FinER-ABSA: A Benchmark for Implicit and Explicit Entity Recognition and Aspect-Based Sentiment Analysis in Financial News

Pachara Akkanwanich[1,†], **Pavorn Thongyoo**[1], **Mahannop Thabua**[2,†], **Konlakorn Wongpatikaseree**[1], **Natthawut Kertkeidkachorn**[3]
[1]Mahidol University International College, [2]Mahidol University, Thailand,
[3]Japan Advanced Institute of Science and Technology, Japan
{pachara.akk, pavorn.tho, mahannop.tha}@student.mahidol.edu
konlakorn.won@mahidol.ac.th, natt@jaist.ac.jp

Abstract

Many approaches to English financial text analysis still rely on keyword or rule-based extraction, with limited trust in sentiment models despite advances in contextual understanding. Past studies have explored concepts such as aspect-based sentiment analysis and named entity recognition, yet none address how entities appear implicitly through context rather than direct mentions, or provide a dataset that brings these elements together. This gap limits how well models capture the links between entities, aspect, and sentiment. We introduce **FinER-ABSA**, a benchmark that integrates implicit and explicit entity recognition with aspect-based sentiment in financial text. Experiments on seven open-weight large language models under zero- and few-shot settings show that even the best systems still miss key aspects of implicit reasoning. In the few-shot case (K=3), *Llama-3.3-70B* reached an F1 of 0.7623 for implicit entities, suggesting that while models can detect signals, their consistency remains far from the level of reliability required for financial analysis or decision-making. These insights emerge only through **FinER-ABSA**, which makes such gaps measurable and advances financial Natural Language Processing (NLP) toward deeper contextual understanding and enables systems that better extract comprehensive insights from market-moving information in an industry where such precision is critical.

Keywords: Financial NLP, Implicit Entity Recognition, Aspect-Based Sentiment Analysis, Human-in-the-Loop Annotation, Evaluation Benchmark

1. Introduction

Language shapes how markets move, yet many systems that monitor financial or public opinion such as those employed by **Brandwatch**[1] and **Refinitiv**[2] continue to rely largely on extracting explicit mentions through keyword- or lexicon-based methods. In practice, these keyword rules must be continually updated to account for new product names, abbreviations, or misspellings. When companies are referred to indirectly through their products, executives, or industry roles, these systems frequently fail to detect them. The result is extensive manual work and a higher risk of error, which can prove costly in real-world financial analysis.

However, this reliance on surface forms is not inevitable. Advances in open-weight large language models such as Meta's **LLaMA** (Touvron et al., 2023) and Alibaba's **Qwen** (Team, 2023) now enable systems to interpret entities and sentiment from context rather than exact word matches, allowing them to reason over meaning rather than form.

A simple example illustrates this distinction:

- Shares of **Hormel Foods** were down about 7% as the company reported a total volume decline of 3.6%.
- Shares of **the Planters brand owner** were down about 7% as the company reported a total volume decline of 3.6%.

This example is drawn from real financial news. The second sentence is the original, while the first replaces the implicit phrase with its explicit company name to highlight the difference. Both sentences describe the same entity, yet only the first names it directly, while the second conveys it implicitly through context. Without accurate entity attribution, even the most advanced Aspect-Based Sentiment Analysis (ABSA) systems risk producing incomplete insights, which in financial applications could lead to costly misinterpretations or missed opportunities.

Although traditional Named Entity Recognition (NER) systems can easily identify explicit mentions, implicit entities remain a challenge that requires further attention. Current financial domain benchmarks such as **FinEntity** (Tang et al., 2023) and the

[†]Work conducted during an internship at Japan Advanced Institute of Science and Technology.

[1]https://www.brandwatch.com/blog/social-listening-guide/

[2]https://www.refinitiv.com/en/financial-data/news-analytics

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1904–1913
11–16 May 2026.

FiQA 2018 (Macedo Maia et al., 2018) challenge dataset still lack coverage of implicit references, as they mainly focus on explicit mentions. This gap limits the ability to evaluate systems that aim to interpret factual, event-driven sentiment within realistic market narratives.

To address this problem, we introduce **FinER-ABSA**, a dataset that integrates implicit and explicit entity recognition with aspect-based sentiment analysis in financial text. It contains 1,000 Reuters sentences annotated through a human-in-the-loop process that combines model-based prelabeling, human correction, and expert validation. The dataset includes indirect mentions and context-dependent references that remain difficult for rule-based systems. The benchmark results show that even advanced models continue to struggle with implicit reasoning, highlighting a persistent gap between linguistic understanding and financial reasoning in NLP.

