MetaGraph: A Large-Scale Meta-Analysis of GenAI in Financial NLP (2022–2025)

Paolo Pedinotti*, **Peter Baumann, Nathan Jessurun**
Leslie Barrett, Enrico Santus*
Bloomberg
{pbaumann25, njessurun, lbarrett4, esantus}@bloomberg.net

Abstract

Financial NLP has evolved rapidly since late 2022, outpacing narrative surveys. We introduce MetaGraph, a methodology for extracting typed knowledge graphs from scientific corpora using ontology-guided LLM extraction to enable structured, large-scale trend analysis. Applied to 681 papers on GenAI in Finance (2022–2025), MetaGraph reveals three phases: early LLM-driven expansion of tasks and datasets, growing emphasis on limitations and risk, and a shift toward modular, system-oriented methods (e.g., retrieval-augmented designs). We release the resulting resource and artifacts to support reproducible meta-analysis and future monitoring of the field.

1 Introduction

The release of ChatGPT in late 2022 marked a structural shift in NLP, rapidly accelerating the adoption of generative AI (GenAI), particularly large language models (LLMs), in high-stakes domains such as finance. LLMs expanded the scope of Financial NLP beyond traditional supervised pipelines – long dominated by sentiment analysis and structured extraction – toward flexible, generative systems capable of zero-shot reasoning, long-document processing, and multimodal inputs.

This rapid evolution has outpaced traditional literature review methodologies. Existing surveys of Financial NLP either predate the widespread use of LLMs or rely on narrative summaries that struggle to capture quantitative trends, structural shifts, and emerging research patterns at scale. As a result, the field lacks a systematic, data-driven view of how tasks, datasets, models, and methods have evolved in response to GenAI.

To address this gap, we introduce **MetaGraph**, a generalizable methodology for extracting structured knowledge graphs from scientific literature

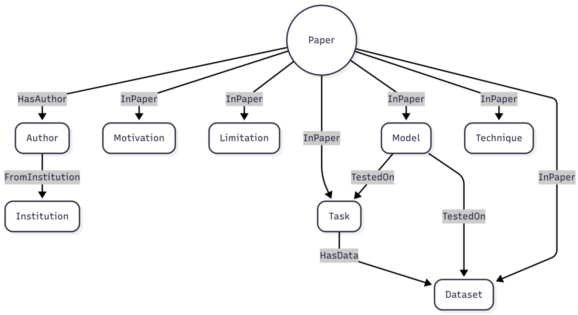

Figure 1: MetaGraph schema: node types and allowed relations (types appear as attributes in the instantiated graph).

using LLMs. MetaGraph combines a manually defined domain ontology with an LLM-based extraction pipeline to transform unstructured papers into a unified, queryable graph capturing research metadata, tasks, datasets, models, techniques, motivations, and limitations. By design, the ontology and prompts are modular, allowing the same framework to be applied beyond Financial NLP. Figure 1 shows the high-level structure of the resulting graph.

We apply MetaGraph to a corpus of 681 Financial NLP papers published between 2022 and 2025, producing a structured, queryable representation of the field. This analysis reveals a research landscape undergoing rapid transformation: the rise of financial question answering, the proliferation of datasets enabled by synthetic generation, increasing attention to model limitations and safety, and a gradual shift from model-centric approaches toward integrated systems combining LLMs with retrieval and other auxiliary components.

Our primary contribution is this large-scale, quantitative mapping of GenAI-driven transformation in Financial NLP. MetaGraph serves as the enabling framework that makes such structured, field-level meta-analysis possible. We do not introduce a novel knowledge graph learning algorithm; rather, our contribution lies in operationalizing ontology-

*Equal contribution. Paolo Pedinotti contributed to this work during his internship at Bloomberg.

Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM), pages 848–861
July 4, 2026 ©2026 Association for Computational Linguistics

guided LLM extraction at scale to produce a reproducible representation of an evolving research domain.

Our contributions are threefold: (i) **Field-level analysis**: We provide a data-driven map of Financial NLP's evolution, tracing systematic shifts in tasks, data sources, model choices, risks, and system design. (ii) **Methodology**: We present MetaGraph, a reusable pipeline for ontology-driven extraction and structured meta-analysis of scientific literature. (iii) **Open resource**: We release the resulting knowledge graph and artifacts[1] (provided on acceptance) to facilitate reproducible meta-analysis and ongoing monitoring of the field.[2]

2 Related Work

Financial NLP Surveys. While the survey by Xing et al. (2018) centered on traditional tasks such as classification, more recent surveys (Li et al., 2023b; Nie et al., 2024; Du et al., 2025) have focused on the transformative impact LLMs have had on financial applications. Yet, these works rely on a traditional narrative review methodology, qualitatively summarizing the applications in the literature.

Our approach diverges fundamentally by employing a bibliometric and holistic analysis of the field. We uncover structural shifts and data-driven trends by quantitatively mapping the research landscape into a knowledge graph, offering a comprehensive view of the impact of LLMs on its evolution. Table 1 summarizes how MetaGraph complements prior surveys along these dimensions.

LLM-Assisted Knowledge Graph Construction. Carta et al. (2023) construct domain-specific knowledge graphs through stepwise prompting strategies, while Funk et al. (2023) and Babaei Giglou et al. (2023) use LLMs to learn hierarchical relations among concepts. Our work takes inspiration from GraphRAG (Edge et al., 2025), which uses LLMs to extract a knowledge graph and enrich it with information at different levels of granularity.

3 Methodology

We introduce **MetaGraph**, a methodology for automatically constructing knowledge graphs from large scientific corpora to support quantitative analysis. By encoding research entities and their relationships in a structured representation, MetaGraph enables analyses that extend beyond frequency-based statistics and uncover relational patterns that are difficult to detect in unstructured text.

Survey	GenAI	Quantit.	Taxon.	KG
Xing et al. (2018)	✗	✗	✗	✗
Li et al. (2023)	✓	✗	✓	✗
Nie et al. (2024)	✓	✗	✓	✗
Du et al. (2025)	✓	✗	✓	✗
Ours (2025)	✓	✓	✓	✓

Table 1: Comparison with Financial NLP surveys.

MetaGraph is designed to be *generalizable* and *reusable*, although in this work we apply it specifically to Financial NLP papers.[3]

3.1 Method and Implementation

MetaGraph builds a typed knowledge graph by (i) defining an ontology; (ii) OCR-ing papers; (iii) extracting schema-constrained records with LLM abstention; (iv) normalizing and resolving entities; (v) inducing taxonomies and enriching metadata; and (vi) instantiating the graph and derived signals (Figure 1).

3.1.1 Ontology Definition.

This stage defines the graph structure that determines the analytical scope of MetaGraph. The expressiveness of the ontology directly governs which research patterns can be queried and analyzed.

Our ontology consists of three components: entity types, attributes, and relationships.

Figure 1 summarizes entity types and allowed relations; co-occurrence is inferred via shared `Paper` links. Graph nodes correspond to the core elements of research papers in the area of NLP, for example Tasks, Models, Methods and Motivations. Narrowing the focus, we investigate topics specifically in Financial NLP allowing a more limited inventory of datasets and tasks.

3.1.2 Corpus Acquisition

We compiled 681 Financial NLP papers (November 2022–April 2025) from ACL Anthology and arXiv:

- **ACL Anthology:** We queried titles and abstracts of conference papers from 2022–2025 using the official Anthology library, applying

[1] The released resource is available at https://zenodo.org/records/16968876.

