a number of advances achieved in a number of tasks such as sentiment analysis (Shah et al., 2023; Malo et al., 2014), question answering (Chen et al., 2021b, 2022),

and financial text summarization (Mukherjee et al., 2022). Several finance-specific LLMs have also been developed to address domain-specific challenges. Notable examples include InvestLM (Yang et al., 2023b), FinGPT (Yang et al., 2023a), PIXIU FinMA (Xie et al., 2023), and Bloomberg GPT (Wu et al., 2023), although access to some of these models is limited, and they are not all available as open-source. While finance-instruction-tuned LLMs have attained increasingly high performance on various financial NLP tasks, it is still unclear whether these models will generalise well to novel tasks not included in the instruction tuning data. The task of financial NLI has been largely overlooked likely due to insufficient resources, yet it has been shown to be informative for evaluating LLMs (Madaan et al., 2024). We aim to fill this gap by creating a benchmark dataset and conducting a comprehensive evaluation across a wide range of models.

While there are no dedicated financial NLI datasets, FinCausal datasets (Mariko et al., 2020, 2022) provide benchmarks for evaluating financial reasoning. However, FinCausal focus solely on causality detection, which is just one type of reasoning. In contrast, NLI task involves various logical relationships that require multiple types of reasoning, with causal reasoning being only one aspect. Our proposed dataset, FinNLI, evaluates more types of relationships just beyond causality such as mathematical, temporal and financial.

Dataset	Domain	Premise Source	Ensures premise diversity	Hypothesis Generation	Ensures hypothesis diversity	Mitigates spurious correlations	Includes challenging instances
SNLI (Bowman et al., 2015)	General	Real-world	No	Human Annotation	No	No	No
MNLI (Williams et al., 2018)	General	Real-world	Yes	Human Annotation	No	No	No
MedNLI (Shivade et al., 2019)	Medical	Real-world	No	Human Annotation	No	No	No
ANLI (Nie et al., 2020)	General	Real-world	No	Human Annotation	No	No	Yes
WANLI (Liu et al., 2022a)	General	Synthetic	No	Synthetic + Human	No	No	Yes
SciNLI (Sadat and Caragea, 2022)	Scientific	Real-world	No	Rule-Based	No	No	No
MSciNLI (Sadat and Caragea, 2024)	Scientific	Synthetic	Yes	Rule-Based	No	No	No
GNLI (Hosseini et al., 2024)	General	Synthetic	Yes	Synthetic	Yes	No	No
<b>Ours (FinNLI)</b>	<b>Financial</b>	<b>Real-world</b>	<b>Yes</b>	<b>Synthetic</b>	<b>Yes</b>	<b>Yes</b>	<b>Yes</b>

Table 1: Qualitative comparison of NLI dataset generation approaches across different methodological aspects.

**Domain-specific NLI** The development of NLI has been significantly advanced by large-scale datasets such as SNLI (Bowman et al., 2015) and MNLI (Williams et al., 2018). However, a common challenge is the performance drop when transitioning from general-domain models to domain-specific applications (Hupkes et al., 2023). This highlights the importance of creating and utilizing domain-specific NLI datasets to ensure models are well-adapted to specialized domains.

Specialized NLI datasets have been developed for various domains, including biomedical (BioNLI (Bastan et al., 2022)), medical (MedNLI (Shivade et al., 2019)), scientific (SciNLI (Sadat and Caragea, 2022), SciTail (Khot et al., 2018)), and legal (ContractNLI (Koreeda and Manning, 2021), LegalNLI (Yang, 2022)) contexts. In this work, we introduce a Financial NLI dataset to evaluate and enhance NLI model performance in the financial domain, where precise inferential reasoning is crucial.

**NLI Dataset Generation** Approaches to NLI dataset construction typically involve either creating both premise and hypothesis (Hosseini et al., 2024), sampling existing premises and generating hypotheses (Williams et al., 2018; Shivade et al., 2019), or sampling both from existing texts (Sadat and Caragea, 2022; Bastan et al., 2022). The generation process can be automated or manual, often followed by human annotation (Liu et al., 2022a). Sampling both premises and hypotheses from existing texts can be limited by rule-based methods (e.g., using linking phrases (Sadat and Caragea, 2022, 2024)), as suitable pairs are not easy to find and automatically sample. Human-generated hypotheses are more costly due to the time required. Recent approaches use LLMs to synthetically generate both premises and hypotheses (Liu et al., 2022a; Hosseini et al., 2024). However, while

this may be effective for general domains, LLMs trained on general text may fail to generate high-quality data that accurately capture the nuances of specialized domains (Xu et al., 2023). To address these limitations, our work samples premise sentences from real-world documents across multiple financial genres to ensure premise diversity and to induce finance-specific knowledge. We ensure hypothesis diversity by employing different writing styles and roles when prompting LLMs (Yu et al., 2023), followed by domain-expert annotation to create a high-quality test set. Table 1 provides a qualitative comparison of our NLI dataset generation approach with others, further comparisons with finance and NLI datasets are in Appendix A.

NLI datasets often contain artifacts—spurious correlations between labels and features that do not represent the underlying task (Stacey et al., 2020; Cosma et al., 2024). These artifacts cause models to learn shortcuts, failing to generalize to new data (Ye et al., 2024). Artifacts can occur in both manually and automatically generated instances (Gururangan et al., 2018; Zhang et al., 2019; Arakelyan et al., 2024; Proebsting and Poliak, 2024). Addressing artifacts involves careful dataset design, filtering (Wu et al., 2022), and adding adversarial samples (Nie et al., 2020; Liu et al., 2022a). In contrast to other approaches, we incorporate z-filtering (Wu et al., 2022) in our dataset generation pipeline to mitigate spurious correlations and leverage feedback from a general-domain NLI model to ensure our resulting dataset includes challenging instances in an adversarial-like setting

### 3 Dataset Construction

We present our data generation framework (Figure 2) which is applicable across various domains. This framework ensures contextually grounded instances by sampling premises from real-world financial texts. It promotes diversity by incorporat-

ing two different LLMs and varied prompts while minimizing artifacts through *z-filtering* (Wu et al., 2022). To enhance the dataset’s difficulty, we leverage feedback from a general domain NLI model and experts to guide model generation. Finally, we maintain high quality by involving multiple finance experts in the annotation process.

### 3.1 Stage 1: Sampling Premises Across Multiple Financial Domains

As the sources of our premise sentences, we use publicly available financial documents: **Annual Reports**<sup>1</sup>, **SEC Filings**<sup>2</sup> and **Earning Calls** (Li et al., 2020). Sampling from diverse financial genres enhances the stylistic variation of the dataset. Annual Reports and SEC Filings exhibit a formal style with prevalence of numerical information (Loukas et al., 2022), while Earning Calls include more colloquial language. Preprocessing is used to filter out invalid sentences, ensuring sample quality (Appendix B).

### 3.2 Stage 2: Generating Hypotheses

Given a premise, we prompt an LLM to generate one entailment, neutral, and contradiction hypothesis. The prompt includes the NLI task definition, label verbalization, domain-specific generation conditions, and exemplars to guide the LLM. Conditions and examples are refined through iterations of steps 1-4 of the pipeline on Figure 2 (full prompt in Appendix G). To encourage stylistic variations, we use conditional prompting with role-playing (Shanahan et al., 2023) and treat writing style as an attribute dimension (Yu et al., 2023). We define a list of finance-related roles as *financial analyst*, *financial reporter*, *finance compliance officer*, and *financial consultant*, and a list of writing styles as *social media*, *news*, *financial textbook*, and *financial reporting*. At each generation step, a role and writing style are randomly selected to condition the generation. We experiment with two LLMs to generate hypothesis-label pairs: GPT-4 (OpenAI, 2023) and Llama 3.1 70B (Dubey et al., 2024). These models belong to different LLM families, vary in size, and perform competitively across a wide range of tasks. This setup allows us to examine potential differences between closed-source and open-source models in generating synthetic instances.

<sup>1</sup><https://www.annualreports.com/>

<sup>2</sup><https://www.sec.gov/>

### 3.3 Stage 3: Automatic Filtering

To minimise spurious correlations that could have appeared during the generation process, we apply the *z-filtering* algorithm from Wu et al. (2022). It is a method to reject samples that contribute to the high spurious correlations between task-independent features of the samples and their labels. The list of task-independent features includes: unigrams and bigrams, lexical overlap between the premise and hypothesis, hypothesis length and premise-hypothesis length ratio, and thenumber and density of financial terms in the hypothesis to capture correlations that may result from the presence of specific financial keywords. We set the parameters of the z-filtering process, the top  $k$  features and sample batch size to 20 and 200 respectively. We observed that a higher value of  $k$  and lower value of batch size lead to a higher number of samples to be removed and lower z-scores. Z-scores and top-k features before and after applying z-filtering are included in Appendix C showing the effectiveness of this our approach in minimizing spurious correlations.

We create the train and development sets by randomly sampling from the filtered instances generated by Llama 3.1 70B with the final optimised prompt<sup>3</sup>. We sample evenly across financial domains and class labels. We use the generated labels as the gold standard for training and development.

