When Machines Disagree: Evidence from Large Language Models

Si Cheng[1] Lin Hu[2] Kun Li[2]

[1]Syracuse University [2]Australian National University

University of Maryland / Singapore Management University /
UBS Quant Investment Forum: AI and Finance

June 2026

Top Generative AI Chatbots (May 2025)

	Generative AI Chatbot	Description	LLMs Used	AI Search Market Share	Estimated Quarterly User Growth
1	**ChatGPT(excluding Copilot)**	General-purpose AI chatbot	GPT-3.5, GPT-4	59.70%	8% ▲
2	**Microsoft Copilot**	General-purpose AI assistant	GPT-4	14.40%	6% ▲
3	**Google Gemini**	General-purpose AI assistant	Gemini	13.50%	5% ▲
4	**Perplexity**	Accuracy-focused AI search engine	Mistral 7B, Llama 2	6.20%	10% ▲
5	**Claude AI**	Business-focused AI assistant	Claude 3	3.20%	14% ▲
6	**Grok**	General-purpose AI search engine	Grok 2, Grok 3	0.80%	12% ▲
7	**Deepseek**	General-purpose AI search engine	DeepSeek V3	0.70%	10% ▲
8	**Komo**	Link-surfacing AI search engine	Not publicly disclosed	0.60%	7% ▲
9	**Brave Leo AI**	Privacy-focused AI assistant	Mixtral 8×7B	0.30%	6% ▲
10	**Andi**	Simplicity-focused AI search engine	Not publicly disclosed	0.20%	4% ▲

- Generative AI, often powered by large language models (LLMs), can perform a wide range of cognitive tasks with growing accuracy.

Use Gen AI to Enhance Financial Decision-Making

- Summarize complex corporate disclosures: e.g., Kim et al. 2024; Wong et al. 2025
- Extract nuanced and hard-to-measure information and facilitate trading: e.g., Bai et al. 2023; Bernard et al. 2024; Chang et al. 2025; Cheng et al. 2025
- Forecast earnings and returns: e.g., Lopez-Lira and Tang 2023; Chen et al. 2024; Chen et al. 2025
- Focus on a single LLM (e.g., ChatGPT) or include a few models for comparison or robustness checks

Use Gen AI to Enhance Financial Decision-Making

- Summarize complex corporate disclosures: e.g., Kim et al. 2024; Wong et al. 2025
- Extract nuanced and hard-to-measure information and facilitate trading: e.g., Bai et al. 2023; Bernard et al. 2024; Chang et al. 2025; Cheng et al. 2025
- Forecast earnings and returns: e.g., Lopez-Lira and Tang 2023; Chen et al. 2024; Chen et al. 2025
- Focus on a single LLM (e.g., ChatGPT) or include a few models for comparison or robustness checks

Research Questions

- Do LLMs exhibit significant disagreement when processing the same information?
  - Six leading LLM providers: ChatGPT, Copilot, Gemini, Claude, LLaMA, and Mistral
- What drives the model disagreement?
  - News and firm characteristics
  - Causal reasoning analysis
- How does this disagreement affect price dynamics, price informativeness, and trading activity?
  - Disagreement + short-sale constraints → lower future returns (Miller 1977)
  - Disagreement → post-earnings-announcement drift (Garfinkel and Sokobin 2006)
  - Disagreement → more trading (Kandel and Pearson 1995)

Research Questions

- Do LLMs exhibit significant disagreement when processing the same information?
  - Six leading LLM providers: ChatGPT, Copilot, Gemini, Claude, LLaMA, and Mistral
- What drives the model disagreement?
  - News and firm characteristics
  - Causal reasoning analysis
- How does this disagreement affect price dynamics, price informativeness, and trading activity?
  - Disagreement + short-sale constraints → lower future returns (Miller 1977)
  - Disagreement → post-earnings-announcement drift (Garfinkel and Sokobin 2006)
  - Disagreement → more trading (Kandel and Pearson 1995)

Research Questions

- Do LLMs exhibit significant disagreement when processing the same information?
  - Six leading LLM providers: ChatGPT, Copilot, Gemini, Claude, LLaMA, and Mistral
- What drives the model disagreement?
  - News and firm characteristics
  - Causal reasoning analysis
- How does this disagreement affect price dynamics, price informativeness, and trading activity?
  - Disagreement + short-sale constraints ➔ lower future returns (Miller 1977)
  - Disagreement ➔ post-earnings-announcement drift (Garfinkel and Sokobin 2006)
  - Disagreement ➔ more trading (Kandel and Pearson 1995)

Large Language Models (LLMs)

- OpenAI (`gpt-4o-mini`): ChatGPT; 8B parameters; data cutoff in October 2023.
- Azure OpenAI (`gpt-4o-mini`): Microsoft Copilot; mirrors OpenAI model, hosted via Microsoft Azure.
- Google (`gemini-1.5-flash`): Google Gemini; 32B parameters; data cutoff in November 2023.
- Anthropic (`claude-3-haiku-20240307`): Claude; 20B parameters; data cutoff in August 2023.
- Meta (`llama3`): 8B parameters; data cutoff in March 2023; open-source model.
- Mistral (`mistral`): 7B parameters; data cutoff in July 2023; open-source model.

Prompt

Start fresh with following instructions:
You are a financial expert with extensive stock recommendation experience. Analyze how this news headline "`[HEADLINE]`" about `[COMPANY]` could affect its stock price.
Important: Even if the headline refers to a future event, provide your analysis based on typical patterns and expected outcomes. Do not ask for clarification or wait for actual results.
You MUST provide your analysis in the exact table format below, regardless of timing or data availability:

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]

Prompt

The response must strictly adhere to the following constraints:

- The `Impact` column must use *exactly one* of three labels: GOOD, BAD, or UNKNOWN. Combinations or other variations (e.g., "somewhat good", "neutral") are explicitly disallowed.
- The `Sentiment Score` is an independent measure of sentiment intensity, on a 0–100 scale: 0–20 for strongly negative sentiment, 21–40 negative, 41–60 neutral/mixed, 61–80 positive, and 81–100 strongly positive.
- The `Timeframe Analysis` reflects immediate market reaction (next day), short-term analyst and investor response (next week), and expected medium-run adjustment to fundamentals (next month).

Prompt Output Examples: Next Day

News headline: *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*.

Provider	Impact	Sentiment	Explanation
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.

