Can generative AI help identify peer firms?

Cao et al.(2025), Review of Accounting Studies

解读人：王梦涵

武汉大学金融系

20251019

Question

- How well **generative AI** can identify **peers** for a given focal firm?

Question

- How well **generative AI** can identify **peers** for a given focal firm?
- Two perspectives to answer this question
  - Accuracy
  - Usefulness

Why interesting?

- Peer identification is important for investors in
  - valuing a stock(eg: comparable company analysis, CCA)
  - evaluating corporate governance(eg: CEO pay)
  - predicting corporate events
- However, identifying peers require tremendous efforts and challenging
  - high cost of acquiring, processing, and integrating info
  - explosion of available data in past two decades

Why interesting?

- Generative AI can collect and aggregate data from a wide range of public sources
  - pre-trained on a large amount of public data–extensive knowledge
  - High efficiency- respond to a user request within seconds
  - freely available to the public: ChatGPT, Bard
- → Generative AI may be able to help identify peer firms

Contribution

- Literature on the impact of generative AI in the capital markets
  - Prior: a tool to summarize or classify user-provided content
  - Ext: focus on the role of generative AI as an information aggregator
- Literature on identifying peer firms
  - Prior(expert-driven apporoach): SIC-too large peer group; stale info; can't handle multi-industry firms
  - Prior(algorithm-driven approach):analyst coverage;the similarity of firms' business descriptions(TNIC)–intensive programming; only 1 type of data source
  - Ext: Generative AI can execute this task

Hyphothesis

- **How good is LLM peer identification?**
  - H1a: LLM peers and peers identified by human experts should overlap to some extent.
  - H1b: LLM peers and Peers identified and alternative established identification systems should overlap to some extent.
  - H1c: Returns of a focal firm and its LLM peer portfolio would have positive correlation
  - H1d: Accounting fundamentals correlation between a focal firm and its LLM peers should be positive
- How useful is this measure for investors/researchers...?
  - H2a: LLM peers are less biased in serving as compensation benchmarks than firm-selected peers
  - H2b: LLM peers can replicate findings from prior research

Data:LLM Peers

- Use the chat-bison-001 version of Bard and collected the LLM peers in July 2023.

We enter the following prompt in Bard to generate a focal firm's peers:

A firm's product market competitors refer to other companies that operate within the same market space, offer similar products or services, fulfill the same customer needs, and compete for market share. For the following tasks, please restrict your knowledge pool to the end of {year}. Among all publicly traded companies in the United States in the year of {year}, please list the major product market competitors of the firm, {cname}. The order of the list of major product market competitors should reflect product market closeness between the focal firm, {cname}, and its competitors, in a descending order.

Data: Sample

Panel A: Collection of peers from Google Bard

	Attrition	Remaining focal firm-years
Total number of focal firm-years in Compustat universe from 1981 to 2023		373,176
Exclude firm-years with error messages	(18,033)	355,143
Exclude firm-years without any fuzzy-matched peers in Compustat	(9,146)	345,997
Exclude firm-years whose identified peers are only themselves	(1,160)	344,837
Focal firm-years in the Bard collection (1981–2023)		344,837
LLM peers for the above focal firm-years (1981–2023)		2,639,157

step1: eg-"I could not provide answers to that question"

step2: Levenshtein distance/length of firm name <0.4

step 3

Panel B: Compustat matching and cleaning

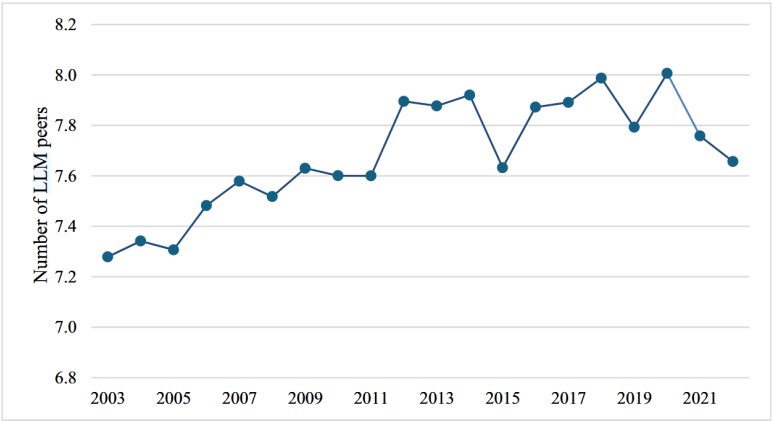
	Attrition	Remaining focal firm-years
Total number of focal firm-years		344,837
Exclude firm-years before 2003 or belonging to 2023	(183,041)	161,796
Exclude foreign firms, financial firms (SIC 6000–6999), and utility firms (SIC 4000–4999)	(84,038)	77,758
Exclude firm-years that miss income before extraordinary items, do not have positive book value of equity or cost of goods sold, or have stock price less than $3 at fiscal year-end or sales less than $10 million	(37,299)	40,459
Exclude firm-years with market capitalization less than $50 million	(2,592)	37,867
Focal firm-years in our main sample (2003–2022)		37,867
LLM peers for the above focal firm-years (2003–2022)		295,270

Wikipedia: founded in 200101 and reached 100,000 webpages in 2003.

