Language Modeling for the Future of Finance: A Survey into Metrics, Tasks, and Data Opportunities

Nikita Tatarinov[†], **Siddhant Sukhani**[†], **Agam Shah**, **Sudheer Chava**

Georgia Institute of Technology {ntatarinov3, ssukhani3, ashah482}@gatech.edu

Georgia Tech Financial Services Innovation Lab

Motivation

Finance Exposes Evaluation Gaps

- **High-stakes domain:** financial NLP systems are used in settings where errors affect risk, compliance, and decision-making.
- **Standard NLP metrics are not enough:** accuracy, F1, and MSE often fail to capture financial usefulness.
- **Real-world evaluation requires more:** robustness, interpretability, latency, reproducibility, and financial impact.

This raises the question:

What does financial NLP reveal about how language models should be evaluated beyond standard benchmarks?

What Existing Surveys Miss

Survey type	Systematic?	Quantitative?	Temporal?
Prior finance NLP	often no	limited	limited
LLM-in-finance	partial	partial	partial
Our survey	✔	✔	✔

We provide the first ***systematic*** survey of finance-domain research in NLP venues, covering 374 screened papers and 221 finance-focused papers.

Our Approach: Systematic Mapping of Financial NLP

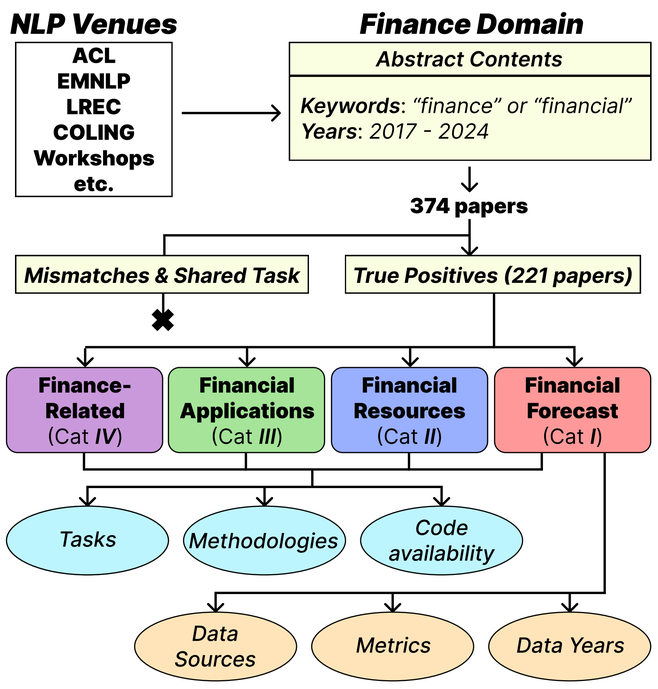


Figure 1: Overview of our paper selection process and analysis dimensions. We collected papers from a broad range of NLP venues using abstract-level keyword filtering, yielding 374 candidates. After removing mismatches and shared task papers, we retained 221 papers, categorized into four groups by their connection to financial tasks.

- **Sources:** 38 NLP conferences/workshops.
- **Scope:** papers from 2017–2024.
- **Analysis:** tasks, methods, metrics, datasets, code, authorship, funding, temporal coverage.

Main Evidence & Findings

What Does Financial NLP Study?

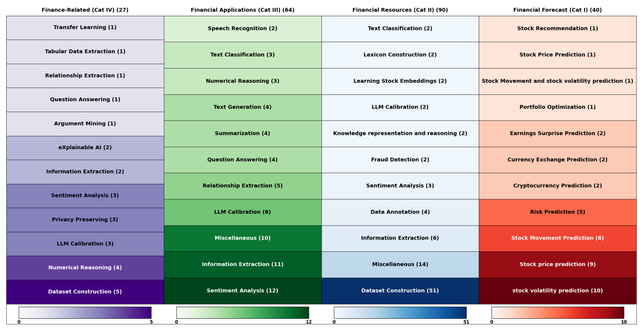


Figure 2: Distribution of primary tasks across categories. Each cell shows the task name and paper count (e.g., "Sentiment Analysis (3)"), with color gradients indicating frequency – darker shades represent more papers. "Miscellaneous" groups tasks that appear only once within Categories II and III.