[2] 58 of the 681 analyzed papers were excluded from redistribution due to CC-BY-NC-ND 4.0 licensing restrictions on arXiv, which prohibit redistribution of derivative content.

[3] Human validation was limited to structured prompt audits on sampled papers and consolidation of taxonomy labels. No large-scale manual correction of extracted entities was performed. The manual validation effort was modest relative to corpus size, and all extraction, normalization, entity resolution, and graph construction steps were fully automated and designed to scale to larger corpora.

Year	Count	Percent (%)
2022	126	18.50
2023	169	24.82
2024	273	40.09
2025	113	16.59

Table 2: Number of papers per year in our corpus.

an initial filter based on finance-related keywords[4] in either the title or abstract. Abstracts were manually screened for relevance to finance as the primary domain.

- **arXiv:** Similarly, we utilized the arXiv API to search 2023–2025 preprints in the Computation and Language (cs.CL) and the Quantitative Finance (q-fin) categories using the same set of finance keywords.

After aggregating and de-duplicating entries from both sources, manual validation resulted in a final corpus of **681 papers**. All papers were obtained in PDF format and processed via Mistral OCR[5] for text extraction. Table 2 reports the number of papers per year.

3.1.3 LLM-based Extraction

We used Gemini 2.5 Flash[6] to extract structured information from papers, selecting it for its strong cost-performance trade-off.[7] The extraction targets standard research entities, including *Tasks*, *Datasets*, *Models*, *Motivations*, *Techniques*, *Limitations*, and *Author* and *Institution* metadata.

We refined prompts via small-sample audits targeting omissions, over-generation, and schema violations; updates emphasized task definitions, abstention, and output constraints (Appendix B). Based on the observed patterns, we updated the prompt instructions (e.g., clarifying task definitions, strengthening abstention rules, and constraining output formats). Refinement stopped once the audit showed no recurring systematic errors. The full set of prompts is provided in Appendix B.

Multi-instance representation. Each `Paper` links to multiple `Task`/`Dataset`/`Model` nodes (many-to-many). Figure 2 illustrates this structure.

Validation. Validation on 12 gold papers found only minor omissions (two tasks, one model) and no hallucinations.

[4]The set of keywords is *financial*, *fintech*, *fraud*, *stock*, *portfolio*, and *finance*

[5]https://mistral.ai/news/mistral-ocr

[6]https://deepmind.google/models/gemini/flash/

[7]https://lmarena.ai/leaderboard

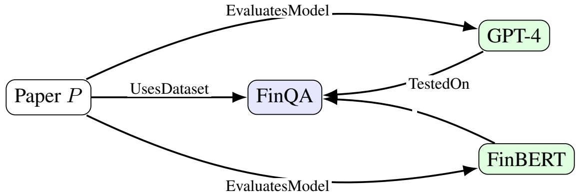

Figure 2: Example of multi-instance representation in the knowledge graph: a paper evaluates multiple models on a dataset (datasets in blue, models in green).

3.1.4 Entity Resolution

We resolved surface-form inconsistencies using normalization followed by embedding-based clustering. Mentions were first lowercased and stripped of punctuation and formatting artifacts. For each entity type independently (e.g., `Dataset`, `Model`), normalized mentions were embedded using OpenAI's `text-embedding-small`. Pairwise cosine similarity was computed, and mentions were merged if cosine similarity was ≥ 0.93, using greedy agglomeration.

The threshold $\tau = 0.93$ was chosen conservatively to avoid merging semantically related but distinct entities. For example, *Finqa* $\rightarrow$ *FinQA*, *FIQA-SA* variants $\rightarrow$ *FIQA-SA*, while *FinQA* and *ConvFinQA* remained distinct.

3.1.5 Taxonomy Induction

Selected entity types were organized into taxonomies using a zero-shot LLM-based categorization procedure. We processed batches of maximum 100 entities at a time. For each batch, the LLM was prompted to propose recurring category types and assign entities accordingly. We manually consolidated categories by merging near-duplicates, removing singletons, and standardizing naming. This final step affected only taxonomy naming accuracy; entity extraction remained fully automated.

Institution type labeling. Institutions were labeled as *academic*, *industry*, or *mixed* using rule-based heuristics applied to affiliation strings and Semantic Scholar metadata[8]. Keywords such as *University* and *College* triggered the *academic* label, while terms such as *Inc.*, *Ltd.*, etc., triggered *industry*. Papers with affiliations spanning categories were labeled *mixed*. The procedure was fully deterministic and did not rely on LLM inference.

3.1.6 Relevance Scoring

To highlight emerging trends, we define a paper-level *relevance score* used to prioritize analyses

[8]https://api.semanticscholar.org/

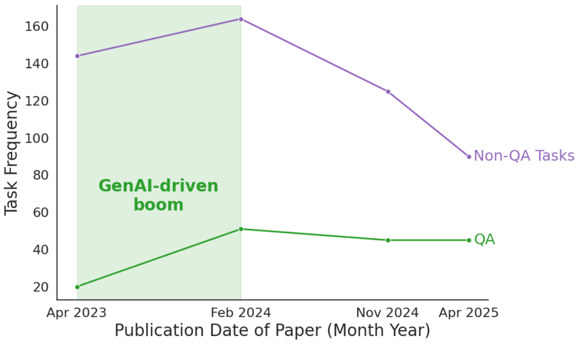

Figure 3: Increasing focus on financial QA. Task frequency here is the number of papers with an instance of the task category.

toward more influential work. The score combines three factors: (i) **institutional centrality**, computed as the PageRank of affiliated institutions in the co-authorship graph; (ii) **productivity**, measured as the number of papers published by the institution; and (iii) **citation normalization**, defined as paper citations normalized by the average citation count of the publication year.

4 Findings and Insights

In this section, we demonstrate the types of analyses enabled by MetaGraph, focusing on how the release of ChatGPT in late 2022 marked a turning point for Financial NLP. For comparability, we partition the corpus into three chronological subsets of approximately equal size: T1 (January 2022–August 2023), T2 (September 2023–July 2024), and T3 (August 2024–April 2025). All temporal analyses in this section use these fixed partitions. The main findings are:

- Task emphasis shifts toward QA variants (Figure 3; Table 3).
- The dataset landscape fragments and data sources diversify (Table 4; Table 5).
- Reported limitations shift from data scarcity toward model- and safety-related concerns (Figure 6; Table 6).
- Methods evolve from prompting-centric approaches toward system-level designs (Figure 7; Table 7).

More details and insights are provided in Appendices A, B, C.

4.1 Release of ChatGPT

Before the release of ChatGPT, the field focused mainly on sentiment analysis, information extraction, and stock prediction. In the period immediately following the release of ChatGPT (November 2022–February 2024), these tasks still constituted 90% of published work, and the most widely used datasets (Table 4a) reflected this focus.

As usage of LLMs matured (February 2024–April 2025, Table 4b), the landscape evolved unlocking new applications and attention toward more complex and reasoning-intensive tasks. **Financial QA** has become the leading focus, rising from 10% to 33% of tasks by 2025, while traditional tasks have steadily declined, as shown in Table 3 for representative tasks (temporal distribution of the full taxonomy in Appendix Table 11).

