Trading on News with AI

MGMT 675: Generative AI for Finance

Kerry Back

Why News Moves Markets

Asset prices reflect expectations. News changes expectations. The trader who **understands news faster and more accurately** captures the profit.

- Earnings surprises, Fed announcements, M&A rumors, FDA decisions, geopolitical shocks
- For decades, humans read the news and placed trades
- Quantitative firms began automating this with simple keyword rules in the 2000s
- **LLMs now understand nuance, context, and implication**—a qualitative leap

The Evolution of Text-Based Trading

Era	Method	Capability
Pre-2010	Keyword matching	Count positive/negative words
2011	Loughran-McDonald dictionary	Finance-specific word lists
2018–19	BERT / FinBERT	Contextual understanding of sentences
2023	GPT-3.5 / GPT-4	Zero-shot reasoning about market impact
2024–25	Fine-tuned SLMs + agents	Domain-optimized, multi-step analysis

Tetlock (2007) showed media content predicts stock returns. Loughran & McDonald (2011) showed generic sentiment dictionaries fail on financial text. LLMs resolve both problems.

The Models

FinBERT: The First Financial Sentiment Model

FinBERT (Araci, 2019) = BERT pre-trained on financial text and fine-tuned for 3-class sentiment (positive / negative / neutral).

Architecture

- 110M parameters (BERT-base)
- Pre-trained on financial news and 10-K filings
- Fine-tuned on Financial PhraseBank (4,840 labeled sentences)
- Open source: `ProsusAI/finbert` on HuggingFace

Performance

- ~87% accuracy on Financial PhraseBank
- 14 pp improvement over vanilla BERT
- Inference in ~5–10ms on GPU
- Became the standard baseline for all subsequent financial NLP

BloombergGPT and FinGPT

BloombergGPT (2023)

- 50B parameters; ~$10M training cost
- Trained on Bloomberg's FinPile (363B tokens) + 345B tokens of general text
- State-of-the-art on financial NLP benchmarks
- **Proprietary**—weights not released

FinGPT (2023)

- Open-source (AI4Finance Foundation)
- Fine-tunes Llama, Falcon, etc. with LoRA
- Training cost: **under $300**
- Sentiment analysis, trading signals
- GitHub: `AI4Finance-Foundation/FinGPT`

BloombergGPT proved the concept; FinGPT democratized it

GPT-4 and General-Purpose LLMs

General-purpose LLMs (GPT-4, Claude, Gemini) can classify financial sentiment **with zero training data**—just a well-crafted prompt.

Strengths

- Understands complex constructs: "despite strong revenue, guidance was weak" → **negative**
- Handles sarcasm, hedging, implicit sentiment
- Flexible: sentiment + event classification + summarization in one model

Weaknesses

- Latency: 500ms–5s per API call
- Cost: \$30/M input tokens (GPT-4 Turbo)
- Non-deterministic outputs
- Dependency on external API
- Regulatory concerns: data leaves your infrastructure

What the Research Shows

Can ChatGPT Forecast Stock Prices?

Lopez-Lira & Tang (2023). Accepted at the *Journal of Finance*.
https://arxiv.org/abs/2304.07619

- 67,586 headlines for 4,138 companies (Oct 2021–Dec 2022)
- GPT-4 achieves ˜90% hit rate on initial market reactions
- Long-short strategies based on GPT-4 sentiment:
  - Overnight news: **Sharpe ratio 2.97**
  - Intraday news: **Sharpe ratio 2.63**
- Stronger for small-cap stocks (less analyst coverage)
- Forecasting ability **increases with model size**—financial reasoning is an "emerging capability" of larger LLMs
- Returns **decline over time** as LLM adoption rises → markets becoming more efficient

LLMs vs. Traditional Sentiment Analysis

Kirtac & Germano (2024), *Finance Research Letters*. 965,375 U.S. financial news articles, Jan 2010–Jun 2023.

