Discussion of

“When Machines Disagree: Evidence from Large Language Models”

Si Cheng · Lin Hu · Kun Li

Byoung-Hyoun Hwang (NTU)

June 2026

What does this paper do?

same news headline → **same prompt**

You are a financial expert with extensive stock recommendation experience.

Analyze how this news headline ‘‘[HEADLINE]’’ about [COMPANY] could affect its stock price.

Timeframe	Impact	Sentiment Score (0–100)	Explanation (< 30 words)
Next Day	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]
Next Week	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]
Next Month	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]

What does this paper do?

same news headline → **same prompt** → **different LLMs** → **different assessments**

Timeframe	Provider	Impact	Sentiment	Explanation
Next Day	Anthropic	GOOD	70	Shift to licensing partnership seen as a positive strategic move.
	Azure OpenAI	GOOD	70	Market may react positively to strategic shift and potential for increased revenue.
	Google	BAD	35	Initial negative reaction due to loss of direct revenue stream.
	Meta	GOOD	75	AMC's focus on licensing partnership may lead to short-term optimism.
	Mistral	GOOD	75	The sale may provide cash, improving short-term financials.
	OpenAI	BAD	40	Investors may react negatively to the divestment news.

What does this paper find?

Average LLM sentiment predicts next-day returns

+0.13%

in next-day return
for a one-standard-deviation increase
in average LLM sentiment rank

The signal appears short-lived

Most of the predictive content incorporated into prices quickly, regardless of whether the prompt asks about the next day, week, or month.

What does this paper find?

Disagreement in LLM sentiment also predicts returns

-0.12%

in returns over days t+2 to t+5
for a one-standard-deviation increase
in cross-LLM disagreement

The effect is stronger among

- opaque firms
- firm facing high operating uncertainty

What does this paper find?

Disagreement in LLM sentiment also predicts returns but less when GPT is down

normal periods	outage periods
six-model dispersion is a stronger proxy for market belief dispersion	ChatGPT is missing from investors' toolkits
-0.16%	**-0.05%**

in returns over days t+2 to t+5
for a one-standard-deviation increase in cross-LLM disagreement

How does the paper fit into industry trends?

75% of UK financial-services firms report using AI, and another 10% plan to adopt it within three years (Bank of England and FCA, 2024).

69% of wealth managers say AI helps uncover hidden investment opportunities, and **62%** say AI is becoming essential for evaluating market risks (Natixis, 2025).

Among investment professionals who use GenAI, 49% use it daily and 42% use it weekly (CFA Institute, 2025).

Morgan Stanley reports that 98% of financial-advisor teams have adopted its OpenAI-powered assistant (Morgan Stanley, 2024).

How does the paper fit into the academic literature?

The literature is moving from (1) LLMs as a potential tool to (2) LLMs as market participants and (3) impact of LLMs on market participants

(1) Simulating investors and markets	(2) Financial signal extraction	(3) Impact of LLMs on market participants
Can LLM agents trade, interact, and generate market dynamics similar to real-world investors?	Could we use LLMs to parse news, disclosures, or analyst text to predict outcomes?	Are investors using LLMs and what's the impact?
Bhagwat et al. (2026)	Lopez-Lira & Tang (2023)	
Lopez-Lira (2025)	Chen et al. (2024)	
Henning et al. (2025)		
del Rio-Chanona et al. (2025)		

How does the paper fit into the literature?

Blankespoor, Croom, and Grant (2026)
- Uses 400,000+ brokerage GenAI chatbot queries.
- Investors use GenAI to interpret news and screen stocks.
- GenAI use shifts toward monitoring and firm-specific interpretation.
- Implication: GenAI already important part of retail investors' "workflow."

How does the paper fit into the literature?

He (2026)
- Uses **ChatGPT outages**.
- Studies retail herding and systemic risk.
- Evidence suggests GenAI may homogenize investor beliefs.
- Implication: Common AI tools may make investor behavior more correlated.

Cheng, Lin, and Zhao (2025)
- **ChatGPT outages** reduce trading volume and lowers (excess) volatility.
- Effects are stronger around fresh corporate news.
- Implication: ChatGPT affects actual trading and price informativeness.

How does the paper fit into the literature?

Chang, Dong, Martin, and Zhou (2025)
- Before ChatGPT, short sellers exploit AI sentiment; retail investors do not.
- After ChatGPT, retail trading aligns more with AI sentiment.
- **ChatGPT outages** weaken this alignment.
- Implication: GenAI may narrow retail investors' information disadvantage.

Even-Tov, Lourie, Munevar, and Nekrasov (2025)
- Uses Italy's **temporary ChatGPT ban**.
- During the ban, retail investors trade fewer assets, initiate fewer new positions.
- Implication: GenAI expands retail investors' information-processing capacity.