Our contributions are as follows.

1. **A new dataset** combining *implicit and explicit* entity recognition with *aspect-based sentiment analysis* in financial text, closing a gap in the way the market language is represented.
2. **A human-in-the-loop annotation setup** with measured inter-annotator-agreement.
3. **Benchmarks on seven open-weight LLMs** that highlight where models still fall short in implicit reasoning over financial narratives.

The full dataset and experimental code are available at our github[3].

The rest of the paper is organized as follows. Section 2 reviews past NER and ABSA datasets in the financial domain. Section 3 describes the FinER-ABSA dataset and Section 4 explains our annotation framework. Section 5 presents our experiments and results. Section 6 concludes with key findings and future directions.

2. Related Work

Named Entity Recognition (NER) and Aspect-Based Sentiment Analysis (ABSA) form the foundation of financial language understanding. While early work in general domain NER and ABSA established methods for extracting entities, aspects, and sentiments from text (Lample et al., 2016) (Pontiki et al., 2014), financial applications require more precise reasoning over factual events and company relationships rather than subjective tone.

In the financial domain, several datasets have addressed parts of this problem, but remain incomplete when viewed together. **FinEntity** (Tang et al., 2023) adapts NER to finance by tagging *explicit company mentions* and labeling their sentiment polarity, yet it omits aspect information and ignores cases where firms are referenced implicitly. **SEntFiN 1.0** (Du et al., 2023) adds structured *entity–aspect–sentiment* triplets but still limits coverage to explicit entities. **FiQA 2018** (Macedo Maia et al., 2018) introduced an early framework linking entities, aspects, and sentiment intensity, but its pre-LLM-era design and focus on explicit opinions limit its relevance to modern, context-aware financial sentiment analysis. **ACOS** (Cai et al., 2021) explores joint extraction of aspect, category, opinion, and sentiment and considers *implicit aspects* but operates in the general domain rather than finance.

Across these corpora, *implicit entity* references are largely overlooked, despite the critical role financial news plays in shaping markets. Gold-standard datasets that combine NER and ABSA for this domain remain scarce (Kasai et al., 2019; Yaseen and Langer, 2021). Without such resources, it is difficult to train or benchmark models that capture how entities, events, and sentiment interact in real-world financial narratives.

In summary, prior financial datasets address only parts of the problem, leaving implicit reasoning between entities, aspects, and sentiment underexplored. Our motivation is to fill this gap by unifying *implicit and explicit* entity recognition with *aspect-based sentiment* analysis in financial news (see Table 1), enabling systematic evaluation of model performance. FinER-ABSA further serves as a benchmark for tasks requiring models to reason over how entities, aspects, and sentiments interact within factual, market-relevant narratives, rather than relying solely on lexical cues.

3. Dataset

3.1. Overview

This section introduces our dataset, explaining the conceptual basis and annotation criteria for explicit and implicit entities, aspects, and sentiment, and presenting statistics on its composition.

We selected Reuters as our source because it has long been a cornerstone resource in Natural Language Processing (NLP), underpinning several influential and publicly available corpora such as **Reuters-21578**[4], **Reuters Corpus Volume 1 (RCV1)** (Lewis et al., 2004), **Volume 2 (RCV2)**[5], and the **Thomson Reuters Text Research Collection (TRC2)**[5]. These corpora have been widely

[3] https://github.com/JAIST-KnOWLab/FinER-ABSA

[4] https://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html

[5] https://trec.nist.gov/data/reuters/reuters.html

Dataset	Entity		Aspect		Sentiment	Data Size	Source
	Explicit	Implicit	Coarse-Grained	Fine-Grained			
FinEntity	✓					979	Reuters Financial News
SEntFiN 1.0	✓				✓	10,753	The Economic Times Headlines
FiQA 2018	✓			✓	✓	1,174	Microblogs + News Headlines
ACOS			✓	✓	✓	6,362	Amazon + Yelp Reviews
FinER-ABSA	✓	✓	✓	✓	✓	1,000	Reuters Financial News

Table 1: Comparison of FinER-ABSA with cornerstone financial-domain datasets in terms of entity, aspect, sentiment, and dataset properties.

used in benchmark studies and shared tasks, including **CoNLL 2003** (Tjong Kim Sang and De Meulder, 2003) and, more recently, **FinEntity** (Tang et al., 2023), all within the domain of Named Entity Recognition (NER) and related NLP tagging tasks.

