Keyu Wang

(+49) 1525 6272814 | keyu.wang@student.uni-tuebingen.de | website | Google Scholar | GitHub | LinkedIn

Education

University of Tuebingen — Tuebingen, Germany
M.Sc. in Machine Learning — *Oct. 2024 – Present*

Current GPA: 1.6 (German scale; 1.0 = best, 5.0 = worst)
Relevant Coursework: Non-convex Optimization for Deep Learning, Statistical Machine Learning, Probabilistic Machine Learning, The Science of Machine Learning Benchmarks, etc.

Southeast University — Nanjing, China
B.Eng. in Artificial Intelligence — *Sep. 2020 – Jun. 2024*

Cumulative GPA: 3.79/4.0
Relevant Coursework: Optimization, Probability Statistics & Stochastic Processes, Distributed Systems, Machine Learning, Natural Language Processing, Reinforcement Learning, etc.

Publications

[1] **Keyu Wang***, Tian Lyu*, Guinan Su, Lu Yin, Marco Canini, Jonas Geiping, and Shiwei Liu. "When Fewer Layers Break More Chains: Layer Pruning Harms Test-time Scaling in LLMs." **COLM 2026**.

[2] Di He, Songjun Tu, **Keyu Wang**, Lu Yin, and Shiwei Liu. "One LR Doesn't Fit All: Heavy-Tail Guided Layerwise Learning Rates for LLMs." **ICML 2026**.

[3] Xinyuan Song, **Keyu Wang**, Pengxiang Li, Lu Yin, and Shiwei Liu. "Demystifying the Roles of LLM Layers in Retrieval, Knowledge, and Reasoning." **Technical Report 2025; ICASSP 2026**.

[4] **Keyu Wang**, Guilin Qi, Jiaqi Li, and Songlin Zhai. "Can Large Language Models Understand DL-Lite Ontologies? An Empirical Study." **Findings of EMNLP 2024.**

[5] **Keyu Wang**, Site Li, Jiaye Li, Guilin Qi, and Qiu Ji. "An Embedding-based Approach to Inconsistency-tolerant Reasoning with Inconsistent Ontologies." **IJCKG 2023.**

Research Experience

Max Planck Institute for Intelligent Systems & ELLIS Institute — Tuebingen, Germany
Research Assistant, WEI Lab | Advisor: Dr. Shiwei Liu — *Mar. 2025 – Present*

Measuring Residual Connections for Computational Depth: Lead & first author; paper in prep.

- Investigated how layer normalization and residual connection design affects computational depth.
- Conducted controlled pretraining across depth-width configurations and normalization, residual, and depth-mixing architectures, including Pre-LN, Peri-LN, LNS, KEEL, AttnRes, HC, and mHC, etc.
- Developed a unified analysis framework using controlled layer perturbations and representation dynamics to diagnose depth utilization and explain the scaling behaviors.

Layer Pruning Harms Test-time Scaling in LLMs: Lead & first author; **COLM 2026**

- Investigated how layer pruning affects reasoning under sequential and parallel test-time scaling, and whether supervised fine-tuning can recover the lost capabilities.
- Highlighted that even mild pruning degrades test-time scaling by inducing recurring loops, reducing solution diversity, and weakening self-correction; standard SFT only partially recovers these losses.

Layerwise Learning Rates for LLMs: Co-author; **ICML 2026**

- Developed an adaptive pretraining scheme with layerwise learning rates to promote balanced training across layers, faster convergence with improved generalization. *Contribution*: pretraining experiments.

Demystifying the Roles of LLM Layers: Co-author; **Report 2025**

- Conducted a systematic analysis of layer roles in retrieval, knowledge, and reasoning across evaluation protocols, task categories, and model families. *Contribution*: evaluation methodology design.

Max Planck Institute for Intelligent Systems & ELLIS Institute Tuebingen, Germany
Master's Student, SEAL Lab | Advisor: Dr. Jonas Geiping *Nov. 2025 – May 2026*

Hierarchical KV Cache Compression for Efficient Reasoning in LLMs Lead

- Developed a hierarchical KV-cache compression method that recursively summarizes token chunks into multi-level representations under a fixed cache budget while preserving performance on commonsense and knowledge-reasoning tasks.

Southeast University Nanjing, China
Undergraduate Student, COIN Lab | Advisor: Prof. Dr. Guilin Qi *Feb. 2022 – Jun. 2024*

Evaluating LLMs with Description Logics: Lead & first author; **EMNLP 2024 Findings**

- Evaluated LLMs' understanding of Description Logics across six tasks: syntax checking, instance checking, subsumption checking, functionality checking, query answering, and satisfiability checking.
- Found that LLMs capture formal syntax and the model-theoretic semantics of concepts and roles, but struggle with negative inclusions and deduction over large instance sets.

Inconsistency-tolerant Reasoning in Description Logics: Lead & first author; **IJCKG 2023**

- Proposed a monotonic selection relation for ranking maximal consistent subsets of inconsistent belief bases, extending from propositional logic to description logics using semantic embeddings.
- Proved key logical properties required for rational inference and empirically demonstrated an improvement of more than 10% in reasoning performance.

Service

Conference Reviewer: COLM 2026, CPAL 2026

Skills

Programming: Python (PyTorch), C/C++.

Expertise: LLM Pretraining, Post-training (SFT and RL), Model Evaluation, Efficient Inference

Languages: Chinese (Native), English (Fluent, IELTS 7.5/C1).