LLMs have transformed **the way** researchers approach financial problems. This shift moves from narrow, task-specific pipelines to flexible, generative systems that bridge previously isolated tasks. Between April 2023 and February 2024, the average number of tasks per paper rose from 1.36 to 1.9. Traditional tasks such as sentiment analysis and information extraction are now often used as intermediate steps in broader systems, such as RAG and financial agents.

Data Sources and Datasets. Datasets evolved as well. On the one hand, QA benchmarks now lead the field, overtaking traditional datasets (Table 4). On the other hand, we witnessed an expansion and diversification of the data sources used to generate QA datasets. Recent papers increasingly mention multimodal and structured inputs – such as tables, charts, audio, and analyst commentaries – alongside core sources such as news and company reports (Table 5). This expansion has been supported by synthetic data generation, which reduced the need for expert annotations. The share of synthetic or human-in-the-loop datasets nearly tripled as LLMs became data generators, from 5% in April 2023 to almost 15% by November 2024 (Krumdick et al., 2024; Guo and Yang, 2024; Liu et al., 2025; Li et al., 2023a). Data trends for tasks are plotted in Figure 8. Most new datasets target QA tasks, while the development of datasets for other tasks, such as sentiment analysis, has slowed.

4.2 A Growing Awareness

LLMs have lowered key barriers to both adoption and data processing. On the one hand, they remove data format constraints – enabling the processing of unstructured data. On the other hand, they support synthetic data generation, helping mitigate challenges such as cost, scarcity, and domain bias (Table 8). We show how limitations have changed over time in Figure 6. As data constraints

Task	T1	T2	T3
Financial Sentiment Analysis	15.67	15.77	9.78
NER	7.21	8.67	5.32
Stock Price Change Prediction	12.19	7.63	8.20
Retrieval Enhanced QA	5.47	7.80	12.81
Numerical QA	5.47	7.45	9.50
Long Document QA	3.23	4.51	7.34
Financial Consulting QA	2.24	4.16	6.04
Claim Verification	1.24	1.21	3.31

Table 3: Distribution (in % of papers) of selected Financial NLP tasks across time periods (T1: Jan 2022–Aug 2023, T2: Sep 2023–Jul 2024, T3: Aug 2024–Apr 2025).

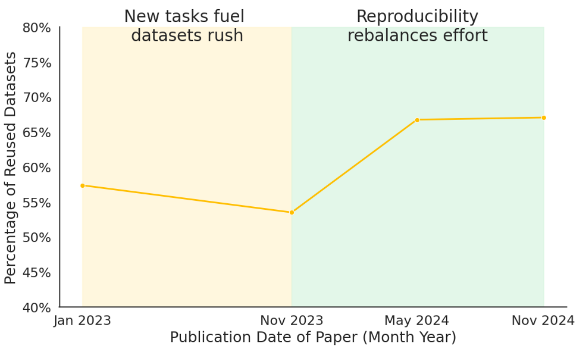

Figure 4: Temporal evolution of the proportion of datasets with references to prior literature.

eased, research attention increasingly shifted toward model-level challenges – particularly reasoning, interpretability, efficiency, and safety. We observed growing concerns around bias, privacy risks, and potential misuse (Table 6).

This shift toward critical reflection is evident in the evolution of research motivations, which increasingly convey a more cautious stance toward LLMs. By 2024, critical themes such as robustness, efficiency, reasoning, and RAG appeared in nearly 18% of papers – twice the share observed in early 2023 (see Table 9). This marks a shift from earlier studies, which primarily focused on leveraging LLMs through zero-shot learning and fine-tuning.

This marks a move from standalone *LLMs* to *system-oriented design*. Prompting strategies have evolved as well (Table 10): the progression from in-context learning to augmented methods such as chain-of-thought, retrieval-based prompts, and self-criticism reflects a move away from relying solely on the model's few-shot capabilities toward more deliberate prompt enrichment aimed at reducing errors.

4.3 From Models to Systems

In the wake of GenAI's rise, researchers initially focused on adapting general-purpose LLMs to Financial NLP tasks through prompt engineer-

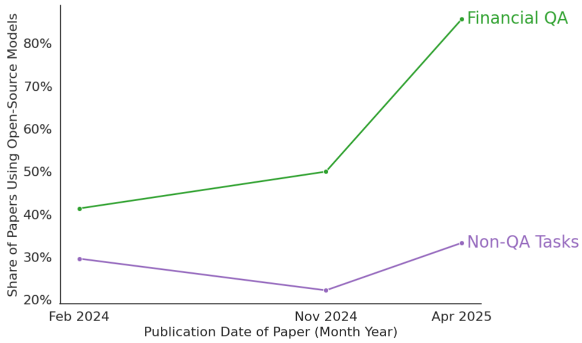

Figure 5: Share of papers using open-source models by task and timeframe.

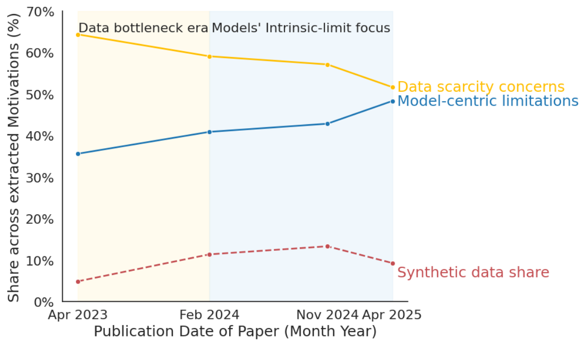

Figure 6: Reported limitations by period. Synthetic data share complements data scarcity concerns.

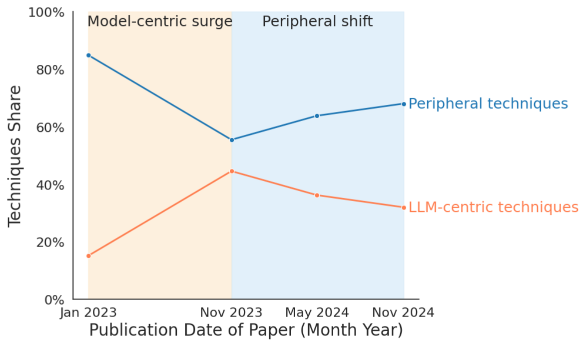

Figure 7: Technique evolution over time.

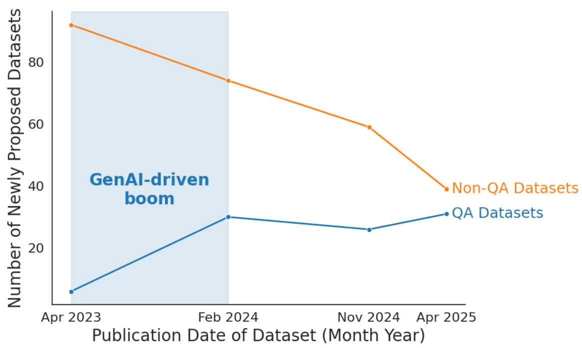

Figure 8: New datasets by period.

Dataset	Task	Freq.
FPB (Malo et al., 2014)	SA	29
FinQA (Chen et al., 2021)	QA	19
FIQA-SA (Maia et al., 2018)	SA	15
ConvFinQA (Chen et al., 2022)	QA	13
RefInd (Kaur et al., 2023)	RE	7

(a) November 2022 – February 2024

Dataset	Task	Freq.
ConvFinQA (Chen et al., 2022)	QA	13
FPB (Malo et al., 2014)	SA	13
FinQA (Chen et al., 2021)	QA	12
FIQA-SA (Maia et al., 2018)	SA	7
FinanceBench (Islam et al., 2023)	QA	7

(b) February 2024 – April 2025

Table 4: Top datasets in Financial NLP by usage across two time periods. (QA: Question Answering, SA: Sentiment Analysis, RE: Relation Extraction).

ing – especially zero-shot and in-context learning – which quickly gained momentum across applications. This was often complemented by post-training methods such as instruction tuning to further specialize models for the financial domain (Table 10).