Building on this legacy, the **FinER-ABSA** dataset extends Reuters news coverage with new annotations that capture both implicit and explicit entity references, providing a resource for more comprehensive evaluation of entity and sentiment understanding in financial text.

3.2. Data Collection

We obtained a news corpus of 12,808 articles from Event Registry API[6], ranging from January 2014 to October 2024, filtered to include only Reuters Financial News. Sentences were extracted solely from the article body, omitting metadata such as article tags (e.g., `{Location}, {Date} (Reuters) –`) and Reuters-specific stock tickers (e.g., `2330.TW`), which are quirks unique to Reuters, removed to better resemble general financial news text and ensure consistency, as not all articles contain these elements.

Sentence extraction followed four primary guidelines:

1. The sentence must be verifiably sourced from Reuters, confirmed by accessing the original article and recording its URL.

2. The sentence must contain exactly one entity referring to a publicly listed company, verified through article context (typically the headline) and by documenting its stock ticker using external sources such as Yahoo Finance[7] or Investopedia[8].

3. If the entity was implicit, the surrounding context had to point clearly and exclusively to a single company.

4. The sentence must include exactly one aspect and one sentiment label.

[6] https://www.newsapi.ai/

[7] https://finance.yahoo.com

[8] https://www.investopedia.com

3.3. Entity Overview

Our definition of an entity differs slightly from conventional NER annotation schemes. Instead of the broad *Person–Organization–Location* categories, we focus solely on a subset of Organization that is objectively verifiable through a single, standardized identifier. An *Entity* is defined as any publicly traded firm with an associated stock ticker, following the conventions of financial data providers such as S&P Capital IQ[9] and LSEG Workspace for Students[10]. This scope allows us to extract 431 unique entities spanning multiple global exchanges, including those in the United States, United Kingdom, Canada, India, Australia, Japan, China, South Korea, and Thailand, as reported through Reuters Financial News. Each entity is annotated as the main company referred to in the sentence, regardless of delisting or merger status, while rebranded companies are labeled under their most recent name as of the annotation period.

3.3.1. Explicit Entity

Explicit Entities refer to companies that are directly mentioned in text, whether through their full names, abbreviations, or commonly used shorthand forms. For example:

> *Shares of Netflix slipped 0.6% during Disney's presentation.*
>
> *GM Chief Executive Mary Barra has promised investors the Detroit automaker will make money selling electric cars by 2021.*

Both sentences clearly identify the company at the center of the event. Even when multiple organizations appear in the same sentence, annotation focuses on the primary entity whose action or outcome drives the sentence's meaning. In the first example, although both *Netflix* and *Disney* are mentioned, the event (*slipped 0.6%*) pertains to *Netflix*, making it the focal entity. Likewise, company names expressed through abbreviations (e.g., *GM*) are annotated as explicit entities when unambiguously linked to a listed firm.

[9] https://www.capitaliq.spglobal.com/

[10] https://www.lseg.com/en/data-analytics/products/workspace/workspace-for-students

Sentence	Entity	Aspect	Sentiment	Comment
Although an enforcement action against a bank for anti-monetary laundering policies and procedures is unfortunately not that uncommon, it is unique for Wells	Wells Fargo & Co (WFC)	Regulations and Compliance	Negative	Explicitly states "Wells", which refers to Wells Fargo & Co.
The maker of Absolut vodka and Jameson whiskey in particular has felt the pressure as retailers and wholesalers in the United States cut back on pricier spirits stock.	Pernod Ricard SA (RI)	Industry Trends	Negative	Entity is inferred through product mentions ("Absolut vodka" and "Jameson whiskey"), which are owned by Pernod Ricard SA, an implicit reference.
In Britain, fully electric vehicles accounted for 7% of new cars sold in 2020, though their market share was temporarily boosted by collapsing sales of internal combustion engine (ICE) vehicles during the pandemic.	*None*	*None*	*None*	"ICE" could be a false positive for Intercontinental Exchange Inc., but no valid entity is present under our definition.

Table 2: Example sentences illustrating explicit, implicit, and no-entity cases in FinER-ABSA.

3.3.2. Implicit Entity

Following Perera et al. (Perera et al., 2015, 2016), *Implicit Entities* are entities that are referenced but not explicitly named in the sentence. We include only cases where the reference clearly and unambiguously points to a single company, verified through manual review of the full article. Ambiguous or speculative references were excluded.