### 3.4 Stage 4: Expert Curation

Evaluating the quality of open-ended data generation across various prompt parameters is a challenging task due to the large number of possible prompt combinations (Long et al., 2024; Chung et al., 2023). To address this, we combine feedback from a general domain NLI model and human experts to identify label inconsistencies and challenging reasoning patterns. These insights are then used to iteratively refine the generation prompt. As a general domain NLI model we chose RoBERTa-Large NLI<sup>4</sup>, trained on MNLI and SNLI as one of top performing models on this task. Through multiple iterations, we refine the prompt to include instructions for mathematical/quantitative reasoning, temporal reasoning, financial knowledge reasoning, and tricky linguistic constructs, along with repre-

<sup>3</sup>We use Llama 3.1 70B model to generate the training and development set as its license permits the use of the generated artefacts for fine-tuning downstream models <https://llama.meta.com/llama3/license/>

<sup>4</sup><https://huggingface.co/joeddav/xlm-roberta-large-xnli>



Split	Size	Label Distribution (E/N/C)
Train	16,200	5,400 / 5,400 / 5,400
Dev	1,800	600 / 600 / 600
Test	3,304	1,361 / 995 / 948

Table 2: Dataset Split and Label Distribution

sentative examples from each reasoning pattern. We provide examples of instances with challenging reasoning patterns in Appendix 12.

We create a domain- and class-balanced test set by sampling instances where the predicted label matches the generated label, as well as instances where they differ. This ensures a mix of correct and incorrect predictions, balancing both challenging and core cases. By doing so, we capture not only “hard” examples but also fundamental NLI reasoning patterns, allowing for a more effective evaluation of the generation model (Bowman and Dahl, 2021).

### 3.5 Stage 5: Human Annotation

We provide the samples for annotation in 4 rounds (R) and at each round collect feedback from the expert annotators on the quality of the instances. We provide annotators with instances generated from GPT-4 in R1 and R2 while Llama generated instances in R3 and R4. We incorporate this feedback by updating the prompt conditions and in-context examples to improve the quality of subsequent generations after each annotation round. The annotators are provided with a premise and hypothesis sentence and asked to provide an appropriate label, a confidence level (*high* or *low*), an optional flag to mark the premise or hypothesis sentence as incomplete/nonsensical and an optional comment field. The annotators do not revise the premise or hypothesis sentences (Appendix H). Each premise-hypothesis pair is reviewed by 3 professional annotators with finance backgrounds and we assign the majority label as the gold label. In very rare cases (0.15%) where all the annotators assign different labels, we do not assign a gold label. We exclude any instances where *any* of the annotators mark the premise or hypothesis as incomplete/nonsensical.

## 4 Dataset Analysis

We provide an analysis of the annotation outcomes and the resulting characteristics of FinNLI<sup>5</sup>.

<sup>5</sup>For information on the data license and dataset access, please contact the authors via email.

### 4.1 Synthetic Label Accuracy

**Per Annotation Round.** We conduct annotations over four rounds, altering either the prompt or the model between rounds. Rounds R1 and R2 utilize GPT-4, while R3 and R4 use hypotheses generated by Llama 3.1 70B. Prompt adjustments occur between R1 and R2, as well as between R3 and R4. In the first round of annotation (R1), domain experts evaluated 1,462 GPT-4-generated instances. In 65.18% of the cases, the label assigned by the annotators matched the generated label. After refining the prompt, the second round (R2) showed improvement, with 76.47% out of 595 instances displaying label agreement, highlighting the value of incorporating expert feedback. In annotation rounds three (R3) and four (R4), annotators were presented with examples generated by Llama 3.1 70B. In R3, 64.57% of the 621 instances had labels that matched the generated ones. The decrease in agreement relative to R2 can be attributed to the use of a GPT-4-optimized prompt with Llama, revealing model-specific prompt sensitivity (Lu et al., 2024). After refining the prompt for Llama in R4, label agreement improved to 76.84% across 626 instances. We use the prompt from R4 to generate the training and development sets. 14.27% of the examples where the general-domain NLI model assigns a similar label to the generated label are relabelled by annotators, compared to 46.88% for instances with differing labels. This finding supports the rationale to sample from both correctly and incorrectly predicted instances, as the former likely includes ‘easy’ examples and the latter contains both ‘hard’ and ‘ambiguous’ cases.

The overall agreement rate for the FinNLI dataset is noticeably lower than the 84.97%<sup>6</sup> agreement reported for general-domain synthetic NLI datasets such as GNLI (Hosseini et al., 2024). Although, we note that Hosseini et al. (2024) prompt-tune a generator LLM with an existing NLI dataset while our approach relies solely on expert-curated prompt conditions and in-context exemplars to steer the generator. We also hypothesize that LLMs may be more efficient data generators for general domains than in specialised settings (Xu et al., 2023).

**Per class.** The breakdown by class label shows significant differences. For contradiction hypothe-

<sup>6</sup>We obtain the overall agreement rate by averaging the agreement rate for *unanimous* labels - 89.53% and majority labels - 80.41% reported in Hosseini et al. (2024)

Domain	#Words		% Sentences		Word	
	Prem.	Hyp.	Prem.	Hyp.	Overlap	Agrmt.
REPORTS	25.0	20.6	95.2	98.8	41.62%	76.14%
SEC	23.2	20.3	96.8	98.9	45.26%	76.20%
CALLS	19.5	17.5	96.0	99.2	27.92%	77.62%
<b>Overall</b>	<b>22.6</b>	<b>19.5</b>	<b>96.0</b>	<b>98.9</b>	<b>38.25%</b>	<b>76.65%</b>

Table 3: Dataset Statistics for FinNLI by genre. *#Words* is the mean number of tokens in the premise and hypothesis. *%-Sentences* – proportion of sentences parsed with an ‘S’ root using the Stanford PCFG Parser (Klein and Manning, 2003). *Word Overlap* – mean proportion of tokens shared between the premise and hypothesis. *Agrmt.* is the % of examples where the generated label matches the label assigned by annotators.

ses, both LLMs achieve high agreement rates (GPT-4 - 88.83% and Llama - 90.86%). For entailment hypotheses, Llama achieves a higher agreement rate of 75.46% compared to 53.55% for GPT-4. In contrast, for neutral hypotheses, GPT-4 achieves a higher agreement rate of 81.07% compared to 66.52% for Llama. Qualitative observations show that many LLM-generated entailment hypotheses are relabelled to neutral due to models appending additional information not present in the premise, resulting in lower agreement rates for entailment.

## 4.2 Dataset Statistics

The domain-experts annotated a total of 3,360 examples, out of which 3,304 instances (98%) are marked as valid examples. For the valid instances, we achieved a Fleiss- $\kappa$  score of 88.31%, indicating an almost perfect agreement. Table 2 outlines the size of the train, development and testing sets, as well as label distribution across the dataset splits. The training and development sets are uniformly balanced across labels, whereas the test set shows a slight skew toward the entailment label. The slight imbalance in the test set is a result of reassigning labels as part of the manual annotation process.

## 4.3 Linguistic Analysis

We provide a detailed linguistic analysis of the FinNLI dataset in Table 3 and comparison to other NLI datasets in Appendix A. Premise sentences are generally longer (mean 22.6 words, max 105 words) than hypothesis sentences (mean 19.5 words, max 84 words). Premises from Annual Reports and SEC Filings tend to be lengthier, reflecting the formal nature of financial reporting, whereas premises from Earning Calls are shorter, indicative of more colloquial language use. Annual

Reports and SEC Filings often include tables and forms, which can result in incomplete premise sentences when sampled. However, our pre-processing pipeline filters out these invalid sentences, with approximately  $\approx 96\%$  of the premise sentences containing ‘S’-root parses, indicating they are syntactically complete. We observe a higher token overlap between the hypothesis and premise sentences in Annual Reports and SEC Filings compared to Earning Calls. Annual Reports and SEC Filings often include specific financial jargon that must be reflected in the hypothesis to remain relevant, whereas utterances in Earning Calls are more open to interpretation.

## 5 Models

We evaluate performance of various models on FinNLI test set. We use the average macro F1 as our main metric due to the test set class imbalance.

### 5.1 Zero-Shot NLI Baselines

We evaluate two state-of-the-art general-domain NLI models: **BART-Large-MNLI** (Lewis et al., 2020a), trained on MNLI (Williams et al., 2018) and **DeBERTa-V3-NLI** (He et al., 2021), trained on both MNLI and SNLI (Bowman et al., 2015).

### 5.2 PLM Baselines

We fine-tune the following general-purpose and domain-specific pre-trained language models (PLMs) on the FinNLI training set: **RoBERTa-Base**, **RoBERTa-Large** (Liu et al., 2019), and **FiLM** (Choe et al., 2023), which is a RoBERTa-Base model further pre-trained on financial corpora. Additionally, we fine-tune **BART-Large-MNLI** (Lewis et al., 2020a) to explore whether training on both MNLI and FinNLI yields any benefits. We conduct a grid search for hyperparameters and choose the best model on the development set. The best model is trained over three independent runs with different random seeds<sup>7</sup>.

### 5.3 LLM Baselines

We investigate the performance of general-purpose and domain-specific LLMs across a wide range of model sizes and prompt settings on the FinNLI test set. For the general-purpose LLMs, we evaluate **Llama-2-7b-chat-hf** (Touvron et al., 2023b),

<sup>7</sup>Full details on the experimental setup and hyperparameter tuning are provided in Appendix D.

