# Zhijin Guo

zhijin.guo97@gmail.com
https://eng.ox.ac.uk/people/zhijin-guo/
https://zhijinguo.github.io/

**Visa Status:** Global Talent Visa holder (no sponsorship required) — Available to start work immediately

## Summary

Postdoctoral Researcher at the University of Oxford with 4+ years' experience in fine-tuning large language models, embedding relational data, constructing knowledge graphs, developing retrieval-augmented generation systems, and interpreting model behavior. Proven ability to adapt machine-learning and NLP techniques from social-network and recommender domains into systematic-trading signal frameworks and bespoke quality-control metrics for quantitative finance.

## Work Experience

| | |
|---|---|
| Present | **Postdoctoral Researcher** Natural Language Processing, **University of Oxford**<br>Supervisor: *Janet B. Pierrehumbert, Xiaowen Dong* |
| | **Quantitative Research Consultant** LLMs for Finance, **Predictive Labs Ltd** |
| 2023 – 2024 | **Research Assistant** Twitter Data Analysis, **University of Bristol** |
| 2021 – 2023 | **Teaching Assistant** Introduction to Data Analytics (NLP), **University of Bristol** |

## Skills

| | |
|---|---|
| Programming & Data Science | Python (6+ years), Java (4 years) |
| Machine Learning / NLP / LLM | Training & fine-tuning Transformers and encoder-decoder models; open-source LLMs (LLaMA, SBERT, DeepSeek, ALBERT) |
| Knowledge Graph & RAG | Designing Retrieval-Augmented Generation workflows with embedding models, vector search and Knowledge Graph construction |
| Libraries & Frameworks | Hugging Face Transformers; SBERT; spaCy; NLTK ; Pytorch; DGL |
| Unstructured Data Processing | Twitter, Reddit, emails and other relevant text/graph data structure |
| API Application | Data extraction via Twitter API, Reddit API, and various databases |
| Version Control & Data Access | Git workflows and handling structured (SQL) and semi-structured (JSON/CSV/TXT) data |

## Education

| | |
|---|---|
| 2021 – 2024 | **Ph.D. Computer Science & Natural Language Processing, University of Bristol**<br>Supervisor: *Nello Cristianini, Martha Lewis, Edwin Simpson*<br>Thesis: *Embedding Relational Data: Analysis, Methods, Applications* |
| 2019 – 2020 | **M.Sc. Advanced Computer Science, The University of Sheffield**<br>Thesis: *A Tool to Correct Errors Made by Grammar Inference Algorithm.* |
| 2015 – 2019 | **B.Sc. Software Engineering, Northeastern University**<br>Thesis: *Design and Implementation of ERP System for Iron and Steel Enterprises.* |

## Selected Projects

2025 **QuantNLP: Turning Financial Text into Trading Signals**

- Real-time news aggregation from APIs like Yahoo Finance, Reuters, Polygon, etc.
- Alpha generation via automated feature discovery and predictive modeling, with a strong focus on NLP for sentiment analysis, event detection, and topic modeling.
- Market backtesting with real data.

2024-2025 **Bridging Social Structure and Discourse**

- **Challenge**: Reddit communities were exhibiting escalating **polarization**, making proactive moderation impossible and increasing user churn.
- **Action**: Built a hybrid **RAG** pipeline that (1) retrieves user-language features via embeddings, (2) fine-tunes **LLAMA** and **BERT** models (using Hugging Face Transformers) for sentiment and persuasion-signal prediction, and (3) trains a continuous-time **GNN** to model interaction dynamics and assign continuously-changing polarization weights.
- **Result**: Forecast community fragmentation weeks ahead with 70% accuracy, analyzed and identified the top linguistic and interaction factors driving polarization, enabling moderators to implement targeted interventions.
- **Finance Application**: Emulated rolling-window retraining and signal forecasting workflows common in quantitative trading, with QA metrics like rolling MAE and drift detection to ensure robust next-period market-prediction signals.

