Shaoyang Xu

Tianjin University, No. 92 Weijin Road, Xuefu Street, Nankai District, Tianjin
(+86) 15222735523 | syxu@tju.edu.cn | Aug '00 | Github | Google Scholar

Education

Tianjin University (985 Project) *Tianjin, China*
Master (Recommendation for Admission), Computer Science *Sept 2022 - Jan 2025 (expected)*

Supervisor: Prof. Deyi Xiong
GPA: 90.6/100 (rank: 5/20)
Research Interests: Multilinguality, Knowledge, and Reasoning of Large Language Models

Soochow University (211 Project) *Soochow, China*
Bachelor, Artificial Intelligence *Sept 2018 - Jun 2022*

Supervisor: Prof. Peifeng Li & Dr. Feng Jiang
GPA: 89.4/100 (rank: 5/70)
Courses: Machine Learning (98), Neural Network Principle (93), Literature Reading and Technical Writing (95), Deep Learning Application Practice (93), Pytorch Programming (93), Advanced Mathematics (95), Linear Algebra (98), etc.

Publications

Language Representation Projection: Can We Transfer Factual Knowledge across Languages in Multilingual Language Models?
Shaoyang Xu, Junzhuo Li, Deyi Xiong
EMNLP 2023 (short paper, main)

Exploring Multilingual Concepts of Human Values in Large Language Models: Is Value Alignment Consistent, Transferable and Controllable across Languages?
Shaoyang Xu, Weilong Dong, Zishan Guo, Xinwei Wu, Deyi Xiong
Under Review, EMNLP 2024 (Meta: 4 , Overall: 3.5/3/2.5 , Soundness: 3.5/3.5/2.5)

DCIS: Efficient Length Extrapolation of LLMs via Divide-and-Conquer Scaling Factor Search
Lei Yang, **Shaoyang Xu**, Deyi Xiong
Under Review, EMNLP 2024 (Meta: 3, Overall: 3/3/2.5, Soundness: 3/3/2)

FuxiTranyu: A Multilingual Large Language Model Trained with Balanced Data
Haoran Sun, Renren Jin, **Shaoyang Xu**, Leiyu Pan, Menglong Cui, Jiangcun Dui, Deyi Xiong, etc.
Under Review, EMNLP 2024 Industry Track (6/6/7)

Mitigating Privacy Seesaw in Large Language Models: Augmented Privacy Neuron Editing via Activation Patching
Xinwei Wu, Weilong Dong, **Shaoyang Xu**, Deyi Xiong
ACL 2024 (long paper, findings)

ConTrans: Weak-to-Strong Alignment Engineering via Concept Transplantation
Weilong Dong, Xinwei Wu, Renren Jin, **Shaoyang Xu**, Deyi Xiong
Under Review, NeurIPS 2024

Topic Segmentation via Discourse Structure Graph Network
Shaoyang Xu, Feng Jiang, Peifeng Li
Journal of Chinese Information Processing 2021

Completed Research Projects

Exploring Abstract Concepts in Multilingual LLMs (1st author, submitted to EMNLP 2024) *Oct 2023 - Feb 2024*

Research Question: **Do LLMs encode abstract concepts similarly to human beings in multiple languages, and how are these concepts represented, consistent and transferred across languages?**
Method: Proposing a framework to explore the existence of multilingual abstract concepts in LLMs and perform cross-lingual analysis on them.
Experiments: Conducting experiments on 7 abstract concepts related to human values, across 16 languages and 3 LLM families, each exhibiting monolingual, bilingual, and multilingual properties, respectively.
Conclusion: **Empirically substantiating the existence of multilingual abstract concepts in LLMs**, and **identifying 3 interesting cross-lingual traits of these concepts** arising from language resource disparities: cross-lingual inconsistency, distorted linguistic relationships, and unidirectional cross-lingual transfer between high- and low-resource languages.

Cross-Lingual Knowledge Transfer (1st author, accepted to EMNLP 2023) *Mar 2023 - Jun 2023*

Research Question: **Are knowledge and linguistic capabilities of LLMs decoupled, and can knowledge be transferred across languages?**
Method: **Proposing a method that enables LLMs to "think" in English while answering in non-English**. This involves two language representation space projection: the first one projects non-English representations into English, while the second one performs a back-projection.
Experiments: Conducting experiments on 2 multilingual factual knowledge probing benchmarks, across 53 languages and 44 knowledge types.
Conclusion & Analysis: **Improving factual knowledge retrieval accuracy and facilitating knowledge transfer across languages**. & Performing interpretable analyses from the perspective of representation space and knowledge neurons.

Zero-Shot Multilingual Machine Translation (1st author) *Nov 2022 - Feb 2023*

Preliminary Experiments: Conducting an in-depth analysis of multilingual machine translation models (encoder-decoder), unveiling the presence of **inconsistent distribution patterns in representations between English and non-English sentences** at the encoder side. This phenomenon could serve as a source of off-target issues.
Method: **Proposing a novel module to disentangle language-specific information from semantic information**. After decoupling, only the language agnostic semantic information from the encoder is preserved and sent to fine-tune the decoder.
Result: **Improving zero-shot Translation BLEU score from 4.52 to 10.83 on OPUS100 dataset**. However, the performance remains below that of English-pivot translation (14.61), indicating room for further improvement.

Current Research Projects/Interests

Pluralistic Culture Alignment of LLMs *Sept 2024 - Now*

Research Question: Existing work has confirmed the cultural dominance of LLMs. Can we align LLMs with pluralistic culture values, primarily based on the knowledge already embedded in the models?

Cross-Lingual Intelligence Transfer *May 2024 - Now*

Research Question: One of the differences between LLMs and humans is that LLMs demonstrate a strong coupling between intelligence and linguistic ability. Can we data-efficiently transfer LLMs' intelligence from English to other languages with minimal loss of intelligence?

Work Experience

Large Language Model and Multimedia Technology Department, Kuaishou Technology *Beijing, China*

LLMs Algorithm Intern *May 2024 - Sept 2024*

Executing a technical roadmap including data construction, SFT, reward modeling, and DPO to enhance the role-playing capabilities of LLMs. Building an evaluation pipeline with benchmarks such as MMLU, GSM8K, and IFEval to assess the general capabilities of trained models.

Awards and Honors

2019	1st Student Scholarship, Academic Excellence Award	*SUDA*
2020	2nd Student Scholarship, Merit Students Award, 3rd Prize of CCSP2020 (East China Division)	*SUDA*
2021	1st Student Scholarship, Merit Students Award, 2nd Prize of National LanQiao Cup	*SUDA*
2022	Excellent Undergraduate Thesis / 1st Student Scholarship	*SUDA / TJU*
2023	2nd Student Scholarship, Advanced Individual Award	*TJU*

Skills

Basic Programming	Python, Shell, LaTeX
Model Training	Pytorch, Transformers, LLaMA-Factory, DeepSpeed
Languages	Mandarin, English (CET-6 Score: 553)