Na Li Ph.D.

n.li1@elsevier.com OR **nali.cosmic@gmail.com**
(+31) 687176834
Google Scholar LinkedIn GitHub

Interests

Information Retrieval | **NLP** | **Agentic AI** | **Machine Learning** | **Data quality control**

Work experiences

Postdoctoral Researcher | Discovery Lab, Elsevier 2025 – present

- Developed an cost-efficient **LLM-based annotation pipeline** for content alignment between scientific articles and source code.
- Built scalable **text processing pipelines** for automated transformation, parsing and segmentation for articles collected from Elsevier journals.
- Applied **advanced large language models (e.g., GPT, BERT, Roberta)** to produce high-quality relevance annotations for paper–code pairs.

Doctoral Researcher | University of Amsterdam 2020 – 2025

- Built and deployed a full-stack search engine, including a **heterogeneous data integration pipeline**, a vector indexer, and a multi-type retriever to enable cross-source search. `https://github.com/nali001/research_asset_search`
- Developed **natural language-based code retrieval methods** using **dense retrieval models and LLMs**. `https://github.com/nali001/notebook_search_docker`
- Developed auxiliary NLP functionalities such as **query reformulation and document summarization** to enhance searching system's usability.
- Designed and implemented an **active learning-based data quality control framework** to improve annotation accuracy and reduce labeling costs in large-scale datasets. `https://github.com/nali001/al_dqc`
- Developed algorithms to address **data imbalance** and **limited labeled data** issues, enhancing model robustness and reliability.

Education

University of Amsterdam Amsterdam, Netherlands
Ph.D. in Computer Science 2020 - 2025

Beihang University Beijing, China
M.S. in Information and Communication Engineering 2017 - 2020

Beihang University Beijing, China
B.S. in Electronic and Information Engineering 2013 - 2017

Skills

Core Strengths: Research and analytical skills in ML, DL, LLM, IR and NLP; Full-stack software development experience in transforming state-of-the-art NLP and LLM research into real-world web applications; Experience in the data engineering pipeline, including data collection, cleaning, quality control, and labeling.

Team work: Work experience in Elsevier, Amsterdam office, with close collaboration with data scientists and software engineers in Elsevier.

NLP, LLM & GenAI: Language models (BERT, SciBERT, CodeBERT); Named entity recognition (DyGIE, DyGIE++, SciDeBERTa). GenAI models (GPT, Claude, Llama).

Programming: Python, JS, C, Matlab.

DL & ML models: Transformer, LSTM, CNN, MLP, XGBoost, Decision tree, Bagging, Boosting, KNN, LR.

ML & DL tools: PyTorch, scikit-learn, transformers, Pandas, Numpy, spaCy, networkx.

Agentic AI: Langchain, prompt engineering.

Full-stack software: HTML/CSS, React, Django, Git, Docker, cloud server deployment.

Data Management: Data crawling, indexing, vector database (FAISS), database (PostgreSQL, Elasticsearch), metadata consolidation.