Large Language Models and Return Prediction in China

Lin Tan, Huihang Wu, Xiaoyan Zhang

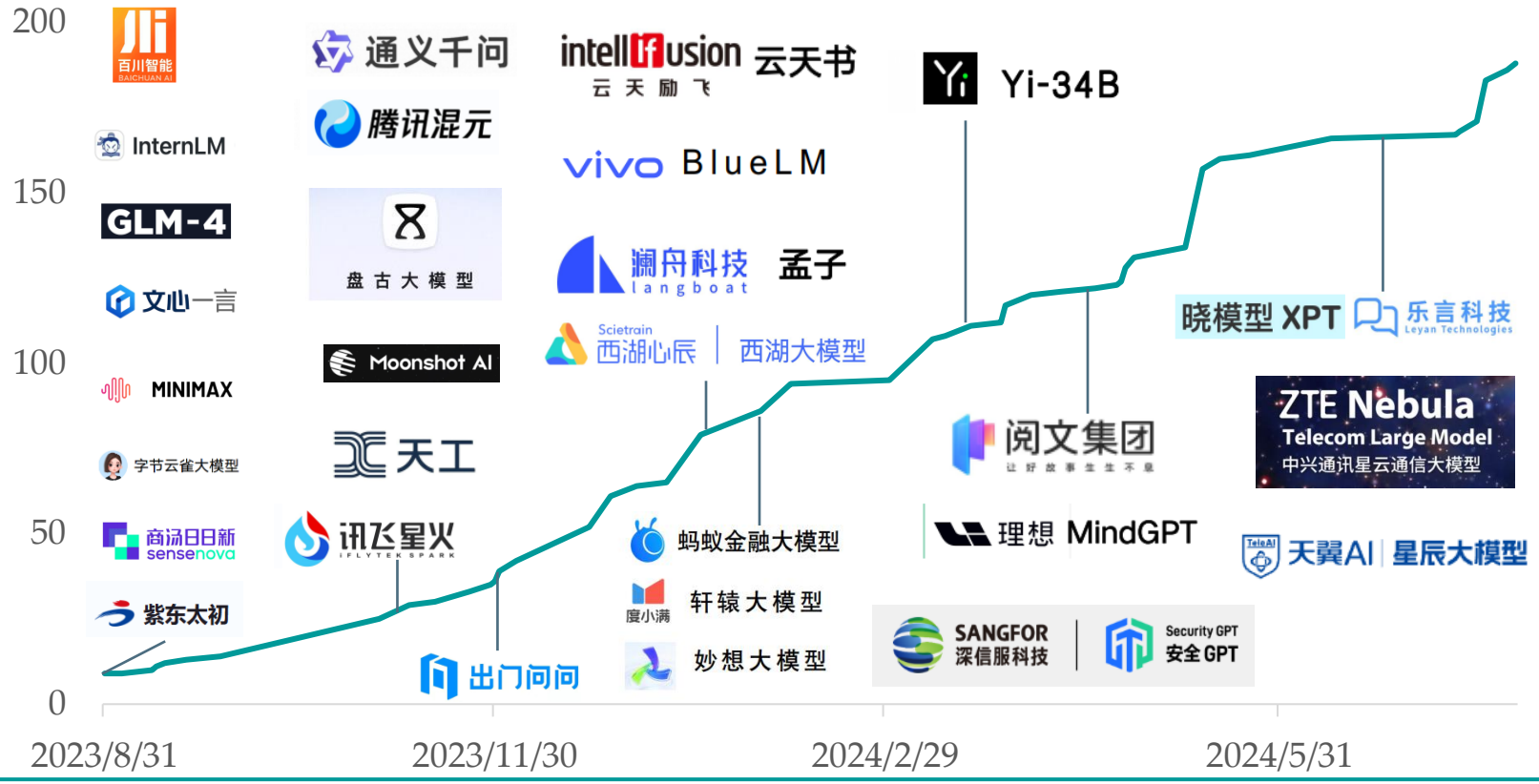
PBC School of Finance, Tsinghua University

ABFER 2024

LLMs in China

"China will have their own LLM that's different from the rest of the world."