Prompt Output Examples: Next Week

News headline: *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*.

Provider	Impact	Sentiment	Explanation
Anthropic	BAD	40	Investors may be concerned about the financial impact of the joint venture sale.
Azure OpenAI	BAD	40	Analysts might express concerns over the implications of reduced ownership and control.
Google	UNKNOWN	50	Analysts will assess licensing deal's profitability; mixed reactions possible.
Meta	BAD	30	Analysts may question the impact of this shift on AMC's revenue streams.
Mistral	UNKNOWN	50	Uncertainty over future earnings potential from the licensing model.
OpenAI	UNKNOWN	50	Analysts will assess the implications of the licensing shift.

Prompt Output Examples: Next Month

News headline: *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*.

Provider	Impact	Sentiment	Explanation
Anthropic	UNKNOWN	50	Long-term implications depend on the success of the new licensing model.
Azure OpenAI	UNKNOWN	50	Long-term effects depend on execution of licensing strategy and market response.
Google	GOOD	70	If licensing proves more efficient/profitable, stock could rebound.
Meta	UNKNOWN	50	The long-term effects of this change depend on how it affects AMC's global expansion plans.
Mistral	BAD	30	Long-term partnership stability and growth prospects are uncertain.
OpenAI	GOOD	70	Potential for improved cash flow and reduced risk could enhance long-term outlook.

Data

- News headlines: from major news agencies and financial news websites (Lopez-Lira and Tang 2023)
- CRSP: daily and monthly stock data
- COMPUSTAT: quarterly and annual financial statement data
- I/B/E/S: analyst forecast
- TAQ: intraday transactions
- RavenPack: news sentiment
- Sample: common stocks listed on the NYSE/NYSE MKT/NASDAQ from 2023 to 2024
- 3,895 unique stocks, 2,447 stocks per month

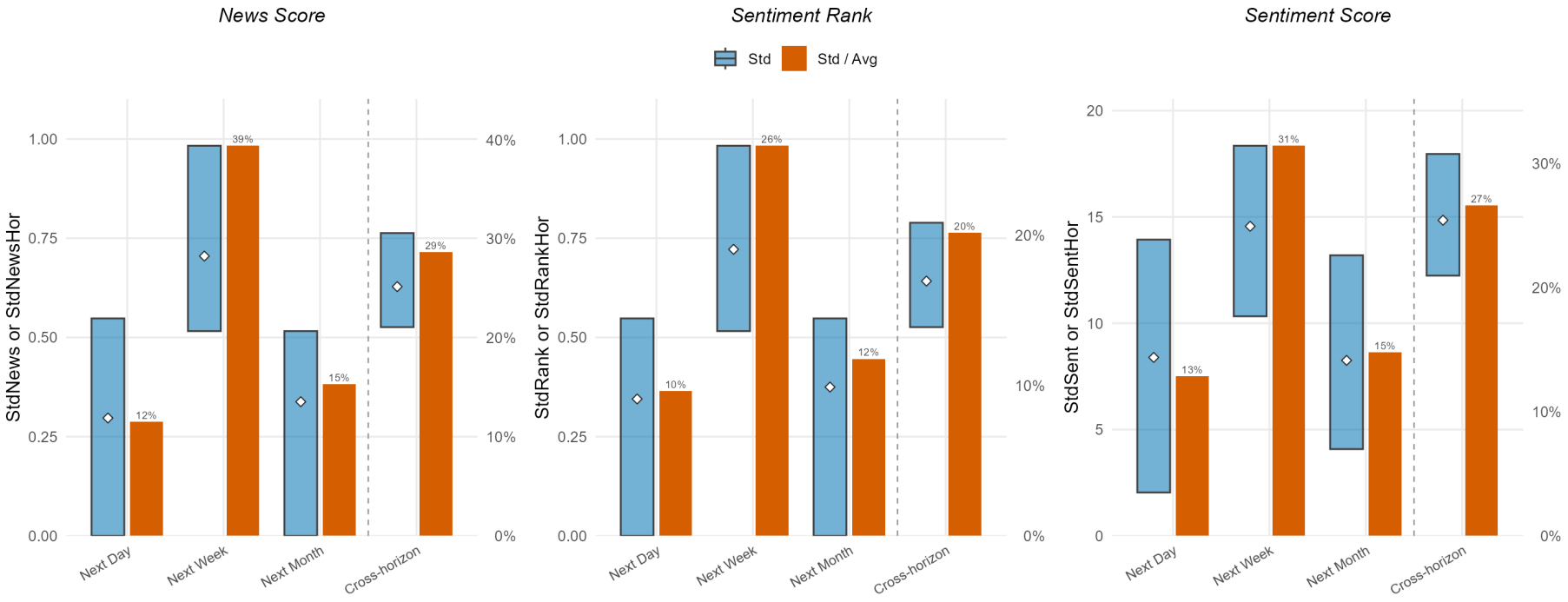
Main Variables

- LLM-based news sentiment measures
  - *News Score*: $= 1$ for `BAD`, 2 for `UNKNOWN`, and 3 for `GOOD`
  - *Sentiment Rank*: $= 1$ if *Sentiment Score* ≤ 20, 2 if $20 <$ *Sentiment Score* ≤ 40, ..., and 5 if > 80
  - *Sentiment Score*: the raw LLM output, ranging from 0 to 100
- Cross-provider dispersion: for each news at each prediction horizon, compute the standard deviations of news sentiment across six providers (*StdNews*, *StdRank*, and *StdSent*)
- Cross-horizon dispersion: for each news-provider pair, first compute the standard deviations of news sentiment across three prediction horizons, then average across providers for each news (*StdNewsHor*, *StdRankHor*, and *StdSentHor*)

Main Variables

- LLM-based news sentiment measures
  - *News Score*: $= 1$ for `BAD`, 2 for `UNKNOWN`, and 3 for `GOOD`
  - *Sentiment Rank*: $= 1$ if *Sentiment Score* ≤ 20, 2 if $20 <$ *Sentiment Score* ≤ 40, ..., and 5 if > 80
  - *Sentiment Score*: the raw LLM output, ranging from 0 to 100
- Cross-provider dispersion: for each news at each prediction horizon, compute the standard deviations of news sentiment across six providers (*StdNews*, *StdRank*, and *StdSent*)
- Cross-horizon dispersion: for each news-provider pair, first compute the standard deviations of news sentiment across three prediction horizons, then average across providers for each news (*StdNewsHor*, *StdRankHor*, and *StdSentHor*)

Summary Statistics: Cross-Provider and Cross-Horizon Dispersion



- *StdNews* for next-day, next-week, and next-month predictions account for 12%, 39%, and 15% of the respective sample means.

