

Salsabila Zahirah Pranida

📍 Indonesia | 📞 +62 853-2672-5956 | ✉ irasalsabila@gmail.com

🌐 Salsabila Zahirah Pranida | 🗣 irasalsabila | 🌐 Website

PROFILE

Researcher and academic mentor specializing in low-resource and dialectal Natural Language Processing, with expertise across text and speech technologies. Experienced in dataset development, model fine-tuning, evaluation, and tech-focused mentoring. Proficient in computer systems, programming environments, and educational technology, with a strong commitment to advancing diversity in language technology.

EDUCATION

Mohamed bin Zayed University of Artificial Intelligence (MBZUAI)

2024 - 2026

M.Sc in Natural Language Processing

Abu Dhabi, UAE

- Research Focus: Computational Social Science, Multilingual and Dialectal, Low-Resource Language
- Thesis: Indodialect—A Multi-Dialect Conversational Dataset for Indonesian Local Languages

Universitas Islam Indonesia

2018 - 2022

S.Kom (Bachelor in Informatics)

Indonesia

- Final Project: Customer Satisfaction Sentiment Analysis on Courier Services Using BiLSTM and BiGRU

PROJECTS

IndoDialect—A Multi-Dialect Conversational Dataset for Indonesian Local Languages

2025-2026

- Spearheaded the creation of the first large-scale conversational dialect benchmark for Indonesian local languages, curating 14.2K multi-turn dialogues across 11 dialects (Javanese, Sundanese, Minangkabau, Dayak) spanning 3 major islands.
- Architected a dual-stream human curation pipeline (parallel translation and culturally grounded dialogue simulation) with rigorous perturbation validation (99.6% quality score) and expert native-speaker evaluations (>95% cultural authenticity and coherence).
- Benchmarked 10+ models across 3 tasks, proving fine-tuned encoders (IndoBERT_{LARGE} F1 $\geq$ 0.96) outperform frontier LLMs in dialect fidelity; research supported by the **OpenAI Mini-Grant** and **Cohere Labs Catalyst Grant Program**.

Visiting Project: Persona and Gender Effects in Indonesian Toxicity Classification

2025-2026

- Evaluated 10 Indonesian-capable LLMs across 4 persona prompting conditions (Neutral, Cis-Female, Cis-Male, LGBT) on 7,524 IndoDiscourse posts comprising 24,655 gender-disaggregated human judgments.
- Disentangled model steerability from perspective adoption using crossed mixed-effects logistic regression and 29-annotator permutation tests, proving that persona prompts alter calibration without selectively increasing demographic label agreement.
- Conducted semantic clustering ($k = 30$) and sensitivity analysis across 5 agreement metrics and 4 aggregation conventions; research conducted under Dr. Daniel Hershcovich at the University of Copenhagen (supported by **DDSA Visit Grant**).

ASR Under Noise: Exploring Robustness for Sundanese and Javanese 🗣 🌐

2025

- Benchmarked 4 Whisper architectures (Tiny, Medium, Large-v3, Large-v3-Turbo) across clean and noisy conditions using 120+ hours of speech across 10 SNR levels (−20 dB to +20 dB) augmented with 24 AudioSet environmental noise classes.
- Developed NoiseTrain and SpecAugment adaptation pipelines, cutting low-SNR Word Error Rate (WER) substantially (e.g., Sundanese Medium WER reduced from 199.1% to 56.2%, Javanese Large-v3 down to 41.4%).
- Conducted granular character- and word-level error analysis (CER/WER) identifying language-specific phonetic and morphological degradation; accepted to WinLP 2025 at EMNLP 2025 and awarded *Best Paper Award 2025*.

Culturally-Nuanced Story Generation for Reasoning in Low-Resource Languages 🗣 🌐

2024

- Constructed the first culturally grounded StoryCloze benchmark containing 3.3K narratives across 12 cultural domains in Javanese and Sundanese, comparing native writing, cultural localization, and LLM generation.
- Built an automated XLM-R filtering pipeline over 12K synthetic stories, producing synthetic training data that matched native cultural authenticity (>90% cultural relevance) and outperformed machine-translated baselines.
- Fine-tuned 6 regional and open LLMs (Qwen 2.5, Llama 3.1, SEA-LION, SahabatAI) via QLoRA for cloze classification and generation, boosting narrative commonsense reasoning performance; accepted to MRL 2025 at EMNLP 2025.