**Quantifying Compositionality in Data Embedding**

- **Challenge**: State-of-the-art **Transformer** and graph models lacked measurable limits on context-driven meaning shifts, undermining trust in novel expression generalization.
- **Action**: Designed a two-step evaluation: (1) applied Canonical Correlation Analysis to quantify linear alignment between known entity attributes and their embeddings; (2) reconstructed embeddings for unseen attribute combinations across **SBERT**, **GPT**, **LLAMA**, and **Knowledge Graph** embeddings (TransE/DistMult), and measured L2 loss, cosine similarity, and retrieval accuracy.
- **Result**: Demonstrated additive compositionality during training, with improved generalization in Multi-BERT and a 1.5× rise in attribute–embedding correlation in graph models., and tracked that deeper transformer layers peaked at 94% compositional signal before tail-off, guiding optimal layer selection for downstream tasks.
- **Finance Application**: Developed rigorous embedding-quality metrics analogous to quant-finance signal validation, guiding model-selection and improving reliability of text-driven systematic strategy inputs.

## Selected Projects (continued)

2023-2024 **Medical Influencer Social Network Analysis**

- **Challenge**: Health misinformation spread unchecked during COVID-19, complicating public-trust efforts by authorities.
- **Action**: Compiled tweets from the top 100 medical influencers by "Influencer Score," then built a **few-shot**, **multi-label** classifier (50 classes) by fine-tuning **ALBERT**; used **BERTopic** for thematic extraction.
- **Result**: Achieved 70% F1 and a 15% boost in multi-label classification, uncovered the top misinformation topics and incorporated them into a dynamic knowledge graph to reveal influencer–audience discourse networks.
- **Finance Application**: Applied LLM-augmented knowledge-graph and RAG pipelines to financial texts, enabling discovery of hidden supply-chain links, co-investor networks, and merger signals for event-driven trading strategies.

2022 **EXTRACT: Explainable Transparent Control of Bias in Embeddings**

- **Challenge**: **Knowledge-graph** embeddings risked leaking protected attributes (e.g., gender, age, occupation) from behavioral data, raising **privacy** and **fairness** concerns.
- **Action**: Engineered a **vector-based knowledge base** by constructing and embedding a knowledge graphs. Developed EXTRACT, a suite that applies CCA to pinpoint bias-leakage sources, decomposes embeddings into private-attribute vectors via linear-system solving, and integrates four transparent mitigation methods to strip unwanted signals without degrading model utility.
- **Result**: Demonstrating robust recommending performance alongside bias control, highlighting the trade-off between accuracy and privacy.
- **Finance Application**: Adapted bias-mitigation techniques to financial-entity embeddings, ensuring fair, regulatory-compliant representations for risk models and portfolio construction without sacrificing signal fidelity.

## Selected Publications

### Bibliography


1. Z. Guo, E. Simpson, and R. Bernardi, "Medfluencer: A network representation of medical influencers' identities and discourse on social media," in *epiDAMIK 2024: The 7th International Workshop on Epidemiology meets Data Mining and Knowledge Discovery at KDD 2024*.
2. Z. Guo, Z. Li, B. Tyler, X. Dong, and J. Pierrehumbert, "Bridging social structure and discourse," Preparing to submit to the ACL Rolling Review (ARR), 2025.
3. Z. Guo, C. Xue, Z. Xu, *et al.*, "Quantifying compositionality in classic and state-of-the-art embeddings," Under review by the the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2025.
4. Z. Guo, Z. Xu, M. Lewis, and N. Cristianini, "Extract: Explainable transparent control of bias in embeddings," in *AEQUITAS 2023: AEQUITAS 2023 First AEQUITAS Workshop on Fairness and Bias in AI| co-located with ECAI 2023*, 2023.
5. Z. Xu, Z. Guo, and N. Cristianini, "On compositionality in data embedding," in *International Symposium on Intelligent Data Analysis*, Springer, 2023, pp. 484–496.