Determinants of News Sentiment Dispersion

Dep. Var. =	StdRank			StdRankHor
	Next Day	Next Week	Next Month	
	Model 3	Model 6	Model 9	Model 12
Complexity	-1.247***	-0.375***	1.611***	-0.956***
	(-13.44)	(-4.96)	(20.04)	(-21.62)
AvgRank	-0.436***	0.506***	0.219***	0.006***
	(-100.70)	(74.53)	(21.03)	(2.62)
Overnight	0.227***	0.213***	-0.111***	0.066***
	(8.53)	(8.76)	(-5.10)	(5.21)
Controls	Y	Y	Y	Y

- *Complexity*: a composite measure based on word count, the Fog index, and the percentage of complex words

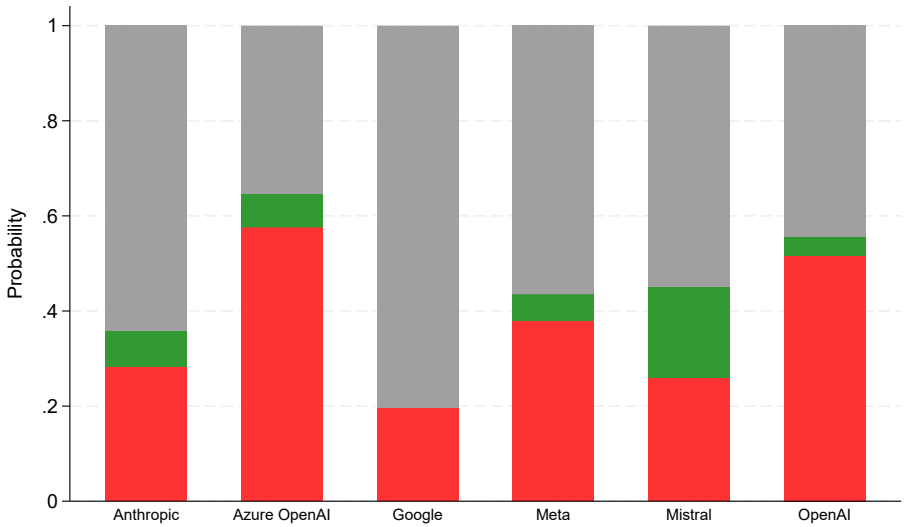
Cause–Effect Chain Examples: Next Day

News headline: *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*. ▸ Next Week ▸ Next Month

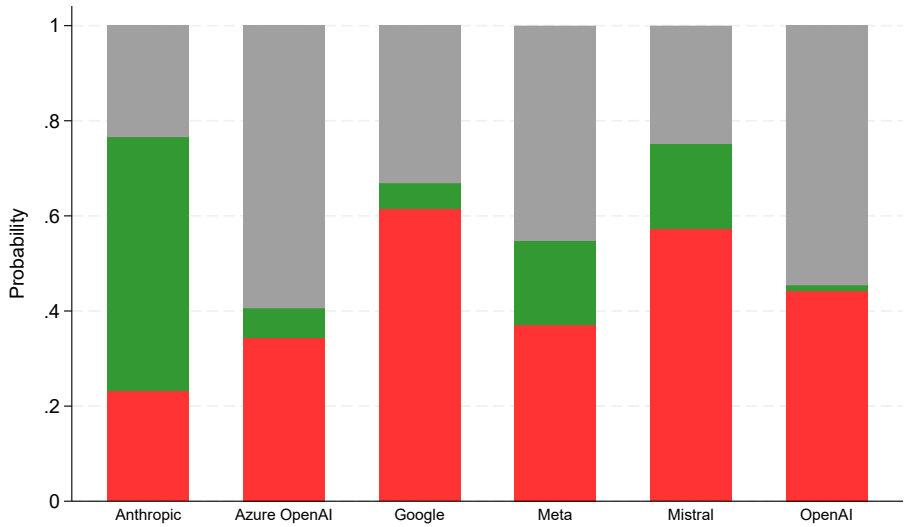
Provider	Impact	Cause → Effect
Anthropic	GOOD	Licensing shift → Positive market reaction
Azure OpenAI	GOOD	Strategic shift → Positive market reaction
Google	BAD	Loss of revenue stream → Negative market reaction
Meta	GOOD	Licensing shift → Positive market reaction
Mistral	GOOD	Stake sale → Positive short-term financials
OpenAI	BAD	Divestment news → Negative investor reaction

- Extract directed acyclic graphs (DAGs) that capture the stated cause and effect from model explanations ▸ Prompt
- Divergence in reasoning and identified effects ▸ Similarity

