# David Anugraha

Email: david.anugraha@gmail.com
Website: davidanugraha.github.io

---

**RESEARCH INTERESTS**

My current research interests primarily lie in the areas of Language Models, Multilingual NLP, and Low-Resource NLP. I am also interested in the generalization and robustness of AI systems, including multi-modal and agentic systems. Additionally, I have also been involved in data-driven decision-making in various research areas, including parallel and distributed databases, as well as drug discovery using machine learning.

**EDUCATION**

**B.Sc (Hons), University of Toronto**, Toronto, Canada — June 2024
Computer Science Specialist (AI), Statistics Major, Chemistry Minor — GPA: 3.97/4.0

**Relevant Courses:** Neural Network and Deep Learning, Artificial Intelligence, Computer Vision, Probabilistic Machine Learning, Advanced Probability, Stochastic Processes, Time Series Analysis.

**AWARDS**

Dean's List Scholar (University of Toronto) — 2020 - 2024
Later Life Learning Scholarship (University of Toronto) — 2020 - 2022
University of Toronto Excellence Award (UTEA) (declined) — 2023

**PUBLICATIONS**

**David Anugraha***, Garry Kuwanto*, Lucky Susanto, Derry Tanti Wijaya, Genta Indra Winata. 2024
MetaMetrics-MT: Tuning Meta-Metrics for Machine Translation via Human Preference Calibration
**Equal Contribution*
*Proceedings of the Ninth Conference on Machine Translation, USA. Association for Computational Linguistics.*
**Winner of Metric Shared Task**

Genta Indra Winata*, Frederikus Hudi*, Patrick Amadeus Irawan*, **David Anugraha***, Rifki Afina Putri*, and 46 other authors.
WorldCuisines: A Massive-Scale Benchmark for Multilingual and Multicultural Visual Question Answering on Global Cuisines
**Equal Contribution*
*arXiv preprint arXiv:2410.12705 (to appear in NAACL Main 2025)*

Genta Indra Winata*, **David Anugraha***, Lucky Susanto*, Garry Kuwanto*, Derry Tanti Wijaya. 2024
MetaMetrics: Calibrating Metrics For Generation Tasks Using Human Preferences
**Equal Contribution*
*arXiv preprint arXiv:2410.02381 (to appear in ICLR 2025)*

Aditya Khan, Mason Shipton, **David Anugraha**, Kaiyao Duan, Phuong H. Hoang, Eric Khiu, A. Seza Doğruöz, and En-Shiun Annie Lee. 2024
URIEL+: Enhancing Linguistic Inclusion and Usability in a Typological and Multilingual Knowledge Base
*Oral at International Conference on Computational Linguistics (COLING 2025)*

**David Anugraha**, Genta Indra Winata, Chenyue Li, Patrick Amadeus Irawan, En-Shiun Annie Lee. 2024
ProxyLM: Predicting Language Model Performance on Multilingual Tasks via Proxy

Models
*arXiv preprint arXiv:2406.0933 (to appear in NAACL Findings 2025)*

Eric Khiu, Hasti Toossi, **David Anugraha**, Jinyu Liu, Jiaxu Li, Juan Armando Parra Flores, Leandro Arcos Roman, A. Seza Dogruöz, En-Shiun Annie Lee. 2024
Predicting Machine Translation Performance on Low-Resource Languages: The Role of Domain Similarity
*Findings of the Association for Computational Linguistics: EACL 2024*

## TALKS

Toronto Machine Learning Summit 2024
*ProxyLM: Predicting Language Model Performance on Multilingual Tasks via Proxy Models*

## EXPERIENCE

**Research Engineer**, Markham, Canada — June 2024 - Present
Distributed Data Storage and Management Lab at Huawei Canada

- Team lead on researching the application of LLMs and LMMs to databases for semantic operations, with a particular focus on GaussDB.
- Reduced completion time and resource usage during query execution for multiple users' workloads by at least 25% in low-resource settings.
- Explored data-driven cost query estimation models to optimize query execution by 12%.
- Researching efficient distributed sorting and windowing algorithms for a mix of batch and streaming execution.

**Research Assistant**, Toronto, Canada — August 2023 - Present
Advised by Prof. Annie En-Shiun Lee

- Led and managed teams, collaborating with external partners to research multilinguality and multicultural NLP systems, focusing on efficient and robust methods for low-resource NLP tasks.
- Published papers at top-tier *CL conferences and presented at the Toronto Machine Learning Summit (TMLS) in 2024.