Model	Size	Setup	Annual Reports	SEC Filings	Earning Calls	Overall
DeBERTaV3-NLI	304M	ZERO-SHOT	68.83 $\pm$ 0.0	68.52 $\pm$ 0.0	71.08 $\pm$ 0.0	69.54 $\pm$ 0.0
BART-LARGE-MNLI	406M	ZERO-SHOT	69.96 $\pm$ 0.0	70.24 $\pm$ 0.0	<b>75.95 <math>\pm</math> 0.0</b>	72.09 $\pm$ 0.0
ROBERTA-BASE	255M	FINETUNED	70.32 $\pm$ 0.1	69.14 $\pm$ 1.2	69.84 $\pm$ 0.2	69.78 $\pm$ 0.5
FiLM	255M	FINETUNED	70.43 $\pm$ 0.6	69.57 $\pm$ 0.2	70.24 $\pm$ 0.8	70.11 $\pm$ 0.01
ROBERTA-LARGE	355M	FINETUNED	<b>75.11 <math>\pm</math> 0.1</b>	73.98 $\pm$ 1.2	74.52 $\pm$ 0.2	<b>74.57 <math>\pm</math> 0.5</b>
BART-LARGE-MNLI	406M	FINETUNED	73.91 $\pm$ 0.8	<b>74.14 <math>\pm</math> 0.6</b>	74.35 $\pm$ 0.9	74.18 $\pm$ 0.2

Table 4: Average Macro F1 Scores (%) of zero-shot general-domain NLI models and fine-tuned PLMs on the FinNLI test set across each financial domain. Results are averaged over three random seeds, with the standard deviation reported. Best performance is highlighted in **bold**.

**Phi-3.5-mini-instruct** (Abdin et al., 2024), **Llama-3.1-8B-Instruct** (Dubey et al., 2024), and **Llama-3.1-70B-Instruct** (Dubey et al., 2024). For the finance-domain LLMs, we evaluate **FinMA-7B-NLP**, **FinMA-30B** (Xie et al., 2023) - Llama-2 7B and 30B respectively instruction-tuned on financial NLP tasks. FinMA models were chosen due to superior performance across diverse financial NLP tasks (Xie et al., 2024) and its open-source availability. We use several prompt setups (Appendix G): Zero-Shot (ZS); Zero-Shot expanded with Annotation Guidelines (ZS-AG); Few-Shot with Annotation Guidelines (FS-AG) (2 examples per class from different financial genres); and Chain-of-Thought (CoT) (Wei et al., 2022).

## 6 Results

### 6.1 Zero-shot and Fine-tuned PLM Baselines

Table 4 presents the results of experiments with zero-shot NLI and fine-tuned PLM Baselines.

**General-domain NLI models exhibit a sizeable performance drop on financial NLI.** BART-Large-MNLI and DeBERTa-V3-NLI achieve accuracy scores of 89.9% (Lewis et al., 2020a) and 91.1%<sup>8</sup>, respectively, on MNLI. However, their accuracy drops significantly to 70.94% and 68.31% on FinNLI, highlighting the impact of domain shift on performance in out-of-domain settings. The general-domain NLI models perform better on Earnings Calls compared to Annual Reports and SEC Filings, likely due to the MNLI training data, which includes telephone transcripts and speeches, aligning more closely with the informal language used in Earnings Calls (Williams et al., 2018).

**Fine-tuning PLMs on FinNLI improves performance over zero-shot NLI models.** Despite the FinNLI test set consisting of GPT-4 and Llama3.1-generated instances while the training set consist-

ing of only Llama3.1-generated instances, we observe performance improvements post fine-tuning. Fine-tuning improves performance by 5% in F1 score when comparing the zero-shot DeBERTa-V3-NLI to the fine-tuned RoBERTa-Large on the FinNLI test set, both having similar parameters. BART-Large-MNLI shows a 2% improvement in performance after fine-tuning on FinNLI. However, combining MNLI and FinNLI offers no benefits compared to training solely on FinNLI. Larger fine-tuned PLMs outperform smaller models, suggesting that FinNLI contains challenging examples that require greater model capacity for accurate inference. Domain-specific pre-training offers only marginal gains over general-purpose PLMs. FiLM (Choe et al., 2023), a RoBERTa-Base model pre-trained on financial documents,<sup>9</sup> performs slightly better than the standard RoBERTa-Base.

### 6.2 LLM baseline results

Figure 3 shows the results of LLM experiments.

**Llama 3.1 70B achieves the best performance on FinNLI, but significant room for improvement remains.** Llama 3.1 70B achieves the highest F1 score of 78.62% on the FinNLI test set and is the only LLM among those evaluated that outperforms a finetuned RoBERTa-Large model, which attains an F1 score of 74.57% ( $\sigma = 0.5\%$ ). Phi 3.5 achieves a comparable performance of 74.74% ( $\sigma = 0.3\%$ ), with this difference being too small to conclude any significant advantage over RoBERTa-Large. All other LLMs are outperformed by the RoBERTa-Large model finetuned on the FinNLI training set, indicating that PLMs are still highly competitive on domain-specific tasks. While the best performing F1 score is 78.62%, we observe that this is notably lower than other LLMs’ performance on general-domain NLI tasks (Wang et al.,

<sup>8</sup>Accuracy score reported from <https://huggingface.co/cross-encoder/nli-deberta-v3-large>

<sup>9</sup>Including 301M tokens from SEC Filings and 1.1B tokens from Earning Calls (Choe et al., 2023)

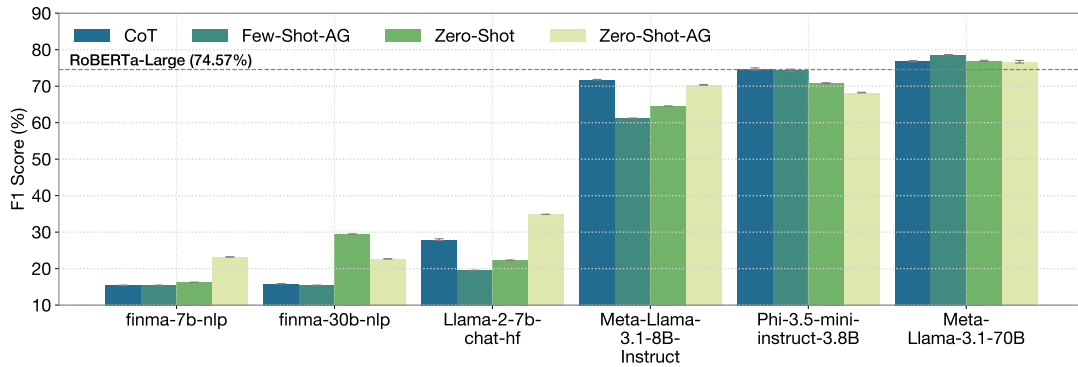


Figure 3: Average macro F1 scores (%) for various LLMs across different prompting setups evaluated on the FinNLI test set. Prompts with **AG** have the class label definitions used in the annotation guidelines included in the prompt. The error bars represent the standard deviation across 3 independent runs. The dotted grey line at 74.57% marks the performance of RoBERTa-Large, the best performing fine-tuned PLM on FinNLI test set.

2024)<sup>10</sup> suggesting that FinNLI may contain examples that exhibit complex reasoning patterns or contain ‘ambiguous’ instances that pose a challenge to current state-of-the-art LLMs. To investigate this, we conduct an error analysis on the instances misclassified by Llama 3.1 70B in Section 7.3.

**Small Language Models (SLMs) outperform significantly larger LLMs on FinNLI.** Phi-3.5-mini-instruct, a 3.8 billion parameter LLM, achieves an F1 score of 74.74%, outperforming Llama 3.1 8B, a significantly larger LLM that achieves an F1 score of 71.79%. This underscores the potential of smaller language models trained on highly curated datasets in bridging the performance gap to larger LLMs (Abdin et al., 2024).

**LLMs instruction-tuned on financial domain tasks perform worse than general-domain LLMs.** FinMA 7B and FinMA 30B - instruction-tuned on financial NLP tasks, achieve F1 scores of 23.20% and 29.46% respectively. In contrast, the base model for FinMA 7B, Llama-2-7B-chat achieves a higher F1 score of 34.87%. This suggests that instruction-tuning may have negatively impacted the model’s ability to generalise to other tasks outside the instruction-tuning data. This finding aligns with previous studies which show that domain-specific instruction tuning can degrade general performance on financial textual-analysis tasks (Xie et al., 2024) and knowledge-based evaluation benchmarks (Lee et al., 2024b). Possible reasons for this could include limited training sets, suboptimal training parameters, or training schedules (Dong et al., 2024; Lee et al., 2024a).