Through manual inspection of financial news, we identify several recurring patterns of implicit references, including ownership or product associations (e.g., "the iPhone maker"), management-based mentions (e.g., "Buffett's conglomerate"), and market position cues (e.g., "the largest U.S. bank").

The resulting dataset reflects the natural frequency and coverage of real financial reporting. Companies such as *Apple Inc.* and *The Walt Disney Company* appear more often due to their higher event volume, resulting in a long-tail distribution of entity mentions that mirrors real market attention and challenges models to handle both frequent and low-resource entities effectively.

	Explicit	Implicit	Total
Entity Count	785	215	1000
Percentage	78.5%	21.5%	100%

Table 3: Entity Type Distribution

The final FinER-ABSA dataset consists of 785 sentences with explicit entities and 215 sentences with implicit entities, totaling 1,000 sentences. We also include five special test cases containing words that resemble entity names seen in the dataset but do not correspond to any real entity as defined in our annotation guidelines. These cases are designed to test whether a model can truly distinguish valid entities from lookalikes and correctly recognize when no extractable entity is present.

3.4. Curated Aspect

We began by manually annotating a sample of 100 sentences without a predefined aspect list to allow categories to emerge naturally from the data. Following this initial stage, the emerging categories were cross-referenced with the aspect schema of **FiQA 2018** (Macedo Maia et al., 2018) to ensure alignment with prior work. Relevant FiQA aspects were adopted where applicable, while new categories were introduced for concepts not covered in FiQA. Low-frequency or overlapping aspects were merged through iterative refinement, resulting in a balanced schema of 36 aspects that reflects the real-world distribution of financial topics. The final schema includes both coarse-grained and fine-grained categories without a fixed hierarchy. Examples of aspects and corresponding sentences are shown in Table 4.

3.5. Fact-based Sentiment

We adopt a fact-based interpretation of sentiment, where polarity is determined by the directional implication of an event associated with the aspect rather than by subjective opinion or emotional tone. In cases where the aspect explicitly reflects market perception or stakeholder attitude, emotional polarity may still apply. We also treat a high proportion of "Neutral" sentiment as uninformative in financial contexts. Even statements with balanced or

Coarse-Grained Category	Final Aspect	Rationale
Financial Performance	Revenue	Top-line indicator reflecting total company income before expenses.
Financial Performance	Profit	Bottom-line measure of net income after costs, taxes, and deductions.
Product Event	Product Launch	Confirmed rollout of a product or service, usually supported by concrete company action.
Product Event	Planned Product Deployment	Forward-looking plans, speculative mentions, or investor hype without execution.
Market & Perception	Stock Price Movement	Observed market reaction, typically a stock price change directly tied to company news.
Market & Perception	Investor Sentiment	General sentiment, not necessarily reflected in price action.

Table 4: Examples of aspect categories and their rationales in FinER-ABSA.

seemingly neutral language can often imply a directional outlook when viewed through the associated aspect.

> Su said last month the company expects to collect roughly $4.5 billion worth of AI chip revenue this year.

Traditional sentiment models might label this statement as neutral, yet in financial interpretation, the projection of $4.5 billion in revenue clearly conveys a positive outlook for AMD. Such cases highlight the need for aspect-aware, fact-based sentiment reasoning in financial text.

FinER-ABSA's sentiment distribution is shown in Table 5.

	Positive	Negative	Neutral	Total
Sentiment Count	504	317	179	1000
Percentage	50.4%	31.7%	17.9%	100%

Table 5: Sentiment Label Distribution of Entities.

4. Annotation Process

4.1. Overview

This section describes the annotation and verification process, the guidelines provided to the annotator and expert validator, and the analysis of agreement across annotation stages.

We employ a Human-in-the-Loop (HITL) framework with human validation. Similar approaches have been shown to improve annotation efficiency and quality (Klie et al., 2020, 2021; Weber and Plank, 2023). To avoid the pitfalls identified by Schroeder et al. (2025), we did not fully rely on LLM outputs; instead, they were used solely for pre-annotation followed by manual verification.

4.2. Annotation Guideline

The annotation process began with batch prompting **GPT-4 Turbo** using a structured, instruction-based zero-shot template to pre-label the dataset at scale. Specifically, the model was instructed to identify the primary entity (a publicly traded company with a stock ticker), assign an aspect category from a predefined list of 36 financial aspects, and determine the sentiment polarity (positive, negative, or neutral) based on the factual context of the sentence. The prompt included guidelines for handling both explicit and implicit entity references, instructing the model to use contextual inference when the company was not directly named. Responses were required to follow a fixed output format specifying the entity name, ticker, stock market, sentiment, and aspect category. Each sentence was processed independently in a batch prompting setup.

