Wasi Uddin Ahmad

Senior Research Scientist, NVIDIA
(+1)434-202-9102 ⋄ wasicse90@gmail.com
https://wasiahmad.github.io

Professional Summary & Vision

AI researcher with 10+ years of machine learning experience, including 5 years in Generative AI, advancing LLM post-training techniques and developing autonomous agents for complex coding workflows.

Industrial Experience

NVIDIA, Santa Clara, CA 05.2024 - Present
Senior Research Scientist, Nemotron Research

- Core contributor to the ecosystem of pipelines that power NVIDIA's state-of-the-art reasoning models.
- Orchestrated the creation of advanced synthetic trajectory data to distill agentic capabilities into the NVIDIA Nemotron-3 model family (Nano, Super, and Ultra).
- Co-led the development of the largest reasoning-based code distillation dataset (OpenCodeReasoning), significantly boosting competitive coding benchmarks for the Nemotron-3 family.
- Directed the curation of the largest open-source instruction tuning dataset for code (OpenCodeInstruct) at launch, serving as a core component in the post-training of Nemotron-Nano-2.

AWS AI Labs, Santa Clara, CA 10.2021 - 05.2024
Applied Scientist, Amazon Q Developer

- Co-led repository-level code generation for Amazon Q by designing a lightweight, client-side retrieval mechanism that enabled high-context code synthesis with minimal latency.
- Co-led the development of code embedding models (CodeSage), achieving state-of-the-art code retrieval performance and driving their integration into the Amazon Titan Text Embeddings suite.
- Collaborated on the fine-tuning of LLMs for fill-in-the-middle (FIM) code generation, contributing to the successful launch and integration of the feature within Amazon Q.
- Contributed to large-scale fine-tuning and optimization of internal LLMs, facilitating the successful production launch of multi-billion parameter models for Amazon Q.
- Established industry-standard benchmarks, including MBXP and CrossCodeEval, to evaluate multilingual LLMs on complex algorithmic and repository-level code generation tasks.
- Evaluated core architectural strategies for Amazon Q, conducting extensive research into attention mechanisms, positional encodings, and transformer architectures to optimize model performance.

Meta AI, Menlo Park, CA 06.2020 - 09.2020
Research Intern, Language and Translation Technology

- Designed and implemented a syntax-aware training objective to enhance multilingual encoders with language-agnostic structural knowledge, yielding significant improvements in cross-lingual generalization across multiple benchmarks.

Yahoo Research, Sunnyvale, CA 06.2019 - 09.2019
Research Intern, Ad Quality Science

- Developed a neural keyphrase generation framework integrating salient sentence selection with joint extraction/generation modules, achieving SOTA performance through a novel coverage attention mechanism and optimized multi-task training.

Microsoft AI & Research, Redmond, WA 06.2018 - 09.2018
Research Intern, Business Applications Group

- Designed an end-to-end architecture that integrates extractive span selection and generative answer synthesis into a single QA framework, attaining SOTA performance on SQuAD and MS MARCO through optimized multi-task training.

Walmart Labs, Reston, VA 06.2016 - 08.2016
Research Intern, Wireless Fraud Prevention

- Developed high-precision ensemble models (Random Forests, XGBoost) for anomaly detection in high-dimensional wireless transaction data, optimizing performance under extreme class imbalance using SMOTE, cost-sensitive learning, and feature-importance analysis.

REVE Systems, Dhaka, Bangladesh 02.2013 - 10.2013
Software Development Engineer, Mobile VoIP Solution

- Worked on P2P VoIP client for Android, implementing low-latency communication frameworks and background services to ensure reliable real-time call management.
- Improved full-duplex audio performance by implementing advanced noise cancellation techniques, significantly enhancing call fidelity and speech clarity in high-noise environments.

Selected Publications (Top 5 by Citations) [Google Scholar]

(* indicates equal contribution)

1. **Ahmad, W. U.***, Chakraborty, S.*, Ray, B., & Chang, K. W. (2021). Unified Pre-training for Program Understanding and Generation. In Proceedings of NAACL-HLT.
2. **Ahmad, W. U.**, Chakraborty, S., Ray, B., & Chang, K. W. (2020). A Transformer-based Approach for Source Code Summarization. In Proceedings of ACL.
3. Bhattacharjee, A.*, Hasan, T.*, **Ahmad, W. U.**, Samin, K., Islam, M. S., Iqbal, A., Rahman, M. S., & Shahriyar, R. (2022). BanglaBERT: Language Model Pretraining and Benchmarks for Low-Resource Language Understanding Evaluation in Bangla. In Findings of the ACL: NAACL.
4. Parvez, R. M., **Ahmad, W. U.**, Chakraborty, S., Ray, B., & Chang, K. W. (2021). Retrieval Augmented Code Generation and Summarization. In Findings of the ACL: EMNLP.
5. Athiwaratkun, B., Gouda, S. K., Wang, Z., Li, X., Tian, Y., Tan, M., **Ahmad, W. U.**, & Others. (2023). Multi-lingual Evaluation of Code Generation Models. In Proceedings of ICLR.

Education

Ph.D. in **Computer Science** 09.2017 - 09.2021
University of California, Los Angeles (UCLA)
Advisor: Dr. Kai-Wei Chang

Master of **Computer Science** 08.2015 - 08.2017
University of Virginia (UVA)
Advisor: Dr. Kai-Wei Chang

B.Sc. in **Computer Science and Engineering** 01.2008 - 02.2013
Bangladesh University of Engineering and Technology (BUET)