<sup>10</sup>General-domain NLI datasets are likely to be included in LLM pretraining data and as such this may not be a fair evaluation

**CoT prompting improves performance for small to mid-sized LLMs on FinNLI.** We observe that using CoT and asking an LLM to provide an explanation for its prediction, leads to performance improvements over zero-shot prompting for small to mid-sized models on FinNLI. Specifically, CoT yields an approximate F1Score improvement of 7% for Llama 3.1 8B, 5.6% for Llama-2-7b-chat-hf, and 3.9% for Phi-3.5-mini-instruct. However, for the larger Llama 3.1 70B model, CoT does not provide significant performance improvement over zero-shot. Score). The FinMA models do not exhibit improvements with CoT as they fail to return any meaningful responses. We observe that smaller LLMs are more sensitive to different prompts on the NLI task, and CoT helps enhance their reasoning. In contrast, larger LLMs like Llama 3.1 70B already demonstrate strong reasoning capabilities and are less affected by prompt variations.

## 7 Impact of Generation Design Decisions

We discuss the impact of different generation design decisions on model performance.

### 7.1 GPT-4 vs Llama-Generated Test Sets

We compare the average macro F1 scores of the best-performing zero-shot NLI model, fine-tuned PLM and LLM on the GPT-4-generated instances (2,057) and Llama-generated instances (1,247), averaged over three independent runs. BART-Large-MNLI performs significantly better on the Llama test set (80.43%) than on the GPT-4 test set (65.19%) suggesting that GPT-4-generated instances are more challenging for zero-shot NLI models. Fine-tuned RoBERTa-Large achieves higher performance on the GPT-4 test set (77.48%)



compared to the Llama test set (71.90%). Fine-tuning appears to help the model adapt to the nuances of GPT-4-generated data but reduces its effectiveness on Llama-generated instances. Llama 3.1 70B exhibits consistent performance across both test sets (77.28% on GPT-4 and 78.84% on Llama)<sup>11</sup>. The slight advantage on the Llama test set could be attributed to self-preference bias, where an LLM’s performance is favourable to its own output as compared to other outputs (Panickssery et al., 2024).

## 7.2 Prompt Roles and Writing Styles

We investigate how different prompt roles and writing styles used to generate hypotheses affect model performance. For the best-performing fine-tuned PLM (RoBERTa-Large) and LLM (Llama 3.1 70B), we observe no significant differences in F1 scores across the various roles or styles, indicating robustness to stylistic variations. However, the zero-shot NLI model, BART-Large-MNLI, shows considerable performance discrepancies. It achieves higher F1 scores on more formal writing styles such as *financial textbook* (78.32%) and *financial reporting* (77.40%) compared to less formal styles like *social media* (60.70%) and *financial news* (73.82%). This suggests that zero-shot NLI models are more sensitive to stylistic variations (Belinkov et al., 2019). In contrast, LLMs like Llama 3.1 70B gain robustness to different writing styles due to pretraining on diverse domains (Dubey et al., 2024). Similarly, the fine-tuned PLM gains robustness by training on the FinNLI, enabling better generalization across stylistic variations.

## 7.3 High vs Low Confidence Annotations

We define high-confidence instances as examples that receive a unanimous label and all annotators mark their confidence level as high for that instance. Low-confidence instances are examples where at least one annotator marks with low confidence or instances that do not have a unanimous label. Out of 3,304 annotated examples, 2,401 examples (72.67%) are classified as high-confidence and 903 examples (27.33%) marked as low-confidence. Llama 3.1 70B achieves an F1 score of 82.51% on the high-confidence instances compared to 67.20% on the low-confidence instances. This suggests that LLMs similarly struggle on examples that human

annotators demonstrate low confidence or high disagreement, possibly the ambiguous cases.

## 8 Error Analysis

Reviewing the frequency of model’s errors across the classes, we observe a notable trend of misclassifying neutral instances as entailment (Appendix D). To gain a deeper understanding, we manually reviewed  $\approx 10\%$  (60 out of 683) of the misclassified instances from the best-performing model, Llama 3.1 70B (Appendix E.2). Our analysis shows that 56% of misclassifications are due to **inference errors** (clear model mistakes), while 44% stem from **ambiguous instances**, i.e., phrases or words that allow more than one interpretation (Sandri et al., 2023; Jiang and de Marneffe, 2022). For inference errors, the model often assumes *implicit information* (26%) not contained in the premise, making logical inferences that, while plausible, are unfaithful to the premise (Zhou et al., 2023). Additionally, the model struggles with instances requiring *financial reasoning* (16%), such as knowledge of financial terms or standard procedures, which domain-expert annotators easily pick up. *Tricky linguistic instances* (Nie et al., 2020) (8%) such as counter-speech (Gligoric et al., 2024), words play, and those requiring *mathematical reasoning* (6%) still pose a challenge to the model. Regarding ambiguous instances, 22% of misclassifications are due to *uncertainty in sentence meaning* that may result in multiple interpretations. Interestingly, a significant proportion (22%) of premise sentences sampled from *Earnings Calls* contain *anaphora* and *deixis* (Poesio and Artstein, 2005), typical of informal conversations —expressions that require contextual information to interpret. Such instances lead to ambiguity, as it is often unclear whether the entity introduced in the hypothesis is mentioned in the premise, resulting in *missing context*.

## 9 Conclusion

We introduced the FinNLI benchmark to evaluate and improve NLI models in the financial domain. Our evaluations show that general-domain NLI models perform worse on financial NLI, LLMs perform comparably to fine-tuned PLMs, and instruction-tuned financial LLMs lack generalizability. This reveals significant room for improvement in financial NLI, necessary for better understanding and reasoning with financial text, ultimately enhancing downstream applications.

<sup>11</sup>A paired t-test confirms that the higher performance on the Llama test set is statistically significant ( $t = 50.59$ ,  $p = 0.0004$ ).

## 10 Limitations

Our NLI task setup employs a single gold label classification, and we have yet to explore more advanced setups, such as those involving explanation rationales or generating probability distributions over the label space that align with human judgment. Due to limited domain expertise, we could only employ three annotators per instance.

We did not explore adversarial settings, which could provide insights into model robustness against attacks. Future work will explore domain adaptation methods and using FinNLI as an intermediate task to enhance performance on related tasks. Currently, the FinNLI benchmark is limited to English, restricting its applicability to non-English financial texts. We also note that the generated premises may contain fictional information and should not be used to train models that learn factual data.

## 11 Ethical Considerations

The financial documents utilized in the FinNLI dataset, including SEC filings, annual reports, and earnings call transcripts, are publicly available and do not contain sensitive, personally identifiable information (PII). Earnings call transcripts have been anonymized. However, it is important to acknowledge that financial texts and model-generated outputs may contain inherent biases. Despite our efforts, we cannot guarantee the complete absence of biases and hallucinations in the synthetic data produced. These limitations should be considered when interpreting the results and applications of our research. This dataset should not be used directly for tasks like fact-checking and question answering, as we cannot guarantee the accuracy of the information.

## Acknowledgments

We would like to thank Arturo Oncevay, Toyin Aguda, and Keshav Ramani for insightful discussions. We thank our colleagues from Annotation Center of Excellence, who collaborated with us to refine annotation guidelines and produce high quality annotations for this research paper.

**Disclaimer** This paper was prepared for informational purposes by the Artificial Intelligence Research group of JPMorgan Chase & Co. and its affiliates “JP Morgan”) and is not a product of the Research Department of JP Morgan. JP Morgan

makes no representation and warranty whatsoever and disclaims all liability, for the completeness, accuracy or reliability of the information contained herein. This document is not intended as investment research or investment advice, or a recommendation, offer or solicitation for the purchase or sale of any security, financial instrument, financial product or service, or to be used in any way for evaluating the merits of participating in any transaction, and shall not constitute a solicitation under any jurisdiction or to any person, if such solicitation under such jurisdiction or to such person would be unlawful.

## References

- Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, Alon Benhaim, et al. 2024. [Phi-3 technical report: A highly capable language model locally on your phone](#). *Preprint*, arXiv:2404.14219.
- D Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.
- 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.
- Mohaddeseh Bastan, Mihai Surdeanu, and Niranjan Balasubramanian. 2022. Bionli: Generating a biomedical nli dataset using lexico-semantic constraints for adversarial examples. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5093–5104.
- Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. [Don’t take the premise for granted: Mitigating artifacts in natural language inference](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.
- Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. " O’Reilly Media, Inc."
- Samuel 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.

