Friedrich-Alexander-Universität
Faculty of Engineering

Four papers accepted at EMNLP 2026

Intelligent Language Systems Lab @ FAU Erlangen-Nürnberg

MAIN CONFERENCE · CL JOURNAL

Instructional Text Across Disciplines: A Survey of Representations, Downstream Tasks, and Open Challenges Toward Capable AI Agents

Abdulfattah Safa, Tamta Kapanadze, Arda Uzunoğlu, ***Gözde Gül Şahin***

What still stands between today's LLMs and robust agent pipelines? A systematic review of 181 papers on instructional and procedural text across NLP, robotics, computer vision and business automation, mapping the representations, resources and downstream tasks of the field, and the open challenges that remain.

Computational Linguistics 52(2): 759–829 · doi.org/10.1162/COLI.a.616

MAIN CONFERENCE

Are Non-English Papers Reviewed Fairly? Language-of-Study Bias in NLP Peer Reviews

Ehsan Barkhordar, Abdulfattah Safa, Verena Blaschke, Erika Lombart, Marie-Catherine de Marneffe, ***Gözde Gül Şahin***

Across 15,645 reviews from major NLP venues, papers focused on non-English languages face up to 40x higher language-of-study bias rates than English-focused work. We release LOBSTER, a human-annotated dataset for this form of bias, and show that LLMs detect it at 87.4 macro F1, making automated screening a realistic step toward fairer peer review.

arxiv.org/abs/2604.07119

SYSTEM DEMONSTRATIONS

TreeThink: A Modular Tree Search Library for Mathematical Reasoning with LLMs

Burak S. Akbudak, Zeynel A. Uluşan, Can S. Erer, ***Gözde Gül Şahin***

An open-source Python library that combines LLM generation with symbolic tree search for neural theorem proving. Search algorithms, policies, evaluators and environments are decoupled into reusable components, with unified REPL clients for Lean 4, Rocq and Isabelle/HOL and fully asynchronous, vLLM-backed execution.

arxiv.org/abs/2607.11258 · github.com/GGLAB-KU/treethink · pypi.org/project/treethink

INSIGHTS FROM NEGATIVE RESULTS IN NLP · WORKSHOP

No Optimal Language Set Exists for Multilingual Instruction Tuning: Insights from a Linguistically-Informed Study

Gürkan Soykan, ***Gözde Gül Şahin***

Does it pay off to pick instruction-tuning languages by typology, family or geography? Across models and tasks, no linguistically-informed selection strategy consistently beats random choice, and gains flatten once the language set grows, a negative result worth reporting for anyone budgeting multilingual tuning data.

arxiv.org/abs/2410.07809

Most of this work was carried out at KUIS AI Lab and Koç University, İstanbul.

We thank our colleagues and collaborators there for the environment that made it possible.