# A Comprehensive Overview of Large Language Models

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

[a] *University of Engineering and Technology (UET), Lahore, Pakistan*
[b] *The Chinese University of Hong Kong (CUHK), HKSAR, China*
[c] *University of Technology Sydney (UTS), Sydney, Australia*
[d] *Commonwealth Scientific and Industrial Research Organisation (CSIRO), Sydney, Australia*
[e] *King Fahd University of Petroleum and Minerals (KFUPM), Dhahran, Saudi Arabia*
[f] *SDAIA-KFUPM Joint Research Center for Artificial Intelligence (JRCAI), Dhahran, Saudi Arabia*
[g] *The University of Melbourne (UoM), Melbourne, Australia*
[h] *Australian National University (ANU), Canberra, Australia*
[i] *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 existing 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 not only provide 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

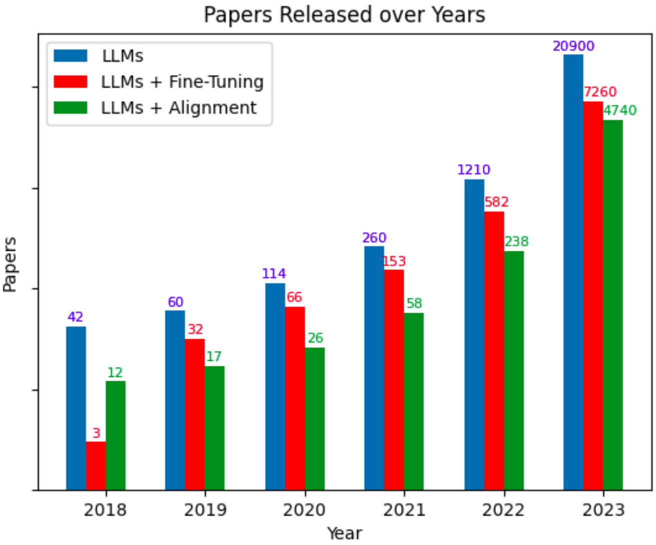

Figure 1: The trend of papers released over years containing keywords "Large Language Model", "Large Language Model + Fine-Tuning", and "Large Language Model + Alignment".

[*] 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)

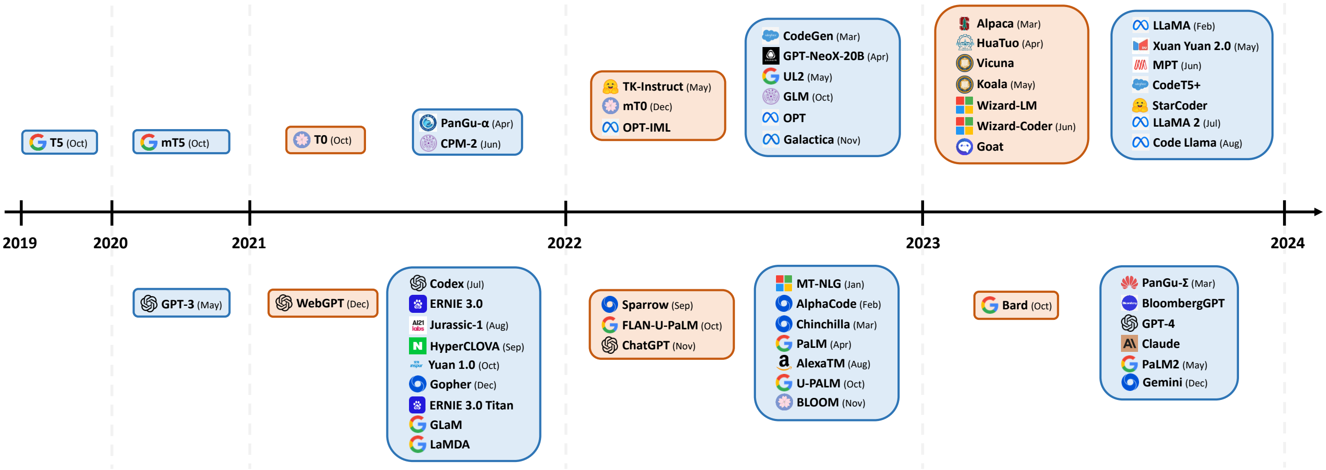

Figure 2: Chronological display of LLM releases: blue cards represent ‘pre-trained’ models, while orange cards correspond to ‘instruction-tuned’ models. Models on the upper half signify open-source availability, whereas those on the bottom half are closed-source. The chart illustrates the increasing trend towards instruction-tuned models and open-source models, highlighting the evolving landscape and trends in natural language processing research.