Researchers began to move beyond model-centric approaches as the limitations of reasoning, safety, interpretability, and scalability became more apparent. Over time, these techniques were increasingly complemented by *system-level innovations* that integrate LLMs into broader frameworks (Figure 7).

The most prominent of these is RAG (Zhang et al., 2023; Xue et al., 2024; Li et al., 2024; Yepes et al., 2024; Chen et al., 2024; Zhao et al., 2024, among others), which has become a cornerstone of the field. Examining the evolution of RAG (Table 7), we find it mirrors the dataset trend: the spectrum of source types and data formats has widened, knowledge bases have grown, and the size of retrieved context has expanded from single sentences to large document chunks.

4.4 Towards Maturity

As the field matured, researchers began prioritizing shared resources over creating new **datasets**, increasingly relying on established, literature-backed benchmarks (Figure 4), with growing coverage across tasks. A similar trend emerged on the modeling side, as the community increasingly turned to open-source models valued for their transparency, controllability, and adaptability, alongside a shift from rapid expansion toward critical evaluation (Figure 5).

Figure 9 illustrates three key phases: the early dominance of GPT models, the emergence of LLaMA (Touvron et al., 2023), and the current diversification toward a mix of open models – such as Qwen (Bai et al., 2023) and DeepSeek (DeepSeek-AI et al., 2025) – and proprietary ones. Figure 10 shows how model sizes have also changed over time. The field is revisiting cost-performance trade-offs, driven by the financial and computational cost of large models. This shift is reflected in a recent inflection in model size trends.

Data Sources	**T1 (%)**	**T2 (%)**	**T3 (%)**
News	27.48	29.14	25.35
Social Media / Forums	21.85	14.20	14.43
Company Reports	28.15	27.37	28.99
Company Fundamentals & Indicators	11.04	15.09	15.13
Earnings Calls	5.18	7.10	6.58
Analyst Reports	4.73	3.25	3.92
University Textbooks	1.13	0.89	2.38
Financial Analyst Exams	0.45	2.96	3.22
Signals			
Text	80.42	73.72	70.44
Tables	16.08	19.87	20.13
Image	1.40	2.56	5.03
Audio	0.70	1.92	1.89
Other	2.10	1.92	2.52

Table 5: Distribution of data sources and signal types across time periods (T1: Jan 2022–Aug 2023, T2: Sep 2023–Jul 2024, T3: Aug 2024–Apr 2025).

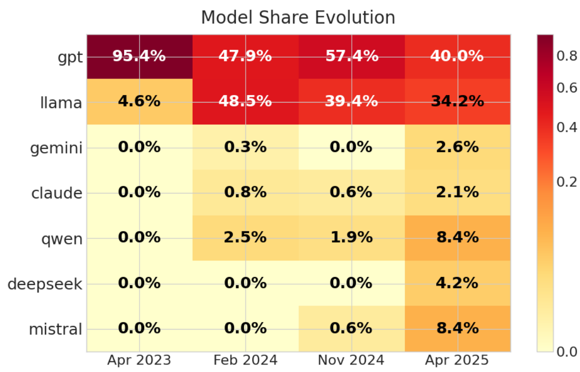

Figure 9: LLMs usage distribution over time.

One Revolution, Two Speeds. GenAI reshaped industry and academia at different paces (Figure 11). We took all the instances of tasks, models, and datasets in our corpus, and computed the relative proportion of financial QA instances, open-model instances, new datasets (datasets created after 2022), and created datasets (datasets created by the same authors who use them). Industry moved faster – dominating financial QA and driving dataset innovation to stay competitive. Academia responded more cautiously, focusing on

Risk / Limitation	T1	T2	T3
Dominant Issues			
Misleading Predictions due to Outdated Data	60	100	119
Misleading Predictions due to Inaccuracies	50	66	75
Lack of Transparency/Explainability	27	66	94
Inability to Generalize Across Financial Tasks	27	58	64
Growing Concerns			
Biases Towards Stocks/Trends/Products	12	30	43
Breaches of Sensitive Data	4	16	25
Susceptibility to Attacks and Misuse	10	12	25
Misinterpretation of Regulatory Text	1	7	21
Gender or Demographic Bias	6	21	17
Generation of Fraudulent Content	3	7	15
Stable Low-Level Issues			
Sensitivity to Data/Market Shifts	25	36	33
Inability to Detect Misinformation	5	7	8
Susceptibility to Corpus Poisoning	0	9	3

Table 6: Classification of LLM risks in finance based on their frequency over time periods (T1: Jan 22–Aug 23, T2: Sep 23–Jul 24, T3: Aug 24–Apr 25).

RAG Configuration	T1	T2	T3
Data Source Size			
Small	38.24	33.33	34.15
Medium	20.59	17.54	14.63
Large	41.18	49.12	51.22
Data Source Type			
Text	66.67	64.71	60.51
Table	18.33	21.57	20.38
Database	5.00	7.84	10.83
Graph	8.33	2.94	5.10
Image	1.67	2.94	2.55
Other	0.00	0.00	0.64
Retrieved Text Granularity			
Chunk	84.85	87.72	95.18
Sentence	15.15	12.28	4.82

Table 7: Distribution of RAG configurations across time periods.

Data	T1	T2	T3
Data-related Limitations			
Costly Human Judging	3.20	3.22	2.83
Insufficient Data Scale/Coverage	12.45	13.00	10.59
Skewed/Imbalanced Classes	3.08	2.29	1.47
Domain/Language Bias	12.03	11.69	9.59
LLM Limitations			
Interpretability Gaps	1.63	2.29	2.04
Weak Reasoning	4.11	4.59	5.24
Cost & Environmental Footprint	2.66	2.79	3.46
Hallucination & Bias	2.18	2.95	3.72
Prompt Sensitivity	1.75	2.84	2.78
Latency / Scalability	2.06	2.18	2.57
Synthetic Data / Label Issues	7.38	4.86	5.14
Capacity Constraints	9.07	9.07	9.70
Gaps: Lab vs. Live	9.43	8.68	10.27
Other (Appendix C.2)	28.81	29.55	30.06

Table 8: Distribution of reported limitations across time periods.

Data	T1	T2	T3
Data			
Data Scarcity & Annotation Cost	32.25	28.82	23.00
Other	37.12	35.81	35.48
Exploiting LLMs			
Zero/Few-Shot Evaluation	4.41	4.37	2.53
Domain-Specific LLM Training	10.44	8.95	11.89
Solving LLM Limitations			
Quantitative Reasoning Gaps	5.10	5.02	6.43
Interpretability & Explainability	3.25	3.71	5.07
Efficiency Constraints	3.02	5.24	5.65
Safety, Robustness, & Fairness	2.78	5.90	7.21
RAG & Retrieval Bottlenecks	1.62	2.18	2.73

Table 9: Distribution of future research directions across time periods.