How does the paper fit into the literature?

I read the paper as being one on

the impact of LLMs on market participants

I think it asks an important question:

As we increasingly rely on LLM, will we disagree more? (some analogy to social media)

Suggestion 1: Validate the measure

It would be useful to validate whether the cross-LLM disagreement positively relates to actual investor disagreement.

For instance, are the prompts realistic?

How close is the paper's prompt to prompts actual investors use?

Use brokerage chatbot queries

Blankespoor, Croom, and Grant (2026)

Survey investors directly

"What prompts do you use when making investment decisions?"

You are a financial expert with extensive stock recommendation experience.

Analyze how this news headline "`[HEADLINE]`" about `[COMPANY]` could affect its stock price.

Timeframe	Impact	Sentiment Score (0–100)	Explanation (< 30 words)
Next Day	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]
Next Week	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]
Next Month	GOOD/BAD/UNKNOWN	[Score]	[Brief reason]

Suggestion 2: Account for other sources of disagreement

The paper currently focuses on:

cross-LLM disagreement

Other sources of disagreement

Cross-provider disagreement is important.

But actual investor disagreement may also come from:

- which model they use
- how they ask the question
- where they ask it
- what the model knows about them

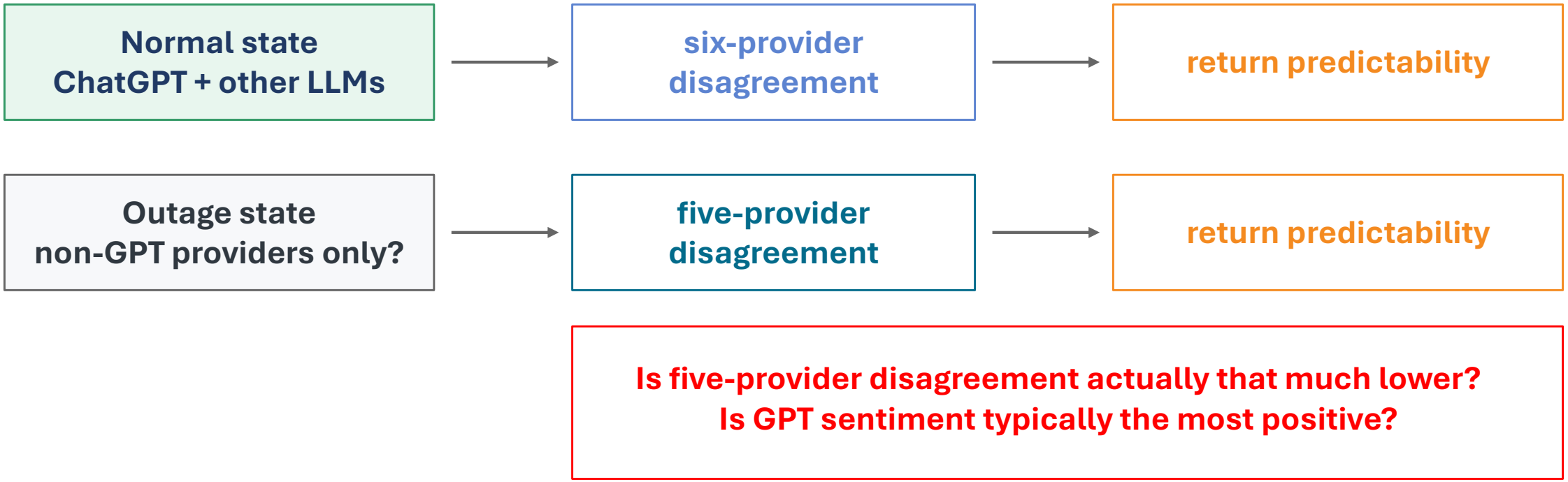
Examining/quantifying these margins could help us better understand where future investor disagreement may come from.

Suggestion 3: ChatGPT outages

The outage test is clever, but I would NOT always compute disagreement across same set of LLMs.

> Currently: *"We always compute disagreement across the same set of LLMs. But when ChatGPT is unavailable, investors cannot use one of the dominant tools, so the usual six-model disagreement measure should be a weaker proxy for the disagreement actually entering market beliefs."*

Instead, do this:



Also, link to existing disagreement measures

How closely does it map to measures of investor disagreement?

Correlate cross-LLM disagreement with existing proxies for investor disagreement

Analyst forecast dispersion

Social-media disagreement

Trading-based disagreement (TAQ)

As five-provider disagreement drops, do the above three measures as well?


THANK YOU!