Method	Sentiment Accuracy	Long-Short Sharpe
Loughran-McDonald dictionary	50.1%	1.23
FinBERT	72.2%	2.07
BERT	72.5%	2.11
OPT (GPT-3 family)	**74.4%**	**3.05**

- OPT produced a **355% cumulative gain** from Aug 2021 to Jul 2023
- The Loughran-McDonald dictionary shows *no significant relationship* with subsequent returns
- Traditional bag-of-words methods are now effectively obsolete for this task

More Key Results

GPT-4 vs. Human Analysts

- Kim, Muhn & Nikolaev (2024), Chicago Booth
- GPT-4 given only anonymized financial statements
- **Outperforms the median human analyst** at predicting earnings direction

Fine-Tuned Small Models

- FinLlama (2024, Imperial College)
- Llama 2 7B fine-tuned with LoRA (4.2M trainable params)
- Outperforms FinBERT by **44.7%** in cumulative returns

Multi-Agent Systems

- MarketSenseAI 2.0 (2025): GPT-4 + RAG multi-agent $\rightarrow$ **125.9% cumulative returns** vs. 73.5% for S&P 100
- FinCon (NeurIPS 2024): manager-analyst LLM hierarchy outperforms deep-RL approaches

How It Works: Implementation

The News Trading Pipeline

- **Data ingestion**: Reuters, Bloomberg, SEC filings, social media
- **Entity resolution**: "Apple" $\rightarrow$ AAPL (not the fruit)
- **Sentiment extraction**: The LLM's core job

- **Signal construction**: Aggregate, weight by source reliability, normalize
- **Trade execution**: Route orders to exchanges
- **The LLM replaces steps 2 and 3**

What the LLM Extracts

Sentiment is more than positive/negative. A production system classifies along multiple dimensions:

Dimension	Question	Example
Polarity	Positive, negative, or neutral?	"Revenue beat estimates" $\rightarrow$ positive
Magnitude	How strong?	"Slight miss" vs. "catastrophic failure"
Relevance	Is this market-moving?	Routine board meeting $\rightarrow$ low
Novelty	Is this new information?	Rehash of known facts $\rightarrow$ low
Direction	Which asset is affected?	"AAPL supplier warns" $\rightarrow$ AAPL

LLMs handle all five dimensions in a single prompt; traditional methods handle only polarity

Small Models vs. Large Models in Production

Fine-Tuned Small Model

- FinBERT, FinLlama, custom BERT
- Inference: **5–10ms** on GPU
- Deterministic; runs on your infra
- Needs labeled training data

General-Purpose LLM

- GPT-4, Claude, Gemini
- Inference: **500ms–5s** via API
- Zero labeled data needed
- Handles nuance; non-deterministic

The Cascade Architecture (Best Practice)

1. **Fast path**: FinBERT processes all headlines in real time (˜5ms). High-confidence → immediate signals.
2. **Slow path**: Low-confidence items routed to GPT-4 (˜1–5s).
3. **Batch path**: End-of-day reprocessing by large model for portfolio review.

Latency and Alpha Decay

Alpha decay = the rate at which a trading signal's profitability diminishes as the market incorporates the information.

Strategy	Latency Budget	Model	Alpha Persists
HFT / market making	<1ms	Keyword lookup	Microseconds
Low-latency systematic	1–100ms	FinBERT	Seconds–minutes
Event-driven	100ms–5s	FinBERT + LLM	Minutes–hours
Daily systematic	Minutes–hours	Full LLM pipeline	Hours–days
Fundamental	Hours–days	Deep LLM analysis	Days–weeks

- You don't need to be the fastest—you need to be **fast enough for the alpha you're targeting**
- Ke, Kelly & Xiu (2020): news sentiment signals retain significant predictive power for 1–3 days

Event-Driven Strategies

Types of News Events

Scheduled Events

- **Earnings releases**: Compare actuals to consensus; analyze management tone on the call
- **Fed / central bank**: Single word changes carry enormous implications ("patient" vs. "data dependent")
- **Economic data**: CPI, jobs reports, PMI
- LLM advantage: pre-compute templates, fill in on release

Unscheduled Events

- **M&A announcements**: Assess deal likelihood and regulatory risk
- **FDA decisions**: Binary, high-impact
- **Geopolitical shocks**: Sanctions, trade policy, elections
- **Activist campaigns**: Detect early signals in 13D filings
- LLM advantage: parse complex, novel situations

Huang, Zang & Zheng (2014): the **tone of an earnings call** often matters more for stock returns than the reported numbers. LLMs excel at tone analysis.

Earnings Call Example

The Prompt

"You are a financial analyst. Read this earnings call excerpt for AAPL. Rate management's tone on a scale from -5 (very bearish) to $+5$ (very bullish). Explain your reasoning. Identify any forward-looking statements that differ from consensus expectations."

