What does Performance Mean for Large Language Models?

Jane Hillston
jane.hillston@ed.ac.uk
University of Edinburgh
Edinburgh, UK

ABSTRACT

In the last decade there has been a significant leap in the capability of foundation AI models, largely driven by the introduction and refinement of transformer-based machine learning architectures. The most visible consequence of this has been the explosion of interest and application of large language models such as ChatGPT. This is one exemplar of how a foundation model trained on a huge amount of data can be specialised for particular task, often by a phase of reinforcement learning with human feedback.

Within the AI community "performance" of such systems is generally taken to mean how well they respond to their users on characteristics such as accuracy, verifiability, and bias. Performance analysis usually considers both the responsiveness of a system to its user and the efficiency and equity of resource use. These foundation models rely on massive amounts of resource but there appears to have been little work considering how to understand the resource use or the trade-offs that exist between how the system responds to users and the amount of resource used.

In this talk I will present initial ideas of what it could mean to develop a framework of performance evaluation for foundation models such as large language models. Such a framework would need to take into consideration the distinct phases of operation for these models, which broadly speaking can be categorised as *training*, *generating* and *fine-tuning*. Evaluating the trade-off between user interests and resource management will require the identification of suitable metrics. The *resources* in these systems can be more than simply compute, storage, bandwidth; data and even human resources also play crucial roles in training and fine-tuning. I will discuss all these topics.

CCS CONCEPTS

• **Hardware** → *Power estimation and optimization*; • **Software and its engineering** → **System modeling languages**; • **Information systems** → **Retrieval efficiency**; **Information extraction**; • **Computing methodologies** → **Natural language processing**; **Natural language generation**; **Machine learning algorithms**; **Model development and analysis**.

KEYWORDS

Performance evaluation, large language models, efficient use of resources, user responsiveness

ACM Reference Format:
Jane Hillston. 2024. What does Performance Mean for Large Language Models?. In *Proceedings of the 15th ACM/SPEC International Conference on Performance Engineering (ICPE '24), May 7–11, 2024, London, United Kingdom.* ACM, New York, NY, USA, **??** pages. https://doi.org/10.1145/3629526.3649130

ICPE '24, May 7–11, 2024, London, United Kingdom

ACM ISBN 979-8-4007-0444-4/24/05.
https://doi.org/10.1145/3629526.3649130