Data: Descriptive statistics

- The mean value of the number of LLM peers for each firm size group

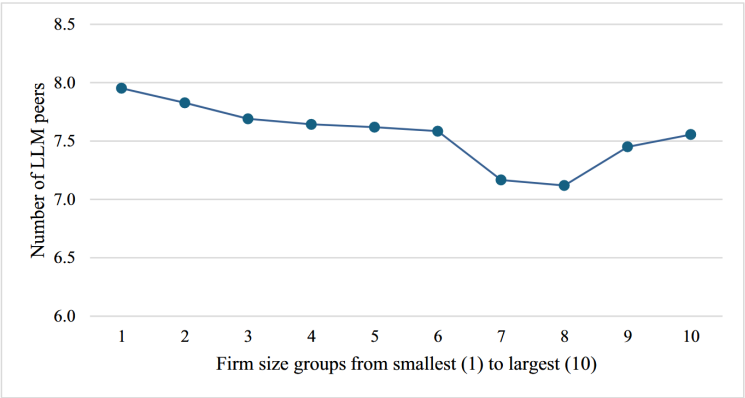
Panel B: Number of LLM peers for a focal firm by year



Data: Descriptive statistics

- The mean value of the number of LLM peers

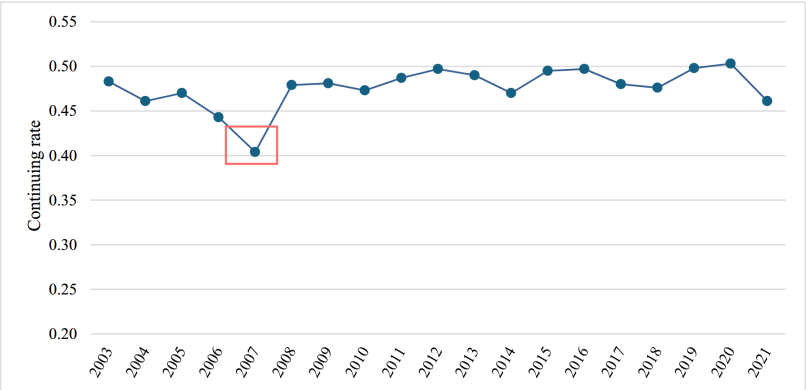
Panel D: Number of LLM peers for a focal firm-year by firm size group



Data: Descriptive statistics

- Continuing rate: the number of overlap peers between the two years divided by the number of peers in year t

Panel A: Continuing rate from current year to next year



Q1: How good is LLM peer identification?–Experts

- Peers identified by experts
  - 3 experts independently identify 3-5 peers for the 40 largest 2021 software firms
  - Average: 4.0 peers; Total: 159 peers
  - Pairwise overlap:**64.8%, 58.5% and 47.8%**
- Overlap between LLM peers and peers identified by experts: **47.8%,42.1%, and 36.5%**
- Overlap between TNIC peers and peers identified by experts: **10.1%, 10.1%, and 11.9%**
- → LLM peer identification contains meaningful signals.

Q1: How good is LLM peer identification?–Alternative systems

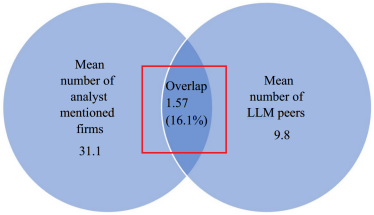
- 3 alternative peer identification systems
  - TNIC peers
  - analyst-mentioned firms
  - firm self-reported compensation benchmarking peers
- LLM peer VS TNIC peers:

Panel B: Subsamples

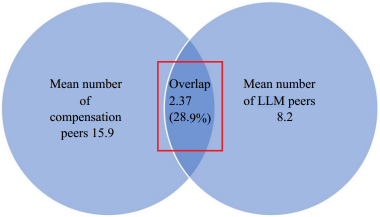
	LLM < TNIC	LLM = TNIC	LLM > TNIC
Number of focal firm-years	25,395	1,120	9,045
Mean number of LLM peers per focal firm	7.167	5.551	9.840
	(7)	(5)	(8)
Mean number of TNIC peers per focal firm	95.984	5.551	3.715
	(46)	(5)	(3)
Percentage of overlap between LLM and TNIC peers	28.4%	25.4%	36.6%

