A Comprehensive Overview of Large Language Models

Humza Naveed[a], Asad Ullah Khan[b,*], Shi Qiu[c,*], Muhammad Saqib[d,e,*], Saeed Anwar[f,g], Muhammad Usman[f,g], Naveed Akhtar[h,j], Nick Barnes[i], Ajmal Mian[j]

[a] *The University of Sydney, Sydney, Australia*
[b] *University of Engineering and Technology (UET), Lahore, Pakistan*
[c] *The Chinese University of Hong Kong (CUHK), HKSAR, China*
[d] *University of Technology Sydney (UTS), Sydney, Australia*
[e] *Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, Australia*
[f] *King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia*
[g] *SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRCAI), Dhahran, Saudi Arabia*
[h] *The University of Melbourne (UoM), Melbourne, Australia*
[i] *Australian National University (ANU), Canberra, Australia*
[j] *The University of Western Australia (UWA), Perth, Australia*

Abstract

Large Language Models (LLMs) have recently demonstrated remarkable capabilities in natural language processing tasks and beyond. This success of LLMs has led to a large influx of research contributions in this direction. These works encompass diverse topics such as architectural innovations, better training strategies, context length improvements, fine-tuning, multi-modal LLMs, robotics, datasets, benchmarking, efficiency, and more. With the rapid development of techniques and regular breakthroughs in LLM research, it has become considerably challenging to perceive the bigger picture of the advances in this direction. Considering the rapidly emerging plethora of literature on LLMs, it is imperative that the research community is able to benefit from a concise yet comprehensive overview of the recent developments in this field. This article provides an overview of the literature on a broad range of LLM-related concepts. Our self-contained comprehensive overview of LLMs discusses relevant background concepts along with covering the advanced topics at the frontier of research in LLMs. This review article is intended to provide not only a systematic survey but also a quick, comprehensive reference for the researchers and practitioners to draw insights from extensive, informative summaries of the existing works to advance the LLM research.

Keywords:
Large Language Models, LLMs, chatGPT, Augmented LLMs, Multimodal LLMs, LLM training, LLM Benchmarking

1. Introduction

Language plays a fundamental role in facilitating communication and self-expression for humans and their interaction with machines. The need for generalized models stems from the growing demand for machines to handle complex language tasks, including translation, summarization, information retrieval, conversational interactions, etc. Recently, significant breakthroughs have been witnessed in language models, primarily attributed to transformers [1], increased computational capabilities, and the availability of large-scale training data. These developments have brought about a revolutionary transformation by enabling the creation of LLMs that can approximate human-level performance on various tasks [2, 3]. Large

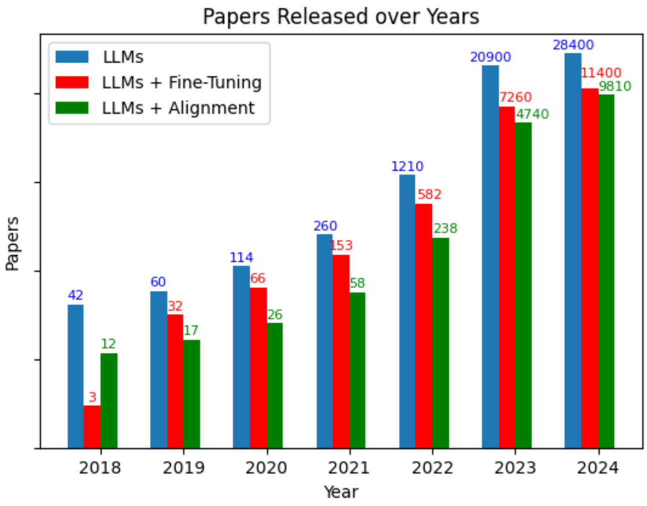

Figure 1: The trend of papers released over the years containing keywords "Large Language Model", "Large Language Model + Fine-Tuning", and "Large Language Model + Alignment".

[*] Equal contribution
Email addresses: humza_naveed@yahoo.com (Humza Naveed), aukhanee@gmail.com (Asad Ullah Khan), shiqiu@cse.cuhk.edu.hk (Shi Qiu), muhammad.saqib@data61.csiro.au (Muhammad Saqib), saeed.anwar@kfupm.edu.sa (Saeed Anwar), muhammad.usman@kfupm.edu.sa (Muhammad Usman), naveed.akhtar1@unimelb.edu.au (Naveed Akhtar), nick.barnes@anu.edu.au (Nick Barnes), ajmal.mian@uwa.edu.au (Ajmal Mian)

