Deriving Strategic Market Insights with Large Language Models: A Benchmark for Forward Counterfactual Generation

Keane Ong[♠◇], **Rui Mao**[♣], **Deeksha Varshney**[♠♡], **Paul Pu Liang**[◇], **Erik Cambria**[♣] and **Gianmarco Mengaldo**[♠*]
[♠]National University of Singapore [◇]Massachusetts Institute of Technology
[♣]Nanyang Technological University [♡]Indian Institute of Technology, Jodhpur
keane.ongweiyang@u.nus.edu; deeksha@iitj.ac.in; {mpegim}@nus.edu.sg;
ppliang@mit.edu; {rui.mao,cambria}@ntu.edu.sg

Abstract

Counterfactual reasoning typically involves considering alternatives to actual events. While often applied to understand past events, a distinct form–forward counterfactual reasoning–focuses on anticipating plausible future developments. This type of reasoning is invaluable in dynamic financial markets, where anticipating market developments can powerfully unveil potential risks and opportunities for stakeholders, guiding their decision-making. However, performing this at scale is challenging due to the cognitive demands involved, underscoring the need for automated solutions. LLMs offer promise, but remain unexplored for this application. To address this gap, we introduce a novel benchmark, FIN-FORCE–**FIN**ancial **FOR**ward **C**ounterfactual **E**valuation. By curating financial news headlines and providing structured evaluation, FIN-FORCE supports LLM based forward counterfactual generation. This paves the way for scalable and automated solutions for exploring and anticipating future market developments, thereby providing structured insights for decision-making. Through experiments on FIN-FORCE, we evaluate state-of-the-art LLMs and counterfactual generation methods, analyzing their limitations and proposing insights for future research. We release the benchmark, supplementary data and all experimental codes at the following link: https://github.com/keanepotato/fin_force

1 Introduction

Counterfactual reasoning–considering alternatives to actual events–allows us to envision possibilities beyond reality (Byrne, 2016). While often used retrospectively to consider *what could have happened*, a distinct forward-looking form–*forward counterfactual reasoning*–focuses on *what could happen next* (Todorova, 2015; Bynum et al., 2023).

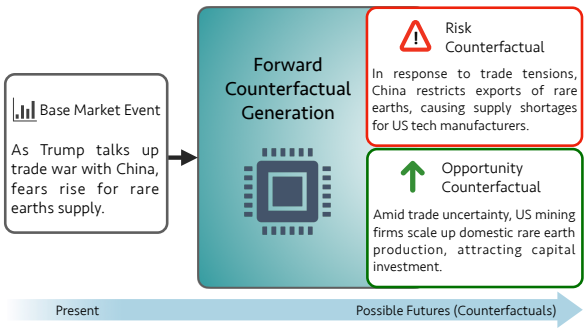

Figure 1: Overview of FIN-FORCE task. Given a financial news headline depicting a market event, an LLM is tasked with generating two forward counterfactuals - an opportunity counterfactual and a risk counterfactual. While the opportunity counterfactual explores how the event can positively shift, the risk counterfactual highlights potential adverse scenarios.

In dynamic financial markets, forward counterfactual reasoning often plays a crucial role in guiding decision-making (Byrne, 2016; Du et al., 2024). To this end, market stakeholders frequently anticipate future developments from current events–not to predict exact outcomes, but to explore plausible future scenarios that inform strategic responses. This allows them to hedge against potential risks and capitalize on emerging opportunities (Greenwood and Shleifer, 2014). Yet, despite its strategic value, forward counterfactual reasoning remains difficult to scale. Specifically, due to its complex cognitive demands–i.e., requiring reasoning across multiple causal relationships (Lebow, 2000)–it is impractical to apply forward counterfactual reasoning on an extensive range of market events or within short timeframes. In other words, stakeholders face constraints in how widely and rapidly they can explore and anticipate future developments from current events. This limits their ability to derive timely foresight for strategic decision-making, leaving them exposed to missed opportunities and strategic missteps (Schoemaker, 1995).

*Corresponding author: mpegim@nus.edu.sg

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 11411–11434
November 4-9, 2025

Addressing this need for scalability, automated tools–such as LLMs–can support forward counterfactual reasoning. Accordingly, LLMs can enable the expansive generation of plausible future developments based on actual events (Du et al., 2025b). However, LLM counterfactual research has often focused on specific applications, such as narrative rewriting or text classification (Wang et al., 2024). The use of LLMs to support forward counterfactual reasoning in finance, or any domain, remains largely unexplored. To bridge this gap, we propose FIN-FORCE – **FIN**ancial **FOR**ward **C**ounterfactual **E**valuation, a novel benchmark to support LLM forward counterfactual generation in finance.

FIN-FORCE comprises news headlines, each describing a distinct market event, which serves as a basis for LLMs to generate two types of forward counterfactuals. i) The *opportunity counterfactual* explores how the market event could positively shift, resulting in favorable implications for market stakeholders. ii) The *risk counterfactual* highlights how the market event could adversely shift, exposing vulnerabilities and negative market implications. While opportunity counterfactuals anticipate potential upside scenarios for stakeholders to capitalize on, risk counterfactuals identify emerging risks for stakeholders to hedge against (Figure 1). By laying the foundation for LLMs to generate these forward counterfactuals, FIN-FORCE supports scalable and automated insights into potential market opportunities and risks before they materialize, enhancing stakeholders' strategic decision-making.

Through experiments, we evaluate a wide range of LLM-based methods on the FIN-FORCE benchmark to provide insights for future model development. Our findings highlight: (1) LLMs, under zero-shot and few-shot prompting, do not perform equally well on FIN-FORCE, with Claude 3.5 Haiku performing better than Qwen 2.5 72B, Llama 4-Maverick, Gemini 2.0 Flash and GPT-4o. (2) The evaluated state-of-the-art (SOTA) counterfactual prompting methods perform poorly, while SOTA sampling-based counterfactual generation achieves the best performance. (3) A self-training paradigm can enable a smaller LLM (i.e. Llama3.1 8B) to outperform all large-scale LLMs (i.e. GPT-4o) under zero-shot and few-shot prompting. (4) The limitations of the different methods via qualitative analysis and sub-task performance, and research directions for tackling these limitations.

We summarize our main contributions as follows. (1) We develop and release FIN-FORCE, a novel benchmark comprising 1368 news headlines that describe market events, to support forward counterfactual generation in finance. (2) We conduct extensive experiments on FIN-FORCE to evaluate a wide-range of methods, offering insights to guide future model development in the NLP community. While both contributions are situated in the financial domain, the underlying task of forward counterfactual generation is, to our knowledge, one of the first task formulations of its kind. The task and its core principles–projecting plausible futures, and directional outcomes (positive and negative)–could be potentially extended to other complex and dynamic domains such as public policy (Tetlock, 2017) or scenario planning (Schoemaker, 1995), to support strategic decision-making.

2 Related Work

Counterfactual Reasoning. Counterfactual reasoning involves considering alternative outcomes based on changes to a given "base" situation (Byrne, 2016). It is used to explain past events, anticipate possible future developments, and support a variety of tasks across domains (Byrne, 2016; Wang et al., 2024). In decision-making contexts, anticipating future developments is particularly valuable in uncertain environments beyond finance, including policy (Tetlock, 2017) and scenario planning (Schoemaker, 1995). While our work centers on the financial domain, the task we develop–scalable counterfactual generation for projecting plausible future developments (i.e. forward counterfactual generation)–could be extended to support strategic decision-making in other complex, uncertain settings.

NLP Counterfactual Generation Benchmarks. Counterfactual generation has been widely studied in NLP, with established benchmarks supporting different applications. TIMETRAVEL focuses on counterfactual story rewriting (Qin et al., 2019), SNLI explores modifications to premises or hypotheses for natural language inference (Kaushik et al., 2019), and COUNTERFACT evaluates whether language models can faithfully update specific factual knowledge (Meng et al., 2022). Despite the progress, existing benchmarks do not consider temporal progression from a base event or forward-looking counterfactuals that

project how events might plausibly unfold from present scenarios. Additionally, leveraging counterfactual generation for practical financial applications remains largely unexplored. To address these gaps, we introduce FIN-FORCE, a novel benchmark for forward counterfactual generation in finance.

LLM Methods for Counterfactual Generation. LLMs' strength in counterfactual generation has spurred methods such as prompt engineering (Nguyen et al., 2024), mask-and-replace (Feng et al., 2024b), anomalous language modeling (Mao et al., 2024), and controlled text generation (Ravfogel et al., 2024). The generalizability of these methods to new tasks remains unexplored, and state-of-the-art LLM paradigms like self-training (Yuan et al., 2024) have yet to be applied to counterfactual generation. Our work extends the literature by evaluating the generalizability of existing LLM counterfactual methods to forward counterfactual generation. We also explore the potential and limitations of self-training when applied to a counterfactual generation setting.

LLM Applications in Financial Tasks. Recent advances in LLMs have enabled a range of promising applications in finance (Du et al., 2025a; Yeo et al., 2025), thanks to LLMs' strong reasoning abilities (Plaat et al., 2024). These include stock prediction (Heng et al., 2025), financial statement analysis (Kim et al., 2024), ESG rating evaluation (Ong et al., 2025a), and greenwashing detection (Ong et al., 2025b). However, the counterfactual reasoning capabilities of LLMs—and specifically their capacity to perform scenario analysis (i.e., exploring possible future risks and opportunities)—remain underexplored. This is despite the central role that scenario analysis plays in the financial sector, including within investment funds that proactively anticipate future market trends based on evolving narratives (Phadnis et al., 2015). By investigating LLMs' abilities in this complex area, we extend research on LLM-based financial applications to advanced reasoning tasks.

3 Benchmark Construction

The construction of FIN-FORCE begins with a broad collection of news data, prior to annotating news that describe market events. These annotated news are then included in the benchmark.