—Sam Altman, AI for Good Global Summit, May 2024.



Why Chinese?

- Importance
  - Substantial information about the second largest economy.
- Uniqueness
  - Ideographic, materially different from phonetics.
  - Difficulty in accurately identify word boundaries based on surrounding context.
    - 上市公司停牌原因说明会延期：上市公司/停牌/原因/说明会/延期 vs. 上市公司/停牌原因/说明/会延期
- Necessity
  - Previous NLP methods that ignore word boundaries, order and cross-word relations miss substantial information in Chinese.
  - LLMs can handle Chinese by learning contextualized representations from character sequences via Transformer.

Why Chinese Stock Market?

- Previous studies focus on the developed markets, but Chinese stock market significantly differs from those developed.
  - Retail investors are prevalent (Jones et al., 2024) and have relatively low financial literacy (Song, 2020; Titman et al., 2022), facing challenges in processing and trading on public news.
  - There could be substantial under-processed information in public news for LLMs to extract.
- Chinese stock market is still in a developing phase, with gradually improving information efficiency (Carpenter et al., 2021).
  - LLMs could help accelerate this process and improve efficiency.
  - For other emerging markets with similar characteristics, there could be important academic, industrial and regulatory implications.

Research Questions

- Can LLMs process public news and predict stock returns in China?
  - By answering this question, we are among the first to use a representative and influential series of Chinese LLMs to effectively extract information from a comprehensive set of Chinese news.

- Do signals from LLMs help the price discovery process and contribute to market efficiency?
  - By answering this question, we are providing novel insights for AI potentially changing the information transmission dynamics, and helping investors (particularly retail investors) making decisions in the future.

Literature Review

- Textual analysis
  - Return forecasting: Tetlock et al. (2008), Loughran and McDonald (2011), Jegadeesh and Wu (2013), Tetlock (2014), Loughran and McDonald (2016), Gentzkow et al. (2019), and Ke et al. (2019).
  - Other applications: Manela and Moreira (2017), Bybee et al. (2023a), Bybee et al. (2023b).

- LLMs provide new possibilities to make use of financial text.
  - U.S. and international markets: various LLMs, such as ChatGPT.
    - Kim and Nikolaev (2023), Lopez-Lira and Tang (2023), Beckmann et al. (2024), Chen et al. (2024a), Chen et al. (2024b).
  - China: mainly into BERT-based models.
    - Jiang et al. (2024), Zhou et al. (2024).

Data

- Sample: January 1st 2008 to December 31st 2023.
- Returns, stock characteristics, accounting data: WIND and CSMAR.
- News data: ChinaScope SmarTag database.
  - 28 million deduplicated raw news articles in Chinese.
    - Covering 5,255 stocks (100% of the A-share stocks).
  - Source: 966 registered internet news providers.
    - Financial medias: Eastmoney, Tonghuashun Finance, Sina Finance, China Securities Net, etc. (47.1%)
    - Government websites: State Council, NDRC and local government official websites. (33.4%)
    - WeChat official accounts: brokerages, funds… (19.5%)
  - For every article: title, full content, source, timestamp, stock id.

Selection of LLMs

- To select representative and influential LLMs, we use 3 criterion:
  - Trained on Chinese texts;
  - Fully open-sourced;
    - Complete weights (learned parameters), and detailed technical documentation of the model structure.
  - Influential, well adopted.
    - Influence in the academia: commonly adopted by existing studies on LLM's finance application.
    - Influence in the industry: approval from the Cyberspace Administration of China.

