AIPerfLLM: 3rd International Workshop on Performance Optimization in the LLM world

Kingsum Chow
School of Software Technology, Zhejiang University
Hangzhou, China
kingsum.chow@gmail.com

Emilio Incerto
IMT School for Advanced Studies Lucca
Lucca, Italy
emilio.incerto@imtlucca.it

Marin Litoiu
York University
Toronto, Canada
mlitoiu@yorku.ca

Zhihao Chang
School of Software, Zhejiang University
Hangzhou, China
changzhihao@zju.edu.cn

Anil Rajput
AMD
Hillsboro, OR, USA
Anil_Rajput@yahoo.com

Khun Ban
Intel Corporation
Seattle, WA, USA
khunban@gmail.com

Daniele Masti
Gran Sasso Science Institute
L'Aquila, Italy
daniele.masti@gssi.it

Zhiheng Lyu
University of Waterloo
Waterloo, ON, Canada
z63lyu@uwaterloo.ca

Abstract

Artificial Intelligence (AI) has been widely adopted in various domains (e.g., computer vision, natural language processing, and reliability analysis). However, its use for performance modeling and evaluation remains limited, and its benefits to the performance engineering field are still unclear. Researchers and practitioners have recently started focusing on methods such as explainable or white-box AI-based solutions in performance engineering, but tools, methodologies, and datasets that enable wider adoption are still lacking. Meanwhile, the rapid rise of large language models (LLMs) such as ChatGPT poses new challenges in performance optimization and cost containment. LLM pre-training is expensive, and the necessary infrastructure also incurs a significant carbon footprint. This workshop aims to bridge research and practice by bringing together academia and industry to share experiences and insights in performance engineering for LLM-based services and AI applications. We target techniques and methodologies to optimize performance while reducing energy consumption and cost.

Notably, this year AIPerfLLM includes a joint panel with WOSP-C, exploring the intersection of AI and Performance Engineering. The panel aims to discuss how AI benefits Performance Engineering and how AI can utilize its methodologies. It fosters a discussion on the reciprocal advantages between these fields.

CCS Concepts

• **Software and its engineering → General**.

ICPE Companion '25, Toronto, ON, Canada

ACM ISBN 979-8-4007-1130-5/2025/05
https://doi.org/10.1145/3680256.3721304

Keywords

Performance engineering, AI, Large language models, Optimization

ACM Reference Format:
Kingsum Chow, Emilio Incerto, Marin Litoiu, Zhihao Chang, Anil Rajput, Khun Ban, Daniele Masti, and Zhiheng Lyu. 2025. AIPerfLLM: 3rd International Workshop on Performance Optimization in the LLM world. In *Companion of the 16th ACM/SPEC International Conference on Performance Engineering (ICPE Companion '25), May 5–9, 2025, Toronto, ON, Canada.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3680256.3721304

1 Organizers and Short Bios

Kingsum Chow (kingsum.chow@gmail.com) is a professor at the School of Software Technology, Zhejiang University. He received his Ph.D. in Computer Science and Engineering at the University of Washington in 1996. Prior to joining Zhejiang University in 2023, Kingsum worked as a chief scientist and senior principal engineer in industry with extensive experience in software-hardware co-optimization over thirty years at Intel and Alibaba. He has delivered keynotes at QCon, appeared in JavaOne keynotes, and authored numerous patents and technical presentations. He has collaborated with many major tech industry groups, including Alibaba, AMD, Amazon, Ampere, Arm, ByteDance, Google, Intel, Microsoft, Oracle, and others.

Emilio Incerto (emilio.incerto@imtlucca.it) is an Assistant Professor in Computer Science at the IMT School for Advanced Studies Lucca. His research focuses on AI-based techniques for performance modeling and control in software systems with stringent extra-functional requirements. He co-Chaired the first two editions of AIPerf workshop at ICPE 2023 and ICPE 2024, and is Posters and Demo Chair of ICPE 2025.

Marin Litoiu (mlitoiu@yorku.ca) is a Professor of Software Engineering at York University. His research interests include software performance engineering, self-adaptive systems, and cloud computing. He co-founded the SEAMS Symposium series (ACM/IEEE

Software Engineering for Adaptive and Self-Managing Systems) and has held key roles at SEAMS, ICPE, and IEEE ACSOS.

Zhihao Chang (changzhihao@zju.edu.cn) is an assistant professor at the School of Software, Zhejiang University. His research includes software-hardware collaborative performance optimization, AI computing acceleration, and sequence representation learning. He has published extensively in VLDBJ, TKDE, ICDE, AAAI, and other venues.