Cause Impact Mapping across Topics



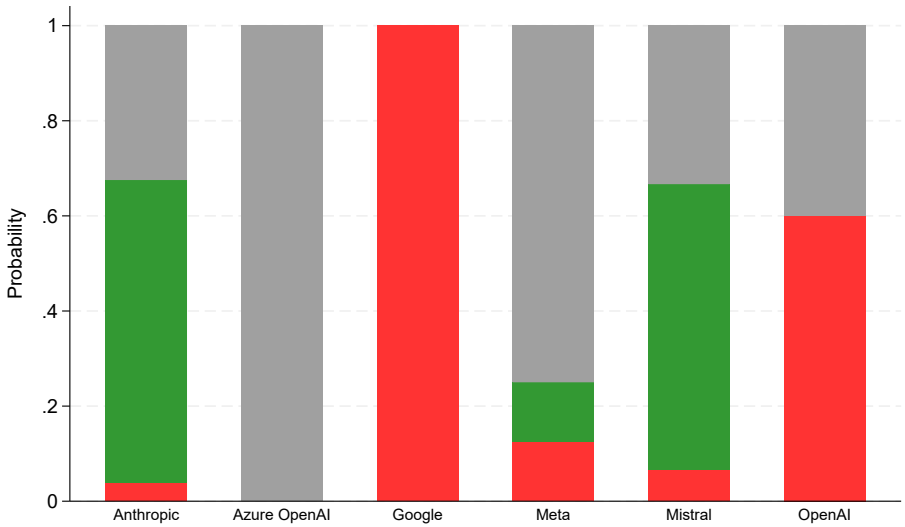
(a) Corporate Governance



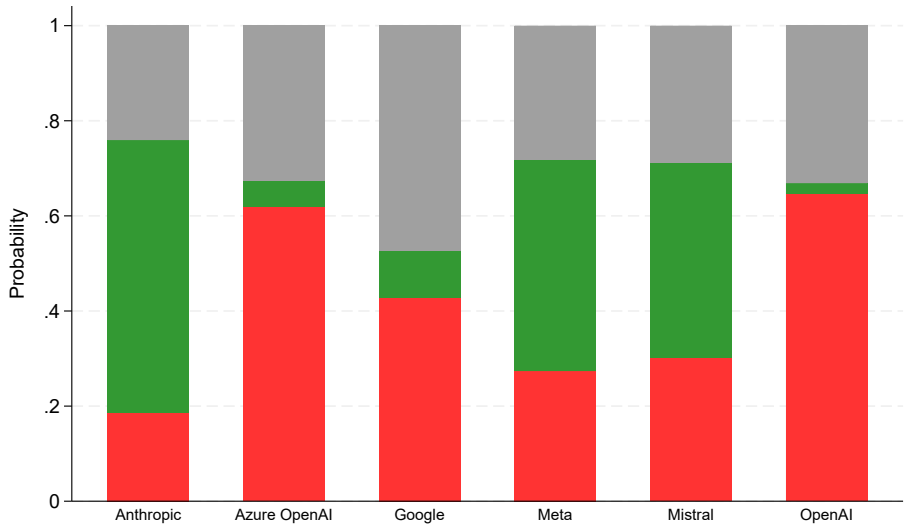
(f) Technological

- Group the extracted causes into six primary topic categories: corporate governance, cost/liquidity, demand/market, macroeconomic, regulation/policy, and technological
- Cross-provider differences within the same cause category: GOOD, BAD, and UNKNOWN ▸ Details

Effect Impact Mapping across Topics



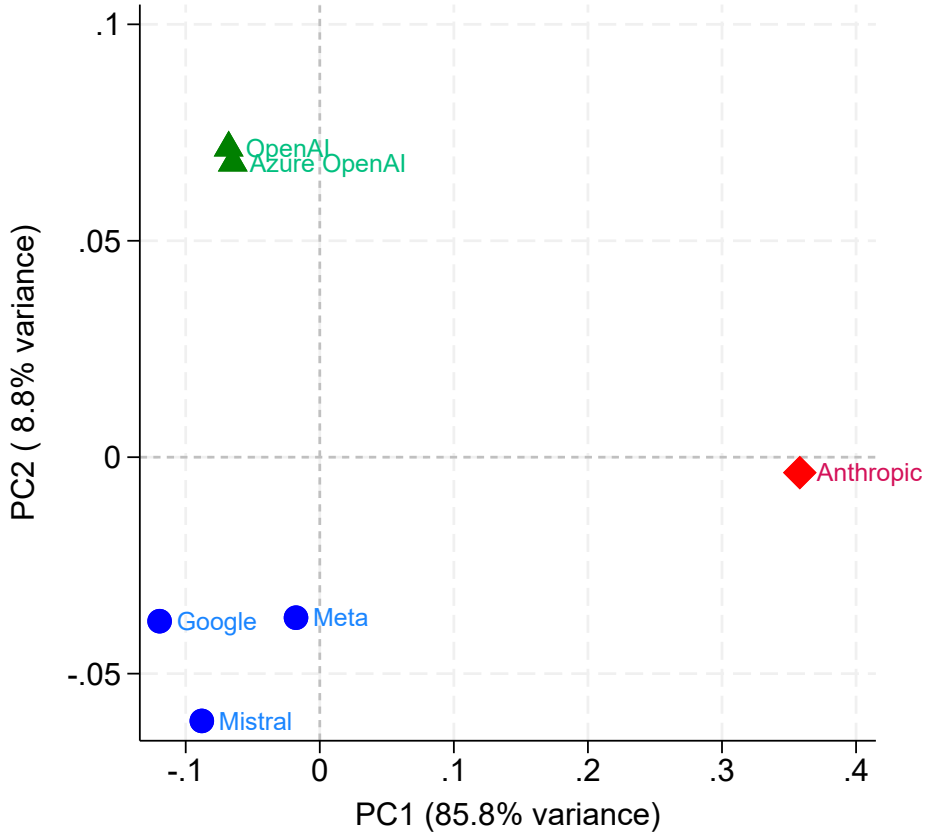
(a) Competitive Position



(e) Revenue/Growth

- Group the extracted effects into six primary topic categories: competitive position, innovation/capacity, market/valuation, profitability/cost, revenue/growth, and risk/volatility
- Cross-provider differences within the same effect category: GOOD, BAD, and UNKNOWN ▸ Details

Provider Comparison: Reasoning Styles



- Proximity indicates greater similarity in extracted cause–effect chains.
- Provider disagreement is not merely noise. ▸ Stability

News Sentiment and Stock Returns

- Daily panel regression: $R_{i,t+1} = \alpha + \beta_1 NewsSent_{i,t} + \varepsilon_{i,t+1}$
- Firm and calendar day fixed effects, two-way clustering

	$R_{i,t+1}$ Model 2	$R_{i,t+2:t+5}$ Model 5	$R_{i,t+2:t+20}$ Model 8
Panel A: Next-Day Predictions			
AvgRank	0.201***	0.024	-0.180
	(5.57)	(0.35)	(-1.49)
Panel B: Next-Week Predictions			
AvgRank	0.287***	0.166**	0.061
	(5.90)	(2.24)	(0.40)
Panel C: Next-Month Predictions			
AvgRank	0.226***	0.001	-0.249
	(2.90)	(0.01)	(-0.95)

News Sentiment Dispersion and Stock Returns

	$R_{i,t+1}$		$R_{i,t+2:t+5}$		$R_{i,t+2:t+20}$	
	Model 3	Model 4	Model 7	Model 8	Model 11	Model 12
StdRank	-0.143	-0.140	-0.336**	-0.329*	-0.904***	-0.896***
	(-1.20)	(-1.16)	(-1.99)	(-1.93)	(-2.92)	(-2.87)
StdRankHor	-0.179	-0.181	-0.854**	-0.859**	-0.883	-0.888
	(-0.87)	(-0.88)	(-2.41)	(-2.42)	(-1.27)	(-1.27)
AvgRank	0.134**	0.142**	-0.134	-0.118	-0.600***	-0.584***
	(2.07)	(2.15)	(-1.15)	(-0.95)	(-3.23)	(-2.94)
CSS		-0.133		-0.266		-0.265
		(-0.44)		(-0.51)		(-0.31)