3.1 News Collection

To collect news headlines that describe financial market events, we queried NewsAPI[1] with market-related keywords (eg., *GDP growth*, *Interest rates*– full-list in Appendix Table 3). These keywords are chosen to capture information generally associated with the financial markets. The duration from which we collected news headlines spans from 1 September 2024 to 18 April 2025. This period was selected to post-date the knowledge cut-offs or release dates of the LLMs evaluated in Section 4, ensuring no overlap with their pretraining data. This improves the robustness of our findings.

3.2 News Annotation Scheme

Given that not all headlines collected from NewsAPI pertain to financial market events, we apply human annotation to select those that clearly do. Accordingly, a headline is selected for the FIN-FORCE benchmark if it meets the following criteria[2], as summarized below:

Event Status: The headline must explicitly describe an ongoing financial market event. An event is defined as a current, factual development representing a specific change or occurrence affecting financial markets or the broader economy. Examples include central bank decisions, regulatory changes, or macroeconomic data releases. Headlines reflecting generic commentary or opinions without concrete developments are excluded.

Material Relevance: The event must be material to market participants, meaning it reflects a development that can potentially influence financial decision-making or market outcomes. This includes broad systemic drivers (e.g., monetary policy changes, macroeconomic indicators) and significant corporate events (e.g., mergers, earnings surprises, corporate restructuring). These event types are described by the market categories in Table 1. To be considered material, the event must explicitly relate to at least one of these market categories[2].

3.3 News Annotation Process

Our annotation process[2] involves 3 annotators and 2 verifiers, all of whom are doctoral or post-doctoral researchers specializing in finance. (1) Training Phase: Annotators and verifiers participate in iterative trial rounds, each consisting of 60 randomly selected samples.

[1] https://newsapi.org

[2] Details of the annotators, annotation instructions, descriptions of the market categories are in Appendix A.

Market Categories	Count
Monetary policy & central banking	109
Corporate strategy & operations	229
Geopolitical & regulatory developments	148
Financial markets & asset performance	574
Supply chain & logistics	24
ESG & sustainability developments	13
Technology & innovation	22
Labour & employment	27
Macroeconomic indicators	186
Banking & financial stability	36

Table 1: FIN-FORCE categories and headline counts. Certain topics appear more frequently due to their natural prevalence in financial news, resulting in a benchmark that better reflects real-world news distributions.

After each round, annotations are reviewed and feedback given, continuing until participants reach $\geq$95% accuracy. (2) Annotation: Once the trial phase is complete, annotators proceed with daily labeling. They are instructed to flag any samples where they are uncertain about the correct label. (3) Disagreement Resolution: Every two days, flagged samples are reviewed in group discussions among the annotators to reach a consensus. If unanimity is not possible, the majority decision is adopted. (4) Verification: Every two days, 25% of the annotated samples from each annotator (excluding those marked as uncertain) are randomly selected and reviewed by the verifiers for correctness and adherence to the guidelines. If more than 5% of the reviewed annotations are found to be incorrect, the entire batch from that two day window is re-annotated. Our final benchmark comprises 1368 news headlines that describe financial market events, with a full breakdown shown in Table 1.

3.4 Supplementary Dataset

To supplement the benchmark, we annotate an additional 2,105 news headlines from 1 Jan 2021 to 20 June 2024, following the same annotation procedures. We use GPT-4o to generate forward counterfactuals for these headlines, creating a supplementary synthetic dataset[3] for the SFT warm-up phase in our self-training paradigm (Section 4).

3.5 Metrics Design

Counterfactuals are traditionally evaluated using *validity*, *similarity*, *diversity*, and *fluency* (Wang et al., 2024). However, not all are relevant to FIN-FORCE, which involves generating risk and opportunity counterfactuals from a market event described by a news headline. *Validity*–whether the counterfactual flips a classifier's label (Mothilal et al., 2020)–is irrelevant, as our task does not involve predefined classifier labels that can be altered. *Similarity*–the minimality of counterfactual edits from the original text (Treviso et al., 2023)–can be counterproductive, as the blanket penalization of edits may discourage changes that introduce meaningful risk or opportunity shifts. *Diversity*–divergence in semantic meaning between counterfactuals (Wang et al., 2024)–overlooks how the counterfactuals must represent distinct market risks and opportunities rather than *any* difference in wording or semantics. *Fluency*–naturalness and grammaticality of counterfactuals, via perplexity (Radford et al., 2019)–is the only metric which remains relevant.

Therefore, for our task, we build on perplexity for fluency evaluation by focusing on Δ Perplexity[4]–the difference between the average perplexity of all counterfactuals and all original headlines–which normalizes for the fluency of the original headlines. In place of *validity*, *similarity*, and *diversity*, we introduce two new metrics–*forward-compatibility* and *directionality*, which are better aligned with FIN-FORCE's objectives. Importantly, these metrics assess whether counterfactuals reflect plausible future developments, not whether they actually occur. This is consistent with real-world strategic analysis, where even future developments that do not materialize can powerfully inform decision-making – provided that they reflect credible future scenarios (Schoemaker, 1995).

Forward-Compatibility[5] assesses whether the counterfactual represents a plausible future development that logically follows from the original market event. It must meet: (i) Consistent progression. It reflects a continuation logically connected the original market event. (ii) Non-contradiction. It does not negate the original event by introducing mutually exclusive scenarios.

[3]Supplementary dataset is also released at https://github.com/keanepotato/fin_force to support self-training model development. Further details of this dataset are in Appendix A.4

[4]Perplexity is computed by GPT-2; the average perplexity of all the original headlines is 267.98, averaged over ten runs.

[5]Prompts for LLM evaluation, details about the human validation study are in Appendix A & B. Further details on the new metrics can also be found in the human study.

Directionality[5] assesses whether the counterfactual shows a clear and meaningful market shift–toward either improvement (opportunity) or deterioration (risk) in market conditions–relative to the original market event. It must meet: (i) Relative Significance. The shift is substantial compared to the original market event. (ii) Logical Soundness. The shift is grounded in logical reasoning that does not reflect economic or financial implausibility. (iii) Scope of Impact. The shift extends beyond isolated parties to affect other market participants. (iv) Financial Consequence. The shift entails financial effects likely to influence market behaviour or decision-making.

Each counterfactual is assessed using an LLM-as-a-judge framework[5] with GPT-4o (Zheng et al., 2023), and is labeled as satisfying *forward-compatibility* or *directionality* only if it meets all their corresponding criteria. The scores in our results–Fwd-Compat. and Dir. (Section 5)–reflect the proportion (in percentages) of counterfactuals meeting the *forward-compatibility* and *directionality* criteria respectively. While LLM evaluation increases impartiality and scalability (Zheng et al., 2023), we acknowledge its limitations. To ensure robustness, we validate the LLM judgments through a human-LLM agreement study[5] with 500 random samples. This study compares LLM labels with those from independent human annotators, showing an average agreement of 81.4% on *directionality* and 89.6% on *forward-compatibility*.

4 Experiments

The following models are tasked with generating a single risk and a single opportunity counterfactual from each headline in FIN-FORCE.

Baseline LLM Prompting[6]**.** Latest LLMs are evaluated under zero and few-shot prompting. To improve experimental robustness, selected models have release or knowledge cut-off dates preceding 1 September 2024 (FIN-FORCE's headlines are collected after this date). LLMs include proprietary–Claude 3.5 Haiku (Anthropic, 2024), Gemini 2.0 Flash (Google, 2024), GPT-4o (OpenAI, 2024), and open-source models–Llama 4 Maverick (Meta, 2025), Qwen 2.5 72B (Yang et al., 2024).

SOTA Counterfactual Generation[6]**.** We evaluate SOTA counterfactual generation algorithms adaptable to our task. LLMs-for-CFs (Nguyen et al., 2024) uses chain-of-thought prompting to identify and replace keywords in the original text. CounterfactualDistil (Feng et al., 2024b) masks topic words and noun phrases in the original text, then prompts an LLM to generate replacements. LM-Counterfactuals (Ravfogel et al., 2024) generates counterfactuals by holding sampling noise fixed across completions using the Gumbel-max trick.

Self-Training Paradigm[6]. As an alternative to prompt-based methods with large-scale LLMs, we adapt a self-training approach–SRLM (Yuan et al., 2024). A smaller LLM–Llama 3.1 8B (Meta, 2024)– is tuned on its own outputs using Direct Preference Optimization (DPO) (Rafailov et al., 2023), after fine-tuning on the supplementary dataset (Section 3.4).

[6]Full details on how we adapted the methods for our task are in Appendix B.2.

5 Results & Discussion

To inform future model development on this benchmark, we analyze each model's performance and compare them. For clarity, we center our discussion on Δ Perplexity and FwdCompat-Dir Avg.–which comprises the average of *forward-compatibility* (Fwd-Compat.) and *directionality* (Dir.) scores (these metrics are defined in Section 3.5).

