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
- Toxicity analysis and bias mitigation for Massively Multilingual Speech models.
- *Supervisors: 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
- *Supervisor: 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)
- *Supervisors: 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 To (English) and BLOOMZ (multilingual) .
- Deliverables: BLOOM+1 (Research Lead), To, PromptSource, BLOOMZ, BLOOM
- *Supervisor: 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
- *Supervisors: 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)

Multilingual AI Safety

[1] Towards Understanding the Fragility of Multilingual LLMs against Fine-Tuning Attacks
S Poppi, **ZX Yong**, Y He, B Chern, H Zhao, OA Yang, J Chi
Preprint. (Collaboration done at Meta.)

[2] Preference Tuning For Toxicity Mitigation Generalizes Across Languages
X Li*, **ZX Yong***, S Bach
EMNLP 2024 Findings.

[3] Aya Model: An Instruction Finetuned Open-Access Multilingual Language Model
A Üstün*, V Aryabumi*, **ZX Yong***, WY Ko*, D D'souza*, G Onilude, Neel Bhandari, S Singh, HL Ooi, A Kayid, F Vargus, P Blunsom, S Longpre, N Muennighoff, M Fadaee, J Kreutzer, S Hooker
ACL 2024. (Role in project: Multilingual safety red-teaming)
Best Paper Award. Work also featured in The New York Times and other media.

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

Low-Resource NLP and Synthetic Data Generation

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

[6] Prompting Multilingual Large Language Models to Generate Code-Mixed Texts: The Case of South East Asian Languages
ZX Yong, R Zhang, JZ Forde, S Wang, A Subramonian, H Lovenia, S Cahyawijaya, GI Winata, L Sutawika, JC Blaise-Cruz, YL Tan, L Phan, R Garcia, T Solorio, AF Aji
EMNLP 2023 CALCS Workshop
Work also featured on WIRED.

[7] BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting
ZX Yong, H Schoelkopf, N Muennighoff, AF Aji, DI Adelani, K Almubarak, MS Bari, L Sutawika, J Kasai, A Baruwa, GI Winata, S Biderman, E Raff, D Radev, V Nikoulina
ACL 2023

Other Publications

[8] CVQA: Culturally-diverse Multilingual Visual Question Answering Benchmark
D Romero, ..., **ZX Yong**, ..., T Solorio, AF Aji (75 authors)
NeurIPS 2024 Datasets and Benchmarks

[9] SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for Southeast Asian Languages
H Lovenia, ..., **ZX Yong**, S Cahyawijaya (61 authors)
EMNLP 2024

[10] A Safe Harbor for AI Evaluation and Red Teaming
S Longpre, ..., **ZX Yong**, ..., P Liang, P Henderson (23 authors)

ICML 2024 (Position Paper)
Oral, 1.6% of Submitted Papers (160/9653)

[11] Representativeness as a Forgotten Lesson for Multilingual and Code-switched Data Collection and Preparation
AS Doğruöz, S Sitaram, **ZX Yong**
EMNLP Findings 2023

[12] Crosslingual Generalization through Multitask Finetuning
N Muennighoff, ..., **ZX Yong**, ..., E Raff, C Raffel (19 authors)
ACL 2023

[13] PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts
S H. Bach*, V Sanh*, **ZX Yong**, ..., MT Jiang, AM Rush (27 authors)
ACL Demo 2023

[14] The Decades Progress on Code-Switching Research in NLP: A Systematic Survey on Trends and Challenges
GI Winata, AF Aji, **ZX Yong**, T Solorio
ACL Findings 2023

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

[16] Multitask Prompted Training Enables Zero-Shot Task Generalization
V Sanh*, ..., **ZX Yong**, ..., T Wolf, AM 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)