- Samuel Bowman and George Dahl. 2021. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4843–4855.
- Chung-Chi Chen, Yu-Min Tseng, Juyeon Kang, Anaïs Lhuissier, Min-Yuh Day, Teng-Tsai Tu, and Hsin-Hsi Chen. 2023. Multi-lingual esg issue identification. In *Proceedings of the Fifth Workshop on Financial Technology and Natural Language Processing and the Second Multimodal AI For Financial Forecasting*, pages 111–115.
- Jifan Chen, Eunsol Choi, and Greg Durrett. 2021a. Can nli models verify qa systems’ predictions? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3841–3854.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan R Routledge, et al. 2021b. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. Convinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292.
- Jaeyoung Choe, Keonwoong Noh, Nayeon Kim, Seyun Ahn, and Woohwan Jung. 2023. [Exploring the impact of corpus diversity on financial pretrained language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2101–2112, Singapore. Association for Computational Linguistics.
- John Chung, Ece Kamar, and Saleema Amershi. 2023. [Increasing diversity while maintaining accuracy: Text data generation with large language models and human interventions](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 575–593, Toronto, Canada. Association for Computational Linguistics.
- Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. Semeval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th international workshop on semantic evaluation (SemEval-2017)*, pages 519–535.
- Adrian Cosma, Stefan Ruseti, Mihai Dascalu, and Cornelia Caragea. 2024. [How hard is this test set? NLI characterization by exploiting training dynamics](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2990–3001, Miami, Florida, USA. Association for Computational Linguistics.
- Guanting Dong, Hongyi Yuan, Keming Lu, Chengpeng Li, Mingfeng Xue, Dayiheng Liu, Wei Wang, Zheng Yuan, Chang Zhou, and Jingren Zhou. 2024. [How abilities in large language models are affected by supervised fine-tuning data composition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 177–198, Bangkok, Thailand. Association for Computational Linguistics.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, et al. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Kristina Gligoric, Myra Cheng, Lucia Zheng, Esin Durmus, and Dan Jurafsky. 2024. [NLP systems that can’t tell use from mention censor counterspeech, but teaching the distinction helps](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5942–5959, Mexico City, Mexico. Association for Computational Linguistics.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. [Debertav3: Improving deberta using electra-style pre-training with gradient-disentangled embedding sharing](#). *Preprint*, arXiv:2111.09543.
- Mohammad Javad Hosseini, Andrey Petrov, Alex Fabrikant, and Annie Louis. 2024. [A synthetic data approach for domain generalization of NLI models](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2212–2226, Bangkok, Thailand. Association for Computational Linguistics.
- Dieuwke Hupkes, Mario Giulianelli, Verna Dankers, Mikel Artetxe, Yanai Elazar, Tiago Pimentel, Christos Christodoulopoulos, Karim Lasri, Naomi Saphra, Arabella Sinclair, et al. 2023. A taxonomy and review of generalization research in nlp. *Nature Machine Intelligence*, 5(10):1161–1174.
- 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.
- Rasmus Jørgensen, Oliver Brandt, Mareike Hartmann, Xiang Dai, Christian Igel, and Desmond Elliott. 2023. Multifin: A dataset for multilingual financial nlp.



- In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 864–879.
- Tushar Khot, Ashish Sabharwal, and Peter Clark. 2018. SciTail: A textual entailment dataset from science question answering. In *AAAI*.
- Dan Klein and Christopher D. Manning. 2003. [Accurate unlexicalized parsing](#). In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430, Sapporo, Japan. Association for Computational Linguistics.
- Yuta Koreeda and Christopher D Manning. 2021. ContractNLI: A dataset for document-level natural language inference for contracts. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 1907–1919.
- Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.
- Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. 2024a. Instruction tuning with human curriculum. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1281–1309.
- Sangmin Lee, Suzie Oh, Saeran Park, Guijin Son, and Pilsung Kang. 2024b. [FINALE : Finance Domain Instruction-Tuning Dataset with High-Quality Rationales via Chain-of-Thought Prompting](#). In *Proceedings of the Eighth Financial Technology and Natural Language Processing and the 1st Agent AI for Scenario Planning*, pages 89–106, Jeju, South Korea.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020a. [BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.
- Patrick Lewis, Myle Ott, Jingfei Du, and Veselin Stoyanov. 2020b. [Pretrained language models for biomedical and clinical tasks: Understanding and extending the state-of-the-art](#). In *Proceedings of the 3rd Clinical Natural Language Processing Workshop*, pages 146–157, Online. Association for Computational Linguistics.
- Jiazheng Li, Linyi Yang, Barry Smyth, and Ruihai Dong. 2020. [Maec: A multimodal aligned earnings conference call dataset for financial risk prediction](#). In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management, CIKM '20*, page 3063–3070, New York, NY, USA. Association for Computing Machinery.
- Xianzhi Li, Samuel Chan, Xiaodan Zhu, Yulong Pei, Zhiqiang Ma, Xiaomo Liu, and Sameena Shah. 2023. Are chatgpt and gpt-4 general-purpose solvers for financial text analytics? a study on several typical tasks. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 408–422.
- Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022a. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847.
- Alisa Liu, Swabha Swayamdipta, Noah A Smith, and Yejin Choi. 2022b. Wanli: Worker and ai collaboration for natural language inference dataset creation. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 6826–6847.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. [Roberta: A robustly optimized BERT pretraining approach](#). *CoRR*, abs/1907.11692.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. On llms-driven synthetic data generation, curation, and evaluation: A survey. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 11065–11082.
- Lefteris Loukas, Manos Fergadiotis, Ilias Chalkidis, Eirini Spyropoulou, Prodromos Malakasiotis, Ion Androutsopoulos, and Georgios Paliouras. 2022. [FiNER: Financial numeric entity recognition for XBRL tagging](#). In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4419–4431, Dublin, Ireland. Association for Computational Linguistics.
- Sheng Lu, Hendrik Schuff, and Iryna Gurevych. 2024. [How are prompts different in terms of sensitivity?](#) In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5833–5856, Mexico City, Mexico. Association for Computational Linguistics.
- Lovish Madaan, David Esiobu, Pontus Stenetorp, Barbara Plank, and Dieuwke Hupkes. 2024. Lost in inference: Rediscovering the role of natural language inference for large language models. *arXiv preprint arXiv:2411.14103*.
- Macedo Maia, Siegfried Handschuh, Andre Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. [Www'18 open challenge: Financial opinion mining and question answering](#). pages 1941–1942.
- Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad



- debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.
- Dominique Mariko, Hanna Abi-Akl, Estelle Labidurie, Stephane Durfort, Hugues De Mazancourt, and Mahmoud El-Haj. 2020. The financial document causality detection shared task (fincausal 2020). In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 23–32.
- Dominique Mariko, Hanna Abi-Akl, Kim Trottier, and Mahmoud El-Haj. 2022. [The financial causality extraction shared task \(FinCausal 2022\)](#). In *Proceedings of the 4th Financial Narrative Processing Workshop @LREC2022*, pages 105–107, Marseille, France. European Language Resources Association.
- Rajdeep Mukherjee, Abhinav Bohra, Akash Banerjee, Soumya Sharma, Manjunath Hegde, Afreen Shaikh, Shivani Shrivastava, Koustuv Dasgupta, Niloy Ganguly, Saptarshi Ghosh, et al. 2022. Ectsum: A new benchmark dataset for bullet point summarization of long earnings call transcripts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10893–10906.
- Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. 2020. Adversarial nli: A new benchmark for natural language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4885–4901.
- Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M Mulvey, H Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *arXiv preprint arXiv:2406.11903*.
- OpenAI. 2023. [Gpt-4 technical report](#). *Preprint*, arXiv:2303.08774.
- Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. [Llm evaluators recognize and favor their own generations](#). *Preprint*, arXiv:2404.13076.
- Massimo Poesio and Ron Artstein. 2005. [The reliability of anaphoric annotation, reconsidered: Taking ambiguity into account](#). In *Proceedings of the Workshop on Frontiers in Corpus Annotations II: Pie in the Sky*, pages 76–83, Ann Arbor, Michigan. Association for Computational Linguistics.
- Grace Proebsting and Adam Poliak. 2024. [Hypothesis-only biases in large language model-elicited natural language inference](#). *Preprint*, arXiv:2410.08996.
- Mobashir Sadat and Cornelia Caragea. 2022. Scinli: A corpus for natural language inference on scientific text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7399–7409.
- Mobashir Sadat and Cornelia Caragea. 2024. [MSciNLI: A diverse benchmark for scientific natural language inference](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1610–1629, Mexico City, Mexico. Association for Computational Linguistics.
- Oscar Sainz, Oier Lopez de Lacalle, Gorka Labaka, Ander Barrena, and Eneko Agirre. 2021. [Label verbalization and entailment for effective zero and few-shot relation extraction](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1199–1212, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Chris Samarin, Wynne Hsu, and Mong Li Lee. 2020. Latent retrieval for large-scale fact-checking and question answering with nli training. In *2020 IEEE 32nd International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 941–948. IEEE.
- Marta Sandri, Elisa Leonardelli, Sara Tonelli, and Elisabetta Jezek. 2023. [Why don’t you do it right? analysing annotators’ disagreement in subjective tasks](#). In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2428–2441, Dubrovnik, Croatia. Association for Computational Linguistics.
- Agam Shah, Suvan Paturi, and Sudheer Chava. 2023. Trillion dollar words: A new financial dataset, task & market analysis. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6664–6679.
- Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.
- Chaitanya Shivade et al. 2019. Mednli-a natural language inference dataset for the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium*. Association for Computational Linguistics, pages 1586–1596.
- Ankur Sinha and Tanmay Khandait. 2021. Impact of news on the commodity market: Dataset and results. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 2*, pages 589–601. Springer.
- Joe Stacey, Pasquale Minervini, Haim Dubossarsky, Sebastian Riedel, and Tim Rocktäschel. 2020. [Avoiding the Hypothesis-Only Bias in Natural Language Inference via Ensemble Adversarial Training](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8281–8291, Online. Association for Computational Linguistics.

