Yong Zheng-Xin

Contact Information

personal website: yongzx.github.io
email: contact.yong@brown.edu

Education

Brown University, Providence, RI
Ph.D. Student, Computer Science 07/2021 - Present
Advisor: Prof. Stephen H. Bach

Minerva University, San Francisco, CA
B.Sc., Computer Science (Major) and Business (Minor) 09/2017 - 05/2021
Advisor: Prof. Patrick D Watson
Major GPA: 4.0/4, Cumulative GPA: 4.0/4

Relevant Employment

Meta AI, Research Scientist Intern
Fundamental AI Research (FAIR) 06/2024 - Present

- Safety red-teaming (toxicity, accent bias, and hallucination) for Massively Multilingual Speech models.
- Research how to best collect data to minimize accent bias for ASR.
- *Mentors: Jean Maillard, Michael Auli & Marta R. Costa-jussà*

Meta AI, Research Collaborator
GenAI Trust (prev. Responsible AI) 07/2024 - 10/2024

- Multilingual LLM safety research on finetuning attacks.
- Deliverable: Explaining cross-lingual generalization of finetuning attacks
- *Mentor: Jianfeng Chi*

Cohere For AI, Research Collaborator
Aya Responsible Release 05/2023 - 02/2024

- Safety red-teaming for multilingual LLM Aya-101.
- Deliverable: Aya Model (co-first author for safety)
- *Mentors: Julia Kreutzer & Sara Hooker*

BigScience, Research Collaborator/Lead
Multilingual Modeling Group 08/2021 - 12/2022

- Led language adaptation research of BLOOM to low-resource languages.
- Helped with data collection for the earliest instruction-following models, namely T0 (English) and BLOOMZ (multilingual) .
- Deliverables: BLOOM+1 (Research Lead), T0, PromptSource, BLOOMZ, BLOOM
- *Mentor: Vassilina Nikoulina*

FrameNet Project, Research Intern
Google Summer of Code 2019 and 2020 06/2019 - 12/2020

- Expanded FrameNet through semi-supervised learning and anomaly detection.
- Investigated cross-lingual alignment of semantic frames in FrameNet graph.
- Deliverables: SDEC-AD, Frame Shift Prediction
- *Mentors: Tiago T. Torrent, Oliver Czulo & Collin F. Baker*

Awards

Best Paper Award at ACL 2024 2024
Best Paper Award at NeurIPS 2023 Socially Responsible Language Modeling (SoLaR) Workshop 2023
3rd Best App Overall at FirstNet Public Safety Hackathon 2018

Best Use of ESRI at FirstNet Public Safety Hackathon	2018
Grand Prize Winner at #100Hacks Hackathon for Puerto Rico	2017
Grand Prize Winner at NCSV Innovate NUS Hackathon	2017
Open Category Grand Prize at SIA Airlines AppChallenge	2017
Global Finalist at Google Science Fair	2016
Outstanding A-Level Cambridge Learner Award for Perfect Score in Maths	2016

Featured Publications (* indicates co-first authorship)

[1] Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks
Samuele Poppi, **Zheng-Xin Yong**, Yifei He, Bobbie Chern, Han Zhao, Aobo Yang, Jianfeng Chi
Preprint.

[2] Preference Tuning For Toxicity Mitigation Generalizes Across Languages
Xiaochen Li*, **Zheng-Xin Yong***, Stephen H. Bach
EMNLP 2024 Findings.

[3] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model
Ahmet Üstün*, Viraat Aryabumi*, **Zheng-Xin Yong***, Wei-Yin Ko*, Daniel D'souza*, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, Sara Hooker
ACL 2024. (Role in project: safety red-teaming)
Best Paper Award. Work also featured in The New York Times and other media.

[4] Low-Resource Languages Jailbreak GPT-4
Zheng-Xin Yong, Cristina Menghini, Stephen Bach
NeurIPS 2023 Socially Responsible Language Modeling Workshop
Best Paper Award. Work also featured in New Scientist and other media.

Other Publications

[5] LexC-Gen: Generating Data for Extremely Low-Resource Languages with Large Language Models and Bilingual Lexicons
Zheng-Xin Yong, Cristina Menghini, Stephen Bach
EMNLP 2024 Findings

[6] CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark
David Romero, ..., **Zheng-Xin Yong**, ..., Thamar Solorio, Alham Fikri Aji (75 authors)
NeurIPS 2024 Datasets and Benchmarks

[7] SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages
Holy Lovenia, ..., **Zheng-Xin Yong**, Samuel Cahyawijaya (61 authors)
EMNLP 2024

[8] A Safe Harbor for AI Evaluation and Red Teaming
Shayne Longpre, ..., **Zheng-Xin Yong**, ..., Percy Liang, Peter Henderson (23 authors)
ICML 2024 (Position Paper)
Oral, 1.6% of Submitted Papers (160/9653)

[9] Representativeness as a Forgotten Lesson for Multilingual and Code-switched Data Collection and Preparation
A. Seza Doğruöz, Sunayana Sitaram, **Zheng-Xin Yong**
EMNLP Findings 2023

[10] Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages
Zheng-Xin Yong, Ruochen Zhang, ..., Alham Fikri Aji (15 authors)
EMNLP 2023 CALCS Workshop
Work also featured on WIRED.

[11] BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting
Zheng-Xin Yong, Hailey Schoelkopf, ..., Vassilina Nikoulina (15 authors)
ACL 2023

[12] Crosslingual Generalization through Multitask Finetuning
Niklas Muennighoff, ..., **Zheng-Xin Yong**, ..., Colin Raffel (19 authors)
ACL 2023

[13] PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts
Stephen Bach*, Victor Sanh*, **Zheng-Xin Yong**, ..., Alexander M. Rush (27 authors)
ACL Demo 2023

[14] The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges
Genta Indra Winata, Alham Fikri Aji, **Zheng-Xin Yong**, Thamar Solorio
ACL Findings 2023

[15] BLOOM: A 176B-Parameter Open-Access Multilingual Language Model
BigScience Workshop (including **Zheng-Xin Yong** and 300+ authors)
Preprint 2023

[16] Multitask Prompted Training Enables Zero-Shot Task Generalization
Victor Sanh*, ..., **Zheng-Xin Yong**, ..., Alexander M. Rush (41 authors)
ICLR 2022
Spotlight, 5% of submitted papers (176/3391)

Selected Talks

Meta AI: *LLM Detoxification Generalizes Across Languages*, Aug 2024.

The Alan Turing Institute, London Data Week: *Multilingual AI Safety*, Jul 2024.

Cohere For AI, C4AI Gatherings: *Aya Multilingual Safety*, Feb 2024.

Cohere For AI, Aya Grande Finale: *Aya Responsible Release*, Feb 2024.

Cohere For AI, Closing the Contribution Chapter: *Malay Ambassador*, Dec 2023.

Selected Press & Media

GPT-4 gave advice on planning terrorist attacks when asked in Zulu (New Scientist, 2023)

ChatGPT Is Cutting Non-English Languages Out of the AI Revolution (WIRED, 2023)

Services

Area Chair:
EMNLP 2023 (Multilingualism and Linguistic Diversity)

Conference/Workshop Reviewer:
COLM 2024
ARR 2021-Present
ACL 2021-2023

Outreach and Mentorship Service Programs:
Deep Learning Indaba Mentorship Service (2022 - 2023)
Brown Computer Science exploreCSR (2022)

Minerva-Masason Mentoring Program (2019 - 2021)