Selection of LLMs

LLMs	Name	Our version	Open-sourced
Individual LLM	BERT	Google Chinese version base BERT	Nov. 2018
	RoBERTa	Chinese adaptation XLM-RoBERTa	Dec. 2019
	FinBERT	Chinese FinBERT Valuesimplex	Oct. 2020
	Baichuan	Most recent open-sourced version: Baichuan2-7B	Jun. 2023
	ChatGLM	Most recent version: ChatGLM3-6B	Mar. 2023
	InternLM	Most recent version: InternLM2-7B	Jun. 2023
Ensemble LLM	Ensemble	We take average of the individual LLMs' signals.	\

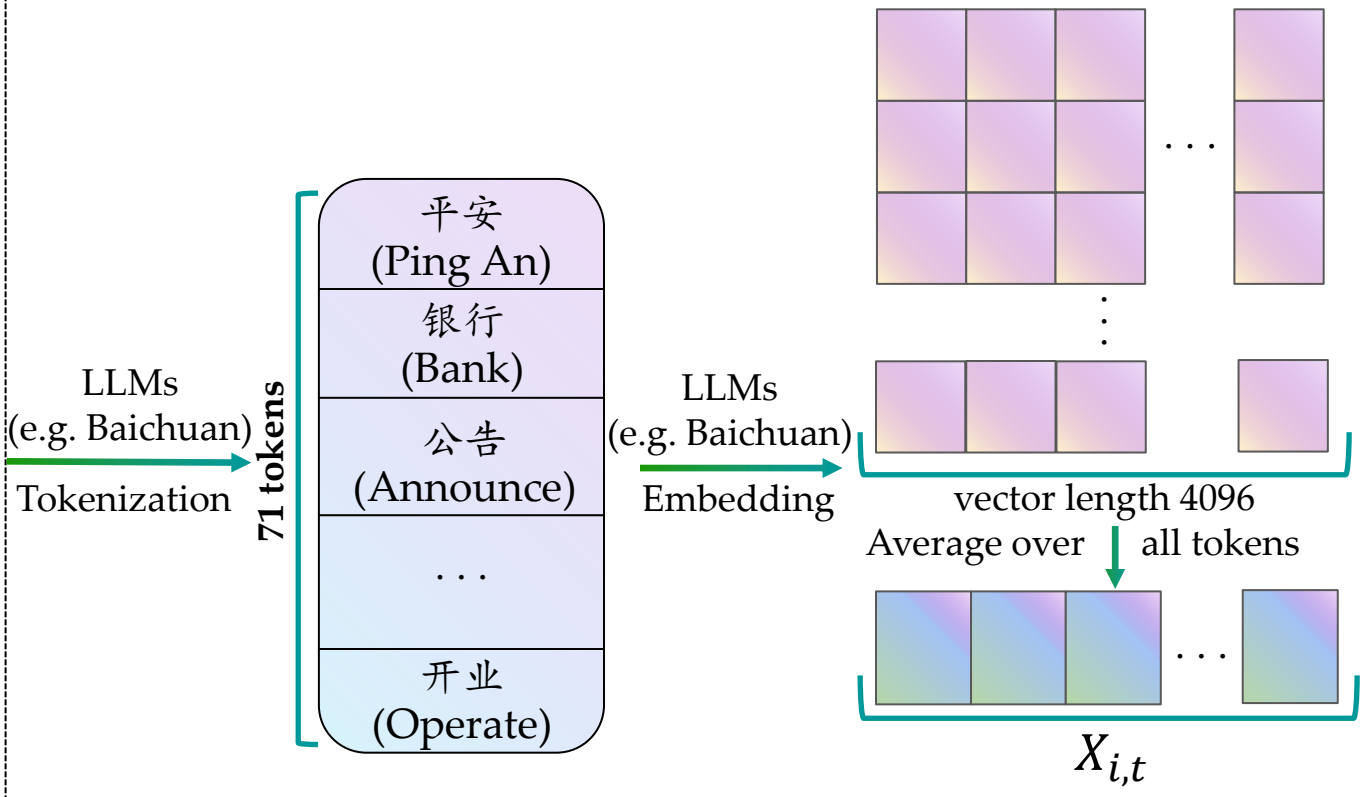
Article-Level Representation

股份行第五家，平安银行理财子公司获批筹建

2019-12-31 15:50:59

来源：新浪财经

平安银行今日公告称，近日收到《中国银保监会关于筹建平安理财有限责任公司的批复》，获准筹建理财子公司平安理财。2018年6月平安银行曾公告称，平安理财注册资本不超过50亿元，注册地未定。这是目前为止，第五家获批筹建理财子公司的股份行，其中3家已经开业。（共计121个字）



- LLM transforms the news into a numerical vector, the "article-level representation", summarizing news content and semantic information.