- Zijun Sun, Chun Fan, Qinghong Han, Xiaofei Sun, Yuxian Meng, Fei Wu, and Jiwei Li. 2020. [Self-explaining structures improve nlp models](#). *Preprint*, arXiv:2012.01786.
- Eugene Sy, Tzu-Cheng Peng, Shih-Hsuan Huang, Heng-Yu Lin, and Yung-Chun Chang. 2023. Fine-grained argument understanding with bert ensemble techniques: A deep dive into financial sentiment analysis. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 242–249.
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajwal Bhargava, Shruti Bhosale, et al. 2023b. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Yuxia Wang, Minghan Wang, and Preslav Nakov. 2024. [Rethinking STS and NLI in large language models](#). In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 965–982, St. Julian’s, Malta. Association for Computational Linguistics.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.
- Adina Williams, Nikita Nangia, and Samuel 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. Association for Computational Linguistics.
- Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhajan Kam-badur, David Rosenberg, and Gideon Mann. 2023. Bloomberggpt: A large language model for finance. *arXiv preprint arXiv:2303.17564*.
- Yuxiang Wu, Matt Gardner, Pontus Stenetorp, and Pradeep Dasigi. 2022. Generating data to mitigate spurious correlations in natural language inference datasets. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2660–2676.
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. The finben: An holistic financial benchmark for large language models. *arXiv preprint arXiv:2402.12659*.
- 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. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 33469–33484.
- Ran Xu, Hejie Cui, Yue Yu, Xuan Kan, Wenqi Shi, Yuchen Zhuang, Wei Jin, Joyce Ho, and Carl Yang. 2023. [Knowledge-infused prompting: Assessing and advancing clinical text data generation with large language models](#). *Preprint*, arXiv:2311.00287.
- Hongyang Yang, Xiao-Yang Liu, and Christina Dan Wang. 2023a. Fingpt: Open-source financial large language models. *FinLLM at IJCAI*.
- Linyi Yang, Eoin Kenny, Tin Lok James Ng, Yi Yang, Barry Smyth, and Ruihai Dong. 2020. Generating plausible counterfactual explanations for deep transformers in financial text classification. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6150–6160.
- Yi Yang, Yixuan Tang, and Kar Yan Tam. 2023b. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*.
- Zhanye Yang. 2022. Legalnli: natural language inference for legal compliance inspection. In *International Conference on Advanced Algorithms and Neural Networks (AANN 2022)*, volume 12285, pages 144–150. SPIE.
- Wenqian Ye, Guangtao Zheng, Xu Cao, Yunsheng Ma, Xia Hu, and Aidong Zhang. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.
- Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Thirty-Seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Guanhua Zhang, Bing Bai, Junqi Zhang, Kun Bai, Conghui Zhu, and Tiejun Zhao. 2019. Mitigating annotation artifacts in natural language inference datasets to improve cross-dataset generalization ability. *arXiv preprint arXiv:1909.04242*.
- Xingxuan Zhang, Jiansheng Li, Wenjing Chu, Junjia Hai, Renzhe Xu, Yuqing Yang, Shikai Guan, Jiazheng Xu, and Peng Cui. 2024. On the out-of-distribution generalization of multimodal large language models. *arXiv preprint arXiv:2402.06599*.
- Wenxuan Zhou, Sheng Zhang, Hoifung Poon, and Muhao Chen. 2023. [Context-faithful prompting for large language models](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14544–14556, Singapore. Association for Computational Linguistics.

Liu Zhuang, Lin Wayne, Shi Ya, and Zhao Jun. 2021. [A robustly optimized BERT pre-training approach with post-training](#). In *Proceedings of the 20th Chinese National Conference on Computational Linguistics*, pages 1218–1227, Huhhot, China. Chinese Information Processing Society of China.

## A Comparison to other Datasets

We provide a detailed comparison of FinNLI to other NLI datasets in Table 5 and other Financial NLP datasets in Table 6. FinNLI is the first NLI dataset specifically focused on the financial domain, filling a significant gap in existing datasets and providing a benchmark for financial reasoning tasks. When comparing FinNLI’s size with other NLI datasets, our evaluation set’s size is comparable to other domain-specific NLI datasets (Table 5) and larger than most existing financial NLP datasets (Table 6). General-domain NLI datasets tend to have larger test sizes, while smaller test sizes are observed in domain-specific datasets reflecting the time and costs incurred in obtaining high-quality expert annotations. Although FinNLI’s training dataset is smaller compared to other datasets, the training and development sets are sufficient for model learning on the task. This is evidenced by model improvements observed when fine-tuned on the FinNLI training set (see results in Section 6). Additionally, our data generation approach is scalable and can generate more instances with additional resources.

FinNLI’s premise and hypothesis pairs are generally longer than those in general-domain datasets, capturing the unique linguistic complexity of financial texts such as Annual Reports and SEC Filings. This length difference may contribute to the performance drop observed in models trained on general-domain NLI datasets when applied to FinNLI (shown in Section 6). The lexical overlap between premise and hypothesis sentences in FinNLI is lower than in general-domain datasets like SNLI and comparable to other domain-specific NLI datasets. This lower overlap indicates that our dataset contains fewer spurious features which models could otherwise exploit to excel on the task.

### A.1 Model Performance Comparison to other Datasets

We compare the performance of RoBERTa-Base and RoBERTa-Large on FinNLI and other NLI datasets, as shown in Table 7. Across the

NLI datasets compared, the performance of both RoBERTa-Large and RoBERTa-Base is lower on FinNLI. This suggests that our dataset is potentially “more difficult” than other NLI datasets, likely due to the specialized language, longer premise-hypothesis pairs and unique reasoning required in the financial domain. However, we note that realistically comparing the difficulty of different NLI datasets can be tricky. This is because various factors, such as how the training data is generated and annotated, the size of the datasets, the distribution of labels, and the evaluation metrics used, can all vary.

Dataset	Domain	Accuracy (%)	Source
<b>ROBERTA-LARGE</b>			
SNLI	General	91.40	(Sun et al., 2020)
MNLI	General	90.20	(Zhuang et al., 2021)
MedNLI	Medical	83.30	(Lewis et al., 2020b)
WANLI	General	75.40	(Liu et al., 2022b)
<b>FinNLI (Ours)</b>	<b>Financial</b>	<b>75.37</b>	
<b>ROBERTA-BASE</b>			
SciNLI	Scientific	78.12	(Sadat and Caragea, 2022)
MSCINLI	Scientific	77.42	(Sadat and Caragea, 2024)
<b>FinNLI (Ours)</b>	<b>Financial</b>	<b>70.69</b>	

Table 7: Comparison of dataset performance across different domains using RoBERTa models.

## B Premise Preprocessing

We observe that directly sampling premise sentences from financial documents without applying any preprocessing results in a high number of invalid sentences. These include: (1) incomplete fragments, possibly caused by parsing errors from the original documents, (2) sentences containing tables, (3) sentences dominated by numerical data, (4) titles and subtitles, and (5) complete but irrelevant sentences, such as pleasantries (e.g., “Thank you for that question” or “Welcome to the call today”) or references to document sections (e.g., “We refer to Section ...”). To mitigate this, we have developed a detailed preprocessing pipeline to filter out such invalid sentences.

Our process begins with NLTK sentence segmentation (Bird et al., 2009) to extract potential sentence sequences. However, this alone does not guarantee that all extracted sentences meet our criteria for valid premises. We identify and remove potential tables by searching for table-related keywords (e.g., “table” or “Table”), long ellipses (e.g., “.....”), extended dash sequences (e.g., “——”), long underscore sequences (e.g., “\_\_\_”), and instances where the number of new lines (e.g., \n) exceeds 10.

Dataset	Domain	#Examples			#Words		% Sentences		Word Overlap
		Train	Dev	Test	Prem.	Hyp.	Prem.	Hyp.	
SNLI (Bowman et al., 2015)	General	550,152	10,000	10,000	14.1	8.3	74.0%	88.9%	52.97%
MNLI (Williams et al., 2018)	General	392,702	20,000	20,000	22.3	-%	91.0%	98.0%	-
ANLI (Nie et al., 2020)	General	162,865	3,200	3,200	-%	-%	-%	-%	-%
WANLI (Liu et al., 2022a)	General	102,885	-	5,000	-%	-%	-%	-%	-%
GNLI (Hosseini et al., 2024)	General	670,739	6,845	490	40.5	11.0	-%	-%	-%
MedNLI (Shivade et al., 2019)	Medical	11,232	1,395	1,422	20.0	5.8	-%	-%	-%
SciNLI (Sadat and Caragea, 2022)	Scientific	101,412	2,000	4,000		25.93	96.8%	96.7%	30.06%
MSciNLI (Sadat and Caragea, 2024)	Scientific	127,320	1,000	4,000	26.84	25.85	94.4%	94.3%	30.29%
<b>FinNLI (Ours)</b>	Financial	16,200	1,800	3,304	22.6	19.5	96.0%	98.9%	38.25%

Table 5: Comparison of key statistics of FinNLI with other related datasets. *#Example* is the number of instances across each data split - we report the human-annotated test size for equal comparison. *#Words* is the mean number of tokens in the premise and hypothesis. ‘%-Sentences’ – proportion of sentences parsed with an ‘S’ root using the Stanford PCFG Parser (Klein and Manning, 2003). *Word Overlap* – mean proportion of tokens shared between the premise and hypothesis.

