

# 【序】LLM综述论文导读第0篇

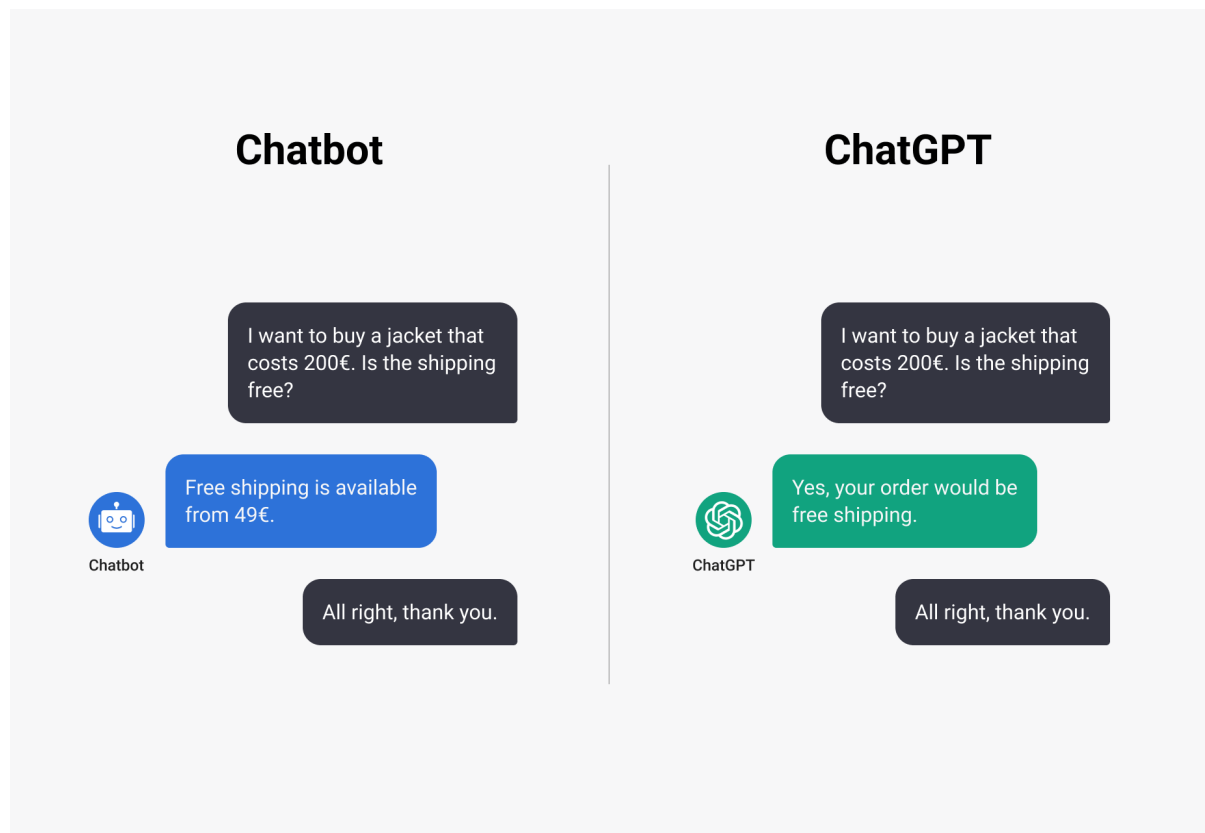
Author: @Sakura

Last Updated: 2024.5.10

This documentation is mainly referred to w. x. Zhao et al., 'A Survey of Large Language Models'. arXiv, Nov. 24, 2023. doi: 10.48550/arXiv.2303.18223.