Q1: How good is LLM peer identification?–Alternative systems

Panel A: LLM peers versus analyst-mentioned peer firms



Panel B: LLM peers versus corporate self-reported compensation peers



Q1: How good is LLM peer identification?–Return

- Return correlation between a focal firm and its LLM peers during year $t+1$

$$Ret_{i,t+1,m} = \alpha_{i,t} + \beta_{i,t} PortRet_{i,t+1,m} + \varepsilon_{i,t+1,m}$$

  - $Ret_{i,t+1,m}$:firm i' s stock return in month m of calendar year $t+1$
  - $PortRet_{i,t+1,m}$: **equal-weighted** return of firm i' s peers in month m of year $t+1$
  - Perform a regression for each focal firm-year

Q1: How good is LLM peer identification?–Return

- Mean value of firm-year betas as well as the t-statistic

Panel A: Robust-regression estimation of firm-year betas

	(1)	(2)	(3)
Variable	$Ret_{t+1,m}$	$Ret_{t+1,m}$	$Ret_{t+1,m}$
LLM PortRet$_{t+1,m}$	0.660***		
	(10.62)		
TNIC PortRet$_{t+1,m}$		0.614***	
		(19.31)	
SIC PortRet$_{t+1,m}$			0.626***
			(18.17)

The difference between LLM Beta and TNIC Beta is 0.046, with a *t*-stat. of 4.93

- LLM peers have stronger return correlations with focal firms than do TNIC or SIC peers.

Q1: How good is LLM peer identification?–Accounting fundamentals

- Accounting fundamentals correlation between a focal firm and its LLM peers during year $t+1$

$$AFP_{i,t+1} = \alpha_t + \beta_t PortAFP_{i,t+1} + \varepsilon_{i,t+1},$$

  - AFP: Sales Growth(SG), gross profit margin (GPM), and profitability (ROA)
  - $AFP_{i,t+1}$:firm i' s accounting fundamentals performance for fiscal year $t+1$
  - $PortAFP_{i,t+1}$: **equal-weighted** peer firms' accounting fundamentals of year $t+1$

Q1: How good is LLM peer identification?–Accounting fundamentals

- Fama–MacBeth approach:

Panel A: Sales growth in the subsequent year (Fama–MacBeth coefficients and *t*-stat.)

	(1)	(2)	(3)
Variables	$Sales\ Growth_{t+1}$	$Sales\ Growth_{t+1}$	$Sales\ Growth_{t+1}$
$LLM\ PortSG_{t+1}$	0.263***		
	(6.35)		
$TNIC\ PortSG_{t+1}$		0.192***	
		(6.83)	
$SIC\ PortSG_{t+1}$			0.020***
			(6.50)

The difference between LLM Beta and TNIC Beta is 0.071, with a *t*-stat. of 2.32

Panel B: Gross profit margin for the subsequent year (Fama–MacBeth coefficients and *t*-stat.)

	(1)	(2)	(3)
Variables	GPM_{t+1}	GPM_{t+1}	GPM_{t+1}
$LLM\ PortGPM_{t+1}$	0.551***		
	(30.08)		
$TNIC\ PortGPM_{t+1}$		0.475***	
		(11.24)	
$SIC\ PortGPM_{t+1}$			0.473***
			(22.26)

The difference between LLM Beta and TNIC Beta is 0.076, with a *t*-stat. of 1.96

Panel C: Profitability for the subsequent year (Fama–MacBeth coefficients and *t*-stat.)

	(1)	(2)	(3)
Variables	ROA_{t+1}	ROA_{t+1}	ROA_{t+1}
$LLM\ PortROA_{t+1}$	0.126***		
	(11.69)		
$TNIC\ PortROA_{t+1}$		0.129***	
		(10.99)	
$SIC\ PortROA_{t+1}$			0.035***
			(5.75)

The difference between LLM Beta and TNIC Beta is −0.003, with a *t*-stat of −0.14

Q2:Usefulness of LLM peers in compensation benchmarking

- Peer-selection pay bias: Firms tend to select peers with higher CEO pay as compensation benchmarks
- **whether LLM peers result in less bias?**
- Data: ISS Incentive Lab(2006-2021), 14103 focal firms-years
- Method: peer-choice probit model(Faulkender and Yang, 2010;Albuquerque et al. , 2013)
  - For each firm-selected peer, we identify the firm (j) with the closest propensity score as the true peer
  - Note: For focal-firm/peer-firm pairs here, the peer candidate can be **any firm listed as a peer** or be **any other focal firm** in the sample year
- LLM bias = (Median CEO pay of LLM peers- Median CEO pay of true peers)/ Median CEO pay of true peers