Dataset	Task	Test Size
FPB (Malo et al., 2014)	Sentiment Analysis	970
FiQA-SA (Maia et al., 2018)	Sentiment Analysis	235
TSA (Cortis et al., 2017)	Sentiment Analysis	561
Headlines (Sinha and Khandait, 2021)	News Headline Classification	2,283
FinCausal (Mariko et al., 2020)	Binary Classification	8,630
FOMC (Shah et al., 2023)	Hawkish-Dovish Classification	496
FinArg-ACC (Sy et al., 2023)	Argument Unit Classification	969
FinArg-ARC (Sy et al., 2023)	Argument Relation Classification	496
MultiFin (Jørgensen et al., 2023)	Multi-Class Classification	690
MA (Yang et al., 2020)	Deal Completeness Classification	500
MLESG (Chen et al., 2023)	ESG Issue Identification	300
<b>FinNLI (ours)</b>	<b>Natural Language Inference</b>	<b>3,304</b>

Table 6: Comparison of FinNLI’s evaluation set with other available Financial NLP textual analysis datasets.

Titles and subtitles are identified by checking if more than 60% of the words in a sentence start with an uppercase letter. Sentences with over 60% numerical characters are also removed, along with any sentences containing URLs.

Additionally, we exclude sentences containing specific keywords, such as [thank you, thanks, greetings, please, see below, continued, check mark]. For Annual Reports and SEC filings, we filter out sentences that do not start with a letter or begin with a lowercase letter. Finally, we limit our selection to sentences that are at least 10 words long but no more than 100 words. To ensure a representative sample, we draw from different length bins to capture a range of premise lengths.

## C Z-Filtering Results

Table 8 shows the top 10 features with the highest *z-statistic* score before and after applying *Z-filtering* (Wu et al., 2022) on the Llama-generated instances. High lexical correlations with the neutral label are observed, with Llama generating hypotheses containing unigrams such as *influence*, *may*, *might*, and *factors*, reported with high *z-scores*.

After applying the algorithm, lower *z-scores* are reported across the features, indicating that task-independent features with high label correlations are filtered out. For instance, the unigram *might* is highly correlated with the neutral class (*z-statistic* = 56.75) but reduces to 25.82 after filtering. However, some correlations remain in the dataset, which models can still exploit, albeit at a reduced level compared to the pre-filtered set. Table 9 presents

Before Z-filtering			After Z-filtering		
Feature	z-statistics	Label	Feature	z-statistics	Label
may	94.69	N	overlap > 0.5	35.18	E
may be	63.66	N	overlap > 0.4	34.26	E
influence	60.64	N	overlap > 0.6	31.61	E
influenced	60.06	N	no	30.62	C
influenced by	59.87	N	'nt	27.74	N
factors	58.23	N	by	26.84	N
flu	58.12	N	might	25.82	C
actors	57.90	N	will be	25.80	N
might	56.75	N	impacted	25.77	N
factor	56.25	N	le	25.76	C

Table 8: Top 10 biased features with the highest *z-statistics* before (left) and after (right) applying *Z-filtering* on the Llama-generated instances. The label *N* indicates the neutral class, *C* indicates the contradiction class, and *E* indicates the entailment class.

the performance scores of XLM-RoBERTa trained on general-domain NLI datasets before and after applying *z-filtering*. The observed performance drop after *z-filtering* suggests that the removed instances were primarily those the model previously classified correctly, likely due to exploiting biases in the training data. Notably, XLM-RoBERTa has been trained on datasets known to contain spurious correlations (Wu et al., 2022; Cosma et al., 2024). Therefore, the performance drop is expected, as removing these artifacts makes the model less reliant on learned biases.



	Before z-filter		After z-filter	
	Acc.	Macro F1	Acc.	Macro F1
Calls	0.83	0.83	0.79	0.77
SEC	0.84	0.84	0.79	0.75
Reports	0.85	0.85	0.82	0.79

Table 9: Performance scores of XLM-RoBERTa-Large trained on general domain NLI datasets, SNLI and MNLI, before and after applying z-filtering.

## D Hyper-parameter Details for Fine-tuning Experiments

We access all models through the HuggingFace library, and conduct training using the PyTorch Lightning framework<sup>12</sup>. Each model is trained for a maximum of 20 epochs. Early stopping is employed with a patience of 5 epochs, based on validation accuracy. We apply an exponential learning rate decay, with the decay rate  $\gamma = 0.95$ . A grid search is performed over the following hyperparameter space: learning rate ( $LR$ )  $\in \{1e-5, 2e-5, 3e-5, 5e-5\}$  and batch size ( $BS$ )  $\in \{24, 32, 64\}$ . The optimal hyperparameter configuration is selected based on the highest validation accuracy, and this configuration is used to retrain the model across three independent runs, each initialized with different random seeds. The best-performing hyperparameters for each model are provided in Table 10.

Model	LR	BS	Max Length
ROBERTA-BASE	$3e-5$	64	200
FILM	$5e-5$	64	200
ROBERTA-LARGE	$1e-5$	32	150
BART-LARGE-MNLI	$1e-5$	32	150

Table 10: Best-performing hyperparameters for each fine-tuned PLM

## E Further Error Analysis

### E.1 Confusion Matrix

Table 11 shows the frequency of misclassifications across entailment (E), contradiction (C), and neutral (N) classes, with a notable trend of misclassifying neutral as entailment. Regardless of which model was used for generation, Llama 3.1 70B often predicts E when the hypothesis expands on the premise without contradiction, making it challenging to determine if the extra information is implied or unrelated. In such cases, we believe

Labels	GPT-4 test.			LLaMA test.		
	E	C	N	E	C	N
E	880	13	60	373	6	29
C	20	539	55	10	289	35
N	209	45	236	195	38	272

Table 11: Confusion matrix for Llama 3.1 70B broken down for GPT-4 and Llama test sets.

LLMs are inferring more than what is logically warranted. Consider example from SEC Filings, a premise “*The Corporate General Partner will continue to maintain compliance with franchise agreements and be economically prudent.*”, and its hypothesis “*The Corporate General Partner will prioritize cost-effective decisions to ensure long-term financial sustainability.*”. While “*economically prudent*” covers a range of financially responsible behaviours, e.g. “*cost-effective decisions*”. The hypothesis indicates that this behaviour is a priority, which is not guaranteed in the premise.

### E.2 Examples demonstrating challenging reasoning patterns

We iteratively refine the prompt exemplars at each generation step by analyzing the predictions of a general-domain NLI model. Table 12 presents examples of the different reasoning patterns we have identified. We use these examples as in-context exemplars to guide the LLM in generating instances that necessitate reasoning for accurate inference. Notably, some of the challenging reasoning patterns resemble revisions made by human annotators in the adversarial NLI setting (Nie et al., 2020).

## F Relabelled instance examples

Examples of instances relabeled by annotators are presented in Table 13. We observe that LLMs tend to append additional information not present in the premise sentence to entailment or contradiction-generated hypotheses, resulting in relabeling to neutral. In some cases, the LLM demonstrates a limited understanding of the NLI task and generates an incorrect label for some of the instances.

<sup>12</sup><https://lightning.ai/docs/pytorch/stable/>

Example	Label	Reasoning
<b>P:</b> We earned \$535,000, \$606,000, and \$794,000 of management fees under the agreement for the years ended December 31, 2014, 2013, and 2012. <b>H:</b> The management fees earned under the agreement in 2014 were less than those earned in 2013.	Contradiction	Mathematical Reasoning
<b>P:</b> Our net equity investment in these VIEs was 808.3 million as of December 31, 2018, and 606.2 million as of December 31, 2017. <b>H:</b> At some point during the year 2018, the net equity investment in these VIEs was either increasing or decreasing from its value at the end of 2017.	Neutral	Temporal Reasoning
<b>P:</b> The Company believes EBITDA is commonly used by financial analysts and others in the industries in which the Company operates and, thus, provides useful information to investors. <b>H:</b> EBITDA is a metric commonly employed by financial industry professionals to assess a company's financial performance, indicating its significance in financial analysis.	Entailment	Financial Knowledge
<b>P:</b> Our case growth for Sysco-branded sales to local customers increased 50 basis points in fiscal 2019. <b>H:</b> There was a modest uptick in our local customer sales last year in 2018.	Neutral	Tricky linguistic construct
<b>P:</b> We've been doubled down in seed for the last four years. <b>H:</b> We have reduced our seed investment substantially over the past four years.	Contradiction	Idiomatic Expression

Table 12: **Examples of challenging reasoning patterns** identified through the prompt iteration process and included in the optimised prompts as in-context exemplars to steer LLM generations towards high-quality examples.

Example	Generated Label	Majority Label
<b>P:</b> If our supply of raw materials is interrupted, our results of operations could be adversely affected. <b>H:</b> If our supply of raw materials is interrupted, our results of operations might be adversely affected.	Neutral	Entailment
<b>P:</b> For the years ended December 31, 2002, 2003 and 2004, our net cash provided by (used for) financing activities was (286.3), 71.1 million and \$(1,686.4) million, respectively. <b>H:</b> The company's cash flow from financing activities fluctuated between 2002 and 2004.	Neutral	Entailment
<b>P:</b> A Look Ahead As we enter 2007, we have identified several priorities by which to grow our business. <b>H:</b> The company has outlined a strategic plan for business expansion in 2007, indicating a proactive approach towards growth and development, <b>which is likely to involve investment in new opportunities, resource allocation, and possibly restructuring of certain business segments.</b>	Entailment	Neutral
<b>P:</b> In November 2016 the LPSC submitted testimony disputing certain aspects of the calculations. <b>H:</b> The LPSC's submission of testimony in November 2016 was the sole factor that <b>led to the immediate acceptance of the calculations without any further review or revision.</b>	Contradiction	Neutral

Table 13: **Examples of relabeled instances:** The first two examples show instances where LLMs generate incorrect labels while the last two examples show instances where LLMs tend to include additional information not in the premise sentence Additional information not in the premise is shown in **red**.