- 1 std.dev. increase in *StdRank* and *StdRankHor* → 0.12% and 0.14% decline in $R_{i,t+2:t+5}$
- 1 std.dev. increase in *AvgRank* → 0.09% increase in $R_{i,t+1}$ ▸ OOS

News Sentiment Dispersion and Stock Returns: ChatGPT Outages

	$R_{i,t+1}$		$R_{i,t+2:t+5}$		$R_{i,t+2:t+20}$	
	Model 2	Model 3	Model 5	Model 6	Model 8	Model 9
StdRank	-0.223*	-0.219	-0.445**	-0.438**	-1.063***	-1.055***
	(-1.65)	(-1.62)	(-2.41)	(-2.35)	(-3.39)	(-3.34)
StdRank × Outage	0.217*	0.217*	0.292*	0.292*	0.427	0.427
	(1.94)	(1.94)	(1.77)	(1.77)	(1.11)	(1.11)
StdRankHor	-0.189	-0.192	-0.868**	-0.872**	-0.896	-0.901
	(-0.93)	(-0.94)	(-2.45)	(-2.46)	(-1.28)	(-1.29)
AvgRank	0.135**	0.143**	-0.133	-0.117	-0.599***	-0.582***
	(2.09)	(2.17)	(-1.14)	(-0.94)	(-3.22)	(-2.93)
CSS		-0.130		-0.268		-0.279
		(-0.43)		(-0.52)		(-0.33)

- The next-week return predictability of *StdRank* declines by 67% during ChatGPT outage periods.

News Sentiment Dispersion and Stock Returns: Pre-LLM Period (Year 2012)

	$R_{i,t+1}$		$R_{i,t+2:t+5}$		$R_{i,t+2:t+20}$	
	Model 3	Model 4	Model 7	Model 8	Model 11	Model 12
StdRank	0.026	0.014	-0.051	-0.058	0.181	0.206
	(0.48)	(0.26)	(-0.57)	(-0.65)	(1.08)	(1.23)
StdRankHor	0.405***	0.436***	0.334	0.352	0.494	0.430
	(2.99)	(3.21)	(1.52)	(1.60)	(1.14)	(0.99)
AvgRank	0.262***	0.239***	0.157***	0.144***	0.261***	0.309***
	(11.33)	(9.44)	(3.84)	(3.13)	(3.79)	(4.14)
CSS		0.645**		0.368		-1.321*
		(2.19)		(0.90)		(-1.69)

- Model disagreement captures belief dispersion arising from the adoption and use of LLMs, not pre-existing investor disagreement.

Earnings Announcement Returns: Small Firms

	$R_{i,t}$			$R_{i,t+1:t+40}$		
	Model 1	Model 4	Model 5	Model 11	Model 14	Model 15
SUE	0.521***	0.735***	0.719***	0.224**	-1.205**	-1.214**
	(9.27)	(3.34)	(3.29)	(2.06)	(-1.99)	(-1.99)
SUE × StdRank		0.069	0.052		0.766	0.742
		(0.43)	(0.32)		(1.53)	(1.49)
SUE × StdRankHor		-0.459	-0.491		1.721**	1.693**
		(-1.38)	(-1.50)		(2.22)	(2.19)
SUE × AnaDisp			0.188*			0.138
			(1.89)			(0.94)
Controls	Y	Y	Y	Y	Y	Y

- Cross-horizon dispersion amplifies price underreaction to earnings news ➔ post-earnings-announcement drift among small firms.

Earnings Announcement Returns: Large Firms

	$R_{i,t}$			$R_{i,t+1:t+40}$		
	Model 1	Model 4	Model 5	Model 11	Model 14	Model 15
SUE	0.517***	0.756***	0.765***	0.123***	0.439*	0.452*
	(15.55)	(5.33)	(5.38)	(2.73)	(1.91)	(1.95)
SUE × StdRank		-0.188*	-0.194*		-0.076	-0.063
		(-1.85)	(-1.86)		(-0.38)	(-0.32)
SUE × StdRankHor		-0.392*	-0.407**		-0.472	-0.448
		(-1.95)	(-2.01)		(-1.49)	(-1.42)
SUE × AnaDisp			0.056			-0.208
			(0.46)			(-1.07)
Controls	Y	Y	Y	Y	Y	Y

- Model disagreement delays the immediate price reaction on the announcement day for large firms.

Additional Analyses

- Heterogeneity analysis: model disagreement has stronger effects for firms with more opaque information environments and greater operating uncertainty.
- Trading volume: model disagreement increases abnormal overall and retail trading volume.
- Headlines vs. full articles

Conclusion

- Significant cross-provider and cross-horizon dispersion in LLM-based news sentiment
- Model disagreement reflects systematic differences in causal reasoning.
- Model disagreement predicts lower future returns, especially for firms with opaque information and high operating uncertainty.
- Model disagreement amplifies post-earnings-announcement drift for small firms and delays immediate price reactions for large firms.
- Model disagreement increases overall and retail trading volume.
- Widespread GenAI use may amplify systematic belief dispersion, increasing information uncertainty and reducing price efficiency.

Appendix

Prompt to Extract DAGs

Identify the causal statement explaining a [Impact]: [Explanation]. Output ONLY a JSON object with this exact format: {"cause":" $\leq$ 5 words>","intermediate_cause":" $\leq$5 words or ">","effect":" $\leq$5 words>"}

Rules:

- cause: the initial trigger or reason
- intermediate_cause: optional middle step (use empty string if none)
- effect: the final outcome or result
- Keep each field to 5 words or less
- Use clear, concise language

Back

Cause–Effect Chain Examples: Next Week

News headline: *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*.