5.1 Baseline LLM Prompting

LLMs with the highest Fwd-Compat-Dir Avg. scores under zero and few-shot settings do not exhibit the strongest general reasoning performance. From Table 2, Claude 3.5 Haiku (64.51%), Claude 3.5 Haiku FS (62.80%), achieve the highest Fwd-Compat-Dir Avg. scores. This is despite models with lower Fwd-Compat-Dir Avg. performance–Gemini 2.0 Flash, GPT-4o, and Llama 4 Maverick–exhibiting stronger results on general reasoning benchmarks (eg., GPQA) (Meta, 2025; Yang et al., 2024). This suggests that FIN-FORCE requires specialized reasoning skills not captured by standard reasoning benchmarks. *Improving* FIN-FORCE *performance will require targeting these specialized skills rather than general reasoning alone.*

LLMs underperform at generating directionally accurate opportunity counterfactuals compared to risk counterfactuals, across zero and few-shot settings. This is shown by the lower *directionality* scores for opportunity compared to risk counterfactuals in Figure 2 for all baseline LLM prompting methods. Error analysis ("Trivial Consequences" in Figure 4) reveals that generated op-

Method	Perplexity ↓	Δ Perplexity ↓	Fwd-Compat. ↑	Dir. ↑	FwdCompat-Dir Avg. ↑
Claude 3.5 Haiku	442.70	+174.72	55.41%	73.61%	64.51%
Claude 3.5 Haiku FS	433.83	+165.85	78.07%	47.54%	62.80%
Gemini 2.0 Flash	459.28	+191.30	53.51%	61.99%	57.75%
Gemini 2.0 Flash FS	432.15	+164.17	66.34%	58.34%	62.34%
Llama4 Maverick	512.01	+244.03	38.93%	50.84%	44.89%
Llama4 Maverick FS	325.36	+57.38	68.92%	48.69%	58.80%
GPT-4o	415.27	+147.29	67.84%	47.22%	57.53%
GPT-4o + FS	327.03	+59.05	**84.47%**	37.76%	61.11%
Qwen 2.5 72B	359.40	+91.42	61.88%	57.02%	59.45%
Qwen 2.5 72B FS	302.34	+34.36	73.32%	51.27%	62.29%
LLMs-for-CFs	550.25	+282.28	54.75%	42.25%	48.50%
CounterfactualDistil	498.31	+230.33	41.23%	14.77%	28.00%
LM-Counterfactuals	**155.84**	**-112.13**	62.06%	68.93%	**65.50%**
SRLM	258.51	-9.47	56.25%	**73.79%**	65.02%

Table 2: Overall evaluation results across all Risk-Opportunity counterfactuals. Best results are bolded, second-best results are underlined. ↓ (lower score is better), ↑ (higher score is better). Fwd-Compat. is *forward-compatibility*, Dir. is *directionality*, FwdCompat-Dir Avg. is the average between *forward-comptability* and *directionality* scores. FS stands for few-shot, with results averaged over 5 random samplings for few-shot examples.

portunity counterfactuals often describe vague or superficial positive market shifts. For example, stating that a firm (Asian Paints) "sees potential in rural markets" without specifying concrete actions that pose an opportunity for market participants. The tendency to default to these superficial shifts suggests that LLMs under zero and few-shot prompting, may lack strong conceptual understanding of meaningful positive market shifts, possibly due to knowledge gaps (Feng et al., 2024a). *To strengthen this conceptual grasp, advanced prompting such as metacognitive prompting can be explored (Wang and Zhao, 2023). This can guide the LLM to explicitly reflect on whether its reasoning exhibits key conceptual principles of a meaningful market shift–i.e. criteria (i) to (iv) of directionality in Section 3.5–thereby improving performance.*

Few-shot prompting does not improve performance across all metrics. From Figure 3, all LLMs improve on Δ Perplexity and Fwd-Compat. but show reduced Dir. performance with few-shot compared to zero-shot prompting. This divergence highlights that few-shot prompting cannot optimize performance across all metrics. *Structured prompting strategies–i.e. least-to-most prompting (Zhou et al., 2022)–may help by decomposing the task into intermediate steps, wherein each metric can be explicitly optimized. This can deliberately guide the model through the reasoning process and reduce the risk of overlooking key criteria.*

5.2 SOTA Counterfactual Generation

SOTA counterfactual prompting strategies underperform significantly, while sampling-based generation performs markedly better. Despite leveraging specialized prompting strategies for counterfactual text classification and QA, LLMs-for-CF and CounterfactualDistil exhibit among the weakest Δ Perplexity scores (+282.28 and +230.33) and FwdCompat-Dir Avg. scores (48.50% and 28.00%). In contrast, LM-Counterfactuals, which utilizes sampling noise, achieves the best performance (Δ Perplexity -112.13; FwdCompat-Dir Avg. 65.50%). *These results suggest that the counterfactual prompting strategies, while effective for other counterfactual tasks that rely on minimal edits or label flipping, do not transfer well to* FIN-FORCE*, which demands more substantive reasoning. However, the sampling-based method achieves better generalization for this more complex task.*

CounterfactualDistil's masking strategy undermines contextual consistency. Error analysis ("Fantastical Edits" in Figure 4) shows that masking and replacing key topic words and noun phrases often disrupts the original headline's narrative. Forced to regenerate content without access to the masked terms, the model often produces counterfactuals that lack contextual fidelity with the original headline, introducing unrelated and disparate market shifts. This compromises both Fwd-Compat. and Dir. scores. *These findings provide insight for future method development on* FIN-FORCE *– preserving the original headline's context may be important for achieving counterfactuals with high contextual fidelity, which translates to stronger Fwd-Compat. and Dir. performance.*

Despite preserving the original headline's context, LLMs-for-CF produces counterfactuals

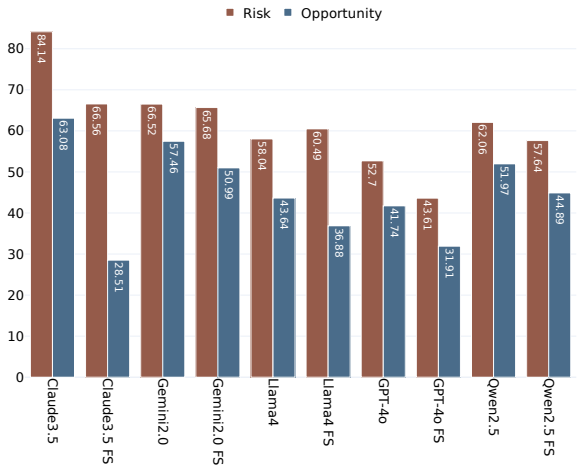

Figure 2: *Directionality* (Dir.) scores between risk and opportunity counterfactuals for different baseline LLM prompting methods under zero and few-shot settings.

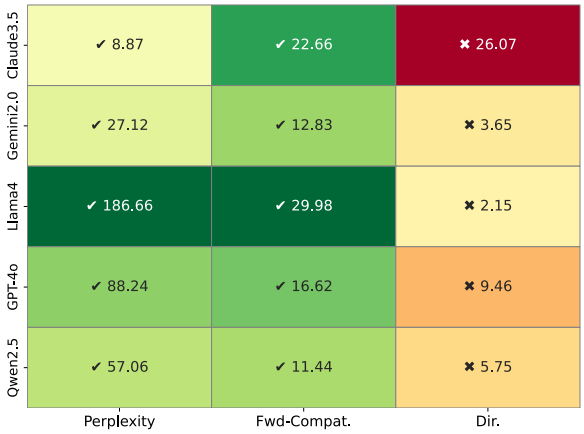

Figure 3: Absolute performance changes with few-shot relative to zero-shot prompting for different LLMs. ✓ indicates improvement; ✗ indicates degradation.

that lack meaningful positive or adverse market developments, as reflected by its poor Dir. score (third-lowest, at 42.25%). Unlike CounterfactualDistil, LLMs-for-CF reasons over the complete, unmasked headline to infer words to replace, generating counterfactuals that have better contextual consistency with the original headline. However, by restricting changes to word-level substitutions without broader narrative changes (e.g. adding new clauses), the generated counterfactuals often reflect shallow semantic changes from the original headline (i.e. "Similar Semantics" in Figure 4). This restricts the counterfactuals from introducing concrete market shifts that reflect meaningful risk or opportunity, leading to a weaker Dir. score. *These findings suggest that narrative-level rewriting, not just token-level replacement, is required for meaningful and directionally valid counterfactuals.*

Word replacement-based methods often compromise the fluency of counterfactual text. The LLMs-for-CF and CounterfactualDistil methods, which focus on replacing key terms in the original headline, yield among the weakest Δ Perplexity (+282.28 and +230.33). Error analysis ("Fluency Lapse" in Figure 4) reveals that direct word replacements often produce awkward phrasing or contradictions, as the surrounding sentence is not adapted to accommodate the word changes. This indicates that beyond limiting semantic depth, word replacement strategies also impair fluency and coherence. *Thus, while methods that involve broader narrative rewriting is needed to produce meaningful and directionally valid counterfactuals, they are equally important for ensuring coherence and fluency.*

Controlling random sampling noise across counterfactual generations shows promise for enhancing performance. LM-Counterfactuals controls sampling noise during generation, such that completions primarily reflect prompt changes rather than stochastic variance in the token sampling process. This may help the model follow prompt instructions more consistently (though the exact underlying mechanism requires further study), thereby generating counterfactuals that reflect the best Δ Perplexity and FwdCompat-Dir Avg scores. Nonetheless, error analysis ("Financially Unrealistic" in Figure 4) shows that the counterfactuals occasionally reflect financially unrealistic reasoning. This affects the validity of their *directionality*, reducing Dir. scores. For example, overly optimistic extensions of negative headlines– i.e. A firm (CVS) planning to hire workers immediately after major layoffs and a potential company breakup. *These findings suggest that controlling sampling noise is a promising direction for counterfactual generation in* FIN-FORCE, *given that reducing stochastic variance has been effective for improving performance. However, further gains may depend on enhancing the financial realism of generated counterfactuals. For this, domain adaptation via preference tuning on financial data (Rafailov et al., 2023), can strengthen the financial reasoning of these generations.*

5.3 Self-Training

Self-training achieves competitive results on FIN-FORCE. SRLM achieves the second highest Δ Perplexity (-9.47) and FwdCompat-Dir Avg.