The initial annotations were manually reviewed and corrected by a coauthor pursuing a minor in finance. The annotator was provided with a finalized aspect list, each accompanied by a one-sentence definition, and used a Python-based interface that displayed one sentence at a time along with its pre-labeled entity and aspect. The annotator was instructed to revise model-generated labels critically rather than accept them directly.

To promote consistency and minimize subjectivity, the guideline included examples of explicit, implicit, and negative cases. All disagreements were manually reviewed to ensure they reflected genuine interpretation differences rather than formatting inconsistencies or model hallucinations.

4.3. Pre-annotation Agreement

To quantify the extent of human correction required, we compare the LLM pre-annotations against the final human-corrected labels. As shown in Table 6,

agreement reached 52.9% for aspects and 72.3% for sentiment, with Cohen's (Cohen, 1960) κ scores of 0.506 and 0.579, respectively. We note that because the annotator revised the LLM outputs rather than labeling independently, these figures reflect correction rates rather than independent inter-annotator agreement.

4.4. Annotator–Expert

To further ensure quality, the corrected annotations were validated by an independent expert, a senior undergraduate majoring in finance with no prior involvement in annotation or modeling. Since the expert annotated independently without access to prior labels, this comparison serves as our primary measure of inter-annotator agreement. Agreement between the annotator and expert reached 67.2% for aspects and 73.0% for sentiment, with Cohen's κ scores of 0.654 and 0.548, respectively, indicating moderate to substantial agreement according to the interpretation of Landis and Koch (Landis and Koch, 1977).

4.5. Fleiss' κ

Overall agreement across all three stages, measured using Fleiss' (Fleiss, 1971) κ, was 0.522 for aspects and 0.492 for sentiment (Table 6).

Label Type	LLM–A	Cohen's κ	A–E	Cohen's κ	Fleiss' κ
Aspect	52.9%	0.506	67.2%	0.654	0.522
Sentiment	72.3%	0.579	73.0%	0.548	0.492

Table 6: Agreement between LLM, Annotator (A), and Expert (E) stages in the Human-in-the-Loop process.

5. Experiments and Results

5.1. Task Design

In our experiment, we evaluated various LLM models by asking them three questions:

1. What is the entity in this sentence?
2. What is the aspect in this sentence?
3. What is the sentiment expressed in this sentence?

Each model is provided with predefined lists of entities, aspects, and sentiment categories appearing in the dataset. We benchmark the models on both the full dataset and the subset containing only implicit cases, and analyze their performance under zero-shot and few-shot settings. During evaluation, the models are prompted to extract all three variables in a single query. Performance is measured using the **macro-F1** score for each variable to capture balanced accuracy across classes.[11] To ensure consistency, we apply strict correctness criteria. Predictions that contain hallucinated information are marked as incorrect, and partial matches such as shortened company names or division names are also treated as errors. This decision reflects a practical requirement for financial applications, where models must produce standardized, conforming extractions across the dataset. Even small deviations in entity naming can lead to misinterpretation and result in decisions that cause substantial financial losses.

5.2. Implementation Details

We benchmarked the performance of seven large language models (LLMs) spanning two prominent open-weight model families, Llama and Qwen, across a range of model sizes to enable a comprehensive comparison. For the Llama family, we evaluated three variants: Llama-3.2-3B, Llama-3.1-8B, and Llama-3.3-70B, representing small, medium, and large-scale models, respectively. For the Qwen family, we included four models: Qwen2.5-1.5B, Qwen2.5-7B-1M, Qwen2.5-14B-1M, and Qwen2.5-32B.

All models were selected to represent diverse architectures, parameter scales, and training origins, allowing analysis of how model family and scale affect performance. In particular, we include both Llama and Qwen series to observe whether differences in pretraining data distribution and alignment practices across regions influence downstream financial reasoning performance.

5.3. Dataset Evaluation

Table 7 summarizes zero-shot and few-shot ($K=3$) results for all models on the 1,000-sentence FinER-ABSA dataset, using macro-averaged F1 scores for entity, aspect, sentiment, and a joint metric requiring all three to be correct.

Across both zero-shot and few-shot settings, entity extraction remains the strongest task overall. The highest entity F1 score of 0.8084 from Qwen2.5-32B in the zero-shot setting shows that large models can reliably detect company mentions when given a fixed list of entities. At first glance, this level of performance seems to correspond with the share of explicit entities in the dataset, but a closer look reveals the main sources of error. Most errors come from:

[11] For each variable (entity, aspect, sentiment), we treat the task as multi-class classification. Precision and recall are computed per class, averaged into a per-class F1, and then macro-averaged across all classes to weight each class equally regardless of frequency.