Provider	Impact	Cause → Effect
Anthropic	BAD	Financial concern → Negative investor sentiment
Azure OpenAI	BAD	Financial concern → Negative investor sentiment
Google	UNKNOWN	Licensing shift → Mixed market reaction
Meta	BAD	Analyst skepticism → Negative investor sentiment
Mistral	UNKNOWN	Licensing shift → Mixed market reaction
OpenAI	UNKNOWN	Licensing shift → Mixed market reaction

Back

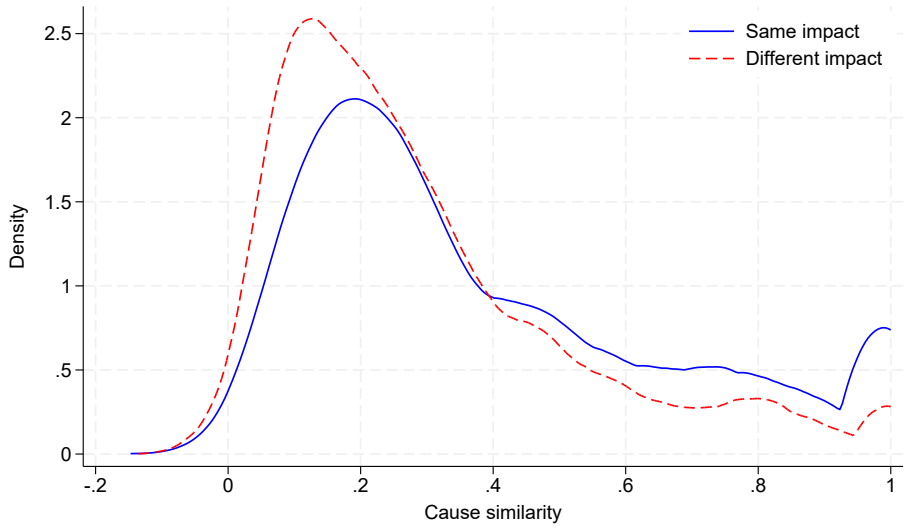
Cause–Effect Chain Examples: Next Month

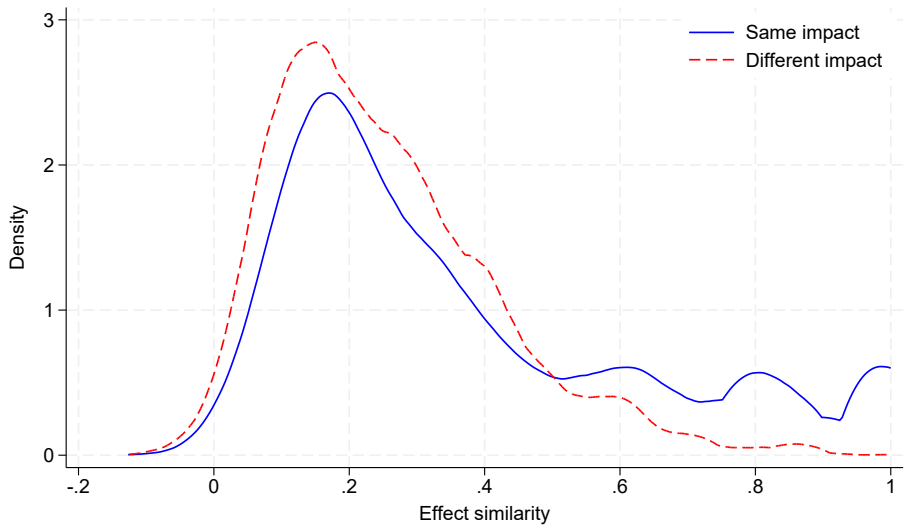
News headline: *"AMC sells stake in Saudi Arabia joint venture, shifts to licensing partnership"*.

Provider	Impact	Cause → Effect
Anthropic	UNKNOWN	Licensing shift → Mixed market reaction
Azure OpenAI	UNKNOWN	Licensing shift → Mixed market reaction
Google	GOOD	Licensing shift → Positive market reaction
Meta	UNKNOWN	Licensing shift → Long-term effects
Mistral	BAD	Uncertain outlook → Investor caution
OpenAI	GOOD	Improved cash flow → Positive long-term outlook

