

Yong Zheng-Xin

RESEARCH INTERESTS	I work on safety alignment, synthetic data generation, and low-resource NLP. My broad research vision is to develop LLMs that are safe and helpful for everyone around the world.	
EDUCATION	Brown University , Providence, RI Ph.D. Student, Computer Science Advisor: Prof. Stephen H. Bach	07/2021 - Present
	Minerva University , San Francisco, CA B.Sc., Computer Science (Major) and Business (Minor) Advisor: Prof. Patrick D Watson Major GPA: 4.0/4, Cumulative GPA: 4.0/4	09/2017 - 05/2021
WORK EXPERIENCE	Meta — Fundamental AI Research (FAIR) <i>Research Scientist Intern</i>	06/2024 - Present
	<ul style="list-style-type: none">• Measuring intrinsic toxicity in speech models.• Minimizing accent biases in Massively Multilingual Speech models.• <i>Mentors: Jean Maillard, Michael Auli & Marta Costa-Jussà</i>	
	Meta — GenAI Safety Alignment <i>Research Collaborator</i>	07/2024 - 10/2024
	<ul style="list-style-type: none">• Safety research on multilingual LLMs' jailbreaks.• Deliverable: Mechanistic explanations for cross-lingual finetuning attacks• <i>Mentor: Jianfeng Chi</i>	
	Cohere For AI — Aya Responsible Release Team <i>Research Collaborator</i>	05/2023 - 02/2024
	<ul style="list-style-type: none">• Safety red-teaming for multilingual LLM Aya-101.• Deliverable: Aya Model (co-first author for safety)• <i>Mentors: Julia Kreutzer & Sara Hooker</i>	
SELECTED PUBLICATIONS (* INDICATES CO-FIRST AUTHORSHIP)	BigScience — Multilingual Modeling Group <i>Research Lead/Collaborator</i>	08/2021 - 12/2022
	<ul style="list-style-type: none">• Led language adaptation research of BLOOM to low-resource languages.• Deliverables: BLOOM+1 (Research Lead), PromptSource, To, BLOOMZ, BLOOM• <i>Mentor: Vassilina Nikoulina</i>	
	Google Summer of Code — FrameNet Project <i>Research Intern</i>	06/2019 - 12/2020
	<ul style="list-style-type: none">• Expanded FrameNet and investigated cross-lingual semantic frame alignment.• Deliverables: SDEC-AD, Frame Shift Prediction• <i>Mentors: Tiago T. Torrent, Oliver Czulo & Collin F. Baker</i>	
	Safety Alignment	
	[16] 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 <i>In submission</i>	

- [15] Preference Tuning For Toxicity Mitigation Generalizes Across Languages
Xiaochen Li*, **Zheng-Xin Yong***, Stephen H. Bach
EMNLP 2024 Findings
- [14] 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, Fred-
die Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee,
Julia Kreutzer, Sara Hooker
ACL 2024
Best Paper Award. Featured on: [The New York Times](#) | [The Washington Post](#).
- [13] Low-Resource Languages Jailbreak GPT-4
Zheng-Xin Yong, Cristina Menghini, Stephen Bach
NeurIPS 2023 Socially Responsible Language Modeling Workshop
Best Paper Award. Featured on: [New Scientist](#) | [The Register](#).

Low-Resource NLP (Synthetic Data Generation and Multilingual LLMs)

- [12] 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
- [11] 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
Featured on: [WIRED](#).
- [10] BLOOM+1: Adding Language Support to BLOOM for Zero-Shot Prompting
Zheng-Xin Yong, Hailey Schoelkopf, ..., Vassilina Nikoulina (15 authors)
ACL 2023

OTHER PUBLICATIONS

- [9] 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
- [8] SEACrowd: A Multilingual Multimodal Data Hub and Benchmark Suite for South-
east Asian Languages
Holy Lovenia, ..., **Zheng-Xin Yong**, Samuel Cahyawijaya (61 authors)
EMNLP 2024
- [7] 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)
- [6] 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
- [5] Crosslingual Generalization through Multitask Finetuning
Niklas Muennighoff, ..., **Zheng-Xin Yong**, ..., Colin Raffel (19 authors)
ACL 2023

	<p>[4] PromptSource: An Integrated Development Environment and Repository for Natural Language Prompts Stephen Bach*, Victor Sanh*, Zheng-Xin Yong, ..., Alexander M. Rush (27 authors) <i>ACL Demo 2023</i></p> <p>[3] 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 <i>ACL Findings 2023</i></p> <p>[2] BLOOM: A 176B-Parameter Open-Access Multilingual Language Model BigScience Workshop (including Zheng-Xin Yong and 300+ authors) <i>Preprint 2023</i></p> <p>[1] Multitask Prompted Training Enables Zero-Shot Task Generalization Victor Sanh*, ..., Zheng-Xin Yong, ..., Alexander M. Rush (41 authors) <i>ICLR 2022</i> Spotlight, 5% of submitted papers (176/3391)</p>
AWARDS	<p>Outstanding Reviewer Award at EMNLP 2024 2024</p> <p>Best Paper Award at ACL 2024 2024</p> <p>Best Paper Award at NeurIPS 2023 Socially Responsible Language Modeling (SoLaR) Workshop 2023</p> <p>3rd Best App Overall at FirstNet Public Safety Hackathon 2018</p> <p>Best Use of ESRI at FirstNet Public Safety Hackathon 2018</p> <p>Grand Prize Winner at #100Hacks Hackathon for Puerto Rico 2017</p> <p>Grand Prize Winner at NCSV Innovate NUS Hackathon 2017</p> <p>Open Category Grand Prize at SIA Airlines AppChallenge 2017</p> <p>Global Finalist at Google Science Fair 2016</p> <p>Outstanding A-Level Cambridge Learner Award for Perfect Score in Maths 2016</p>
SELECTED TALKS	<p>Meta AI: LLM Detoxification Generalizes Across Languages, Aug 2024.</p> <p>The Alan Turing Institute, London Data Week: Multilingual AI Safety, Jul 2024.</p> <p>Cohere For AI, C4AI Gatherings: Aya Multilingual Safety, Feb 2024.</p> <p>Cohere For AI, Aya Grande Finale: Aya Responsible Release, Feb 2024.</p> <p>Cohere For AI, Closing the Contribution Chapter: Malay Ambassador, Dec 2023.</p>
SELECTED PRESS & MEDIA	<p>GPT-4 gave advice on planning terrorist attacks when asked in Zulu (New Scientist, 2023)</p> <p>ChatGPT Is Cutting Non-English Languages Out of the AI Revolution (WIRED, 2023)</p>
SERVICES	<p>Area Chair/Program Committee: EMNLP 2023 (Multilingualism and Linguistic Diversity)</p> <p>Conference/Workshop Reviewer: EMNLP 2024 (Outstanding Reviewer) COLM 2024 ARR 2021-Present (including all *ACL conferences) ACL 2021-2023</p> <p>Outreach and Mentorship Service Programs: Deep Learning Indaba Mentorship Service (2022 - 2023)</p>

Brown Computer Science exploreCSR (2022)
Minerva-Masason Mentoring Program (2019 - 2021)

**CONTACT
INFORMATION**

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