Model	Zero-shot				Few-shot (K=3)			
	Ent	Asp	Sent	Joint	Ent	Asp	Sent	Joint
Llama-3.2-3B	0.3703	0.1670	0.1812	0.0440	0.6989	0.2930	0.2643	0.1290
Llama-3.1-8B	0.5643	0.3560	0.3396	0.1210	0.7375	0.3960	0.3750	0.1830
Llama-3.3-70B	0.7550	**0.4545**	0.3971	**0.2941**	0.7623	0.4792	0.4177	**0.3218**
Qwen2.5-1.5B	0.3465	0.1530	0.1705	0.0230	0.4435	0.2540	0.2459	0.0560
Qwen2.5-7B-1M	0.6370	0.3705	0.4351	0.1310	0.7518	0.4103	0.4701	0.1890
Qwen2.5-14B-1M	0.7980	0.4487	**0.4875**	0.2160	**0.8405**	**0.4919**	**0.5206**	0.2300
Qwen2.5-32B	**0.8084**	0.4310	0.4036	0.2240	0.8389	0.4856	0.4429	0.2490

Table 7: Zero-shot vs few-shot (K=3) performance on the FinER-ABSA dataset. Scores are macro-averaged F1 for Entity, Aspect, and Sentiment; "Joint" counts a prediction as correct only if all three match.

- Hallucinations, where the model generates entities not mentioned in the text
- Near-entity extractions, such as predicting "TD Wealth" instead of "Toronto-Dominion Bank"
- Context-window leakage, where the model repeats extractions from earlier sentences instead of re-evaluating the current one.

Still, this high F1 score suggests that modern models can extract entities without relying purely on rule-based matching, reinforcing our motivation for an LLM-based extraction approach.

Aspect extraction continues to be the weakest part of the task. The best model, Qwen2.5-14B-1M in the few-shot setting, achieves a macro-F1 of around 0.49, while smaller models remain closer to 0.20. This gap suggests that aspect identification requires deeper reasoning about events or business factors that are not explicitly stated in the text. Larger models show moderate improvement, but even they struggle to select the correct aspect when the sentence requires understanding financial relationships such as between "revenue" and "profit." Smaller models also tend to under-extract rather than over-generate, often skipping uncertain cases instead of making random guesses. This pattern is reflected in their high precision but low recall scores.

Sentiment extraction sits in the middle. Most models achieve macro-F1 scores around 0.40 to 0.52, meaning they can usually identify whether a sentence is positive or negative but struggle with neutral tone or mixed expressions. The improvements from few-shot prompting are smaller here compared with entity or aspect extraction. This is likely because the sentiment design in our dataset is fact-based rather than emotion-based, which differs from the type of sentiment that many models are trained to recognize.

When the three tasks are combined, performance drops sharply. Even the best few-shot model, Llama-3.3-70B, reaches a joint F1 of only 0.3218, which is far lower than its individual scores. This drop occurs because each prediction must be correct simultaneously for entity, aspect, and sentiment, and errors from any one component quickly affect the others. For example, the model may correctly identify the company but miss the aspect, or find the right aspect but assign the wrong sentiment polarity. These results highlight how difficult it remains for current models to reason about relationships between entities, events, and opinions within a single sentence.

Larger models consistently outperform smaller ones in both zero-shot and few-shot settings. The improvement is especially clear when comparing 1.5-billion-parameter models with 32-billion-parameter ones. Bigger models not only recall more entities but also generate more coherent aspect and sentiment predictions. However, scaling alone does not solve the deeper reasoning problem. The models still fail to maintain internal consistency and often miss the connection between what event is being described and how it affects the company's sentiment.

Overall, few-shot prompting improves results across all dimensions. It helps the model better understand task structure and reduces random hallucinations. Yet, the gap in aspect and joint extraction remains large, suggesting that the main challenge lies in reasoning rather than exposure. Future work should explore connecting external financial knowledge or evaluating reasoning-oriented models on this dataset, resources that were not available to us during our benchmarking.

