Global Business Networks

作者：Christian Breitung , Sebastian Müller　　阅读：程心烨

1. What are the research questions ?

• How to construct global business networks that accurately capture firms' economic links using AI and advanced embedding models?

• Can these global business networks effectively predict the lead-lag effect in global stock returns and target firms in M&A deals?

• How to mitigate the look-ahead bias introduced by LLMs when using embedding models for historical business description analysis?

• How to fine-tune language models to distinguish between different types of business relations in the networks?

2. Why are the research questions interesting?

• Traditional industry classifications fail to reflect within-sector heterogeneity, making it hard to identify real firm links.

• Existing business network research mostly focuses on the U.S. market; a global network fills the need for cross-country economic link analysis.

• Solving LLMs' look-ahead bias improves the reliability of AI-based textual analysis.

• Distinguishing relation types helps tailor research, adding practical value.

3. **What is the paper's contribution?**

• Pioneering Global Time-Varying Business Networks.First to construct time-varying global networks with OpenAI embedding models and an open-source model.

• Validated Practical Utility. It systematically demonstrates the networks' value in two core financial scenarios—generating significant seven-factor alphas in stock lead-lag strategies and accurately predicting M&A targets.

• Shows fine-tuning an open-source model can distinguish relation types with 85.73% accuracy.

4. **What hypotheses are tested in the paper? List them explicitly.**

• H1: Global business networks built with embedding models better identify value-relevant economic links.

• H2: Masking firm-specific details reduces LLM look-ahead bias, leading to more reliable network performance in lead-lag effect and M&A prediction.

• H3: Fine-tuning a language model with actual business relations enables accurate classification of competitor, supplier, and customer links.

(a) Do these hypotheses follow from and answer the research questions?

• Yes.

(b) Do these hypotheses follow from theory or are they otherwise adequately developed?

• Yes. Hypotheses follow from finance theory and prior literature.

5. Sample: comment on the appropriateness of the sample selection procedures.

• The sample covers 68,402 stocks (67 countries, 2000–2021), including U.S. and international data—comprehensive for global network research.

• Limitations noted ,e.g., underrepresentation of small stocks, lower coverage in Asian/African markets due to language barriers, are transparent and do not undermine the core analysis .

6. Dependent and independent variables: comment on the appropriateness of variable definition and measurement.

• Independent variables:Textual similarity, cosine similarity of embeddings. Masked similarity.

• Dependent variables:Lead-lag effect. M&A likelihood, Dummy variable.

• All variables align with research goals and use established measurement methods in finance/NLP.

7. Regression/prediction model specification: comment on the appropriateness of the regression/ prediction model specification.

• M&A prediction: Logistic regression controlling for industry, country, and firm fundamentals .Relation classification: Multiclass Longformer model with oversampling for balanced classes.

• All models are theoretically grounded and robust to alternative specs.

8. What difficulties arise in drawing inferences from the empirical work?

•Look-ahead bias residual\Small stock underrepresentation\Poor coverage in non-English markets means inferences about these regions are less reliable.

9. Describe at least one publishable and feasible extension of this research.

• Fix GPT-3 tokenizer gaps (10% Chinese firm coverage) with multilingual models + industry dictionaries.

Can generative AI help identify peer firms?

作者：Yi Cao, et al.　　阅读：程心烨

1. What are the research questions ?

• Can generative AI (LLMs) effectively identify product market competitors or peer firms for a given focal firm?

• How does the performance of LLM-identified peers compare with that of peers identified by human experts and established peer identification systems?

• Are LLM-identified peers useful in practical scenarios?

2. Why are the research questions interesting?

• Peer identification is crucial for investors but traditionally complex, time-consuming, and resource-intensive. Generative AI's rise brings a potential low-cost, efficient solution.

• Existing peer identification methods have limitations. Testing LLM's ability to overcome these flaws fills a gap in current research.

• Verifying LLM's practical utility in real-world scenarios helps bridge the gap between AI technology and capital market applications, providing actionable insights for investors, regulators, and researchers.

3. **What is the paper's contribution?**

• It join the emerging literature on generative AI's impact on capital markets and distinguishing itself by focusing on information aggregation.

• It proposes a new, convenient peer identification approach. Unlike traditional methods, LLM-based identification is accessible to anyone with internet access, reducing information costs.

•It demonstrates LLM's effectiveness. It also proves LLM's utility in compensation benchmarking and hypothesis testing.

4. **What hypotheses are tested in the paper? List them explicitly.**

• H1: LLM-identified peers have a higher overlap with peers identified by human experts and established systems than random chance.

•H2: LLM-identified peers have stronger correlations with the focal firm in subsequent-year stock returns, sales growth, and gross profit margin than peers identified by TNIC or SIC.

• H3: LLM-identified peers exhibit higher homogeneity than peers identified by TNIC or SIC.

• H4: LLM-identified peers have less bias as compensation benchmarks than firm-selected peers.

• H5: Using LLM-identified peers can effectively support researchers' hypothesis testing .

(a) Do these hypotheses follow from and answer the research questions?

• Yes. The hypotheses directly address whether LLM identifies peers effectively (H1-3) and its practical utility (H 4-5).

(b) Do these hypotheses follow from theory or are they otherwise adequately developed?

• Yes. The hypotheses are grounded in theoretical logic.Generative AI's information aggregation ability implies it should identify more accurate peers (H1-3); LLM's objectivity suggests less bias in compensation benchmarking (H4); and LLM's accuracy supports reliable hypothesis testing (H5).

5. Sample: comment on the appropriateness of the sample selection procedures.

• The sample selection is highly appropriate. The main sample covers 2003-2022. 2003 is chosen because Wikipedia reached 100,000 pages then, ensuring sufficient training data; 2023 is excluded

due to incomplete data.Foreign firms, financial/utility firms are excluded; small firms with missing data or low market value are excluded to ensure data reliability and reduce noise from abnormal observations.Align LLM peers with Compustat firms, minimizing errors from name mismatches.

6. Dependent and independent variables: comment on the appropriateness of variable definition and measurement.

• independent variables are measured by number/rank or per literature. Dependent variables use Beta, sales growth etc. , and absolute deviations. All align with literature, objective and reliable.

7. Regression/prediction model specification: comment on the appropriateness of the regression/ prediction model specification.

• The paper's models fit research goals well. Time-series regression for return correlation, Fama-MacBeth cross-sectional regression for accounting correlation, non-regression tests for homogeneity, all consistent with literature.

8. What difficulties arise in drawing inferences from the empirical work?

• LLM's "black box" problem.This limits inferences about the mechanism of LLM's peer identification.

• Potential sample selection bias. The main sample excludes small firms due to data constraints, restricting the conclusion's external validity.

• The sample only includes U.S. publicly listed firms. LLM's training data for non-U.S. markets may be insufficient.

9. Describe at least one publishable and feasible extension of this research.

• Exploring the effectiveness of generative AI in identifying peers for firms in emerging markets and analyzing the impact of local data availability on LLM performance.