Technique	T1	T2	T3
Core Prompting			
Zero-Shot	69	126	162
Few-Shot	20	98	74
Chain-of-Thought	13	61	94
Augmentation Strategies			
RAG (Retrieval-Augmented)	39	65	92
Decomposition	23	52	58
Self-Criticism	4	13	30
High-Level Methods			
Ensembling	17	14	13
Agents	9	25	40

Table 10: Frequency counts of prompt engineering techniques across time periods.

established tasks and open-source models, with a stronger emphasis on transparency and reproducibility. This is likely due to academia's structural constraints, which prioritize transparency, reproducibility, and the use of publicly available data and models – factors that inherently slow down the adoption of cutting-edge approaches. In contrast, industry has largely traded off transparency in favor of rapid experimentation, leveraging proprietary data and closed-source LLMs to push forward advanced use cases such as financial QA.

4.5 Looking Ahead

Financial NLP is entering a new phase, driven not only by LLMs but by a deeper understanding of their strengths and limitations. As techniques such as RAG and open-source fine-tuning become standard (grey line in Figure 12), multimodal models and small language models (green line) are gaining traction. New trends are also emerging (blue line in Figure 12), most notably multi-agent systems, which range from simple expert–critic setups to more complex architectures. At the same time, the gap between academic research and real-world financial practice remains an open question, as the field shifts its focus from question answering toward reasoning-oriented systems.

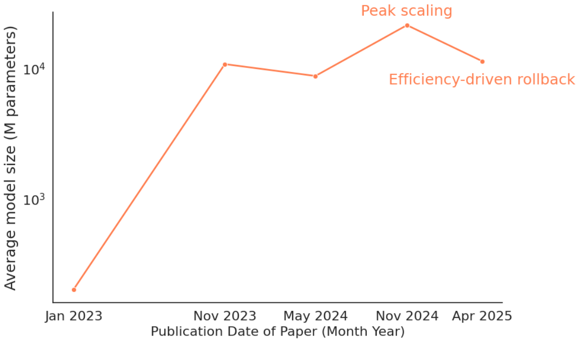

Figure 10: Sizes of open-source LLMs over time.

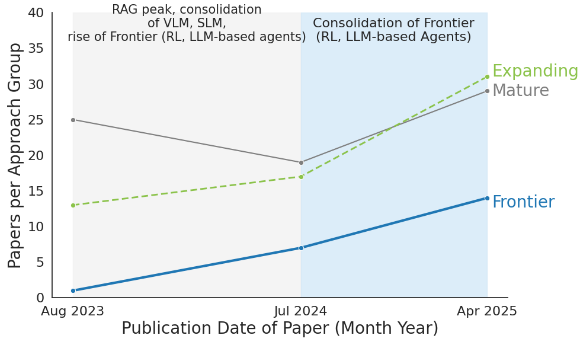

Figure 12: Latest trends in Financial NLP.

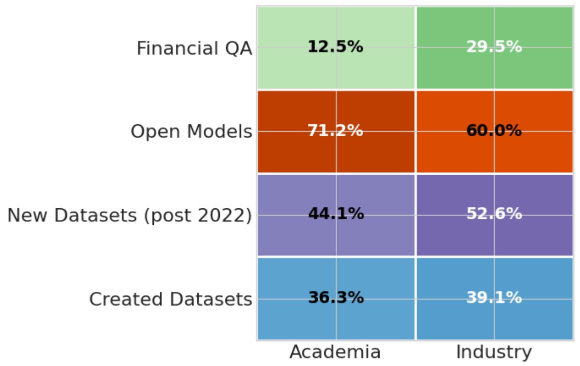

Figure 11: Academia vs. Industry: proportions of instances of tasks, models, and datasets.

5 Conclusion

We presented a structured, quantitative meta-analysis of GenAI-driven transformation in Financial NLP, based on a corpus of 681 papers from 2022–2025. Our analysis reveals systematic shifts in tasks, datasets, risks, and architectural paradigms, marking a transition from model-centric experimentation to system-oriented design.

MetaGraph serves as the enabling framework that operationalizes ontology-guided LLM extraction at scale, producing a reproducible and queryable representation of the field. Beyond Financial NLP, this approach illustrates how structured knowledge graph construction can support data-driven monitoring of rapidly evolving research domains.

6 Limitations

- Our approach relies on a *manually defined ontology*, which introduces an inductive bias in how entities and relations are categorized. While this provides structure and interpretability, it may also limit flexibility and overlook alternative or emergent conceptualizations.

- Despite continuous human validation and refinement, the *entity extraction and taxonomy induction processes remain based on LLMs*, which are inherently susceptible to hallucinations, inaccuracies, and bias. These limitations may affect both the precision and completeness of the extracted knowledge.

- The initial selection of papers was based on heuristic keyword search. This approach may have missed some papers that should be considered part of Financial NLP.

Acknowledgments

We thank David Rosenberg for his detailed feedback, which substantially improved this paper.

References

Hamed Babaei Giglou, Jennifer D'Souza, and Sören Auer. 2023. Llms4ol: Large language models for ontology learning. In *The Semantic Web – ISWC 2023*, pages 408–427, Cham. Springer Nature Switzerland.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *Preprint*, arXiv:2309.16609.

Salvatore Carta, Alessandro Giuliani, Leonardo Piano, Alessandro Sebastian Podda, Livio Pompianu, and Sandro Gabriele Tiddia. 2023. Iterative zero-shot llm prompting for knowledge graph construction. *Preprint*, arXiv:2307.01128.

Yuemin Chen, Feifan Wu, Jingwei Wang, Hao Qian, Ziqi Liu, Zhiqiang Zhang, Jun Zhou, and Meng Wang. 2024. Knowledge-augmented financial market analysis and report generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1207–1217, Miami, Florida, US. Association for Computational Linguistics.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. FinQA: A dataset of numerical reasoning over financial data. In *Proceedings* *of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

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

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitansky, Robert Osazuwa Ness, and Jonathan Larson. 2025. From local to global: A graph rag approach to query-focused summarization. *Preprint*, arXiv:2404.16130.

Maurice Funk, Simon Hosemann, Jean Christoph Jung, and Carsten Lutz. 2023. Towards ontology construction with language models. In *Joint proceedings of the 1st workshop on Knowledge Base Construction from Pre-Trained Language Models (KBC-LM) and the 2nd challenge on Language Models for Knowledge Base Construction (LM-KBC) co-located with the 22nd International Semantic Web Conference (ISWC 2023), Athens, Greece, November 6, 2023*, volume 3577 of *CEUR Workshop Proceedings*. CEUR-WS.org.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *ArXiv*, abs/2312.10997.

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

Pranab Islam, Anand Kannappan, Douwe Kiela, Rebecca Qian, Nino Scherrer, and Bertie Vidgen. 2023. Financebench: A new benchmark for financial question answering. *Preprint*, arXiv:2311.11944.

Simerjot Kaur, Charese Smiley, Akshat Gupta, Joy Sain, Dongsheng Wang, Suchetha Siddagangappa, Toyin Aguda, and Sameena Shah. 2023. Refind: Relation extraction financial dataset. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 3054–3063, New York, NY, USA. Association for Computing Machinery.

Michael Krumdick, Rik Koncel-Kedziorski, Viet Dac Lai, Varshini Reddy, Charles Lovering, and Chris Tanner. 2024. BizBench: A quantitative reasoning benchmark for business and finance. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8309–8332, Bangkok, Thailand. Association for Computational Linguistics.

