Module 8: AI in Trading and Markets

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
- 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

Loughran & McDonald (2011) showed generic sentiment dictionaries fail on financial text. LLMs resolve this problem.

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)
- 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
- Became the standard baseline for financial NLP

BloombergGPT and FinGPT

BloombergGPT (2023)

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

FinGPT (2023)

- Open-source (AI4Finance Foundation)
- Fine-tunes Llama, Falcon, etc. with LoRA
- Training cost: **under $300**

BloombergGPT proved the concept; FinGPT democratized it.

GPT-4 and General-Purpose LLMs

General-purpose LLMs can classify financial sentiment **with zero training data** — just a well-crafted prompt.

Strengths

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

Weaknesses

- Latency: 500ms–5s per API call
- Cost: \$30/M input tokens (GPT-4 Turbo)
- Data leaves your infrastructure

Fine-Tuning Concepts

- **Fine-tuning**: Take a pre-trained model, continue training on domain-specific data
- **LoRA** (Low-Rank Adaptation): Update only a small fraction of model weights — dramatically reduces cost and GPU requirements
- **Small language models** (1–10B parameters): privacy (data stays local), speed (5–10ms inference), specialization

Why this matters for trading: **FinBERT processes headlines in milliseconds; GPT-4 takes seconds.**

What the Research Shows

Can ChatGPT Forecast Stock Prices?

Lopez-Lira & Tang (2023). Accepted at the *Journal of Finance*.

- 67,586 headlines for 4,138 companies (Oct 2021–Dec 2022)
- Long-short strategies: overnight **Sharpe ratio 2.97**, intraday **Sharpe ratio 2.63**
- Forecasting ability **increases with model size** — financial reasoning is an "emerging capability" of larger LLMs

LLMs vs. Traditional Sentiment Analysis

Kirtac & Germano (2024), *Finance Research Letters*. 965,375 U.S. financial news articles.

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

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

How It Works: Implementation

The News Trading Pipeline

- **Data ingestion**: Reuters, Bloomberg, SEC filings
- **Entity resolution**: "Apple" $\rightarrow$ AAPL
- **Sentiment extraction**: The LLM's core job
- **Signal construction**: Aggregate, weight, normalize
- **Trade execution**: Route orders
- **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

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

The Cascade Architecture

Fine-Tuned Small Model

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

General-Purpose LLM

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

Best Practice: Cascade

1. **Fast path**: FinBERT processes all headlines (˜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.

Latency and Alpha Decay

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

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

You don't need to be the fastest — you need to be **fast enough for the alpha you're targeting**.

Event-Driven Strategies

Types of News Events

Scheduled Events

- **Earnings**: Compare actuals to consensus; analyze management tone
- **Fed / central bank**: Single word changes carry enormous implications
- **Economic data**: CPI, jobs reports, PMI

Unscheduled Events

- **M&A**: Assess deal likelihood and regulatory risk
- **FDA decisions**: Binary, high-impact
- **Geopolitical**: Sanctions, trade policy, elections

The **tone of an earnings call** often matters more for stock returns than the reported numbers. LLMs excel at tone analysis.

Earnings Call Analysis

The Prompt

"Rate management's tone from -5 (very bearish) to $+5$ (very bullish). Explain your reasoning. Identify forward-looking statements that differ from consensus."

- The LLM processes 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

This is where LLMs provide the greatest edge over traditional NLP.

Beyond Company Sentiment: Geopolitical Risk

The Propagation Problem

Most LLM trading research focuses on **company-level** sentiment. The real competitive advantage: understanding how **macro events propagate across sectors**.

- "Russia invades Ukraine" — which sectors benefit? Which suffer?
- 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

Measuring Geopolitical Risk

Caldara-Iacoviello GPR Index

- Published in *AER* (2022)
- Counts articles in 10 newspapers across 8 categories
- Available for 44 countries since 1900

BlackRock BGRI

- Neural network NLP on brokerage reports + news
- For each risk, identifies the **3 most sensitive assets**
- The industry gold standard

Knowledge Graphs: Mapping Shock Propagation

A **knowledge graph** maps entities (firms, sectors, commodities) and their relationships. Combined with an LLM, it enables multi-hop reasoning about how shocks propagate.

FinDKG (Li et al., 2024)

- Fine-tuned LLM builds a **dynamic** knowledge graph from financial news
- Outperforms thematic ETFs at identifying sector themes
- Open source

Supply Chain Mapping

- LLMs extract multi-tier supplier networks from text
- Map 3,000+ firms and 11,000+ supply links
- Predict how disruptions cascade through supply chains

Industry Adoption and Risks

Who Is Doing This?

Hedge Funds

- **Bridgewater**: $2B AI fund using OpenAI, Anthropic, Perplexity
- **Two Sigma**: NLP for 10+ years
- **Numerai**: Crowdsourced AI fund, $550M AUM

Platforms

- **Bloomberg**: BloombergGPT, terminal-integrated NLP
- **RavenPack**: Structured sentiment signals
- **Kensho**: Acquired by S&P Global for $550M

Permutable AI: LLM-based trading live Oct 2024 — **20.6% return, Sharpe ratio 2.85**.

Cautionary Tales

AP Twitter Hack (2013)

- Fake tweet: "Explosions at White House"
- S&P 500 lost **$136 billion** in seconds
- Recovered within 6 minutes

Adversarial Attacks

- Imperceptible headline changes trick LLM trading systems
- Alpha Arena: ChatGPT suffered a **63% loss**
- IMF warned AI trading increases volatility

The competitive edge is real, but so are the risks.

The Crowding Problem

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

- Lopez-Lira & Tang document declining returns as LLM adoption rises — consistent with the Efficient Market Hypothesis
- More sophisticated signals (tone analysis, cross-sector propagation) 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** to **depth of understanding**.

Exercises

Exercise 1: Headline Sentiment Classification

1. Collect 20 recent financial headlines
2. For each, have Claude classify:
   - Sentiment (positive / negative / neutral)
   - Magnitude (strong / weak)
   - Relevance (high / low)
3. Compare to actual stock price movements over the following day
4. Calculate accuracy
5. Submit: spreadsheet with headlines + classifications + prices + accuracy

Exercise 2: Earnings Call Analysis

1. Download an earnings call transcript (e.g., from Seeking Alpha)
2. Ask Claude to:
   - Rate management tone (-5 to $+5$)
   - Identify forward-looking statements that differ from consensus
   - Flag hedging language
3. Compare to the stock's post-earnings move
4. Submit: transcript excerpt + Claude's analysis + stock movement + reflection

Exercise 3: FinBERT Comparison (Bonus)

1. Run the same 20 headlines through FinBERT (via HuggingFace)
2. Compare FinBERT's classifications to Claude's
3. Which handles nuance better? Where does each fail?
4. Submit: side-by-side comparison + analysis

Summary

What We Know

- LLMs dramatically outperform dictionaries (50% $\rightarrow$ 74% accuracy)
- GPT-4 can match or beat human analysts
- Fine-tuned small models compete at a fraction of cost

What to Watch

- Alpha decay as adoption increases
- Adversarial attacks and manipulation
- Crowding risk: correlated AI trades $\rightarrow$ systemic instability

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

References

Key References