Q2:Usefulness of LLM peers in compensation benchmarking

- Peer-choice probit model

Panel A: Probit model on a focal firm's selection of compensation peer

Variables	Coefficient	*z*-stat
Two-digit industry	1.101***	314.64
Three-digit industry	0.514***	121.98
Sales within 50%–200%	0.505***	209.53
Assets within 50%–200%	0.298***	128.62
Market cap within 50%–200%	0.178***	80.70
S&P Mid Cap 400	0.233***	62.22
S&P 500	0.343***	137.05
CEO is chair	0.070***	27.89
CEO is not chair	0.003	1.00
Number of peers	0.005***	160.02
Talent flows	1.077***	140.74
Pseudo-R^2	26.6%	
N	19,032,429	

Q2:Usefulness of LLM peers in compensation benchmarking

Size decile	LLM peers	Firm-selected peers	Difference	*T*-test	Wilcoxon test
1	0.159	0.172	−0.013	−0.90	−2.24**
2	0.120	0.145	−0.025	−2.40**	−3.66***
3	0.070	0.120	−0.050	−4.85***	−5.36***
4	0.097	0.120	−0.023	−2.20**	−2.37**
5	0.132	0.162	−0.030	−3.60***	−2.93***
6	0.141	0.163	−0.022	−2.15**	−2.28**
7	0.088	0.119	−0.031	−3.35***	−3.08***
8	0.073	0.087	−0.014	−1.85*	−0.26
9	0.097	0.106	−0.009	−1.20	0.14
10	0.103	0.141	−0.038	−3.40***	−1.51
Full sample	0.108	0.133	−0.025	−7.95***	−2.36**

- LLM peers can be useful for investors in evaluating CEO pay

Q2:Usefulness of LLM peers in testing research hypotheses

- Brown et al. (2024): analysts' information environment for a firm deteriorates after its peer firm is delisted following a merger or acquisition
- Data: Focal firm-years
  - with LLM, TNIC, and GICS peers at the end of year t
  - any of its LLM peers is delisted in year t +1 following a merger or acquisition
- Design

$$ForecastError/ForecastDispersion_{i,t} = \alpha + \beta PostMerger_{i,t} + \sum_k \gamma_k X_{i,t,k} + \epsilon$$

  - $PostMerger = 0$at pre-event year$t$,$PostMerger = 1$at post-event year$t+1$

Q2:Usefulness of LLM peers in testing research hypotheses

Table 10 Informational Effects of Peers being Delisted after a Merger or Acquisition

	Forecast Error			*Forecast Dispersion*		
	LLM	TNIC	GICS	LLM	TNIC	GICS
PostMerger	0.093***	0.050*	0.019	0.001**	0.000	−0.001**
	(3.02)	(1.69)	(0.72)	(2.28)	(0.94)	(−2.12)
Size	−0.433***	−0.434***	−0.418***	−0.004***	−0.002***	−0.003***
	(−33.69)	(−31.77)	(−64.24)	(−26.01)	(−20.28)	(−45.97)
MtoB	−0.012***	−0.016***	−0.013***	−0.000***	−0.000***	−0.000***
	(−4.75)	(−6.59)	(−9.87)	(−3.78)	(−8.54)	(−9.26)
Numest	−0.058***	−0.048***	−0.061***	−0.000***	−0.000***	−0.000***
	(−19.09)	(−16.72)	(−38.02)	(−9.83)	(−11.84)	(−20.62)
D_Diff	0.004***	0.005***	0.003***	0.000	0.000***	−0.000
	(2.88)	(5.75)	(5.51)	(−0.89)	(3.27)	(−1.37)
Vol	0.412***	0.427***	0.409***	0.004***	0.003***	0.004***
	(28.33)	(29.13)	(52.66)	(25.68)	(23.98)	(46.58)
Ret	0.235***	0.325***	0.240***	0.002***	0.002***	0.002***
	(13.25)	(15.70)	(29.11)	(13.24)	(10.36)	(27.22)
R^2	26.2%	27.0%	27.0%	15.3%	14.1%	14.7%
N	6,969	7,065	27,046	6,969	7,065	27,046

Ideas

- In China: less information disclosure; less Chinese corpus used for LLM(ChatGpt) training
- Compare the performance of different LLMs
- Use LLM as information aggregator
  - Risk Rating: ESG、Credit risk
  - ...