Empirical Modelling

- For the first signal, news tone, estimate logistic model:

$$E\left(y_{i,t+1}\middle|x_{i,t}\right) = \frac{e^{x'_{i,t}\beta}}{1 + e^{x'_{i,t}\beta}} \quad (1)$$

  - $y_{i,t+1}$: positive return dummy, takes the value of 1 when the next day's return $r_{i,t+1} > 0$, otherwise it's 0.
  - $x_{i,t}$: LLM's article-level representation for stock *i* on day *t*.

- For the second signal, return forecast, estimate linear model:

$$E\left(r_{i,t+1}\middle|x_{i,t}\right) = x'_{i,t}\theta \quad (2)$$

  - $r_{i,t+1}$: return of stock *i* on day *t+1*.
  - $x_{i,t}$: LLM's article-level representation for stock *i* on day *t*.

Estimation Method

- We set training sample 2008-2018, testing sample 2019-2023.
  - Expanding training window approach: for testing year 2019, 2008-2018 are training sample; for testing year 2020, 2008-2019 are training sample; …
- In training sample
  - Estimate β by minimizing the cross-entropy loss function with L2 penalty.
  - Estimate θ by minimizing the MSE loss function with L2 penalty.
- In testing sample
  - Construct predicted news tone: $\hat{y}_{i,t+1} \equiv \frac{e^{x'_{i,t}\hat{\beta}}}{1+e^{x'_{i,t}\hat{\beta}}}$.
  - Construct return forecast: $\hat{r}_{i,t+1} \equiv x'_{i,t}\hat{\theta}$.

Out-of-Sample Fitness

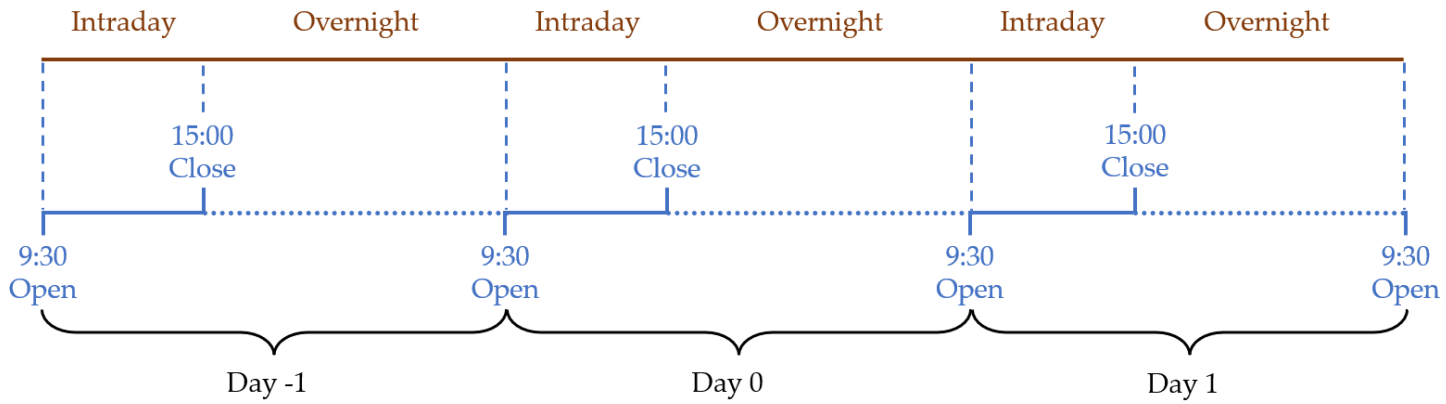
- Out-of-sample accuracy of the news tone
  - Accuracy: percentage of correct direction predictions.
  - Model performs better when the accuracy is higher.

BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	InternLM	Ensemble
52.71%	52.77%	52.63%	52.37%	51.93%	52.08%	52.74%

- Out-of-sample correlation between real returns and forecasts
  - Model performs better when the correlation is higher.

