13 March 2025 · CC-BY 4.0

Peer Review

Review of: "A Primer on Large Language Models and their Limitations"

Sebastian Spethmann[1]

[1] Independent researcher

This review provides a comprehensive introduction to large language models (LLMs) and their limitations. It covers the fundamentals of LLM technology, including their architecture, pre-training methods, and fine-tuning techniques. The document discusses different types of LLMs, such as decoder-only, encoder-decoder, and encoder-only models, as well as their applications and training strategies. The authors also explore the orchestration of LLMs with other technologies and show how these models can be integrated into traditional information retrieval systems and knowledge representation techniques to enhance their capabilities.

The paper discusses important concepts such as self-supervised learning, contextual learning, and prompt engineering. It explains various pre-training objectives such as masked language modelling (MLM) and causal language modelling (CLM) and discusses various approaches to fine-tuning, including supervised fine-tuning, reinforcement learning with human feedback (RLHF), and parameter-efficient fine-tuning (PEFT). The paper also describes the challenges and risks associated with LLMs, such as catastrophic forgetting, model collapse, jailbreak attacks, and hallucinations, and provides insights into possible strategies for risk mitigation.

The review offers unique insights into the rapidly developing field of LLMs. It highlights the trend towards open-source models and the potential impact of new technologies such as Liquid Foundation Models (LFMs) and specialised hardware such as Groq's Language Processing Unit (LPU). The authors also offer a thought-provoking perspective on the concept of 'hallucinations' in LLMs, arguing that these are better understood as a form of 'bullshit' in the philosophical sense, rather than real hallucinations. This classification emphasises the importance of critically evaluating LLM results and the need for continued research to address these limitations. The article concludes by emphasising the importance of

doi.org/10.32388/OWY7M9

continuously evaluating and adjusting LLM tools to ensure their effectiveness and reliability in different applications.

Limitations:

The paper provides a comprehensive overview of LLMs, but it does not include empirical research or experiments. This limits the ability to draw definitive conclusions about the effectiveness of the methods discussed.

The paper briefly mentions the issue of bias in LLMs but could have addressed it in more detail. LLMs can maintain or even amplify societal biases present in their training data, leading to discriminatory outcomes based on race, gender, disability, nationality, or religion. The review does not extensively address privacy concerns related to the vast amounts of data used to train LLMs, including issues of data ownership, consent for data use, and potential misuse of personal information. Additionally, the paper does not examine the ethical and legal concerns surrounding the use of copyrighted material in training data and the potential of LLMs to generate content that violates intellectual property rights.

It also could have explored the ethical implications of unequal access to LLM technology and its potential to exacerbate existing societal inequalities, such as disparities in education, economic opportunity, and technological infrastructure. Finally, the paper does not discuss the potential psychological and social impacts of widespread interaction with AI language models, including concerns about AI dependency or the blurring of lines between humans and AI.

Declarations

Potential competing interests: No potential competing interests to declare.

doi.org/10.32388/OWY7M9