NLP+Code: Code Intelligence in Language Models

Terry Yue Zhuo, Qian Liu, Zijian Wang
Wasi Uddin Ahmad, Binyuan Hui, Loubna Ben Allal

🌐 `https://code-lm.github.io`

Language models (LMs) like GPT and Claude have shown impressive abilities in a range of natural language processing (NLP) tasks. Among these tasks, code understanding and generation have quickly become one of the most popular applications of LMs, given its nature of executable logic forms. However, there is a practical understanding of how programming knowledge can be combined with natural language to automate software development. Moreover, recent studies also empirically demonstrate that code can be a better form for complex reasoning and agentic task automation, but they do not indicate their significance. In this tutorial, we deem such superior capabilities brought by code modeling as *Code Intelligence*, and aim to provide a coherent overview of recent advances in this topic. We will start by first providing preliminaries of training foundation models on code and their common practices. We will then focus on downstream tasks in the domain of code and their evaluations. Then, we will cover how code can contribute to advancements in general tasks, and the opportunities of future research on Code Intelligence.

Terry Yue Zhuo, Ph.D student, Monash University & CSIRO's Data61
Email: `terry.zhuo@monash.edu`
Website: `https://terryyz.github.io/`
Terry Yue Zhuo is a Ph.D. student at Monash University and a researcher at CSIRO's Data61. His main research interests are code reasoning, code generation, and LMs for software engineering. Terry is currently supported by Data61 PhD Scholarships, IBM PhD Fellowship Awards, and Google Research Scholar Program. He is an active contributor to the BigCode organization and has been involved in or led various projects like StarCoder, StarCoder2, OctoPack, Astraios, BigCodeBench, and BigCodeArena. He has served multiple times as Area Chair for ACL Rolling Review and now serves as a Senior Area Chair for EMNLP 2025.

Qian Liu, Research Scientist, ByteDance
Email: `qian.liu@bytedance.com`
Website: `https://siviltaram.github.io/`
Qian Liu is a research scientist at ByteDance. Before joining ByteDance, he was a joint Ph.D. candidate at Beihang University and Microsoft Research Asia. His

1

research interests encompass code generation and language models. He has published several papers at top conferences, with notable works including StarCoder, OpenCoder and RegMix. Qian Liu has received several awards such as the KAUST AI Rising Star in 2024, and was nominated for the Baidu Scholarship in 2020. Additionally, he was one of the co-founders of the MLNLP community, a renowned NLP community in China. He has served multiple times as an Area Chair for ACL, EMNLP, and ICLR.

Zijian Wang, Research Scientist Manager, Meta Superintelligence Labs
Email: `zijianwang@meta.com`
Website: `https://zijianwang.me/`
Zijian Wang is a research scientist manager at Meta Superintelligence Labs. Previously, he was an applied scientist manager at AWS AI Labs building models for Amazon Q Developer. His research focuses on building better generative models for code, especially on training, evaluating, and deploying these models at scale. Zijian is an Area Chair of ARR, a lead organizer of Deep Learning for Code (DL4C) workshop at ICLR 2023, ICLR 2025, and NeurIPS 2025, and a co-organizer of LLM4Code at ICSE 2025, a top venue in software engineering.

Wasi Uddin Ahmad, Senior Research Scientist, NVIDIA
Email: `wasiuddina@nvidia.com`
Website: `https://wasiahmad.github.io/`
Wasi Uddin Ahmad is a senior research scientist in the conversational AI research team at NVIDIA. His current research aims to enhance the capabilities of Code LMs in areas such as competitive programming challenges, complex reasoning tasks, and detailed explanation generation, through the use of synthetic data. Prior to his role at NVIDIA, Wasi was at AWS AI Labs, working on code generation for Amazon Q Developer. Wasi obtained a Ph.D. in Computer Science at the University of California Los Angeles. Wasi has published more than 30 research articles in leading NLP, ML, and AI conferences and regularly serves as a program committee member for these venues.

Binyuan Hui, Senior Research Scientist, Meta Superintelligence Labs
Email: `binyuan.hby@alibaba-inc.com`
Website: `https://huybery.github.io/`
Binyuan Hui is a senior research scientist at Alibaba Qwen team, where he leads the development and open-sourcing of the Qwen-Coder series, focusing on enhancing the coding and agent capabilities of large language models (LLMs). Binyuan has made contributions to the open-source code community, including work on projects like StarCoder2, OctoPack, and OpenHands. He has published over 20 papers in

2

top NLP and AI conferences and has served multiple times as an Area Chair for major venues such as ACL and EMNLP.

Loubna Ben Allal, Research Engineer, Hugging Face
Email: `loubna@huggingface.co`
Website: `https://loubnabnl.github.io/`
Loubna Ben Allal is a research engineer at Hugging Face in the Open Science team, leading efforts on training small language models and creating high-quality pre-training datasets like Cosmopedia and FineWeb-Edu. She was previously a member of the BigCode core team, where she worked on The Stack dataset, the largest open dataset of source code, and co-developed the StarCoder and StarCoder2 models for code generation. Loubna has published several key papers in top AI venues, including NeurIPS and ICLR, and frequently presents her work at global conferences.

3