Language Models (LLMs) have emerged as cutting-edge artificial intelligence systems that can process and generate text with coherent communication [4], and generalize to multiple tasks [5, 6].

The historical progress in natural language processing (NLP) evolved from statistical to neural language modeling and then from pre-trained language models (PLMs) to LLMs. While conventional language modeling (LM) trains task-specific models in supervised settings, PLMs are trained in a self-supervised setting on a large corpus of text [7, 8, 9] with the aim of learning a generic representation that is shareable among various NLP tasks. After fine-tuning for downstream tasks, PLMs surpass the performance gains of traditional language modeling (LM). The larger PLMs bring more performance gains, which has led to the transitioning of PLMs to LLMs by significantly increasing model parameters (tens to hundreds of billions) [10] and training dataset (many GBs and TBs) [10, 11]. Following this development, numerous LLMs have been proposed in the literature [10, 11, 12, 6, 13, 14, 15]. An increasing trend in the number of released LLMs and names of a few significant LLMs proposed over the years are shown in Fig 1 and Fig 2, respectively.

The early work on LLMs, such as T5 [10] and mT5 [11] employed transfer learning until GPT-3 [6] showed LLMs are zero-shot transferable to downstream tasks without fine-tuning. LLMs accurately respond to task queries when prompted with task descriptions and examples. However, pre-trained LLMs fail to follow user intent and perform worse in zero-shot settings than in few-shot. Fine-tuning them with task instructions data [16, 17, 18, 19] and aligning with human preferences [20, 21] enhances generalization to unseen tasks, improving zero-shot performance significantly and reducing misaligned behavior.

In addition to better generalization and domain adaptation, LLMs appear to have emergent abilities, such as reasoning, planning, decision-making, in-context learning, answering in zero-shot settings, etc. These abilities are known to be acquired by them due to their gigantic scale even when the pre-trained LLMs are not trained specifically to possess these attributes [22, 23, 24]. Such abilities have led LLMs to be widely adopted in diverse settings including, multi-modal, robotics, tool manipulation, question answering, autonomous agents, etc. Various improvements have also been suggested in these areas either by task-specific training [25, 26, 27, 28, 29, 30, 31] or better prompting [32].

The LLMs abilities to solve diverse tasks with human-level performance come at a cost of slow training and inference, extensive hardware requirements, and higher running costs. Such requirements have limited their adoption and opened up opportunities to devise better architectures [15, 33, 34, 35] and training strategies [36, 37, 21, 38, 39, 40, 41]. Parameter efficient tuning [38, 41, 40], pruning [42, 43], quantization [44, 45], knowledge distillation, and context length interpolation [46, 47, 48, 49] among others are some of the methods widely studied for efficient LLM utilization.

Due to the success of LLMs on a wide variety of tasks, the research literature has recently experienced a large influx of LLM-related contributions. Researchers have organized the LLMs literature in surveys [50, 51, 52, 53], and topic-specific surveys in [54, 55, 56, 57, 58]. In contrast to these surveys, our contribution focuses on providing a comprehensive yet concise overview of the general direction of LLM research. This article summarizes architectural and training details of pre-trained LLMs and delves deeper into the details of concepts like fine-tuning, multi-modal LLMs, augmented LLMs, datasets, evaluation, applications, challenges, and others to provide a self-contained comprehensive overview. Our key contributions are summarized as follows.

- We present a survey on the developments in LLM research providing a concise comprehensive overview of the direction.
- We present extensive summaries of pre-trained models that include fine-grained details of architecture and training details.
- We summarize major findings of the popular contributions and provide a detailed discussion on the key design and development aspects of LLMs to help practitioners effectively leverage this technology.
- In this self-contained article, we cover a range of concepts to present the general direction of LLMs compre-