BERT	FinBERT	RoBERTa	Baichuan	ChatGLM	InternLM	Ensemble
1.62%	1.94%	1.80%	1.99%	1.52%	1.73%	1.95%

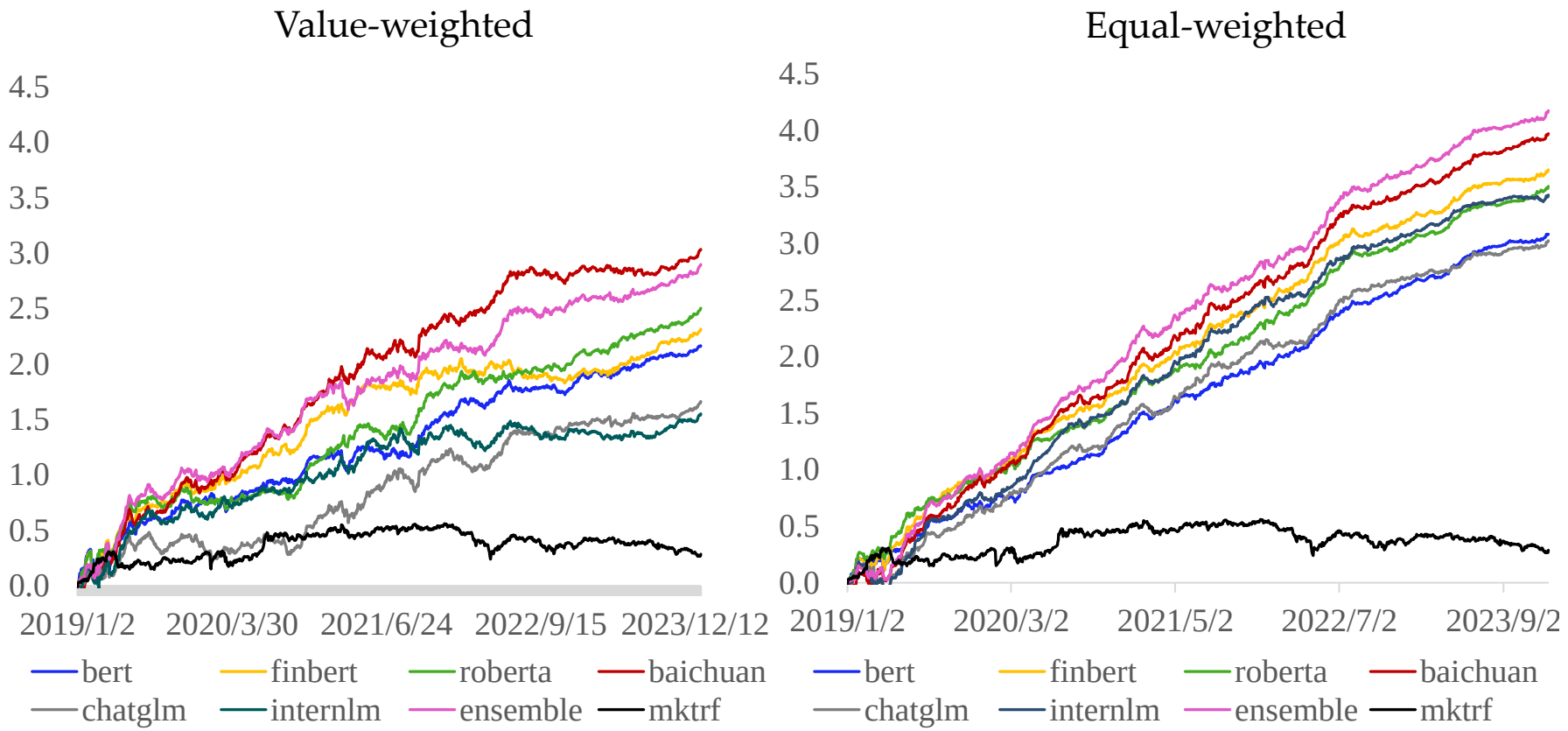
Q1. LLM's Return Predictive Power



- Portfolio Sorting
  - Open-to-open timeline.
    - Allow for the timeliest use of most news arriving overnight.
  - For news that occur on day 0:
    - Compute predicted news tone and return forecasts for day 1;
    - Build positions at the open on day 1, hold for 1 day;
      - Long top 10% stocks, and short bottom 10% stocks.
    - Rebalance on daily basis.

