GIOIA ZHENG

M.Sc. Computer Science · NLP Evaluation · Information Retrieval · Reproducible ML
Darmstadt, Germany · gioia.zheng.stud@gmail.com · +39 377 851 8303 · github.com/GioiaZheng · gioiazheng.github.io

Research-oriented computer science student building reproducible NLP and retrieval evaluation systems in Python. Experience with BM25, cross-encoder reranking, Transformers, statistical evaluation, failure analysis, and research engineering.

EDUCATION

TU Darmstadt
M.Sc. Computer Science · Enrolled, Winter Semester 2026/27

Sapienza University of Rome
B.Sc. Applied Computer Science and Artificial Intelligence · Expected Dec 2026

RESEARCH AND ENGINEERING EXPERIENCE

Undergraduate Thesis Research — RAG Pipeline Evaluation
Sapienza University of Rome · Aug 2026–Present
Supervisor: Prof. Marco Raoul Marini · Co-tutor: Valerio Venanzi

- Evaluated retrieve-rerank-generate pipelines across MS MARCO, TREC-DL, SciFact, and NFCorpus.
- Found 72/323 NFCorpus queries with no relevant document in the top 100, including 48 still unresolved at depth 1,000; distinguished depth-recoverable misses from persistent retrieval failures.
- Compared ranking improvements with fixed candidate recall to identify when reranking cannot recover missing evidence.

NLP Algorithm Engineer Intern
Microsoft · Remote · Feb–May 2026

- Implemented BM25 retrieval and cross-encoder reranking for an MS MARCO question-answering pipeline in Python.
- Extended evaluation to TREC-DL 2019/2020 and built reproducible experiment, metric-validation, and query-level analysis workflows.

SELECTED RESEARCH PROJECTS

msmarco-genqa Python · BM25 · Cross-Encoder · T5

- Built a reproducible retrieve-rerank-generate pipeline over 8.8M MS MARCO passages; cross-encoder reranking improved TREC-DL 2019 MRR@10 from 0.5471 to 0.8787 and nDCG@10 from 0.4239 to 0.7210.
- Added paired-bootstrap confidence intervals, grounding audits, automated tests, CI checks, manifests, and output hashes.

RAG Observatory Python · SciFact · Trace Diagnostics

- Built a local-first trace framework for retrieved evidence, selected context, generated answers, evaluation signals, provenance, and failure labels.
- Demonstrated on a controlled 20-query SciFact study that increasing retrieval depth from 1 to 5 changed retrieval hits from 13/20 to 14/20 and retrieval-noise labels from 7/20 to 20/20.

OPEN SOURCE CONTRIBUTIONS

- Merged MTEB contributions: symmetric STS pair counting (#4958), GermanGovService v2 (#5323), and evaluation-request workflow documentation (#4861).
- Merged fixes for cached M-BEIR instruction lookup in Pyserini (#2655) and model language-scope display in the MTEB leaderboard frontend (#31).

TECHNICAL SKILLS

Programming: Python, SQL, Bash, Git; working knowledge of Go, JavaScript, and Java
NLP and ML: PyTorch, Transformers, sentence-transformers, cross-encoders, T5, scikit-learn, BM25, FAISS
Evaluation: MRR, nDCG, Recall, paired bootstrap, confidence intervals, ablations, retrieval-depth and failure analysis
Engineering: Linux, Docker, GitHub Actions, pytest, SQLite, REST APIs, reproducible configurations and CI
Languages: Chinese (native), Italian (bilingual), English (professional), German (beginner)