- [C.1] Salsabila Zahirah Pranida*, Antonia Karamolegkou*, Andrea Morales-Garzón, and Daniel Hershcovich. 2026. **Whose Toxicity? Persona Conditioning and Model–Annotator Label Agreement by Gender in Indonesian Toxicity Classification**. *First workshop on Pluralistic AI & NLP: Diversity-aware, Sociotechnical, Responsible Alignment (PANDORA 2026)*, Budapest, Hungary.
- [S.2] Jauza Akbar Krito*, Ayu Widowati Dwi Hapsari*, Salsabila Zahirah Pranida*, and Dwi Satya Palupi. 2026. **Multitask Learning and Scaling Analysis for Crystallographic Classification from XRD Patterns**. *Under review at the 13th International Conference on Physics and Its Applications (ICOPIA 2026)*, IOP Conference Series: Journal of Physics.
- [C.1] Salsabila Zahirah Pranida, Tamar Solorio, and Fajri Koto. 2026. **INDODIALECT: A Multi-Dialect Conversational Dataset for Indonesian Local Languages**. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing (EMNLP 2026)*. Association for Computational Linguistics.
- [C.2] Muhammad Dehan Al Kautsar, Salsabila Zahirah Pranida, Bilal Elbouardi, and Fajri Koto. 2026. **CULTURETALK-ID: A Multi-Task Dialogue Benchmark for Cultural Commonsense in Indonesian Local Languages**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2026*. Association for Computational Linguistics.
- [C.3] Samuel Cahyawijaya, Peerat Limkonchotiwat, ..., Salsabila Zahirah Pranida, ..., Cho Chan Myei Oo, Hee Ming Shan. 2026. **Anthropogenic Regional Adaptation in Multimodal Vision-Language Model**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2026*. Association for Computational Linguistics.
- [C.4] Saeed Almheiri*, Bilal Elbouardi*, Salsabila Zahirah Pranida*, ..., Suma Bhat, Fajri Koto. 2026. **Multilingual Idioms in Sentences and Conversations Across High-, Medium-, and Low-Resource Languages**. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12363–12389, San Diego, California, United States. Association for Computational Linguistics.
- [C.5] Dilara Torunoğlu-Selamet, Dogukan Arslan, Salsabila Zahirah Pranida, ..., Zahra Saaberi, Zhuohan Xie. 2026. **A Parallel Cross-Lingual Benchmark for Multimodal Idiomaticity Understanding**. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, (pp. 9434–9448). European Language Resources Association (ELRA).
- [R.1] Tyler A. Chang, Catherine Arnett, ..., Salsabila Zahirah Pranida, ..., Jiayi Wang, David Ifeoluwa Adelani. 2025. **Global PIQA: Evaluating Commonsense Reasoning Across 100+ Languages and Cultures**. *arXiv preprint arXiv:2510.24081*.
- [C.6] Salsabila Zahirah Pranida*, Rifo Ahmad Genadi*, Muhammad Cendekia Airlangga*, and Shady Shehata. 2025. **ASR Under Noise: Exploring Robustness for Sundanese and Javanese**. In *Proceedings of the 9th Widening NLP Workshop*, pages 87–99, Suzhou, China. Association for Computational Linguistics. **Best Paper Award at WinLP 2025**
- [C.7] Salsabila Zahirah Pranida*, Rifo Ahmad Genadi*, and Fajri Koto. 2025. **Culturally-Nuanced Story Generation for Reasoning in Low-Resource Languages: The Case of Javanese and Sundanese**. In *Proceedings of the 5th Workshop on Multilingual Representation Learning (MRL 2025)*, pages 369–384, Suzhou, China. Association for Computational Linguistics.
- [C.8] Samuel Cahyawijaya, Holy Lovenia, ..., Salsabila Zahirah Pranida, ..., Börje F. Karlsson, Peerat Limkonchotiwat. 2025. **Crowdsourcing, Crawl, or Generate? Creating SEA-VL, a Multicultural Vision-Language Dataset for Southeast Asia**. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18685–18717, Vienna, Austria. Association for Computational Linguistics.
- [J.1] Salsabila Zahirah Pranida and Arrie Kurniawardhani. 2022. **Sentiment Analysis of Expedition Customer Satisfaction using BiGRU and BiLSTM**. *Indonesian Journal of Artificial Intelligence and Data Mining*, Vol. 5, No. 1, pages 44–53.
- [C.9] Salsabila Zahirah Pranida*, Muhammad Rizhan Ridha*, Dimas Danu Budi Pratikto*, Ari Farhan Nurihsan*, Salma Aufa Azaliarahma, Dhomas Hatta Fudholi*, and Ari Sujarwo. 2019. **Jendela Jogja: Portal Promosi dan Pemasaran Pariwisata Yogyakarta**. In *Proceedings of the Seminar Nasional Aplikasi Teknologi Informasi (SNATI 2019)*, Yogyakarta, Indonesia.

Authors marked with * contributed equally.