Anil Rajput (Anil_Rajput@yahoo.com) is an AMD Fellow in Software System Design focusing on datacenter and cloud performance optimizations. Previously at Intel for more than 20 years, he led platform design, scripting languages, and contributed to developing representative benchmarks (e.g., SPECjbb2005, SPECjvm2008, SPECjEnterprise2010, SPECpower_ssj2008). He guides graduate students and mentors high-school science fair participants in Oregon, USA.

Khun Ban (khunban@gmail.com) is an Intel cloud performance architect with over two decades of enterprise software development experience. His current focus is on Open-Source Relational Databases for performance optimization. He received his B.S. in Computer Science and Engineering from the University of Washington.

Daniele Masti (daniele.masti@gssi.it) received his PhD in Systems Science from IMT School for Advanced Studies Lucca in 2021. He is currently a PostDoc at Gran Sasso Science Institute. His research lies at the intersection of control theory and machine learning, with the aim of bridging the gap between the two.

Zhiheng Lyu (z63lyu@uwaterloo.ca) is an M.Math student at the University of Waterloo. He obtained his B.Eng from the University of Hong Kong. He has contributed to LLM interpretability research at UCB and ETH Zürich, and interned at Megvii's R-face Institute on advanced CV model training systems. His publications appear in EMNLP, NIPS, and UAI.

2 Main Objectives

Artificial Intelligence (AI) has been widely adopted in mainstream domains, yet its role in performance evaluation and modeling remains under-explored. Traditional AI tools are often used as black-box solutions not tailored to performance engineering, leading to models that demand extensive time, data, and expert interpretation.

Simultaneously, the rise of Large Language Models (LLMs) has brought new challenges in terms of infrastructure cost, energy usage, and specialized skills required. For instance, pre-training GPT-3 (behind ChatGPT) reportedly cost around 1,287,000 kWh in dynamic computing, generating a notable carbon footprint and high hardware expenses. These considerations underscore the urgent need to develop systematic performance engineering approaches that balance efficiency, scalability, and sustainability.

This workshop aims to bridge the gap by convening researchers and industry practitioners to share techniques and insights on applying AI methods (including specialized or explainable AI approaches) for performance engineering of LLMs and similar large-scale systems. The objective is to identify best practices, new tools, and open research directions that facilitate optimized performance while reducing resource consumption.

3 Workshop Format and Topics

Format. The workshop will be a half-day event, consisting of:

- Invited talks
- Work-in-progress presentations
- Fully refereed papers
- A panel discussion with industry and academia representatives

Topics of interest include (but are not limited to):

- Optimizing LLM Workloads on Traditional and New Architectures
- Hardware-Assisted LLM Systems
- LLM Optimization at Scale
- Code Generation Optimization for Modern Hardware
- Data-driven Model Identification for Performance Evaluation
- White-box Performance Modeling
- Datasets and Benchmarks for AI-driven Performance Models
- Explainability and Robustness in AI-Based Performance Engineering
- AI for Automated Performance Anomaly Detection and Self-Optimization
- AI Models for Performance Tasks Automation, including Auto-Scaling

Target audience.

- Researchers in software-hardware performance optimization for LLMs
- Practitioners solving runtime performance problems in LLM deployments
- Researchers and Practitioners in general performance optimization, modeling, and control of large-scale ICT systems

Pre-requisite knowledge. Background in software performance concepts, basic familiarity with AI/ML techniques, and general understanding of large language models or HPC systems.

4 Organizer Past Experience with Similar Events

The main organizer and collaborators have delivered the following workshops and tutorials in the past:

- *AIPerf + Performance Optimization in the LLM World*, ICPE 2024/05, half-day, 20 attendees
- *Runtimes in the Cloud 3*, HPCA 2020/02, full-day, 20 attendees
- *Runtimes in the Cloud 2*, ISCA 2019/06, full-day, 20 attendees
- *Runtimes in the Cloud*, ISCA 2018/06, full-day, 30 attendees
- *Scaling Software Performance & Software Performance in the Cloud*, PNSQC 2017/10, half-day, 50 attendees
- *Software Performance Analytics in the Cloud*, ICPE 2017/04, full-day tutorial, 20 attendees
- *Applying Analytics to Data Center Performance*, CMG Performance and Capacity Conference 2015/11, half-day, 30 attendees

5 Workshop Website

For further details, please visit:

https://sites.google.com/view/aiprefllm2025