Return Prediction: Raw Returns

- Long minus short portfolio returns, sorted by news tones.



Return Prediction: Alphas

- CH4-adjusted returns (annualized) for portfolios sorted by news tones.

Model	VW						EW					
	Long Leg		Short Leg		Long minus Short		Long Leg		Short Leg		Long minus Short	
	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat	Alpha	*t*-Stat
BERT	9.18%	1.54	-38.22%	-5.28	47.40%	4.97	16.33%	3.74	-49.79%	-9.41	66.13%	9.83
FinBERT	7.67%	1.20	-43.77%	-5.61	51.44%	4.92	24.68%	5.09	-54.26%	-8.47	78.95%	9.68
RoBERTa	14.56%	2.23	-39.89%	-5.33	54.45%	5.22	20.32%	4.61	-54.89%	-9.42	75.22%	10.25
Baichuan	22.92%	3.66	-46.99%	-5.92	69.90%	6.52	32.42%	6.75	-54.68%	-8.64	87.09%	10.91
ChatGLM	13.60%	2.23	-27.90%	-4.13	41.50%	4.23	22.53%	4.96	-44.70%	-8.31	67.23%	9.41
InternLM	11.68%	1.85	-26.46%	-3.70	38.14%	3.83	30.02%	6.93	-45.38%	-7.57	75.41%	10.09
Ensemble	15.08%	2.41	-51.67%	-6.80	66.75%	6.42	30.78%	6.49	-60.54%	-10.03	91.32%	11.76

Word Cloud

- Positive cloud

  - 激励: Stimulus
  - 投资: Investment
  - 同比增长: Year-on-year growth
  - …

- Negative cloud

  - 减持: Shareholder sales
  - 亏损: Loss
  - 披露: Disclosure
  - …

LLM Signals' Information Content

- What's the specific content that LLMs discover from public news?
  - Tetlock et al. (2008) suggest news conveys important fundamental information about firms' future performance.
- Consider firms' future earnings surprises (SUE).

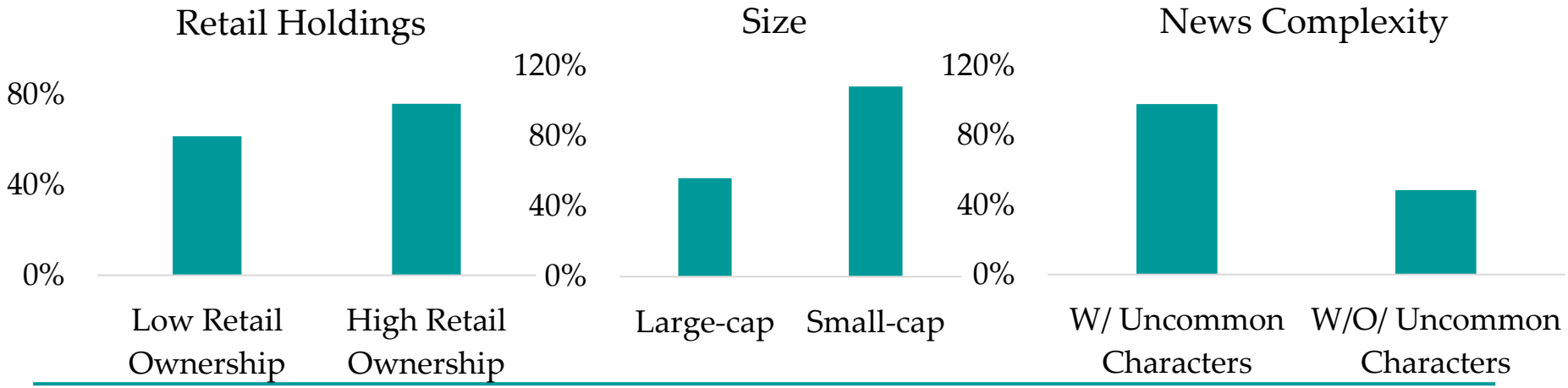
$$SUE_{i,t+1} = a_0 + a_1 LLM\ Signal_{i,t} + a_2' Controls_{i,t} + u_{i,t+1} \quad (3)$$