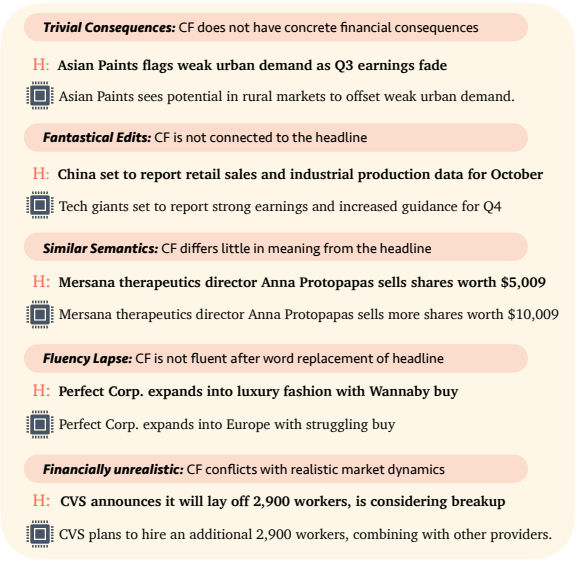

Figure 4: Analysis of prominent error cases. H represents a headline in FIN-FORCE; denotes the erroneous LLM response; CF stands for counterfactual.

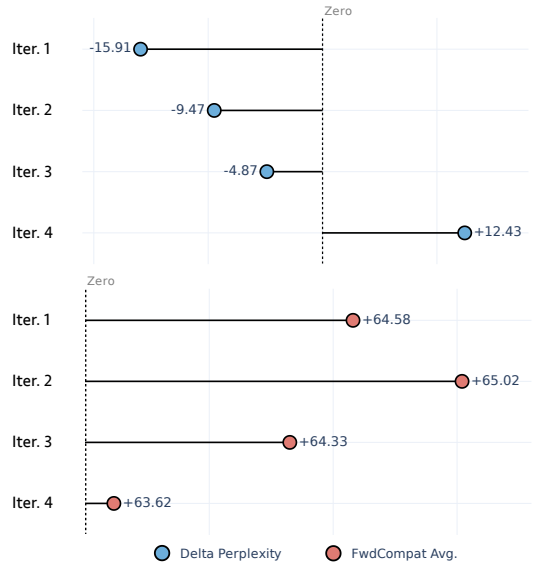

Figure 5: Performance across different iterations of training in the SRLM self-training paradigm. Iter. stands for training iteration.

(65.02%) scores. Notably, SRLM achieves this despite using a relatively simple self-training setup, where the model evaluates its own responses on a basic 10-point scale to construct a preference set for DPO (see Appendix B.2). *This strong performance, achieved with a relatively basic design, suggests considerable potential for improvement through more advanced self-training strategies.*

Self-training reaches saturation quickly, limiting further improvements in performance. In SRLM, the model constructs preference sets from its own outputs and trains on them over multiple iterations. From Figure 5, SRLM reaches its peak FwdCompat Avg. (65.02%) at the second iteration, and best Δ Perplexity (-15.91) at the first iteration, with no further gains beyond these points. This effect, known as saturation, is a common limitation of self-training observed across tasks (Wu et al., 2024; Yuan et al., 2024). However, in FIN-FORCE, this limitation may be exacerbated by the added complexity of balancing multiple objectives–i.e. *forward-compatibility* (Fwd-Compat.), *directionality* (Dir.), and fluency (Δ Perplexity)–which makes it harder to reliably distinguish preferred from rejected outputs when building preference sets. *Hence, future development of self-training methods could focus on sustaining effective learning signals and exploring how* FIN-FORCE*'s multi-objective complexity–by making it more difficult to judge outputs–affects self-training performance.*

5.4 Comparison of Methods

Baseline prompting, self-training, and sampling methods offer distinct trade-offs. Baseline LLM prompting–using LLMs in zero and few-shot settings–delivers competitive performance without fine-tuning and can be deployed in plug-and-play fashion (i.e. directly through APIs without GPU training). In contrast, tuning-based methods (SRLM) and sampling-based algorithms (LM-Counterfactuals) can achieve better performance on FIN-FORCE but involve greater computational complexity (i.e. training pipelines, decoding control) for preference tuning and controlled decoding. *Therefore, each method entails practical trade-offs that extend beyond performance optimization alone, and should be considered in relation to computational cost, complexity, and resource availability.*

From Table 2, SRLM surpasses the performance of all baseline LLM prompting methods on Δ Perplexity (-9.47) and FwdCompat-Dir Avg. (65.02%). This highlights how a self-training method (SRLM) can enable smaller LLMs (Llama 3.1-8B) to achieve strong FIN-FORCE performance, mitigating the need to rely on larger models (i.e. GPT-4o, Llama4-Maverick). With performance no longer a limiting factor, this enables the adoption of smaller models that offer greater controllability and are more computationally efficient at inference time–qualities that are often impractical with larger LLMs. *Therefore, not only does SRLM achieve strong performance, it also enables the develop-*

ment of efficient, controllable smaller language models tailored to FIN-FORCE*, reducing dependence on large-scale LLMs.*

The sampling-based LM-Counterfactuals and SRLM methods are not mutually exclusive, and can potentially be integrated to optimize performance. In fact, combining them can address a key limitation of LM-Counterfactuals noted previously in Section 5.2–generating counterfactuals that occasionally reflect unrealistic financial reasoning. Self-training can help mitigate this limitation by adapting the LM-Counterfactuals model to the financial domain. As part of self-training, an LLM-judge can automatically construct a preference set from LM-Counterfactuals' outputs, selecting responses with sound financial reasoning and rejecting those without. The model can then be optimized on this preference set through DPO, reducing financial reasoning lapses and improving counterfactual generation quality. *Therefore, integrating LM-Counterfactuals with self-training (SRLM) may thus represent a promising direction for improving model performance on* FIN-FORCE.

6 Conclusion

We introduced FIN-FORCE, a novel benchmark for forward counterfactual generation in finance, and evaluated a range of methods to offer insights for future model development. Through experiments, we find that while existing methods offer a starting point, they face limitations in producing fluent, forward-compatible, and directionally valid counterfactuals. Of these methods, sampling-based generation achieved the highest overall performance, self-training enabled smaller models to attain competitive performance, while zero and few-shot prompting for selected LLMs offered a plug-and-play approach with strong results. However, all methods have limitations–errors and inconsistent performance on metrics–highlighting the need for further work. Our work aims to lay the foundation for scalable, automated insights into potential market opportunities and risks for stakeholders.

Limitations

Our benchmark is currently limited to English-language headlines. While this provides broad coverage, we acknowledge that multilingual financial news is important for global market analysis–particularly in regions where English sources are limited. Future work will explore extending the benchmark to include non-English news sources. Additionally, while the benchmark focuses on delivering counterfactual insights for financial stakeholders, counterfactuals have also been applied to improve the explainability and performance of AI models (Treviso et al., 2023; Mothilal et al., 2020), which is an ongoing concern in critical sectors such as finance and healthcare (Mengaldo, 2024; Turbé et al., 2025). To this end, future work may also explore leveraging forward counterfactuals to improve the explainability and performance of financial models (i.e. for stock prediction).

Ethical Considerations

Data collection procedures received approval from our research group's internal ethics review board. We uphold ethical standards by collecting and processing data with strict attention to privacy and confidentiality. Our data sources include publicly available online news obtained via NewsAPI. As these news items may reference companies or individuals, we take care to anonymize sensitive or personal information in the FIN-FORCE benchmark, focusing exclusively on the financial content relevant to our study. The models employed are also publicly accessible and sourced from published research. We comply with the copyright terms set by the respective holders for all data, software packages, and models used. Human annotators operate under rigorous guidelines to ensure objectivity and minimize bias. We are committed to transparency in our methodology and clearly attribute all sources to uphold ethical standards in data usage and sharing. Dataset and code are released publicly for research purposes only, in accordance with relevant copyright requirements.

Acknowledgements

This research/project is supported by the NUS Sustainable and Green Finance Institute (SGFIN), NUS Asian Institute of Digital Finance (AIDF), Ministry of Education, Singapore under its MOE Academic Research Fund Tier 2 (STEM RIE2025 Award MOE-T2EP20123-0005: "Neurosymbolic AI for Commonsense-based Question Answering in Multiple Domains"), MOE Tier 2 Award (MOE-T2EP50221-0006: "Prediction-to-Mitigation with Digital Twins of the Earth's Weather"), MOE Tier 1 Award (MOE-T2EP50221-0028: "Discipline-Informed Neural Networks for Interpretable Time-Series Discovery"), and by the RIE2025 Industry

Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026: "Generative AI"), administered by A*STAR, as well as supported by Alibaba Group and NTU Singapore

References

Anthropic. 2024. Claude 3.5 haiku. https://www.anthropic.com/claude/haiku. Accessed: 2025-04-29.

Lucius EJ Bynum, Joshua R Loftus, and Julia Stoyanovich. 2023. Counterfactuals for the future. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14144–14152.

Ruth MJ Byrne. 2016. Counterfactual thought. *Annual review of psychology*, 67(1):135–157.

Kelvin Du, Frank Xing, Rui Mao, and Erik Cambria. 2024. Financial sentiment analysis: Techniques and applications. *ACM Computing Surveys*, 56(9):1–42.

Kelvin Du, Yazhi Zhao, Rui Mao, Frank Xing, and Erik Cambria. 2025a. Natural language processing in finance: A survey. *Information Fusion*, 115:102755.

Kelvin Du, Yazhi Zhao, Rui Mao, Frank Xing, and Erik Cambria. 2025b. A retrieval-augmented multi-agent system for financial sentiment analysis. *IEEE Intelligent Systems*, 40(2):15–22.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024a. Don't hallucinate, abstain: Identifying llm knowledge gaps via multi-llm collaboration. *arXiv preprint arXiv:2402.00367*.

Tao Feng, Yicheng Li, Li Chenglin, Hao Chen, Fei Yu, and Yin Zhang. 2024b. Teaching small language models reasoning through counterfactual distillation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5831–5842, Miami, Florida, USA. Association for Computational Linguistics.

Google. 2024. Gemini 2.0 flash model documentation. https://ai.google.dev/gemini-api/docs/models#gemini-2.0-flash. Accessed: 2025-04-29.