- The LLM can process the entire transcript (8,000–15,000 tokens)
- Detects hedging: "We're *pleased* with results but *expect headwinds* in Q4" $\rightarrow$ net negative
- Traditional keyword methods would flag "pleased" as positive
- The LLM also compares current tone to previous quarter's tone

This is where LLMs provide the greatest edge over traditional NLP

Industry Adoption

Who Is Doing This?

Hedge Funds

- **Bridgewater**: $2B AI fund (Jul 2024) using OpenAI, Anthropic, Perplexity
- **Two Sigma**: NLP for 10+ years; generative AI for 5+ years
- **Man AHL**: ML-driven strategies since 2014
- **Numerai**: Crowdsourced AI fund, $550M AUM, JPMorgan backing

Data Platforms

- **Bloomberg**: BloombergGPT; terminal-integrated NLP
- **RavenPack**: Structured sentiment signals for quant desks
- **Kensho**: Acquired by S&P Global for $550M
- **AlphaSense**: NLP search across filings, calls, and research

Permutable AI (London) launched an LLM-based trading platform live in Oct 2024: **20.6% return, Sharpe ratio 2.85**, with negative correlation to S&P 500 during tariff-driven volatility.

Cautionary Tales

AP Twitter Hack (April 2013)

- Hackers posted fake tweet: "Explosions at the White House"
- NLP-driven algorithms triggered massive sell-off
- S&P 500 lost **$136 billion** in seconds
- Markets recovered within 6 minutes
- Demonstrated fragility of automated news trading

Adversarial Attacks (2025)

- Research shows imperceptible changes in headlines can trick LLM trading systems
- Fake-news fabrication caused severe losses in simulations
- Alpha Arena crypto competition: ChatGPT suffered a **63% loss**
- IMF (Oct 2024) warned AI-driven trading increases volatility

The competitive edge is real, but so are the risks

Beyond Company Sentiment: The Macro Perspective

The Propagation Problem

Most LLM trading research focuses on **company-level** sentiment: "AAPL beats estimates" → buy AAPL. But the real competitive advantage lies in understanding how **macro events propagate across sectors**.

- "Russia invades Ukraine" — which sectors benefit? Which suffer?
- The answer requires multi-hop causal reasoning:
  - **1st order**: Oil prices rise → energy stocks up
  - **2nd order**: Fertilizer costs rise → agriculture costs up
  - **3rd order**: Food prices rise → consumer staples margins squeezed
- This kind of reasoning is where sustainable alpha lives — it is hard to replicate and slow to be arbitraged away

Measuring Geopolitical Risk with NLP

Caldara-Iacoviello GPR Index

- Published in *AER* (2022)
- Counts articles in 10 newspapers across 8 categories (war threats, terror, military buildups, . . .)
- Decomposed into Threats vs. Acts
- Available for 44 countries since 1900
- Rising GPR $\rightarrow$ lower investment, higher uncertainty

BlackRock BGRI

- Neural network NLP on Refinitiv brokerage reports + Dow Jones news
- Two components: **market attention** (NLP score) + **market movement** (asset price pattern)
- For each risk, identifies the **3 most sensitive assets**
- Expert-built "Market-Driven Scenarios" map each risk to sector exposures
- The industry gold standard

Which Sectors Are Most Exposed?

Culver, Niepmann & Sheng (Fed, 2025): NLP on **240,000+ earnings call transcripts** from ˜7,000 US firms to build industry-specific geopolitical risk sentiment indices.

Most Exposed (Negative)

- Finance
- Mining (incl. oil extraction)
- Manufacturing
- Fabricated products, electronic equipment

Least Exposed / Benefiting

- Agriculture
- Pharmaceuticals
- Defense / aerospace
- Insulated from or benefit from geopolitical turmoil

A 1-SD increase in GPR $\rightarrow$ **1.6% decline in investment** for firms in the top quartile of risk exposure

Case Study: Russia-Ukraine War (Feb 2022)

- **Defense**: European defense stocks saw abnormal returns of **up to +12%** within days
- **Energy**: US energy outperformed; European energy initially suffered, then surged
- **Commodities**: Oil, wheat, palladium spiked (Russia/Ukraine = major exporters)
- **Geographic**: Closer countries suffered more; trade linkages explained **2/3 of differential returns** (Federle et al., *JMCB*, 2024)

Amundi / Causality Link (2022)

- NLP processed 50,000+ texts/day in 27 languages with real-time causal extraction
- Ontology covers 3,000+ GICS-aligned industry segments
- Most impacted: automobiles, electric utilities, energy, aerospace/defense
- Before Mar 2022: semiconductor shortage was the main driver; after: war and sanctions

Knowledge Graphs: Mapping How Shocks Propagate

A **knowledge graph** maps entities (firms, sectors, commodities) and their relationships (supplies, competes with, depends on). Combined with an LLM, it enables multi-hop reasoning about shock propagation.