## G LLM Experiments

### G.1 Running Inferences

All the models are accessed from Hugging Face<sup>13</sup>, and inferences are run using vLLM (Kwon et al., 2023) if a model is supported. We run inferences on all the models using full precision, except for Llama 3.1 70B, for which we use a 4-bit quantized version to reduce memory footprint. Additionally, we run the FinMA models with the Guidance<sup>14</sup> library to constrain the output to only the NLI class labels. During the initial runs of the FinMA models (Xie et al., 2023), we observed that the LLMs frequently failed to follow the task instructions and instead generated hallucinated class labels, such as *positive* and *negative*, possibly due to overfitting to sentiment analysis tasks in the instruction-tuned data. We run all inferences using a temperature value of 0 across three independent runs with randomly initialized seeds.

### G.2 Prompts

We provide the prompt templates used in the paper across the generation and evaluation experiments below.

#### 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 (ZS-AG)

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:

<sup>13</sup><https://huggingface.co/>

<sup>14</sup><https://github.com/guidance-ai/guidance>

### Few-Shot w/ Annotation Guidelines (Few-Shot-AG)

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: {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**: Wet 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:



### Chain-of-Thought (CoT) w/ Rationale

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 determine the relationship between the premise and hypothesis and assign one of the following labels:

- **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.

To decide on the correct label, follow these steps:

1. **Understand the Premise**: Carefully read and comprehend the premise to identify the key information it provides.
2. **Analyze the Hypothesis**: Read the hypothesis and determine what it claims or implies. Consider if it adds new information, rephrases the premise, or predicts something based on the premise.
3. **Compare and Reason**: Compare the key information from the premise and hypothesis. Does the hypothesis logically follow from the premise, contradict it, or is it unrelated?
4. **Assign the Label**: Based on your reasoning, choose the appropriate label: 'entailment', 'neutral', or 'contradiction'.
5. **Explain Your Reasoning**: Briefly describe the thought process that led to your decision.

Here are a few examples: {examples}

Given the premise and hypothesis below, determine the relationship between the sentences and return 'entailment', 'neutral', or 'contradiction':

Premise: {premise}

Hypothesis: {hypothesis}

Reasoning: [Describe your reasoning here]

Label:

### Data Generation Prompt

You are now a **{role}** and will be provided with a sentence from a **{financial domain}**. Using only this sentence and your general knowledge about the world:

- Write one **entailment** hypothesis where the truth of the premise guarantees the truth of the hypothesis.
- Write one **neutral** hypothesis where the truth of the premise neither guarantees nor contradicts the truth of the hypothesis.
- Write one **contradiction** hypothesis where the truth of the premise guarantees that the hypothesis is false.

#### Guidelines for Writing High-Quality Hypotheses:

1. **Numerical Reasoning:** If the premise contains numbers, write hypotheses that require mathematical knowledge or quantitative reasoning. This means the relationship between the premise and the hypothesis should involve some form of mathematical deduction. Example: {examples}
2. **Temporal Reasoning:** If the premise includes dates, write hypotheses that require an understanding of time. The hypothesis should involve an inference that requires understanding dates and the relationship between different points in time. Example: {examples}
3. **Financial Knowledge:** If the premise includes financial concepts, generate hypotheses that require familiarity with financial reasoning. Example: {examples}
4. **Linguistic Knowledge:** Try to write hypotheses that require linguistic understanding, such as antonyms, hypernyms, metaphors, or syntactic ambiguity. Example: {examples}

#### Important Writing Considerations:

- Write all hypotheses as clear, declarative sentences. Do not phrase them as questions.
- For an **entailment** label, ensure the hypothesis can be directly and fully verified using only the information provided in the premise. Avoid adding assumptions or speculative language such as "likely," "indicating," or "potential."
- When generating a **contradiction** hypothesis, avoid simply negating the premise. Instead, use financial knowledge and linguistic reasoning to construct a meaningful contradiction.

Given the premise: {premise sentence}

Write three hypotheses using: {writing style}

Return your response in the format below:

- **Entailment:** {entailment hypothesis}
- **Neutral:** {neutral hypothesis}
- **Contradiction:** {contradiction hypothesis}

## H Annotation Instructions for the Financial Natural Language Inference Dataset (FinNLI)

### Annotation Instructions for FinNLI

**Task Overview:** You will annotate a dataset for **Financial Natural Language Inference (FinNLI)**. You will be given pairs of sentences: a **premise** and a **hypothesis**, where:

- The **premise** comes from financial texts (SEC filings, annual reports, earnings call transcripts).
- The **hypothesis** is generated using a Large Language Model.

**Your task is to:**

1. **Verify that both the premise and hypothesis are valid and meaningful sentences** e.g. are not phrase fragments or nonsense, which would prevent us being able to assign a label to the relationship between the premise and the hypothesis. We have taken steps to preprocess the sampled sentences, so we expect invalid sentences to be a low frequency event. If either of the sentences is invalid, you do not need to provide relationship label for them.
2. **Assign one label to each of hypothesis-premise pair.:** Choose one of the three:
  - **Entailment** – A relationship between two statements where the truth of the premise guarantees the truth the hypothesis. The hypothesis follows logically from the information contained in the premise.
  - **Neutral** – A relationship between two statements where the truth of the premise neither guarantees nor contradicts the truth of the hypothesis. It is not possible to determine whether the hypothesis is true or false without further information.
  - **Contradiction** – The truth of the premise guarantees that the hypothesis is false. The hypothesis is logically false from the information contained in the premise.
3. **Indicate your level of confidence in the annotation: Low or High.** Depends on how confident an annotator is while defining the relation between premise and hypothesis.
4. **[Optional] Provide Comments** pertaining to annotating the instance, e.g. provide the rationale for the decision in free text form explaining why the instance is difficult, and/or why the confidence is low. Raise any other concerns you have in this field, e.g. in unlikely cases where you consider the texts harmful or biased.

### Label Definitions and Examples:

**Entailment:** The hypothesis logically follows from the premise.

**Premise:** We may acquire hotels in joint ventures with third parties that could result in conflicts.

**Hypothesis:** The company is considering partnerships with third parties for hotel acquisitions.

**Label:** Entailment

**Neutral:** The premise neither guarantees nor contradicts the hypothesis.

**Premise:** Director stock options granted under the plans expire after seven years and vest fully after six months.

**Hypothesis:** All employee stock options granted under the plans expire after seven years and vest fully after six months.

**Label:** Neutral

**Contradiction:** The truth of the premise guarantees that the hypothesis is false.

**Premise:** The user assumes all risks for any damages or losses arising from any use of this information, except to the extent such damages or losses cannot be limited or excluded by applicable law.

**Hypothesis:** The provider is liable for any damages or losses that the user may incur from using this information.

**Label:** Contradiction

**Handling Invalid Sentences:** If a sentence is not meaningful, classify it as **Premise\_notValid** or **Hypothesis\_notValid**. Common cases include:

1. **Incomplete Sentence:** Sentence is visibly incomplete. *Example:* “Premise: Before joining FMC Technologies, Mr.”
2. **Complex Notation:** Excessive numbers or financial notation making it unreadable. *Example:* “Premise: \$700.0 \$613.7 \$492.1 95 Rest of World 108.6 96.0 93.9”
3. **Document Title or Subtitle:** Appears to be a section title instead of a standalone sentence. *Example:* “Net Cruise Cost less fuel expense.”
4. **Reference to External Text:** Refers to another document section without self-contained meaning. *Example:* “For further discussion of our annual impairment test, see the Critical Accounting Policies and Estimates section.”

#### Edge Cases:

- **Labels Are Not Symmetrical:**

**Premise:** Girl ate an apple.

**Hypothesis:** Girl ate a fruit.

**Label:** Entailment

**Premise:** Girl ate a fruit.

**Hypothesis:** Girl ate an apple.

**Label:** Neutral

- **Use of Domain Knowledge:** Some financial knowledge is allowed.

**Premise:** The Company believes EBITDA is commonly used by financial analysts and provides useful information to investors.

**Hypothesis:** EBITDA is a metric commonly used by financial professionals to assess financial performance.

**Label:** Entailment

- **Handling Questions:**

**Premise:** Jane is coming at 6.

**Hypothesis:** When is Jane coming?

**Label:** Entailment

**Premise:** Jane is coming at 6.

**Hypothesis:** Why isn't Jane coming?

**Label:** Contradiction

#### Final Notes:

- If uncertain, mark annotation confidence as **Low**.
- Avoid unnecessary speculation beyond the provided premise.
- If the annotation process reveals bias or harmful content, leave a comment.