- **Dominant tasks:** question answering, information extraction, sentiment analysis, and forecasting.
- **Forecasting is narrow:** mostly stock movement and volatility prediction.
- **Underexplored:** risk, macroeconomic forecasting, bankruptcy, cryptocurrency, privacy, and explainability.
- Many benchmarks do not capture domain usefulness.

Evaluation Gap

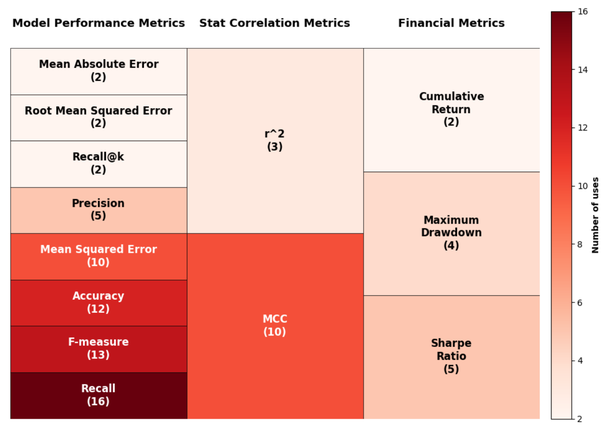


Figure 3: Distribution of evaluation metrics used in **Category I** papers. Most rely on ML-based metrics, while only a few financial metrics appear repeatedly.

- Most forecasting papers rely on ***ML metrics*** such as accuracy, F1, MSE.
- Few repeatedly use ***finance-specific metrics***.
- This weakens comparability with real-world goals.

Robustness Gap

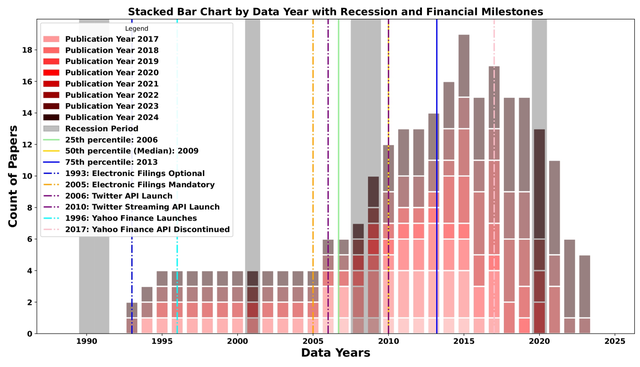


Figure 4: Data year distribution in financial forecasting papers, annotated with major financial events and infrastructure milestones. Highlights underuse of crisis periods despite their importance for model robustness.

- Stable periods dominate evaluation.
- Crisis periods are underused.
- **Robustness should be tested across market regimes.**

Broader Implications

Data Bias and Coverage

- **Language bias:** financial NLP is heavily skewed toward English-language data.
- **Market bias:** many datasets focus on U.S. filings, earnings calls, and financial news.
- **Modality gap:** truly multimodal financial datasets remain limited.
- **Accessibility issue:** paid/proprietary sources limit reproducibility. temporal coverage.

Do We Always Need Larger LMs?

- General-purpose PLMs/LLMs dominate, while custom and finance-specific models are less common.
- But finance needs constraints: ***interpretability, latency, compliance, and auditability***.
- Classical and hybrid methods remain useful for regulated or low-latency settings.

Main Recommendations

- **Move beyond generic metrics:** add finance-specific metrics such as Sharpe Ratio, Maximum Drawdown, and Cumulative Return.
- **Evaluate robustness under crisis periods:** include 2008–2009, COVID-era, and other volatile periods.
- **Document benchmark coverage:** report language, country, time period, data source, and update status.
- **Compare against practical baselines:** include interpretable, classical, and hybrid methods, not only LLMs.
- **Treat finance as measurement beyond capability:** evaluate reliability, reproducibility, deployment relevance, and real-world financial impact.

Finance is a concrete case study for a broader question: how do academic metrics translate to real-world impact? Financial NLP shows that language model evaluation should measure not only predictive performance, but also robustness, reproducibility, interpretability, and practical utility.

Scan to Read Our Paper

Paper

LinkedIn