Jiangtong Li, Yuxuan Bian, Guoxuan Wang, Yang Lei, Dawei Cheng, Zhijun Ding, and Changjun Jiang. 2023a. Cfgpt: Chinese financial assistant with large language model. *Preprint*, arXiv:2309.10654.

Xiang Li, Zhenyu Li, Chen Shi, Yong Xu, Qing Du, Mingkui Tan, and Jun Huang. 2024. AlphaFin: Benchmarking financial analysis with retrieval-augmented stock-chain framework. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 773–783, Torino, Italia. ELRA and ICCL.

Yinheng Li, Shaofei Wang, Han Ding, and Hang Chen. 2023b. Large language models in finance: A survey. *arXiv preprint arXiv:2311.10723*. Accepted at the 4th ACM International Conference on AI in Finance (ICAIF-23).

Jiaxin Liu, Yi Yang, and Kar Yan Tam. 2025. Evaluating and aligning human economic risk preferences in llms. *Preprint*, arXiv:2503.06646.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Pekka Malo, Ankur Sinha, Pekka Korhonen, Jyrki Wallenius, and Pyry Takala. 2014. Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4):782–796.

Yuqi Nie, Yaxuan Kong, Xiaowen Dong, John M. Mulvey, H. Vincent Poor, Qingsong Wen, and Stefan Zohren. 2024. A survey of large language models for financial applications: Progress, prospects and challenges. *Preprint*, arXiv:2406.11903.

Sander Schulhoff, Michael Ilie, Nishant Balepur, Konstantine Kahadze, Amanda Liu, Chenglei Si, Yinheng Li, Aayush Gupta, HyoJung Han, Sevien Schulhoff, Pranav Sandeep Dulepet, Saurav Vidyadhara, Dayeon Ki, Sweta Agrawal, Chau Pham, Gerson Kroiz, Feileen Li, Hudson Tao, Ashay Srivastava, and

12 others. 2025. The prompt report: A systematic survey of prompt engineering techniques. *Preprint*, arXiv:2406.06608.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *Preprint*, arXiv:2302.13971.

Frank Z. Xing, Erik Cambria, and Roy E. Welsch. 2018. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73.

Siqiao Xue, Fan Zhou, Yi Xu, Ming Jin, Qingsong Wen, Hongyan Hao, Qingyang Dai, Caigao Jiang, Hongyu Zhao, Shuo Xie, Jianshan He, James Zhang, and Hongyuan Mei. 2024. Weaverbird: Empowering financial decision-making with large language model, knowledge base, and search engine. *Preprint*, arXiv:2308.05361.

Antonio Jimeno Yepes, Yao You, Jan Milczek, Sebastian Laverde, and Renyu Li. 2024. Financial report chunking for effective retrieval augmented generation. *Preprint*, arXiv:2402.05131.

Boyu Zhang, Hongyang Yang, Tianyu Zhou, Muhammad Ali Babar, and Xiao-Yang Liu. 2023. Enhancing financial sentiment analysis via retrieval augmented large language models. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, ICAIF '23, page 349–356, New York, NY, USA. Association for Computing Machinery.

Yiyun Zhao, Prateek Singh, Hanoz Bhathena, Bernardo Ramos, Aviral Joshi, Swaroop Gadiyaram, and Saket Sharma. 2024. Optimizing LLM based retrieval augmented generation pipelines in the financial domain. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 6: Industry Track)*, pages 279–294, Mexico City, Mexico. Association for Computational Linguistics.

A The Geography of Financial NLP

The map depicted in Figure 13 illustrates the geographical distribution of institutions represented in our corpus. It highlights that Financial NLP research predominantly clusters around three major global hubs. In the United States, research activity is highly concentrated along the Atlantic Coast, with a distinct epicenter in New York City. In East Asia, significant research centers have emerged in major economic and technological hubs, notably within China, Korea, Japan, Hong Kong, and Singapore. Europe, on the other hand, presents a different pattern: research activities are more dispersed, reflecting a fragmented landscape with multiple smaller centers rather than a single dominant hub. This contrasting geography suggests regional differences in collaboration patterns, institutional density, and possibly cultural or economic factors influencing research organization in Financial NLP.

A.1 Financial NLP Graph Sub-trees

We show a snippet of the ontology emphasizing the connections between the papers on financial topics, the models used and datasets used. For readability, we do not include author names and institutions, paper motivations or limitations. Figure 14 shows the subtree.

B Prompts for Graph Extraction and Enrichment

This section details the prompts designed for extracting and enriching the graph structure from the corpus of Financial NLP research papers. The prompts were crafted to ensure clarity, conciseness, and consistency in extracting different entity types and relationships, facilitating accurate and systematic analysis.

B.1 Graph Extraction

For entities such as limitations, motivations, and techniques, a unique standardized prompt format was adopted to maintain consistency across extraction tasks. Given an entity type represented by X (e.g., limitations), the prompt follows the format shown in Listing 1.

A separate prompt, detailed in Listing 2, was developed to specifically extract tasks, datasets, and models. This distinct prompt was necessary to accurately capture relationships among these entities. Other types of relationships, such as entities co-occurring within the same paper, were implicitly inferred from extracted information and thus did not require additional prompting.

B.2 Taxonomy Extraction

To systematically derive taxonomies for specific entity types, we employed an iterative process, where multiple representative samples of entities were provided to the LLM using the prompt shown in Listing 3. This iterative approach allowed for the identification of recurrent patterns and commonly occurring entity types, ensuring robust and meaningful taxonomic categories.

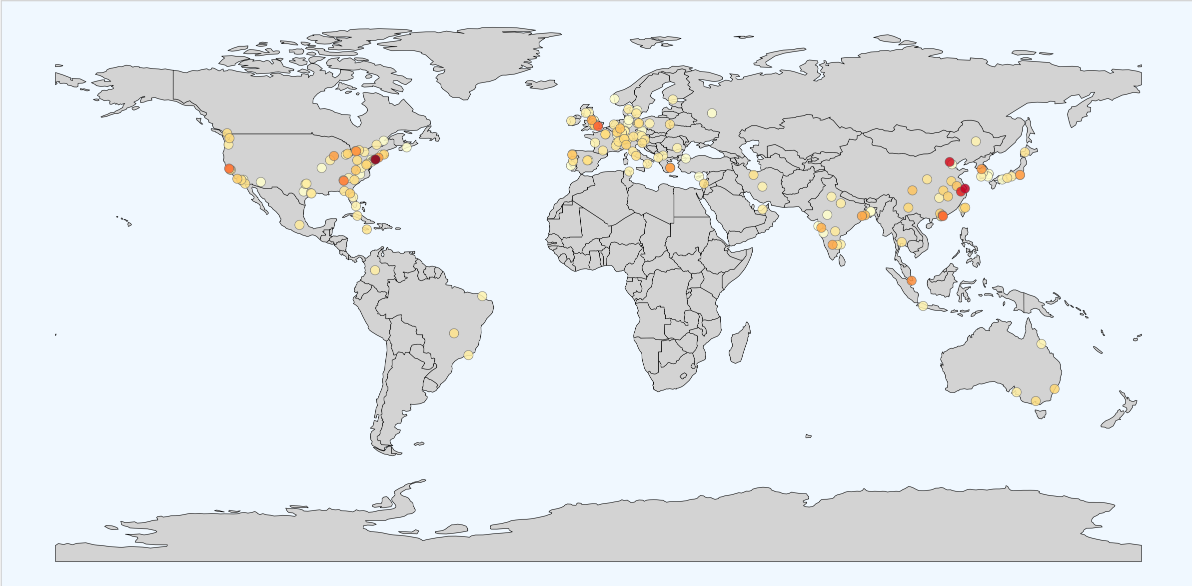

Figure 13: Geography of Financial NLP. The intensity of colors indicates the frequency of contributions involving an institution based in that place.