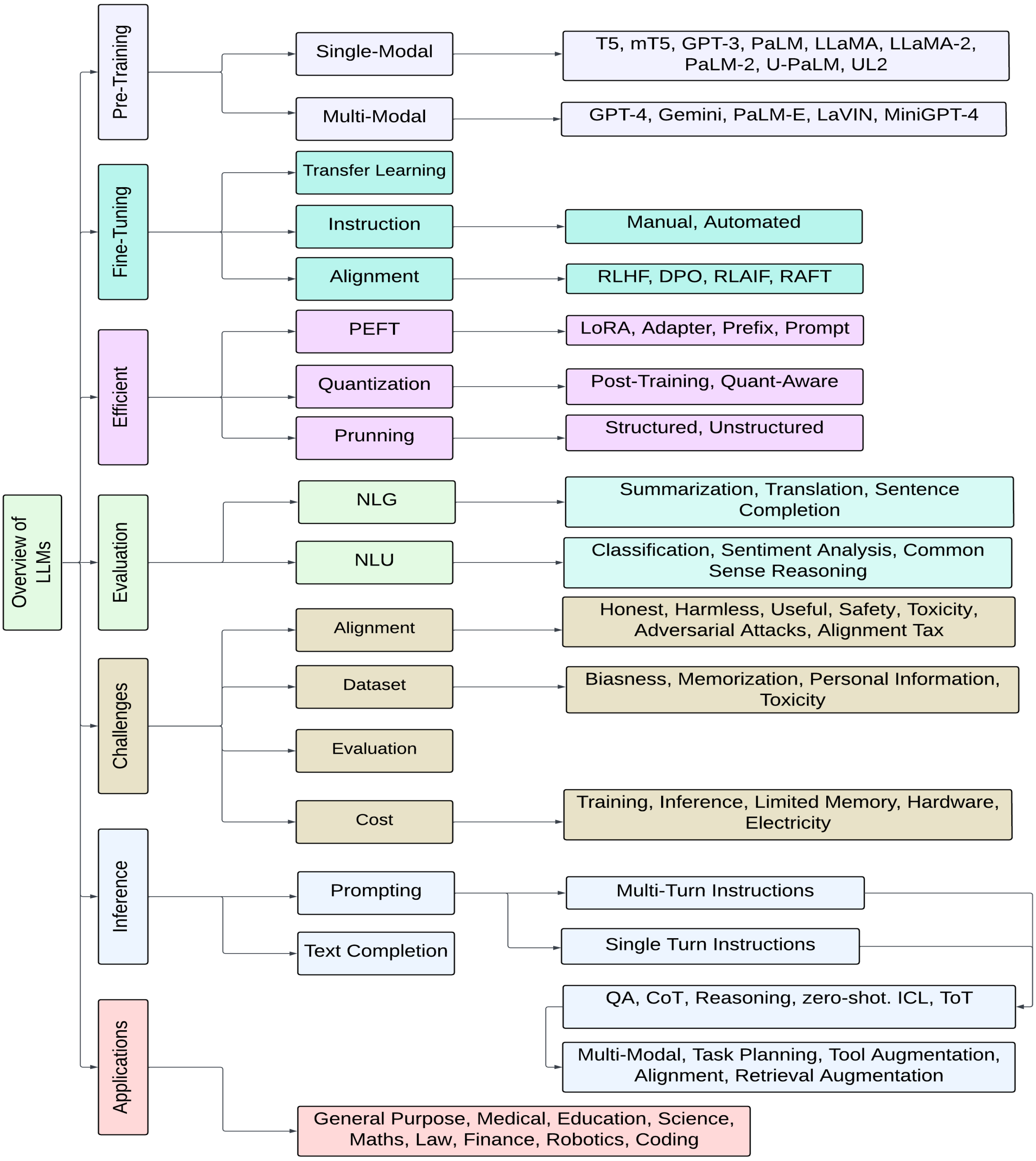
Overview of LLMs
Pre-Training
Single-Modal
T5, mT5, GPT-3, PaLM, LLaMA, LLaMA-2, PaLM-2, U-PaLM, UL2
Multi-Modal
GPT-4, Gemini, PaLM-E, LaVIN, MiniGPT-4
Fine-Tuning
Transfer Learning
Instruction
Manual, Automated
Alignment
RLHF, DPO, RLAIF, RAFT
Efficient
PEFT
LoRA, Adapter, Prefix, Prompt
Quantization
Post-Training, Quant-Aware
Prunning
Structured, Unstructured
Evaluation
NLG
Summarization, Translation, Sentence Completion
NLU
Classification, Sentiment Analysis, Common Sense Reasoning
Challenges
Alignment
Honest, Harmless, Useful, Safety, Toxicity, Adversarial Attacks, Alignment Tax
Dataset
Biasness, Memorization, Personal Information, Toxicity
Evaluation
Cost
Training, Inference, Limited Memory, Hardware, Electricity
Inference
Prompting
Multi-Turn Instructions
Single Turn Instructions
QA, CoT, Reasoning, zero-shot. ICL, ToT
Multi-Modal, Task Planning, Tool Augmentation, Alignment, Retrieval Augmentation
Text Completion
Applications
General Purpose, Medical, Education, Science, Maths, Law, Finance, Robotics, Coding

Figure 3: A broader overview of LLMs, dividing LLMs into seven branches: 1. Pre-Training 2. Fine-Tuning 3. Efficient 4. Inference 5. Evaluation 6. Applications 7. Challenges

multi-modal LLMs, augmented LLMs, LLMs-powered agents, datasets, evaluation, etc.

We loosely follow the existing terminology to ensure a standardized outlook of this research direction. For instance, following [50], our survey discusses pre-trained LLMs with 10B parameters or more. We refer the readers interested in smaller pre-trained models to [51, 52, 53].
The organization of this paper is as follows. Section 2 discusses the background of LLMs. Section 3 focuses on LLMs overview, architectures, training pipelines and strategies, fine-tuning, and utilization in different domains. Section 4 highlights the configuration and parameters that play a crucial role in the functioning of these models. Summary and discussions are presented in section 3.8. The LLM training and evaluation, datasets, and benchmarks are discussed in section 5, followed by challenges and future directions, and conclusion in sections 7 and 8, respectively.