Robin Greenwood and Andrei Shleifer. 2014. Expectations of returns and expected returns. *The Review of Financial Studies*, 27(3):714–746.

Ryan Quek Wei Heng, Edoardo Vittori, Keane Ong, Rui Mao, Erik Cambria, and Gianmarco Mengaldo. 2025. Leveraging LLMs for top-down sector allocation in automated trading. In *Proceedings of ICLR Workshops*.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, Adriane Boyd, and 1 others. 2020. spacy: Industrial-strength natural language processing in python.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data. *arXiv preprint arXiv:1909.12434*.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. 2024. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Richard Ned Lebow. 2000. What's so different about a counterfactual? *World politics*, 52(4):550–585.

Rui Mao, Kai He, Claudia Beth Ong, Qian Liu, and Erik Cambria. 2024. MetaPro 2.0: Computational metaphor processing on the effectiveness of anomalous language modeling. In *Findings of the Association for Computational Linguistics: ACL*, pages 9891–9908, Bangkok, Thailand. Association for Computational Linguistics.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. *Advances in neural information processing systems*, 35:17359–17372.

Gianmarco Mengaldo. 2024. Explain the black box for the sake of science: the scientific method in the era of generative artificial intelligence. *arXiv preprint arXiv:2406.10557*.

Meta. 2024. Llama 3.1 model cards and prompt formats. https://www.llama.com/docs/model-cards-and-prompt-formats/llama3_1/. Accessed: 2025-04-29.

Meta. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. Accessed: 2025-04-29.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. 2020. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 607–617.

Van Bach Nguyen, Paul Youssef, Christin Seifert, and Jörg Schlötterer. 2024. LLMs for generating and evaluating counterfactuals: A comprehensive study. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14809–14824, Miami, Florida, USA. Association for Computational Linguistics.

Keane Ong, Rui Mao, Ranjan Satapathy, Ricardo Shirota Filho, Erik Cambria, Johan Sulaeman, and Gianmarco Mengaldo. 2025a. Explainable natural language processing for corporate sustainability analysis. *Information Fusion*, 115:102726.

Keane Ong, Rui Mao, Deeksha Varshney, Erik Cambria, and Gianmarco Mengaldo. 2025b. Towards robust ESG analysis against greenwashing risks: Aspect-action analysis with cross-category generalization. In *Proceedings of ACL*, pages 14854–14879.

OpenAI. 2024. Hello GPT-4O. https://openai.com/index/hello-gpt-4o/. Accessed: 2024-11-23.

Shardul Phadnis, Chris Caplice, Yossi Sheffi, and Mahender Singh. 2015. Effect of scenario planning on field experts' judgment of long-range investment decisions. *Strategic Management Journal*, 36(9):1401–1411.

Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Bäck. 2024. Reasoning with large language models, a survey. *CoRR*.

Lianhui Qin, Antoine Bosselut, Ari Holtzman, Chandra Bhagavatula, Elizabeth Clark, and Yejin Choi. 2019. Counterfactual story reasoning and generation. *arXiv preprint arXiv:1909.04076*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, and 1 others. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741.

Shauli Ravfogel, Anej Svete, Vésteinn Snæbjarnarson, and Ryan Cotterell. 2024. Counterfactual generation from language models. *arXiv preprint arXiv:2411.07180*.

Paul JH Schoemaker. 1995. Scenario planning: a tool for strategic thinking. *MIT Sloan Management Review*.

Philip E Tetlock. 2017. Expert political judgment: How good is it? how can we know?-new edition.

Mariana Todorova. 2015. Counterfactual construction of the future: Building a new methodology for forecasting. *World Future Review*, 7(1):30–38.

Marcos Treviso, Alexis Ross, Nuno M. Guerreiro, and André Martins. 2023. CREST: A joint framework for rationalization and counterfactual text generation. In *Proceedings of ACL*, pages 15109–15126.

Hugues Turbé, Mina Bjelogrlic, Gianmarco Mengaldo, and Christian Lovis. 2025. Tell me why: Visual foundation models as self-explainable classifiers. *arXiv preprint arXiv:2502.19577*.

Yongjie Wang, Xiaoqi Qiu, Yu Yue, Xu Guo, Zhiwei Zeng, Yuhong Feng, and Zhiqi Shen. 2024. A survey on natural language counterfactual generation. *arXiv preprint arXiv:2407.03993*.

Yuqing Wang and Yun Zhao. 2023. Metacognitive prompting improves understanding in large language models. *arXiv preprint arXiv:2308.05342*.

Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Wei Jie Yeo, Wihan Van Der Heever, Rui Mao, Erik Cambria, Ranjan Satapathy, and Gianmarco Mengaldo. 2025. A comprehensive review on financial explainable AI. *Artificial Intelligence Review*, 58:189.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason E Weston. 2024. Self-rewarding language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 57905–57923. PMLR.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

A Benchmark

A.1 Keywords for Querying Market News

As described in Section 3.1, we collect a broad corpus of news data by querying NewsAPI using relevant keywords. The retrieved headlines are then annotated to select those included in our benchmark. We specify the full list of keywords in Table 3.

Market Keywords
GDP growth
Interest rates
Inflation rate
Unemployment rate
Consumer spending
Disposable income
Stock market performance
Foreign exchange rates
Economic growth forecasts
Investment climate
Cost of capital
Economic stability

Table 3: Full list of market-related keywords used to query NewsAPI for collecting financial news headlines.

A.2 Financial Market Categories

Our benchmark comprises 1368 news headlines that belong to defined financial market categories. Each of these categories describes a specific type of financial market event. We provide a full description of these categories in Table 4.

A.3 Benchmark Samples

To provide a better appreciation of the financial news headlines collected in the FIN-FORCE benchmark, we provide more samples in Table 5. Additionally, we denote the financial market category that each sample belongs to.

A.4 Supplementary Dataset

To generate the forward counterfactuals in the supplementary dataset, we generate a risk and opportunity counterfactual from a single news headline utilizing GPT-4o. We leverage the same implementation of GPT-4o under zero-shot setting as detailed in Appendix B.2 (baseline LLM prompting methods).

A.5 Annotators

A total of five human contributors, all based in Singapore and recruited from reputable research institutions, participated in the benchmark annotation (Section 3.3) and human-LLM agreement study (Section 3.5) as annotators or verifiers. They are affiliated with the Asian Institute of Digital Finance and the National University of Singapore's College of Design and Engineering. All annotators are actively pursuing doctoral or post-doctoral research in finance and possess expertise in financial markets. Their participation in the annotation process forms part of their formal academic and research activities. They were compensated at a rate exceeding the local minimum wage (SGD $15/hour). The annotators adhered strictly to the established annotation scheme and guidelines and provided consent for the dataset's use in research. To ensure objectivity in the human-LLM agreement study for *directionality* and *forward-compatibility* metrics, the annotators were deliberately kept independent from the paper's development.

A.6 News Annotation Instructions

In this section, we present the full annotation instructions provided to annotators for labeling the news headlines in our benchmark. The annotators are first instructed to read the background, followed by the general instructions, event status and material relevance guidelines.

Background: Financial news headlines describe a wide range of information about the economy and financial markets. These headlines can provide actionable intelligence for market stakeholders, helping to inform decision-making. However, the content of financial headlines varies greatly, from opinions and commentary to concrete events and factual developments. In this annotation task, you will analyze financial news headlines and evaluate them according to specific criteria. Your annotations will be used for research purposes.

General Instructions:

1. Please annotate each headline according to whether it describes a financial market event.
2. In order for a headline to qualify as describing a financial market event, it must meet two criteria: *Event Status* and *Material Relevance*.
3. A headline satisfies the *Event Status* criteria if it satisfies all the requirements outlined in the *Event Status Guidelines*. A headline satisfies the *Material Relevance* criteria if it satisfies all the requirements outlined in the *Material Relevance Guidelines*.
4. First, evaluate whether the headline satisfies all the requirements in *Event Status Guidelines*.
5. If the headline satisfies all the requirements in *Event Status Guidelines*, assess whether the headline satisfies all the requirements in the *Material Relevance Guidelines*.
6. Only if the headline meets both criteria–*Event Status* and *Material Relevance*–classify the headline as "TRUE". This denotes that the headline describes a financial market event.
7. If the headline meets neither criteria or only meets one criteria, mark it as "FALSE". This denotes that the headline does not describe a financial market event.
8. For headlines where you are not confident about the annotation, please flag them. These will be discussed and further reviewed by the annotation team.

Event Status Guidelines:

1. The headline must explicitly describe an ongoing financial market event. This is defined as a current, factual development representing a specific change or occurrence affecting financial markets or the broader economy.

2. Examples of financial market events include but are not limited to central bank decisions, earnings reports releases, regulatory changes, geopolitical escalations, or macroeconomic data releases.

3. Exclude headlines that provide only generic commentary or opinions without describing a concrete development.

Material Relevance Guidelines:

1. The headline must describe an financial market event that is material or meaningful to financial market participants, indicating that it can potentially influence financial decision-making or financial market outcomes.

2. Material or meaningful financial market events include, but are not limited to, broad systemic drivers (e.g., monetary policy changes, macroeconomic indicators) and significant company-specific events (e.g., mergers, earnings surprises, corporate restructuring). To guide this assessment, we provide a detailed list of financial market categories that encompass these types of events: *Monetary Policy & Central Banking, Corporate Strategy & Operations, Geopolitical & Regulatory Developments, Financial Markets & Asset Performance, Supply Chain & Logistics, ESG & Sustainability Developments, Technology & Innovation, Labour & Employment, Macroeconomic Indicators, Banking & Financial Stability.* [Note: We provided the annotators with Table 4, which provides definitions of these categories.]

3. The financial market event highlighted in the headline must explicitly relate to at least one of the aforementioned categories to be considered material.

4. If the event explicitly relates to at least one of the categories, please mark the category that the headline is most closely associated with.