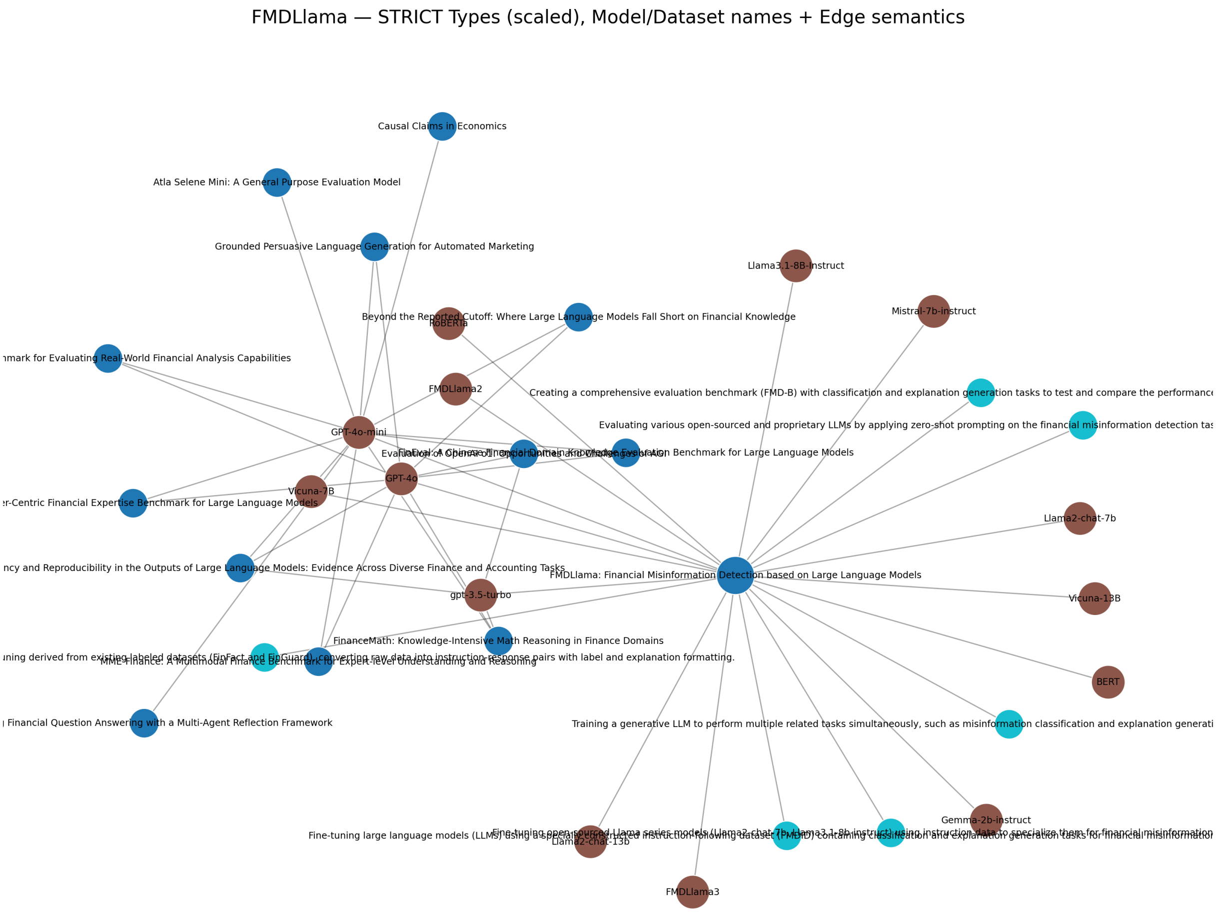

Figure 14: Subtree snippet of the Financial NLP graph structure with selected paper titles, models and dataset descriptions.

```
You will be given the full text of a paper that applies NLP techniques to financial data analysis. Your goal is to identify the {ENTITY TYPE} mentioned in the paper.
You should summarize each {ENTITY TYPE} in a brief and concise description (no more than 50) words.
```

Algorithm 1: Format of the prompt used for extraction of limitations, motivations, and techniques

```
You will be given the full text of an article that applies NLP to financial data. Your goal is to answer with the following information:

- General Reasoning: reason step-by-step to answer the following questions: which are the tasks the paper evaluates the models on? Which are the datasets used for each task?
Use the result of the previous reasoning to output:
- Tasks: a list of tasks the paper evaluates the models on. For each task, you have to specify:
    - Datasets: a list of datasets used to test the models on the task.
    For each dataset, you have to provide:
        - Dataset Reasoning: reasoning step-by-step to extract the following information
        - Dataset Name: the name of the dataset. If the dataset is referred generically, specify "generic"
        - Dataset Created: "yes" if the dataset is created by the authors, "no" otherwise
        - Reference name: name of the paper that proposes the dataset, if the reference paper is cited by the authors
        - Reference Year: Year of the paper that proposes the dataset, if the reference paper is cited by the authors where the dataset was extracted.
        - Signals: List of signals included in the dataset. Can be one or more from [Text, Tables, Image, Audio, Code, Time Series, Video, Charts, Equations]
        - Sources: List of sources from which the dataset is taken. Can be one or more from [News, Social Media/Forums, Company Reports, Company Fundamentals \& Indicators, Earnings Calls, Analyst Reports, University Textbooks, Financial Analyst Exams]
        - Annotation: The way the dataset was annotated. Can be one or more from [Manual, Synthetic, Spontaneous, Human-in-the-loop]
    - Models: a list of models that are evaluated on the task:
        - Models reasoning: reasoning step-by-step to extract the following information
        - Model Position: "main" if it is the model on which the main contribution of the authors is based, "comparison" if the model was used as a comparison.
        - Model Name: the name of the model
        - Reference Name: Name of the paper that proposes the model, if the reference paper is cited by the authors where the model was extracted.
        - Reference Year: Year of the paper that proposes the model, if the reference paper is cited by the authors where the model was extracted.
        - Size: Can be one of small (up to 8B parameters), medium (between 8B and 80B parameters), and large (more than 80B parameters).
        - Parameter Size: Number of parameters (in billions) if it is specified by the authors, None otherwise.
```

Algorithm 2: Format of the prompt used for extraction of tasks, datasets, models, and their relationships

```
You will be provided with a list of descriptions. Each description corresponds to a {ENTITY TYPE} extracted from a Financial NLP paper.
Your goal is to identify recurring patterns in the list of {ENTITY TYPE}. In particular, you should identify types of {ENTITY TYPE} that appear multiple times in the list. Do not try to find types of {ENTITY TYPE} that cover every single example-only identify those that recur with a certain frequency.
```

Algorithm 3: Format of the prompt used for taxonomy extraction

B.3 Entity Classification

To classify entities consistently across different categories, we utilized a universal prompt structure, as illustrated in Listing 4. This approach ensures uniformity and accuracy in the classification of entities based on predefined categories.

C Additional Material for Analysis

This section provides supplementary analyses and detailed data supporting the insights discussed in the main body of the paper. It encompasses comprehensive overviews of task trends, evolving methodologies, and changing attitudes within Financial NLP research, presented through tables and figures to offer clear visual representations of temporal developments.