  - $SUE_{i,t} = \frac{\Delta_{i,t}}{\sigma(\Delta_i)}$, $\Delta_{i,t}$= year-over-year change in quarterly earnings.
- Positive a_1: The higher the signals, the higher the earnings surprises.

Dep.Var	Next-day SUE			
	News Tones		Return Forecast	
	Coef	*t*-Stat	*Coef*	*t*-Stat
LLM Signal	4.63***	8.65	135.31***	6.95
Controls	Yes		Yes	
Adj.R2	5.38%		6.81%	

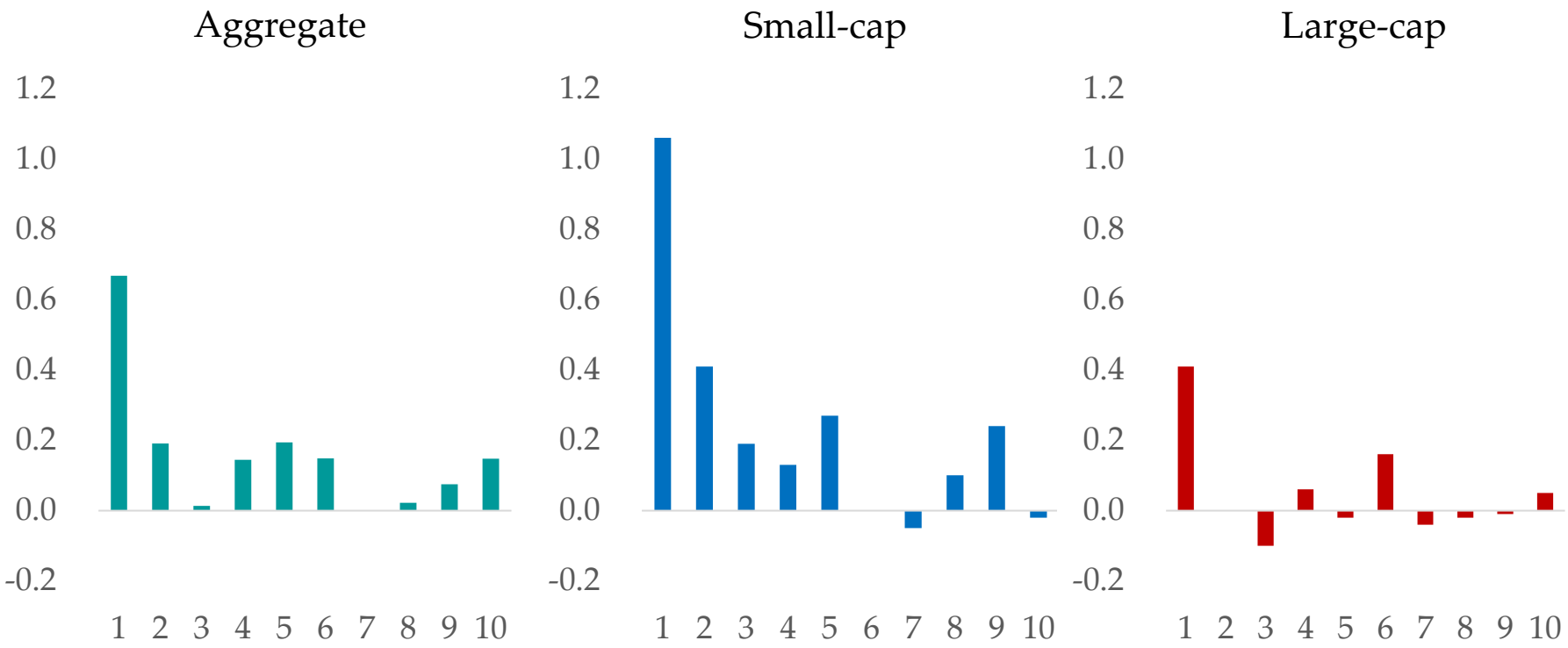
Cross-sectional Evidence

- Economic intuition: even if the fundamentals are revealed, market's ability to incorporate information is subject to information frictions.
- The benefits of LLM-based news processing are likely to be more pronounced when these frictions are higher, e.g.:
  - higher retail inattention creating more room for mispricing;
  - less transparent information environments;
  - news harder to understand.



Assimilation Speed

- How fast does news assimilate into prices?
  - 2 days on aggregate, slower (faster) for smaller (bigger) firms.



LLMs and Trading Dynamics

- When news becomes public, it is likely that sophisticated investors may correctly interpret and swiftly act upon it, whereas less sophisticated investors cannot.
- Consider four investor groups varying in trade sizes: small (<50k), medium (50~200k), large (200k~1 mil), extra-large (>1 mil) trades.
  - Trading direction (order imbalance): $Oib_{i,t}^{G} = \frac{Buy_{i,t}^{G} - Sell_{i,t}^{G}}{Buy_{i,t}^{G} + Sell_{i,t}^{G}}$.

Dep.Var	Next-day Oib(Small)		Next-day Oib(Medium)		Next-day Oib(Large)		Next-day Oib(ExtraLarge)	
LLM Signal	News tone	Return forecast	News tone	Return forecast	News tone	Return forecast	News tone	Return forecast
Coef	-4.62	-67.68	-0.43	112.31	4.43	190.82	6.49	145.93
t-Stat	-8.26	-5.08	-0.82	8.67	5.39	9.61	8.24	7.62