Comparing across model families, Qwen models tend to outperform similarly sized Llama models on entity and sentiment extraction, while Llama-3.3-70B leads on joint and implicit performance. One possible factor is the difference in pretraining data composition, such that Qwen models are trained on multilingual corpora with substantial Chinese-language financial text, which may strengthen general financial reasoning, while Llama models are more heavily weighted toward English. However,

we observe no consistent pattern that definitively indicates that training origin alone determines performance, as model scale remains the dominant factor across both families. Notably, this suggests that exposure to financial text in any language may transfer to English financial reasoning, rather than English-centric pretraining being strictly necessary.

Model	Zero-shot	Few-shot
Llama-3.2-3B	0.2066	0.4105
Llama-3.1-8B	0.5661	0.5207
Llama-3.3-70B	**0.6909**	**0.7550**
Qwen2.5-1.5B	0.0897	0.1027
Qwen2.5-7B-1M	0.3755	0.4216
Qwen2.5-14B-1M	0.5519	0.6249
Qwen2.5-32B	0.6729	0.6886

Table 8: Implicit entity F1 comparison for zero-shot vs few-shot on FinER-ABSA (macro-averaged F1).

5.4. Implicit Sample

To assess performance on implicit entity recognition, we evaluated a subset of sentences where the company was not explicitly named but implied through context (see Table 8).

The results clearly show that current models still struggle with implicit entities. Even the best-performing systems, such as Llama-3.3-70B and Qwen2.5-32B, achieve only 0.6729–0.7550 macro-F1—roughly ten points below their scores on the full dataset that includes explicit mentions. This gap means that even state-of-the-art models correctly infer the hidden company reference in only about three out of four sentences. While this may appear strong in isolation, such inconsistency can still be costly in real-world financial applications, helping explain why practical adoption of LLM-based extraction remains cautious.

At smaller scales, performance collapses completely: Qwen-1.5B reaches only 0.09–0.10, and Llama-3.2-3B performs slightly better at 0.20–0.41 depending on prompting. These models often fail to extract any valid entity when it is not explicitly named, confirming that simple memorization or token-level co-occurrence is insufficient for implicit reasoning.

Overall, these findings confirm that implicit entity recognition remains an unsolved problem. Even the strongest open-weight models fail to generalize beyond explicitly stated names, often struggling to infer the correct company from subtle contextual cues, even when sufficient evidence is available in the text.

6. Conclusion

FinER-ABSA introduces a benchmark that unifies implicit and explicit entity recognition with aspect-based sentiment analysis in financial news. Across seven large language models tested under zero-shot and few-shot settings, results show clear progress but persistent limitations in reasoning about implicit information.

Entity extraction remains the strongest component, reaching up to 0.8405 macro-F1 on the full dataset, while aspect extraction lags behind at around 0.4919, and sentiment at 0.5206. When all three tasks are combined, performance drops sharply, with the best joint F1 reaching only 0.3218 despite much higher individual scores. This indicates that models can identify entities and sentiments independently but still struggle to reason about how they interact.

On the implicit subset, even the strongest models such as Llama-3.3-70B and Qwen2.5-32B achieve only 0.6729–0.7550 macro-F1, roughly ten points below the highest performance on explicit cases. Smaller models collapse to below 0.10, showing that implicit reasoning remains far from solved.

These results confirm that scale and prompting improve recall but do not fully close the reasoning gap. FinER-ABSA thus provides a concrete framework for evaluating this limitation and a resource for future work on models that integrate structured reasoning and domain-specific financial knowledge. The benchmark highlights a core challenge for financial NLP, to move beyond pattern recognition toward genuine understanding of how entities, events, and sentiment connect in market narratives.

7. Future Work

We position FinER-ABSA as a starting point for a broader community effort to address these gaps. Future work could expand the dataset to include more sentences, additional event types, and refined aspect taxonomies, as well as explore the implicitification of explicit entity mentions as a strategy for generating synthetic training data. Another promising direction is the development of hybrid methods that combine LLM capabilities with retrieval-based or symbolic knowledge to better ground predictions in financial context. Additionally, agentic frameworks such as ReAct (Yao et al., 2023) that combine reasoning with tool use may be better suited for implicit entity resolution, as they allow models to iteratively retrieve and verify contextual information rather than relying on a single inference pass.

Potential applications extend beyond benchmarking. FinER-ABSA could support content-based financial recommendation systems, enhance so-

cial listening platforms that still rely on exact string matching to detect company mentions in news and social media, and enable more accurate extraction of actionable insights for financial decision-making.

Acknowledgements

We are grateful to Mahidol University International College for funding the research internship at JAIST and supporting conference participation. We also wish to thank Aye Aye Mar, Vidchaphol Sookplang, Chananchita Sirisinrungrueang, and Nuttapon Lee for their helpful discussions in dataset design and annotation guidelines.