FinDKG (Li et al., ICAIF 2024)

- Fine-tuned LLM builds a **dynamic** knowledge graph from financial news
- Tracks causal effects between markets, persons, and events
- Outperforms thematic ETFs at identifying sector themes
- Open source: `github.com/xiaohui-victor-li/FinDKG`

Supply Chain Mapping

- InterCorpRel-LLM (2025): GNN + LLM maps 3,211 firms and 11,635 supply links
- AlMahri et al. (2024): zero-shot LLM extracts multi-tier supplier networks from public sources
- Hilt & Schwenkler (2024): extracts 4 firm-network types from 40+ years of NYT articles

Building a Proprietary Transmission Map

The sustainable competitive advantage: a proprietary model that maps how different event types affect different sectors, validated against historical shocks.

1. **Knowledge graph backbone** — extract firm/sector relationships from news, filings, and supply chain data
2. **Causal extraction** — LLM identifies cause-effect statements across industries (Causality Link approach: 3,000+ industry segments)
3. **Economic structure** — integrate input-output tables (Acemoglu production networks, BEA I-O data) to trace propagation paths
4. **Historical validation** — backtest against Russia-Ukraine, COVID, US-China trade war, Brexit
5. **Real-time updating** — process news flows continuously to update edge weights

The key gap: no public system yet integrates real-time LLM news analysis with production network models

The Frontier and Its Limitations

What Works Today

- GPR indices (keyword + NLP)
- BlackRock BGRI (NLP + expert scenarios)
- Industry-level sentiment from earnings calls (Fed study)
- Dynamic knowledge graphs (FinDKG)
- Supply chain mapping with LLMs

What Doesn't Work Yet

- LLMs still **struggle with multi-hop economic reasoning** (EconNLI, 2024)
- 2nd/3rd order effects are unreliable without structured knowledge
- Shock magnitude estimation (not just direction)
- Nonlinear amplification in large shocks (ECB WP, 2024)

Combine LLM text understanding with **structured economic models**. LLMs identify *what* happened; production networks and knowledge graphs predict *where* it propagates.

The Big Picture

The Crowding Problem

As more firms adopt the same LLM-based analysis, the alpha from news sentiment decays faster.

- Lopez-Lira & Tang document declining returns as LLM adoption rises—consistent with the Efficient Market Hypothesis
- Simple positive/negative sentiment from headlines has seen significant alpha erosion since 2015
- More sophisticated signals (tone analysis, reading between the lines, cross-referencing sources) retain more alpha
- The arms race: better NLP $\rightarrow$ alpha captured $\rightarrow$ faster decay $\rightarrow$ need even better NLP

The moat is shifting from **speed of access** (everyone sees the same news) to **depth of understanding** (interpreting nuance, context, and implication). This is where LLMs provide genuine differentiation.

Summary

What We Know

- LLMs dramatically outperform dictionary-based methods (50% $\rightarrow$ 74% accuracy)
- GPT-4 can match or beat human financial analysts
- Fine-tuned small models (FinLlama) compete with large LLMs at a fraction of the cost
- Multi-agent systems are the current frontier

What to Watch

- Alpha decay as adoption increases
- Adversarial attacks and market manipulation
- Regulatory response (SEC, ESMA)
- "Monoculture" risk: correlated AI-driven trades causing systemic instability
- Hybrid cascade architectures becoming standard

The value of LLMs in trading is not just speed—it is depth of understanding

References

Key References

More References

Macro & Cross-Sector References

Macro & Cross-Sector References (cont.)

Resources for Further Exploration

Open-Source Code

- FinGPT: github.com/AI4Finance-Foundation/FinGPT
- FinBERT: github.com/ProsusAI/finBERT
- TradingAgents: github.com/TauricResearch/TradingAgents

Books & Data

- Jansen, *ML for Algorithmic Trading* (2020)—3 NLP chapters with notebooks
- CFA Institute, *AI in Asset Management* (2025)—free PDF
- Financial PhraseBank dataset on HuggingFace

- **Surveys**: "The New Quant" (2025, https://arxiv.org/abs/2510.05533); "LLM Agent in Financial Trading" (2024, https://arxiv.org/abs/2408.06361)