WORK EXPERIENCE

- PT Dua Empat Tujuh 

Data Scientist

Aug 2022 - Jul 2024
Indonesia

 - Enhanced benchmarking methodologies using ROUGE and BLEU metrics to assess media analytics functions, including document summarization and keyword extraction.
 - Refined revenue and expense forecasting models for a telecommunications company, improving financial projections, cost estimation, and budgeting accuracy.
 - Delivered a rule-based Named Entity Recognition (NER) system for Indonesians to identify entities such as organizations, landmarks, persons, and time references.
- PT Dua Empat Tujuh 

Data Scientist Intern

Jun 2021 - Jun 2022
Indonesia

 - Developed and fine-tuned a sentiment analysis model for a Media Analytics product using BiLSTM, GRU, BiGRU, and FastText methodologies.
 - Engineered an API to detect sentiment in ambiguous text, incorporating POS tagging techniques for enhanced accuracy.
 - Streamlined the dataset collection process by web scraping over 20,000 tweets for sentiment analysis using Python.
 - Established and standardized dataset tagging guidelines to ensure consistency and precision in analysis.

TEACHING & MENTORING EXPERIENCE

- Techready.id

Machine Learning Mentor (DTS Komdigi)

Jul 2026 - Present
Remote

 - Mentored a cohort of 69 students in the Digital Talent Scholarship (DTS Komdigi x Yandex x Techready.id) program for the course *Linear Models in Machine Learning: Fundamentals and Applications*.
 - Guided students through the end-to-end ML lifecycle, data preprocessing, linear regression, and model deployment strategies.
- DeepLearning.AI

Course Tester

Oct 2022 - Jul 2024
Remote

 - Evaluated upcoming AI and Deep Learning course modules and hands-on coding labs, providing technical feedback to refine instructional clarity and learning effectiveness.
- Bangkit Academy led by Google, Tokopedia, Gojek, & Traveloka

Group Facilitator (Machine Learning Path)

Feb 2022 - Aug 2022
Yogyakarta, Indonesia

 - Selected as one of 50 top-graduate machine learning facilitators to lead and mentor a cohort of 25 students.
 - Allocated 6–7 hours per week providing consultative support, academic guidance, and peer mentoring, achieving a student graduation rate above 85%.
- Universitas Islam Indonesia

Assistant Lecturer

Sep 2019 - Dec 2021
Yogyakarta, Indonesia

 - Assisted primary lecturers in undergraduate courses, including Deep Learning (SIF910), Software Engineering (SIF204), Fundamental of Informatics (SIF103), Indonesian for Scientific Papers (SIF501), and Professional Ethics (SIF601).
 - Supported the teaching process, evaluated assignments, prepared solution sets, and provided constructive feedback to students.
- Laboratorium Terpadu Informatika UII

Laboratory Assistant

Sep 2019 - Jul 2021
Yogyakarta, Indonesia

 - Conducted practical tutorial and lab sessions for Programming & Data Structure (SIF104), Web-based Application Development (SIF302), and Application Development Fundamentals (SIF202).
 - Designed assignment materials, facilitated live coding exercises, and mentored undergraduate students on their final software projects.

ACHIEVEMENTS & AWARDS

- OpenAI Mini-Grant Recipient

2025

Dept. of NLP, MBZUAI

 - Awarded a \$1,500 internal mini-grant from MBZUAI’s NLP Department to support thesis work on multilingual and dialectal language modeling.
- Cohere Labs Catalyst Grant Program Recipient

2025

Cohere

 - Awarded a \$1,000 Catalyst Grant Program by Cohere to support thesis work on multilingual and dialectal language modeling.

DDSA Visit Grant Recipient

Jun – Aug 2025

Danish Data Science Academy

- Awarded a 15,000 DKK DDSA Visit Grant to conduct collaborative research at the University of Copenhagen, working under the supervision of Dr. Daniel Herschovich.

Bangkit 2021 Scholarship - Machine Learning Path

Jan 2021

Google, Tokopedia, Gojek, and Traveloka



- Selected as an awardee for the Bangkit 2021 program, focusing on AI and machine learning applications.
- Graduated as a Distinct Graduate, recognized among the top 10% of participants.
- Collaborated on a team project, **Elevate!**, focusing on product input classification and data processing using BiGRU, achieving 84.7% accuracy.

SKILLS

Programming & Data Science: Python, SQL, Data Structures & Algorithms, Data Mining, Data Visualization

Machine Learning & Deep Learning: PyTorch, TensorFlow, HuggingFace Transformers, Scikit-learn, Keras

Natural Language Processing: Text & Speech Processing, LLM Fine-tuning, Multilingual NLP, Dataset Construction, Evaluation & Benchmarking

Languages: English (Fluent), Indonesian (Native), Javanese (Native)

VOLUNTEER EXPERIENCE

Volunteer

Jan 2025

COLING 2025 – The 31st International Conference on Computational Linguistics



- Supported the registration desk, assisted during poster sessions, and served as a floating volunteer throughout the conference.
- Helped coordinate attendee logistics and provided guidance to participants, contributing to the smooth execution of the event.

Course Tester

Oct 2022 - Jul 2024

DeepLearning.AI



- Reviewed new courses from DeepLearning.AI, providing feedback to improve content quality and user experience.
- Identified and corrected typos in course labs and recommended enhancements to increase clarity and effectiveness.