A.7 Human-LLM Agreement Study

Besides the news annotation, we also conducted a human-LLM agreement study. The goal was to evaluate whether the LLM judge (GPT-4o) aligns with the preferences of proficient human annotators when using the criteria of *forward-compatibility* and *directionality* to evaluate counterfactuals. The human-LLM agreement study did not exactly follow the same News Annotation Process (described in Section 3.3). Specifically, we retained only the (1) Training Phase, and the annotations were conducted over a three-day period. We did not implement disagreement resolution or verification procedures, as the goal was to measure average alignment across the annotators and the LLM, rather than to produce a fully adjudicated ground truth.

In the following, we present the full annotation instructions provided to the annotators for the study. Similar to the news annotation, the annotators are first instructed to read the background, followed by the general instructions, *forward-compatibility* and *directionality* guidelines. A total of 500 counterfactual samples, randomly selected from all experiments in Section 4, were evaluated. Each counterfactual is paired with its original financial news headline, which is also taken into account during evaluation, as we will detail below.

Background: In this annotation task, you will evaluate counterfactuals generated by LLMs according to specific criteria. Each counterfactual must be assessed with consideration of the original financial news headline from which it was generated. Your annotations will be used for research purposes.

General Instructions:

1. For each counterfactual, evaluate it in relation to its original news headline, as required by the *Forward-Compatibility* and *Directionality* criteria.

2. Assess the counterfactual using the two evaluation criteria: *Forward-Compatibility* and *Directionality*.

3. First, determine whether the counterfactual satisfies all the requirements specified in the *Forward-Compatibility Guidelines*. If it does, mark Forward-Compatibility as "TRUE"; otherwise, mark it as "FALSE".

4. Then, assess whether the counterfactual satisfies the requirements in the *Directionality Guidelines*. If it does, mark Directionality as "TRUE"; otherwise, mark it as "FALSE".

5. Record the outcome for each criterion–*Forward-Compatibility* and *Directionality*–separately. The counterfactual should receive a TRUE or FALSE label for each criterion independently of the other.

Forward-Compatibility Guidelines:

1. The counterfactual must represent a plausible future development that logically follows from the market event described in the original news headline. This entails that the counterfactual meets the requirements of *consistent progression* and *non-contradiction*.

2. The counterfactual must have *consistent progression*. The counterfactual must reflect a continuation logically connected to the original market event. For example, if the original headline reports an interest rate hike, a counterfactual about unrelated environmental regulations would not qualify.

3. The counterfactual must have *non-contradiction*. The counterfactual must not negate or reverse the original market event by introducing mutually exclusive scenarios. For example, if the original headline states that a company is exiting a market, a counterfactual suggesting the company is expanding in that same market would be contradictory.

Directionality Guidelines:

1. The counterfactual must reflect a clear and meaningful market shift–either toward improvement (opportunity) or deterioration (risk) in market conditions–relative to the market event described in the original news headline. This entails that the counterfactual meets the requirements of *relative significance*, *logical soundness*, *scope of impact* and *financial consequence*.

2. The counterfactual must have *relative significance.* The counterfactual must reflect a shift that is substantial compared to the original market event. Superficial changes–such as minor wording differences or trivial updates that do not alter the market implications–do not qualify.

3. The counterfactual must have *logical soundness.* The counterfactual must reflect a shift that is grounded in logical reasoning. It cannot reflect implausibility in financial or economic logic.

4. The counterfactual must have sufficient *scope of impact*. The counterfactual must reflect a shift that affects not just isolated parties, but also other market participants. For example, a company initiative to upgrade office equipment would not qualify, while a CEO change that could affect investors, competitors, or market expectations would.

5. The counterfactual must have *financial consequence*. The counterfactual must reflect a shift that entails financial effects likely to influence market behavior or decision-making. Examples include but are not limited to changes affecting revenue, costs, investment, or market valuation.

B Experiments

B.1 LLM-as-a-judge Evaluation

Two of the evaluation metrics for FIN-FORCE, *directionality* and *forward-compatibility* leverage GPT-4o classification under an LLM-as-a-judge framework (Zheng et al., 2023). Each generated counterfactual is presented to the LLM one at a time for evaluation, and separate runs are performed for assessing *forward-compatibility* and *directionality*. We highlight the LLM-as-a-judge prompts utilized for *directionality* in Table 6, and for *forward-compatibility* in Table 7.

B.2 Model Implementation Details

We will detail the implementation of the experimental methods in Section 4. For the baseline LLM prompting methods, we leverage the respective LLM versions: gpt-4o-2024-08-06 for GPT4o (OpenAI, 2024), claude-3-5-haiku-20241022 for Claude 3.5 Haiku (Anthropic, 2024), gemini-2.0-flash for Gemini 2.0 Flash (Google, 2024), Llama 4 Maverick (17Bx128E) for Llama 4 Maverick (Meta, 2025), Qwen 2.5-72B-Instruct for Qwen 2.5 72B (Yang et al., 2024). For these methods, we use the prompt template in Table 8, and include few-shot examples for few-shot setups

while omitting them for zero-shot setups. We provide samples of our few shot examples in Table 9. To ensure consistency, few-shot examples are kept the same across the different LLMs.

For the SOTA counterfactual generation methods, we adapt LLMs-for-CFs (Nguyen et al., 2024), CounterfactualDistil (Feng et al., 2024b), LM-Counterfactuals (Ravfogel et al., 2024) to our task. The original LLMs-for-CF implementation uses chain-of-thought prompting to identify and replace keywords for counterfactual generation, and includes a one-shot example. In our setup, we adopt a similar approach–guiding the model to identify and replace keywords–while also providing a reasoning chain example. We implement this using GPT-4o as the underlying LLM. The prompt template for our implementation of LLMs-for-CF is shown in Table 8.

The original CounterfactualDistil implementation comprises of two main steps. First, it masks the topic word and noun phrases in the original text. Then, it prompts an LLM to generate replacements for the masked spans based on a predefined target label, ultimately producing the counterfactual. In our setup, we use GPT-4o to identify topic words using the prompt in Table 10, and apply the SpaCy library (Honnibal et al., 2020) to extract noun phrases. These text spans are then masked. The masked text is then passed to GPT-4o to generate replacements according to the target risk and opportunity labels. The full prompt template for our implementation of CounterfactualDistil at the final counterfactual generation step is shown in Table 8.

The original LM-Counterfactuals method generates counterfactuals by fixing the sampling noise across both the base and counterfactual generations, using the Gumbel-max trick. It first generates a base completion to recover the sampling noise, which is then held fixed when generating the counterfactual. In our setup, this base completion is formulated as a continuation of the original financial news headline, using the prompt in Table 11. The recovered noise is then reused to generate both risk and opportunity counterfactuals. For this counterfactual generation step, we utilize a similar task prompt to the one used in the baseline LLM prompting methods, as shown in Table 8, without few-shot examples. Following the original implementation, we use the Llama-3-8B-Instruct model from Hugging Face[7], applying 4-bit quantization and keeping the same hyperparameters.

For self-training, we adapt the SRLM framework (Yuan et al., 2024). In the original implementation, an LLM is first fine-tuned on human-annotated alignment data, then used to generate synthetic prompts and responses. The LLM also judges these responses to create a synthetic preference set, which is used to further train the model via Direct Preference Optimization (DPO) (Rafailov et al., 2023). This process is repeated over multiple iterations. In our setup, we train a 4-bit quantized Llama3.1-8B-Instruct model from Hugging Face[8] using LoRA and the Unsloth library. After fine-tuning on our supplementary dataset (Section 3.4), the model is prompted with the prompt shown in Table 12 to generate synthetic financial headlines. It is then prompted with a similar task prompt to the one used in the baseline LLM prompting methods (Table 8), without few-shot examples, to produce risk and opportunity counterfactuals. These counterfactual outputs are judged by the LLM using a point-scale prompt (Table 13), similar to the original implementation, to distinguish chosen and rejected outputs for constructing the DPO preference set. The rest of our implementation follows the original SRLM with two modifications to the hyperparameters. First, due to the lack of multi-source validation data, we use validation loss for early stopping (patience = 2 steps), due to its simplicity and established role in model selection. Second, to accommodate computational constraints, we train the model with LoRA using lora_dropout=0.1, lora_alpha=16, and lora_r=64, and approximate the original effective batch size of 16 by using a batch size of 1 with 16 gradient accumulation steps. Additionally, following the SRLM setup, we train the model for two iterations and report results from the second iteration in Table 2. We continue training the model for a third and fourth iteration to examine the saturation effect, as discussed in Section 5.3.

B.3 Computation and Tools Used for Our Study

This study was conducted with the help of external, publicly available tools (NewsAPI[9], Py-

[7] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[8] https://huggingface.co/unsloth/Meta-Llama-3.1-8B-Instruct-bnb-4bit

[9] https://newsapi.org

torch[10], Huggingface[11], SpaCy (Honnibal et al., 2020), GPT-4o, Llama 3, Llama 4, Claude 3.5 Haiku, Qwen 2.5, Gemini 2.0 Flash), with all experiments run on a single NVIDIA GeForce RTX 4090 GPU. To compute perplexity, we use GPT-2 from the evaluate[12] package. GPT-4o is deployed through the OpenAI API[13], while Claude 3.5 Haiku, Llama 4 Maverick, Qwen 2.5 72B and Gemini 2.0 Flash are deployed through the Openrouter API[14]. We specify the model sizes utilized for our study (model sizes for proprietary models are unavailable): Qwen 2.5 72B (72B), GPT-2 (1.5B), Llama 4 Maverick (400B), Llama3.1-8B-Instruct (8B), Llama3-8B-Instruct (8B).

B.4 Results Breakdown

Since the FIN-FORCE task requires generating both a risk and an opportunity counterfactual for each news headline, the overall results in Table 2 report the average evaluation scores across all generated risk and opportunity counterfactuals. For a more detailed performance breakdown, Table 14 and Table 15 report the evaluation scores for all risk and opportunity counterfactuals, respectively.