**Research Assistant**, Toronto, Canada — August 2023 - September 2024
Advised by Prof. Maryam Mehri Dehnavi

- Focusing on optimization and sparse training, particularly for a paper titled "SLoPe: Double-Pruned Sparse Plus Lazy Low-Rank Adapter Pretraining of LLMs" published under ICLR 2025.
- Fine-tuned LMs and LLMs, including LLaMA and BERT, using various compression techniques such as pruning and quantization, and conducted data analysis on their performance against multiple benchmark evaluations.
- Developed sparse kernels in CUDA to implement sparsity in the weights of language models for more efficient pre-training and inference.

**Assistant Research Engineer**, Markham, Canada — May 2022 – August 2023
Distributed Data Storage and Management Lab at Huawei Canada

- Contributed to the MindPandas project by developing 16 map, reduce, and window operators in both lazy batch and streaming mode, resulting in a 5x increase in performance compared to Pandas and receiving an outstanding team award.
- Conducted research on efficient shuffling algorithms for a potential patent in Huawei's next AI Analytics Engine.
- Maintained and handled 23 issues and requirements from headquarters, researching and implementing possible performance improvements on the MindData codebase.

## PROJECTS

**MindSpore (Open-source deep learning training/inference framework)**

- Designed and implemented support for compressed TFRecord dataset in MindData pipeline, benefiting MindSpore users migrating from TensorFlow for better performance in MindSpore.
- Added documentation to 448 test files and reorganized 284 source files using multiple code check tools to prevent future errors in the CI/CD pipeline.

**Drug Synergy Prediction**

- Designed a preprocessing pipeline for feature extraction from drug and cell data to predict synergy scores for cancer treatment, using DrugComb as the benchmark.
- Implemented a graph-based deep neural network using Torch in Python, improving prediction accuracy by 2x compared to state-of-the-art benchmarks.

**Solubility Prediction**

- Developed a machine learning algorithm to estimate the solubility of compounds in water, a critical task in pharmaceutical chemistry on expediting drug discovery processes.
- Implemented deep neural networks, RandomForest, and XGBoost using TensorFlow in Python, achieving RMSE of 0.81, surpassing results from related papers such as SolTransNet in 2021 with RMSE of 1.141 and Graph Convolutional Neural Network in 2023 with RMSE of 0.86.

**Personalized Education Algorithm**

- Developed an algorithm focused on improving educational strategies and personalized learning that estimates the students' ability level.
- Designed machine learning models using KNN, Rasch model, and neural network in Python, achieving an accuracy rate of 72%.

## PROFESSIONAL SERVICES

**Reviewer for ICLR 2025 and NAACL 2025.**

## SERVICES

**Peer Mentor** May 2020 – September 2021

Innis Mentorship (University of Toronto)

- Mentored and supported a diverse group of international first-year students during their transition to the University of Toronto.
- Provided guidance and assistance by offering accurate information on majors, program prerequisites, and other essential resources.
- Facilitated regular check-ins and maintained open lines of communication, resulting in positive feedback and improved student satisfaction.

**International Committee Council** September 2019 – September 2020

Innis Residence Council (University of Toronto)

- Successfully orchestrated 7 engaging international-themed events at Innis Residence, fostering cultural understanding and fostering a sense of belonging among international students.
- Presented event proposals to the council, leading to their approval and ensuring seamless coordination and execution of each event.
- Demonstrated strong event planning and organizational abilities while creating a welcoming and comfortable atmosphere that enhanced the overall student experience at the residence.

**Math and Chemistry Tutor** January 2020 – June 2020

The Saturday Program

- Achieved an average grade improvement of 10% among students by providing targeted math and chemistry tutoring and personalized support.

- Developed tailored lesson plans and study materials to address individual learning needs.

**ADDITIONAL SKILLS**

**Programming Languages**: Python, R, C, C++, Java, CUDA, SQL, Bash, Assembly.
**Libraries and Frameworks**: Torch, TensorFlow, pandas, NumPy, Spark.
**Applications**: Docker, Kubernetes, PostgreSQL, Vim, Git, SLURM, LaTeX.
**Operating Systems**: Unix, Linux, Mac OSX, Windows.