Limitations

Due to the labor-intensive nature of extracting and annotating this dataset, driven by its complexity and our emphasis on quality, the total data volume is smaller than typical train-test datasets. However, its scale is comparable to FiQA 2018 Task 1, and FinEntity. We position this dataset more as a challenge set for evaluation and benchmarking rather than full-scale training. We encourage future work to build on this annotation template to expand coverage, and suggest combining this dataset with re-annotated versions of SentFin or other similarly scoped corpora. Additionally, each sentence in the dataset contains exactly one target entity, which simplifies the extraction task relative to real-world text where multiple entities may co-occur. However, our results show that models still struggle significantly even under this controlled setting, suggesting that multi-entity scenarios would pose an even greater challenge.

Ethics

All annotations were performed by the authors. One independent external expert participated in the validation process, in accordance with the ACL Ethics Policy. They were clearly informed about the task, contributed voluntarily, and were compensated in line with local research-assistant rates. No personally identifiable information (PII) was used or collected at any stage. Portions of this paper were refined using GPT-5 for stylistic improvement. All ideas, analyses, and final text were reviewed and verified by the authors.

This dataset and study are intended solely for research and business analytics applications, such as improving financial language understanding. They are not to be used for surveillance or any form of social governance.

8. Bibliographical References

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382.

Jungo Kasai, Kun Qian, Sairam Gurajada, Yunyao Li, and Lucian Popa. 2019. Low-resource deep entity resolution with transfer and active learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5851–5861, Florence, Italy. Association for Computational Linguistics.

Jan-Christoph Klie, Michael Bugert, Richard Eckart de Castilho, and Iryna Gurevych. 2020. From zero to hero: Human-in-the-loop entity linking in low resource domains. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6982–6993.

Jan-Christoph Klie, Michael Bugert, Richard Eckart de Castilho, and Iryna Gurevych. 2021. Human-in-the-loop annotation in low-resource domains for entity linking. In *Proceedings of the Workshop on Data Science with Human-in-the-Loop (Language Advances and Challenges, DaSH)*, pages 45–54.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270, San Diego, California. Association for Computational Linguistics.

J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174.

Sujan Perera, Pablo Mendes, Amit Sheth, Krishnaprasad Thirunarayan, Adarsh Alex, Christopher Heid, and Greg Mott. 2015. Implicit entity recognition in clinical documents. In *Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics*, pages 228–238, Denver, Colorado. Association for Computational Linguistics.

Sujan Perera, Pablo N. Mendes, Adarsh Alex, Amit P. Sheth, and Krishnaprasad Thirunarayan.

2016. Implicit entity linking in tweets. In *Proceedings of the Extended Semantic Web Conference (ESWC)*, pages 127–143. Springer.

Hope Schroeder, Deb Roy, and Jad Kabbara. 2025. Just put a human in the loop? investigating llm-assisted annotation for subjective tasks. In *Findings of the Association for Computational Linguistics: ACL 2025*, Vienna, Austria. Association for Computational Linguistics.

Qwen Team. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Hugo Touvron et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Noa Weber and Barbara Plank. 2023. Activeaed: A human in the loop improves annotation error detection. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8830–8843.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*.

Usama Yaseen and Stefan Langer. 2021. Data augmentation for low-resource named entity recognition using backtranslation. In *Proceedings of the 18th International Conference on Natural Language Processing (ICON)*, pages 352–358, National Institute of Technology Silchar, Silchar, India. NLP Association of India (NLPAI).

9. Language Resource References

Hongjie Cai, Rui Xia, and Jianfei Yu. 2021. Aspect-category-opinion-sentiment quadruple extraction with implicit aspects and opinions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 340–350, Online. Association for Computational Linguistics.

Zeyu Du, Yifeng Xing, and Erik Cambria. 2023. Sentfin 1.0: Targeted aspect-based financial sentiment dataset. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

David D. Lewis et al. 2004. Rcv1: A new benchmark collection for text categorization research. *Journal of Machine Learning Research*, 5:361–397.

S. Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, R. McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering (fiqa dataset). In *Companion Proceedings of The Web Conference 2018*.

Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland. Association for Computational Linguistics.

Yixuan Tang, Yi Yang, Allen Huang, Andy Tam, and Justin Tang. 2023. Finentity: Entity-level sentiment classification for financial texts. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 15465–15471, Singapore. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the conll-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147. Association for Computational Linguistics.