Back

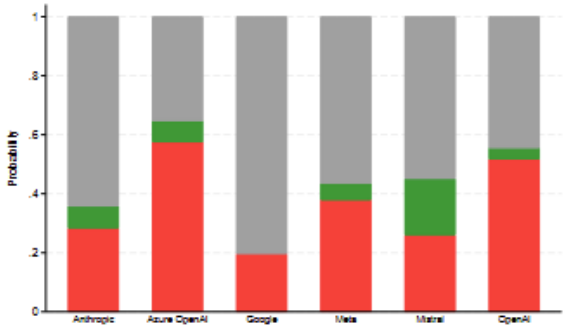
Causes and Effects Similarity



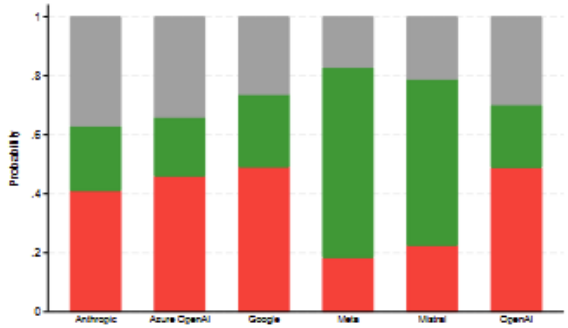


Back

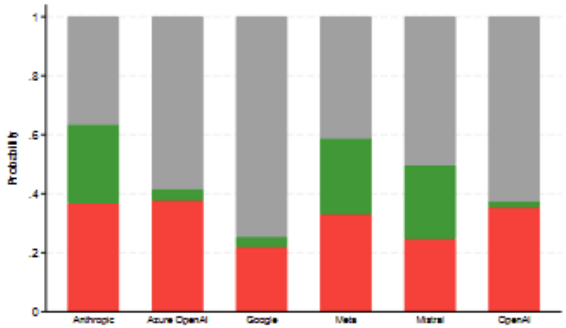
Cause Impact Mapping across Topics



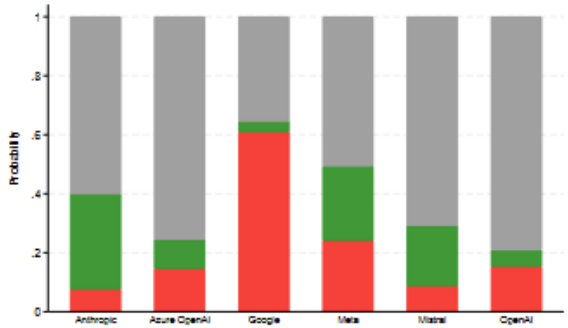
(a) Corporate Governance



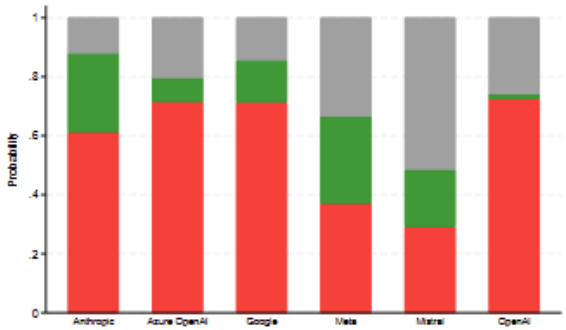
(b) Cost/Liquidity



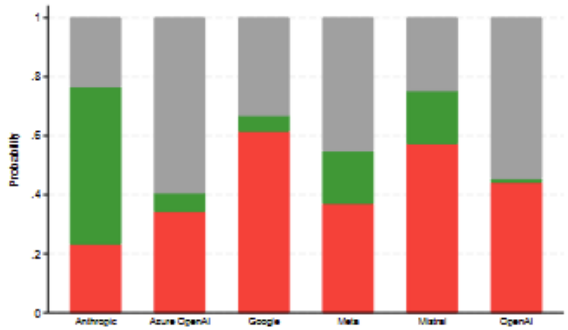
(c) Demand/Market



(d) Macroeconomic



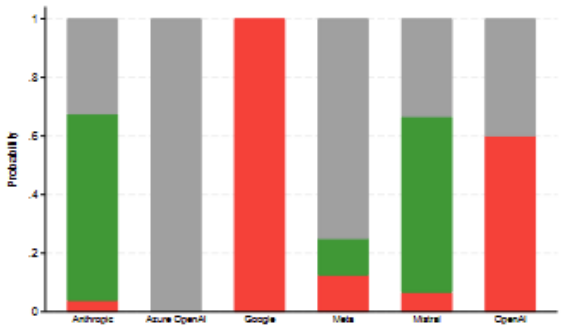
(e) Regulation/Policy



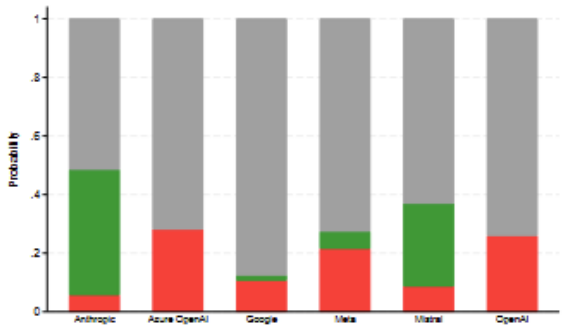
(f) Technological

◂ Back

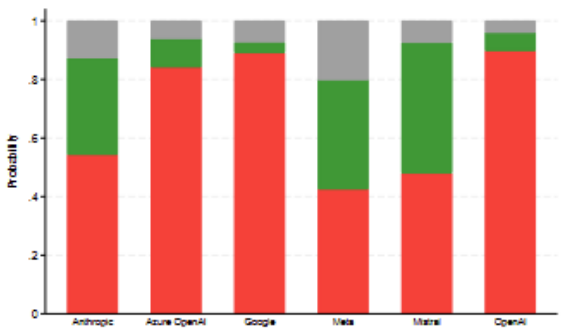
Effect Impact Mapping across Topics



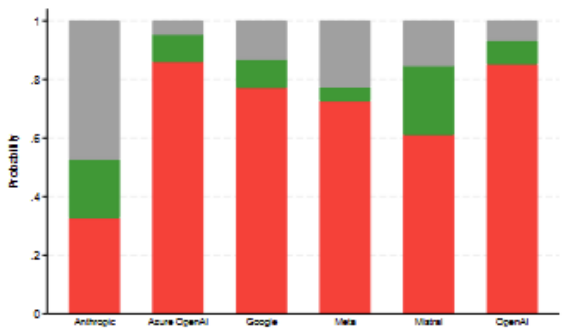
(a) Competitive Position



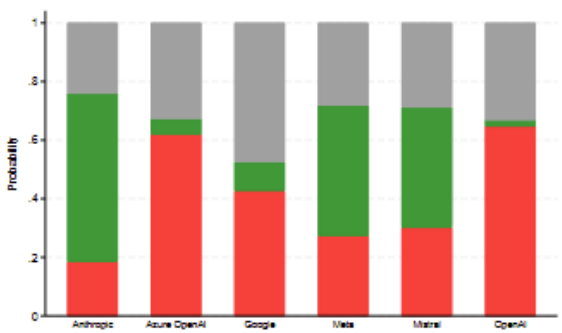
(b) Innovation/Capacity



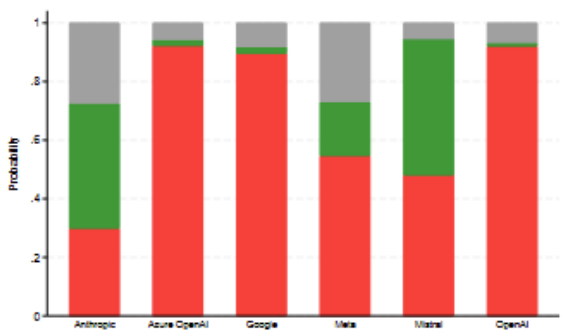
(c) Market/Valuation



(d) Profitability/Cost



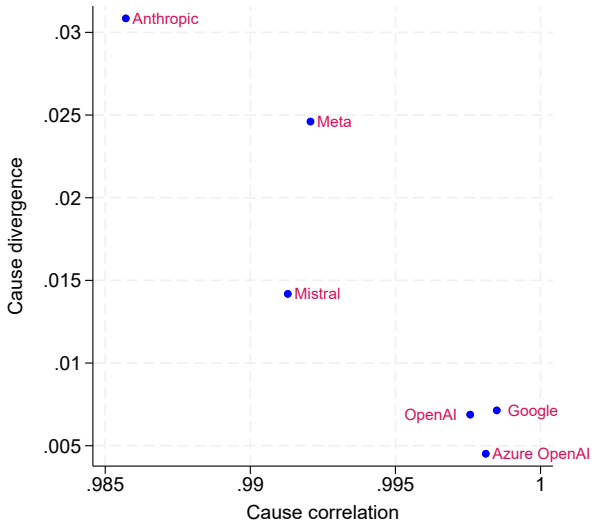
(e) Revenue/Growth

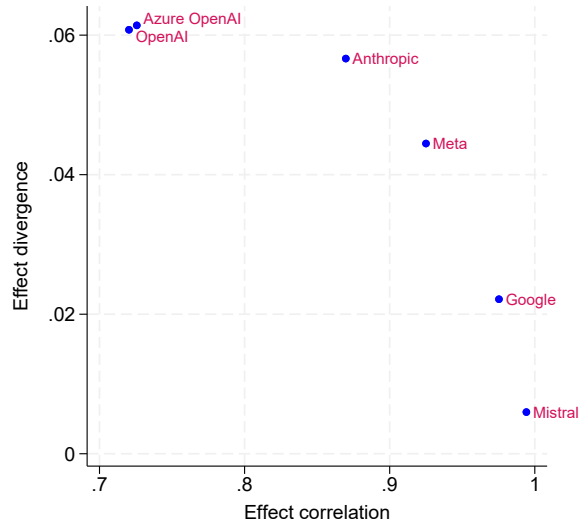