[10]https://pytorch.org
[11]https://huggingface.co
[12]https://huggingface.co/docs/evaluate/en/index
[13]https://platform.openai.com/docs/overview
[14]https://openrouter.ai

Category	Description
Monetary policy & central banking	Involves central bank actions, interest rate decisions, and policy announcements that systematically influence market liquidity and investor sentiment.
Corporate strategy & operations	Pertains to strategic business decisions, restructuring, and operational changes that impact corporate performance and market dynamics.
Geopolitical & regulatory developments	Covers political events, regulatory changes, or international developments that systematically affect market stability and economic policy.
Financial markets & asset performance	Relates to trends, shifts, or events in financial markets that directly drive asset valuations and investment portfolios.
Supply chain & logistics	Encompasses disruptions or improvements in global supply chains and logistics that systematically influence corporate earnings and market risk.
ESG & sustainability developments	Focuses on environmental, social, and governance initiatives that are shaping long-term market trends and systematic risk assessments.
Technology & innovation	Addresses technological breakthroughs and innovations that are driving structural shifts in competitiveness and economic growth.
Labour & employment	Deals with changes in employment trends and labour market dynamics that have systematic implications for consumer spending and economic performance.
Macroeconomic indicators	Involves key economic statistics and trends that provide systematic insights into overall market conditions.
Banking & financial stability	Concerns developments within the banking sector or financial system that influence systemic risk and market confidence.

Table 4: Descriptions of key market categories in the FIN-FORCE benchmark.

News Headline	Market Category
Google, Microsoft Are Spending Massively on AI, Quarterly Earnings Show.	Technology & innovation
UK's largest retailers warn Budget will lead to job cuts.	Labour & employment
Japan approves new climate, energy and industry policies through 2040.	ESG & sustainability developments
U.S. tariffs on steel, aluminum spark strong backlash across Europe.	Geopolitical & regulatory developments
MAS eases monetary policy for the second time this year; lowers core inflation forecast.	Monetary policy & central banking

Table 5: Sample financial news headlines and their associated market categories.

Evaluation	Prompt

Directionality (Opportunity)	You are an expert judge assessing whether a counterfactual generated by an LLM represents a materially favorable market development relative to the original market development described in the news headline.
	The news headline is enclosed between <news> and </news>.
	The counterfactual is enclosed between <response> and </response>.
	
	Evaluation Criterion:
	Your task is to determine whether the counterfactual reflects a clear and meaningful improvement in market conditions for market participants relative to the original market development. Return true only if all the following criteria are met:
	- The counterfactual must clearly describe a market-relevant improvement with a positive impact on market participants more broadly, beyond any isolated party involved.
	- The market-relevant improvement must be clearly tied to financial consequences. General positivity or reputational gain (e.g., "increased trust") is not sufficient unless explicitly linked to tangible market benefits.
	- The improvement must be material relative to the original market development, representing a clear positive shift relative to the original market development, not a superficial or marginal gain.
	- The counterfactual must be logically coherent and causally sound. Improvements based on implausible, inconsistent, or unsupported reasoning should be rejected.
	- The improvement must be clearly stated and well-supported. Vague, implied, or unsubstantiated benefits do not qualify.
	- The counterfactual must introduce or amplify a credible opportunity for market participants–including but not limited to enhanced revenue, improved efficiency, capital access, or favorable positioning–that could reasonably influence decision-making or market behavior.
	
	Return your evaluation in the following JSON format:
	{
	"directionality": {
	"value": <true or false>
	}
	}

Directionality (Risk)	You are an expert judge assessing whether a counterfactual generated by an LLM represents a materially adverse market development relative to the original market development described in the news headline.
	The news headline is enclosed between <news> and </news>.
	The counterfactual is enclosed between <response> and </response>.
	
	Evaluation Criterion
	Your task is to determine whether the counterfactual reflects a clear and meaningful deterioration in market conditions for market participants relative to the original market development. Return true only if all the following criteria are met:
	- The counterfactual must clearly describe a market-relevant deterioration with a negative impact on market participants more broadly, beyond any isolated party involved.
	- The market-relevant deterioration must be clearly tied to financial consequences. General negativity or reputational harm (e.g., "loss of trust") is not sufficient unless explicitly linked to tangible financial consequences.
	- The deterioration must be material relative to the original market development. It should reflect a clear negative shift relative to the original market development, not a superficial or minor setback.
	- The counterfactual must be logically coherent and causally sound. Deteriorations based on implausible, inconsistent, or unsupported reasoning should be rejected.
	- The deterioration must be clearly stated and well-supported. Vague, implied, or unsubstantiated harms do not qualify.
	- The counterfactual must introduce or amplify a credible risk for market participants–including but not limited to uncertainty, volatility, or exposure to future loss–that could reasonably affect decision-making or market behavior.
	
	Return your evaluation in the following JSON format:
	{
	"directionality": {
	"value": <true or false>
	}
	}

Table 6: LLM-judge prompts for evaluating the *directionality* of counterfactuals. Directionality (Opportunity) is used to evaluate opportunity counterfactuals, while Directionality (Risk) is used to evaluate risk counterfactuals.

Prompt

You are an expert judge tasked with assessing the quality of the counterfactual response generated by an LLM, based on provided financial news headlines.
- The news headline is enclosed between <news> and </news>.
- The LLM response (the counterfactual) is enclosed between <response> and </response>.
 	
Your task is to carefully review the LLM-generated counterfactual and assess whether it represents a plausible future development that remains consistent with the original news headline. You should return a structured evaluation using the criterion below.
 	
Evaluation Criterion:
Forward Compatibility, or in other words, does the counterfactual represent a plausible future development from the original market development described in the news headline? Respond with a true or false value. To be considered forward-compatible, the counterfactual must satisfy the following criteria:
- The counterfactual must describe a development that could plausibly take place after the original news event.
- The counterfactual must not cancel out the original market development. This includes but is not limited to introducing mutually exclusive outcomes or implying that the original market development did not happen.
 	
Return your evaluation in the following JSON format:
{
"forward_compatibility": {
"value": <true or false>
}
}

Table 7: LLM-judge prompt for evaluating the *forward-compatibility* of counterfactuals.

Method(s)	Prompt

Claude 3.5 Haiku, Gemini 2.0 Flash, GPT-4o, LLaMA 4 Maverick, Qwen 2.5 72B, LM-Counterfactuals, Self-Training (SRLM)	You are a financial expert tasked with generating minimally edited counterfactuals based on a provided financial headline that describes a market development.
	Your goal is to generate a risk counterfactual and an opportunity counterfactual, following from our requirements below:
	
	Risk Counterfactual:
	- Minimally edit the original market development headline to represent a plausible alternate counterfactual that represents an adverse shift in the market development.
	- The adverse shift should reflect an adverse market outcome or deterioration in market conditions.
	- The alternate counterfactual must be forward-looking, which means that it can plausibly occur after the original market development.
	
	Opportunity Counterfactual:
	- Minimally edit the original market development headline to represent a plausible alternate counterfactual that represents a positive shift in the market development.
	- The positive market shift reflects a beneficial market outcome or improvement in market conditions.
	- The alternate counterfactual must be forward-looking, which means that it can plausibly occur after the original market development.
	
	Input format:
	- A single financial news headline.
	
	Your output must be valid JSON matching this structure. Do not make explicit numeric predictions or quantitative outcomes.
	{
	"Counterfactuals": [
	{
	"original_headline": "Article Headline",
	"opportunity_counterfactual": "Opportunity Counterfactual",
	"risk_counterfactual": "Risk Counterfactual"
	}
	]
	}
	
	{**Few-Shot Examples**}
	
	Input: {**Financial News Headline**}

Table 8: Prompt template used by each method for counterfactual generation – continued on the next page

Method(s)	Prompt

LLMs-for-CF	You are a financial expert tasked with generating minimally edited counterfactuals based on a provided financial headline that describes a market development.
	Your goal is to generate a risk counterfactual and an opportunity counterfactual, following from our requirements below:
	
	Risk Counterfactual:
	- Minimally edit the original market development headline to represent a plausible alternate counterfactual that represents an adverse shift in the market development.
	- The adverse shift should reflect an adverse market outcome or deterioration in market conditions.
	- The alternate counterfactual must be forward-looking, which means that it can plausibly occur after the original market development.
	
	Opportunity Counterfactual:
	- Minimally edit the original market development headline to represent a plausible alternate counterfactual that represents a positive shift in the market development.
	- The positive market shift reflects a beneficial market outcome or improvement in market conditions.
	- The alternate counterfactual must be forward-looking, which means that it can plausibly occur after the original market development.
	
	Please follow these reasoning steps before returning your output:
	1. Identify key phrases or words that signal the core market development in the original headline.
	2a. Change these key phrases or words to construct a forward-looking adverse market counterfactual with minimal changes.
	2b. Change these key phrases or words to construct a forward-looking positive market counterfactual with minimal changes.
	3a. Replace the key phrases and words from step 1 in the original text by the key phrases and words in step 2a, returning a risk counterfactual that reflects an adverse market shift.
	3b. Replace the key phrases and words from step 1 in the original text by the key phrases and words in step 2b, returning an opportunity counterfactual that reflects a positive market shift.
	