C.1 Temporal Evolution of Financial NLP Tasks

Table 11 presents a detailed percentage distribution of Financial NLP tasks across three different timeframes: January 2022–August 2023 (T1), September 2023–July 2024 (T2), and August 2024–April 2025 (T3). This analysis is based on categories derived by prompting Gemini-2.5-Flash to output a taxonomy from samples of task descriptions extracted from the paper, and we used the same model to assign each task to one of the categories.

The data illustrates limited innovation in information extraction: the tasks remain the same since the first period, and the percentage of papers practicing them is decreasing (with the exception of claim verification, which is a recent novelty). We can observe that the sharpest decline concerns financial sentiment analysis, which was the most practiced task in the first period and has experienced a significant drop in the last, reaching a percentage that is, by itself, lower than retrieval-enhanced QA. Finally, we can see that the true core of innovation is concentrated around financial QA. Tasks that were already practiced in the first period have attracted increasing attention over time, and new, more complex forms of QA (such **as** financial consulting and multimodal QA) have become widespread in the latest period.

Task Category & Name	**T1(%)**	**T2(%)**	**T3(%)**
Information Extraction			
Intent Detection	1.99	1.73	1.73
Stance Detection	1.00	1.56	0.58
NER	7.21	8.67	5.32
Relation Extraction	7.21	4.16	4.03
Semantic Annotation	12.69	7.11	4.17
Claim Verification	1.24	1.21	3.31
Event Based Text Annotation	6.47	9.53	6.47
Event Extraction	3.23	2.08	0.86
Argument Mining	0.75	1.56	0.58
Financial Sentiment Analysis			
Financial Sentiment Analysis	15.67	15.77	9.78
Financial Emotion Analysis	0.50	0.87	0.43
Stock Market Prediction			
Stock Price Change Prediction	12.19	7.63	8.20
Stock Volume Prediction	0.25	0.35	0.58
Stock Return Prediction	4.73	3.47	2.59
QA & Specialized Tasks			
Retrieval Enhanced QA	5.47	7.80	12.81
Numerical QA	5.47	7.45	9.50
Conversational QA	2.24	3.12	3.60
Long Document QA	3.23	4.51	7.34
Causal QA	1.99	1.04	2.30
Tabular QA	2.24	3.99	4.46
Financial Consulting QA	2.24	4.16	6.04
Temporal Reasoning QA	0.50	0.52	0.86
Financial Terminology Explanation	0.75	1.21	2.73
Multimodal QA With Images	0.75	0.52	1.73

Table 11: Percentage distribution of Financial NLP tasks across different timeframes (T1: Jan 2022–Aug 2023, T2: Sep 2023–Jul 2024, T3: Aug 2024–Apr 2025).

C.2 Trends in LLM Risks and Limitations

Table 6 categorizes the evolving risks associated with large language models (LLMs) in financial applications, as identified in the reviewed corpus. Recall that these categories are obtained by prompting Gemini-2.5-Flash to output a taxonomy from samples of safety limitations extracted from the paper, and we used the same model to assign the limitations to one of the categories.

Notably, safety concerns have sharply increased from the first to the second period, and **are** still sharply rising today. However, we can distinguish groups with different trends: concerns that were already prevalent, such as inaccuracy and the inability to adapt to new data, have become even more pronounced. At the same time, limitations rarely highlighted by paper authors before have emerged. These include the impact that well-known limitations of LLMs, such as bias, can have in the financial sector, as well as privacy and misuse issues.

C.3 Evolution of Retrieval-Augmented Generation (RAG) Approaches

Table 7 summarizes the distribution of different retrieval-augmented generation (RAG) configurations over the studied timeframes. It should be noted that we are using the distinguishing factors for RAG approaches as proposed in the survey by Gao et al. (2023).

We can observe a clear trend that follows the ex-

```
You will be given a description of an entity extracted from a Financial NLP
    papers. The entity is of type {ENTITY TYPE}. Your goal is to classify the
    entity into one or more of the following categories:

{LIST OF POSSIBLE CATEGORIES}
```

Algorithm 4: Format of the prompt used for entity classification

```
You will receive a JSON array containing every *motivation paragraph* for ONE
research paper in NLP-for-Finance.

**Task (for EACH element, in the same order)**
1. Think step-by-step: does the paragraph primarily
    - show *enthusiasm / experimentation* with LLMs, **or**
    - propose *hybrid / alternative* solutions (i.e. highlight LLM limitations)?
    - If no stance towards LLMs is expressed, set Enthusiasm = "other"
2. Summarise your reasoning in 2-3 sentences.
```

Algorithm 5: Prompt used for detection of the stance towards LLMs from motivations

pansion of data sources within the sector: the documents from which retrieval occurs are becoming increasingly longer, and the landscape of modalities is growing more diverse. In particular, there has been an increased use of structured data sources such as tables and databases. As RAG tasks become more complex and refined, the model is asked to retrieve more and more information, which is then incorporated into the prompt. This testifies to the progressive transformation of information extraction tasks into retrieval tasks.

C.4 Shifts in Data Reuse Practices

Figure 4 illustrates the changing proportions of datasets explicitly referencing prior literature over time, relative to all datasets used in published studies. The trend suggests a shift in data reuse practices: following an initial phase characterized by widespread creation of new datasets, researchers increasingly began to adopt and build upon existing resources.

C.5 Prompt Engineering Techniques

Table 10 highlights the evolving landscape of prompt engineering techniques over three distinct timeframes. The categories are taken from (Schulhoff et al., 2025). We can see an explosion in the use of prompt engineering techniques in the period following the advent of LLMs. While zero-shot remains the most widespread and dominant technique, few-shot has seen a recent decline in usage, being increasingly replaced by newer techniques to improve output accuracy. Chain-of-thought prompting enhances the model's reasoning, improving performance on complex tasks and supporting interpretability. RAG adds contextual information to reduce inaccuracies and provide updated knowledge, while techniques such as self-criticism use another or the same LLM to check the output. These techniques share a common goal: they add complexity and sophistication to the prompting process to mitigate model shortcomings.

C.6 Stance Detection towards LLMs in Papers Motivations

In order to obtain a quantitative insight into how the stance towards LLMs has changed over time, we applied Gemini 2.5 Flash zero-shot on the papers' motivation, using the prompt in Listing 5.

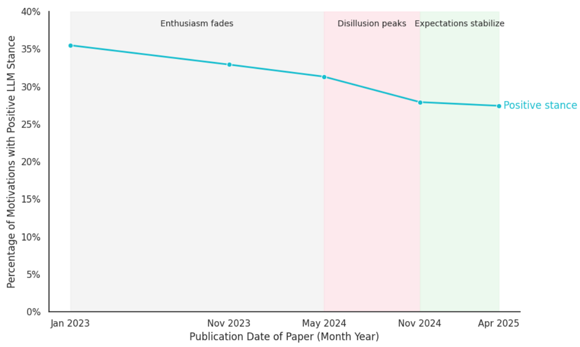

Figure 15: Attitude toward LLMs **has** evolved over time.

Figure 15 provides a quantitative depiction of evolving attitudes toward LLMs over time, as reflected in the motivations stated in their work. We observe that papers framing LLMs as a solution to the limitations of earlier models have steadily declined, giving way to a growing number of studies motivated by the goal of mitigating the limitations of the LLMs themselves.