(f) Risk/Volatility

Back

Provider Comparison: Cause and Effect Stability





- Correlation: the Pearson correlation coefficient between category distributions across adjacent time horizons
- Divergence: the Jensen-Shannon divergence between these distributions

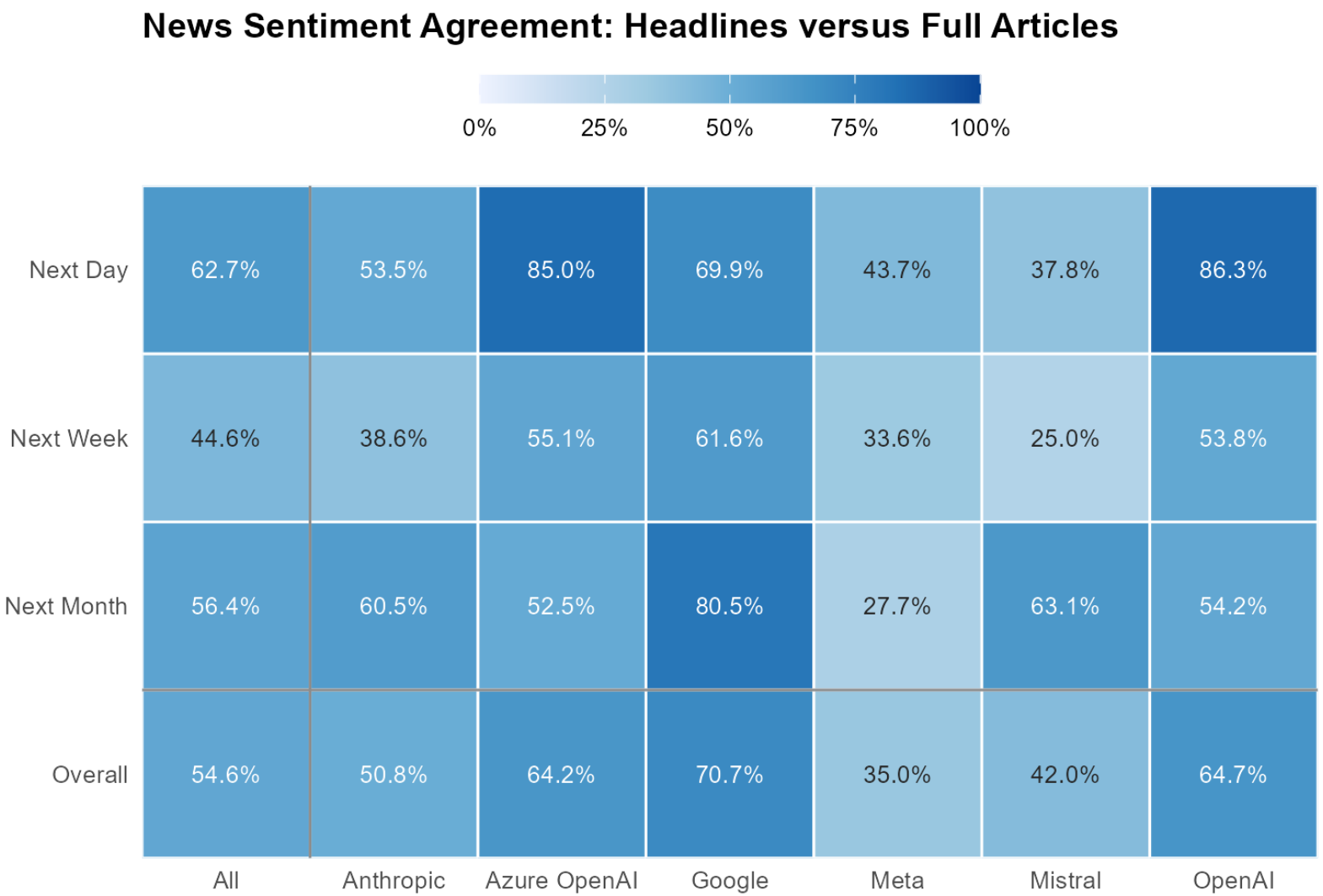
Back

News Sentiment Dispersion and Stock Returns: Year 2024

	$R_{i,t+1}$		$R_{i,t+2:t+5}$		$R_{i,t+2:t+20}$	
	Model 3	Model 4	Model 7	Model 8	Model 11	Model 12
StdRank	-0.230*	-0.220*	-0.454*	-0.459*	-0.884*	-0.856*
	(-1.79)	(-1.72)	(-1.94)	(-1.93)	(-1.80)	(-1.73)
StdRankHor	0.182	0.180	-0.537	-0.536	0.317	0.311
	(0.63)	(0.63)	(-0.98)	(-0.97)	(0.32)	(0.31)
AvgRank	0.068	0.092	-0.223*	-0.235*	-0.685**	-0.620**
	(0.93)	(1.24)	(-1.82)	(-1.74)	(-2.59)	(-2.19)
CSS		-0.414		0.203		-1.102
		(-0.96)		(0.23)		(-0.92)

Back

Additional Analysis: Headlines vs. Full Articles



Back

Additional Analysis: Headlines vs. Full Articles