	Example Reasoning Chain:
	(Original headline)
	"Apple announces expansion of iPhone production in India"
	
	Step 1. Identify key phrases:
	- "announces expansion"
	- "iPhone production"
	- "in India"
	
	Step 2a. Edits for risk:
	- "announces expansion" → "faces delay in expansion"
	
	Step 2b. Edits for opportunity:
	- "announces expansion" → "accelerates expansion"
	
	Step 3a:
	"Apple faces delay in expansion of iPhone production in India"
	Step 3b:
	"Apple accelerates expansion of iPhone production in India"
	
	Final output format:
	{
	"original_headline": "Article Headline",
	"reasoning": "Reasoning Chain",
	"opportunity_counterfactual": "Opportunity Counterfactual",
	"risk_counterfactual": "Risk Counterfactual"
	}

Table 8: Prompt template used by each method for counterfactual generation – continued on the next page

Method(s)	Prompt

CounterfactualDistil	You are a financial expert tasked with generating minimally edited counterfactuals based on a provided financial headline that describes a market development.
	Your goal is to generate a risk counterfactual and an opportunity counterfactual, following from our requirements below:
	
	Risk Counterfactual:
	- Minimally edit the original market development headline to represent a plausible alternate counterfactual that represents an adverse shift in the market development.
	- The adverse shift should reflect an adverse market outcome or deterioration in market conditions.
	- The alternate counterfactual must be forward-looking.
	
	Opportunity Counterfactual:
	- Minimally edit the original market development headline to represent a plausible alternate counterfactual that represents a positive shift in the market development.
	- The positive market shift reflects a beneficial market outcome or improvement in market conditions.
	- The alternate counterfactual must be forward-looking.
	
	You will be given a masked version of the original headline, with placeholders like [MASK] in key slots (e.g., actors, sectors, verbs, or descriptors).
	These masked tokens are meant to be minimally edited while reflecting a directional shift (risk or opportunity).
	Please complete the [MASK] part of the headlines based on the specified opportunity and risk direction, to make it a counterfactual with smooth semantics and clear logic.
	
	Example:
	Input: "[MASK] markets set to [MASK] ahead of [MASK] and [MASK] decisions this week, [MASK] and [MASK] markets [MASK]"
	Output:
	{
	"original_masked_headline": "[MASK] markets set to [MASK] ahead of [MASK] and [MASK] decisions this week, [MASK] and [MASK] markets [MASK]",
	"opportunity_counterfactual": "Japan markets set to rally ahead of Fed and BOJ decisions this week, Australia and China markets reopen",
	"risk_counterfactual": "Japan markets set to plunge ahead of Fed and BOJ decisions this week, Australia and China markets remain closed"
	}

Table 8: Prompt template used by each method for counterfactual generation.

Few-shot examples
Input: Canada unveils multibillion-dollar plan to cut carbon emissions. Output: original_headline: Canada unveils multibillion-dollar plan to cut carbon emissions. risk_counterfactual_scenario: Canada's multibillion-dollar emissions reduction plan faces setbacks due to regulatory challenges. opportunity_counterfactual_scenario: Spurred by initial multibillion-dollar plan's success, Canada further invests in renewable energy to boost economic growth and reduce carbon emissions.
Input: Germany's bond yields set for biggest monthly jump in over a decade - Reuters. Output: original_headline: Germany's bond yields set for biggest monthly jump in over a decade - Reuters. risk_counterfactual_scenario: Germany's bond yields rise more rapidly than expected, raising concerns of potential economic downturn - Reuters. opportunity_counterfactual_scenario: Germany's bond yields stabilize after recent jump, signaling improved investor confidence - Reuters.
Input: Hartree Partners invests in nature-based voluntary carbon offset projects - Reuters. Output: original_headline: Hartree Partners invests in nature-based voluntary carbon offset projects - Reuters. risk_counterfactual_scenario: Hartree Partners' investment in nature-based voluntary carbon offset projects faces regulatory hurdles. opportunity_counterfactual_scenario: Hartree Partners' investment in nature-based voluntary carbon offset projects yields significant environmental returns.
Input: Nikkei rides high while traders wait on US inflation. Output: original_headline: Nikkei rides high while traders wait on US inflation. risk_counterfactual_scenario: Nikkei falters as US inflation data stokes market fears. opportunity_counterfactual_scenario: Nikkei soars further as US inflation shows signs of cooling.
Input: US government debt reaches new milestone. Output: original_headline: US government debt reaches new milestone. risk_counterfactual_scenario: US government debt milestone triggers investor concerns over economic stability. opportunity_counterfactual_scenario: US government debt milestone prompts strategic fiscal policy reforms.

Table 9: Few-shot examples used for counterfactual generation.

Prompt
You are a topic word extractor. Your task is to extract the most relevant topic word from the given text. Input: {Financial News Headline}

Table 10: Prompt for inferring topic words from the news headline, as part of the CounterfactDistil method.

Prompt
News: {Financial News Headline}. You are a financial expert tasked with reasoning about plausible future developments based on this headline. Generate a minimally edited, forward-looking continuation that remains coherent with the original market development.

Table 11: Prompt for generating the base continuation from the news headline, as part of the LM-Counterfactuals method.

Prompt
<task> Come up with one new financial news headline. Write only the financial news headline, with no further text or explanation. The examples below are enclosed in <example></example> tags. </task>

Table 12: Prompt for generating synthetic news headlines, as part of the SRLM method.

Prompt
Review the base news and the corresponding counterfactual scenarios using the additive 10-point scoring system described below. The original news is enclosed between <news> and </news>, and the generated scenario is enclosed between <response> and </response>. Points are accumulated based on the satisfaction of each criterion: - Add 1 point if the risk counterfactual is topically relevant and reflects a modification or extrapolation of the original news. - Add another point if the opportunity counterfactual is also relevant to the original news. - Add 1 point if the risk counterfactual makes minimal edits to the original news (preserving structure and intent). - Add another point if the opportunity counterfactual also uses minimal edits appropriately. - Add 1 point if the risk counterfactual is forward-compatible, describing a plausible future development that does not contradict the original news. - Add another point if the opportunity counterfactual is also forward-compatible. - Add 1 point if the risk counterfactual clearly reflects an adverse market development. - Add another point if the opportunity counterfactual clearly reflects a favorable market development. - Add 1 point if the risk counterfactual is cohesive, meaning it is clear, logical, and internally consistent. - Add another point if the opportunity counterfactual is cohesive in the same way. - If either counterfactual is incoherent, irrelevant, or fails to fulfill any criteria, award 0 points. <news>{news}</news> <response>{response}</response> After examining the news and the counterfactual scenario: - output the score of the evaluation using this exact format: "score: <total points>", where <total points> is between 0 and 10 - Briefly justify your total score, up to 100 words.

Table 13: Prompt for self-judging counterfactual responses to create a synthetic preference set, as part of the SRLM method.

Method	Perplexity ↓	Δ Perplexity ↓	Fwd-Compat. ↑	Dir. ↑	FwdCompat-Dir Avg. ↑
Claude 3.5 Haiku	448.56	+180.58	37.94%	**84.14%**	61.04%
Claude 3.5 Haiku FS	432.73	+164.75	65.90%	66.56%	66.23%
Gemini 2.0 Flash	434.38	+166.40	53.44%	66.52%	59.98%
Gemini 2.0 Flash FS	396.48	+128.51	67.07%	65.68%	66.38%
Llama4 Maverick	504.64	+236.67	32.31%	58.04%	45.18%
Llama4 Maverick FS	341.86	+73.88	65.72%	60.49%	63.10%
GPT-4o	389.76	+121.78	68.57%	52.70%	60.64%
GPT-4o FS	327.14	+59.16	**84.84%**	43.61%	64.22%
Qwen 2.5 72B	352.13	+84.15	63.08%	62.06%	62.57%
Qwen 2.5 72B FS	294.73	+26.75	74.78%	57.64%	66.21%
LLMs-for-CFs	514.41	+246.43	47.81%	47.59%	47.70%
CounterfactualDistil	527.83	+259.85	38.96%	21.71%	30.34%
LM-Counterfactuals	**193.58**	**-74.39**	57.46%	77.56%	**67.51%**
SRLM	253.92	-14.06	44.30%	78.87%	61.59%

Table 14: Evaluation results across all risk counterfactuals. Best results are bolded, second-best results are underlined. ↓ (lower score is better), ↑ (higher score is better). Fwd-Compat. is *forward-compatibility*, Dir. is *directionality*, FwdCompat-Dir Avg. is the average between *forward-comptability* and *directionality* scores. FS stands for few-shot, with results averaged over 5 random samplings for few-shot examples.

Method	Perplexity ↓	Δ Perplexity ↓	Fwd-Compat. ↑	Dir. ↑	FwdCompat-Dir Avg. ↑
Claude 3.5 Haiku	436.83	+168.86	72.88%	63.08%	67.98%
Claude 3.5 Haiku FS	434.93	+166.95	**90.24%**	28.51%	59.37%
Gemini 2.0 Flash	484.18	+216.20	53.58%	57.46%	55.52%
Gemini 2.0 Flash FS	467.82	+199.84	65.61%	50.99%	58.30%
Llama4 Maverick	519.38	+251.40	45.54%	43.64%	44.59%
Llama4 Maverick FS	308.85	+40.87	72.11%	36.88%	54.50%
GPT-4o	440.77	+172.79	67.11%	41.74%	54.43%
GPT-4o FS	326.92	+58.94	84.10%	31.91%	58.01%
Qwen 2.5 72B	366.67	+98.69	60.67%	51.97%	56.32%
Qwen 2.5 72B FS	309.94	+41.96	71.86%	44.89%	58.37%
LLMs-for-CFs	586.10	+318.13	61.70%	36.92%	49.31%
CounterfactualDistil	468.78	+200.81	43.49%	7.82%	25.66%
LM-Counterfactuals	**118.10**	**-149.88**	66.67%	60.31%	63.49%
SRLM	263.12	-4.86	68.20%	**68.71%**	**68.46%**

Table 15: Evaluation results across all opportunity counterfactuals. Best results are bolded, second-best results are underlined. ↓ (lower score is better), ↑ (higher score is better). Fwd-Compat. is *forward-compatibility*, Dir. is *directionality*, FwdCompat-Dir Avg. is the average between *forward-comptability* and *directionality* scores. FS stands for few-shot, with results averaged over 5 random samplings for few-shot examples.