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Adj.R2	5.12%	5.09%	1.81%	1.83%	0.25%	0.24%	0.01%	0.00%

Transaction Costs

- After-fee performances: deduct a stamp fee of 10.0 bps (upon selling) and commission fee of 1.5 bps (upon both buying and selling) for every transaction.

			Long Leg				Long minus Short			
	Group	Holding	Alpha	*t*-Stat	Turnover	Transact Costs	Alpha	*t*-Stat	Turnover	Transact Costs
VW	Small cap	1-day	3.83%	0.49	95.26%	35.72%	49.13%	4.12	92.29%	34.61%
		5-day	16.58%	1.92	19.35%	7.26%	35.57%	4.92	19.05%	7.14%
		10-day	17.60%	2.36	9.70%	3.64%	22.73%	4.03	9.59%	3.60%
	Large cap	1-day	-9.69%	-1.40	89.78%	33.67%	-0.10%	-0.01	90.14%	33.80%
		5-day	5.35%	0.68	18.48%	6.93%	8.68%	1.22	18.66%	7.00%
		10-day	7.27%	1.07	9.31%	3.49%	7.30%	1.27	9.41%	3.53%
EW	Small cap	1-day	12.06%	1.50	95.15%	35.68%	52.11%	4.29	91.99%	34.50%
		5-day	22.47%	2.53	19.34%	7.25%	42.17%	5.26	19.00%	7.13%
		10-day	21.62%	2.82	9.70%	3.64%	27.78%	4.52	9.57%	3.59%
	Large cap	1-day	-4.00%	-0.79	92.28%	34.61%	15.57%	1.96	90.62%	33.98%
		5-day	7.58%	1.01	18.94%	7.10%	16.82%	3.03	18.79%	7.05%
		10-day	8.52%	1.26	9.52%	3.57%	11.50%	2.49	9.48%	3.56%

Conclusions

- Return prediction: using a comprehensive set of Chinese news articles, we find LLMs' signals can predict returns in China.

- LLMs signals contribute to price discovery process.
  - LLM signals contain information on future firms fundamentals.
  - Higher prediction when higher information frictions.
  - News assimilation speed: 2 days in general.
  - Different investors trade oppositely on LLM signals.

- AI could be a useful tool for helping investors process information, particularly for small retail investors and in emerging markets.