Zhiwei He (何志威)

✉ hezw.tkcw@gmail.com @zwhe99 @zwhe99 Google Scholar
https://zwhe99.github.io/

About me

I am a forth-year PhD student (2021 - present) of Department of Computer Science and Engineering at Shanghai Jiao Tong University (SJTU). I am fortunate to be advised by Prof. Rui Wang. Before that, I received the bachelor degree in Software Engineering from South China University of Technology (SCUT). I am currently a research intern at Tencent AI Lab, co-advised by Dr. Xing Wang and Dr. Zhaopeng Tu.

Research Overview

Large and Efficient Reasoning Models

- Underthinking issue in o1-like models [Preprint]
- Overthinking issue in o1-like models [Preprint]
- Rank-sharing LoRA [ICLR 2025]

Autonomous Agent powered by Large Language Models

- Multi-agent debate. [EMNLP 2024]
- Evaluating and improving agent safety. [EMNLP 2024 (Findings)]

Multilinguality & Machine Translation

- Bridging the gap between training signal and real user input. [ACL 2022]
- Human-like translation strategy. [TACL 2023]
- Improving translation with human feedback. [NAACL 2024]
- Cross-lingual consistency for text watermark [ACL 2024 (Oral)]

Education

2021 – present	◇ **PhD, Shanghai Jiao Tong University** Computer Science. Supervisor: Rui Wang
2017 – 2021	◇ **BSc, South China University of Technology** Software Engineering. Ranking: 1/252 | GPA: 3.91

Internship

2021 – present	◇ **Tencent AI Lab** Research Intern Mentors: Xing Wang & Zhaopeng Tu

Awards & Competitions

2022	◇ 1st place in the WMT22 General Translation Task, English to Livonian.
	◇ 2nd place in the WMT22 General Translation Task, Livonian to English.
2018, 2019	◇ First Class Scholarship.

Selected publications

* denotes co-first authors

Preprints

◇ **Thoughts Are All Over the Place: On the Underthinking of o1-Like LLMs**
Yue Wang*, Qiuzhi Liu*, Jiahao Xu*, Tian Liang*, Xingyu Chen*, **Zhiwei He***, Linfeng Song, Dian Yu, Juntao Li, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, Dong Yu
arXiv 2025

◇ **Do NOT Think That Much for 2+3=? On the Overthinking of o1-Like LLMs**
Xingyu Chen*, Jiahao Xu*, Tian Liang*, **Zhiwei He***, Jianhui Pang, Dian Yu, Linfeng Song, Qiuzhi Liu, Mengfei Zhou, Zhuosheng Zhang, Rui Wang, Zhaopeng Tu, Haitao Mi, Dong Yu
arXiv 2024

Journal Article

◇ **Exploring Human-Like Translation Strategy with Large Language Models**
Zhiwei He*, Tian Liang*, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, Xing Wang
TACL 2023

Conference Proceedings

◇ **RaSA: Rank-Sharing Low-Rank Adaptation**
Zhiwei He, Zhaopeng Tu, Xing Wang, Xingyu Chen, Zhijie Wang, Jiahao Xu, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang
ICLR 2025

◇ **R-Judge: Benchmarking Safety Risk Awareness for LLM Agents**
Tongxin Yuan*, **Zhiwei He***, Lingzhong Dong, Yiming Wang, Ruijie Zhao, Tian Xia, Lizhen Xu, Binglin Zhou, Fangqi Li, Zhuosheng Zhang, Rui Wang, Gongshen Liu
EMNLP 2024 (Findings)

◇ **Encouraging Divergent Thinking in Large Language Models through Multi-Agent Debate**
Tian Liang*, **Zhiwei He***, Wenxiang Jiao*, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, Shuming Shi
EMNLP 2024

◇ **Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models**
Zhiwei He*, Binglin Zhou*, Hongkun Hao, Aiwei Liu, Xing Wang, Zhaopeng Tu, Zhuosheng Zhang, Rui Wang
ACL 2024 | 🎖 **Oral presentation**

◇ **Improving Machine Translation with Human Feedback: An Exploration of Quality Estimation as a Reward Model**
Zhiwei He, Xing Wang, Wenxiang Jiao, Zhuosheng Zhang, Rui Wang, Shuming Shi, Zhaopeng Tu
NAACL 2024

Selected publications (continued)

- **Tencent AI Lab-Shanghai Jiao Tong University Low-Resource Translation System for the WMT22 Translation Task**
  Zhiwei He, Xing Wang, Zhaopeng Tu, Shuming Shi, Rui Wang
  WMT 2022 | 🏅 : ranked **1st** (English-to-Livonian) and **2nd** (Livonian-to-English) in General Translation Task

- **Bridging the Data Gap between Training and Inference for Unsupervised Neural Machine Translation**
  Zhiwei He, Xing Wang, Rui Wang, Shuming Shi, Zhaopeng Tu
  ACL 2022