## ChatGPT: A Sensation

Chatbot vs. ChatGPT<sup>1</sup>

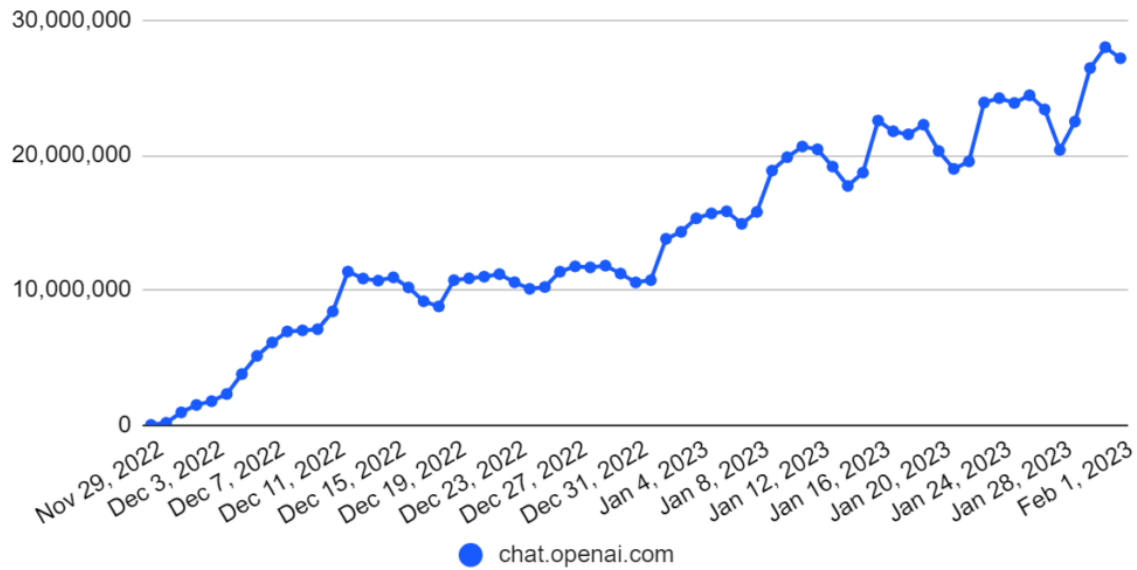


1. <https://micoope.com.gt/?o=chatgpt-the-tech-behind-the-hype-and-what-it-nn-n7AYnRsk>

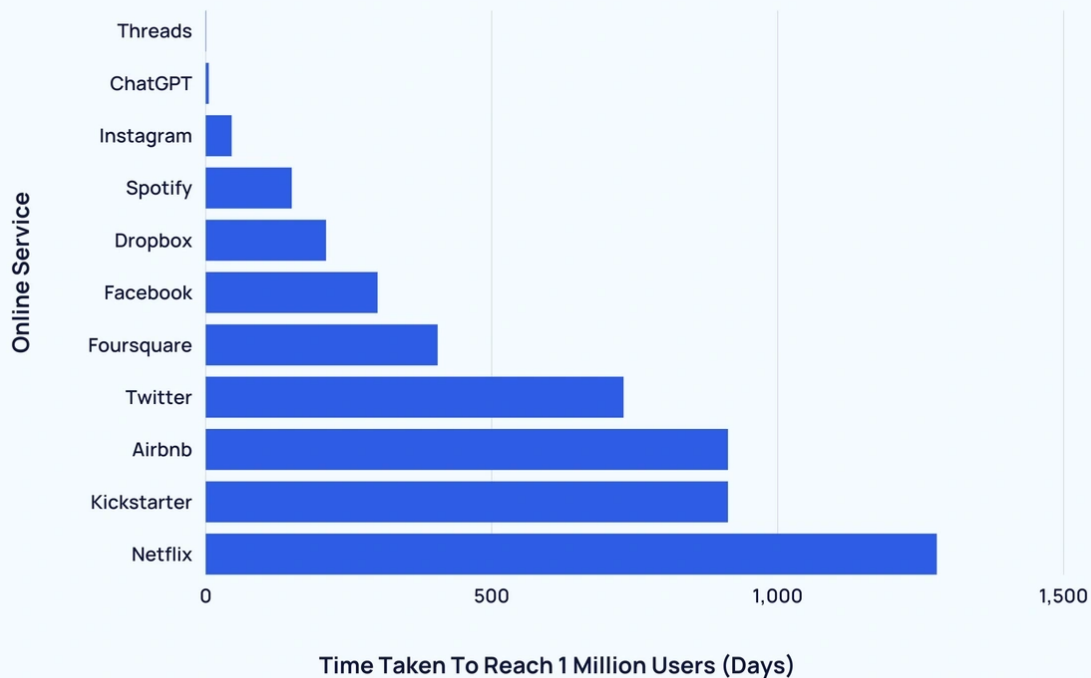
Fastest Ever User Growth<sup>1,2</sup>

# ChatGPT

Daily Visits, Desktop & Mobile Web, Worldwide



## Time taken to reach 1 million users



1. <https://influencemarketinghub.com/ai-marketing-benchmark-report>
2. <https://explodingtopics.com/blog/chatgpt-users>

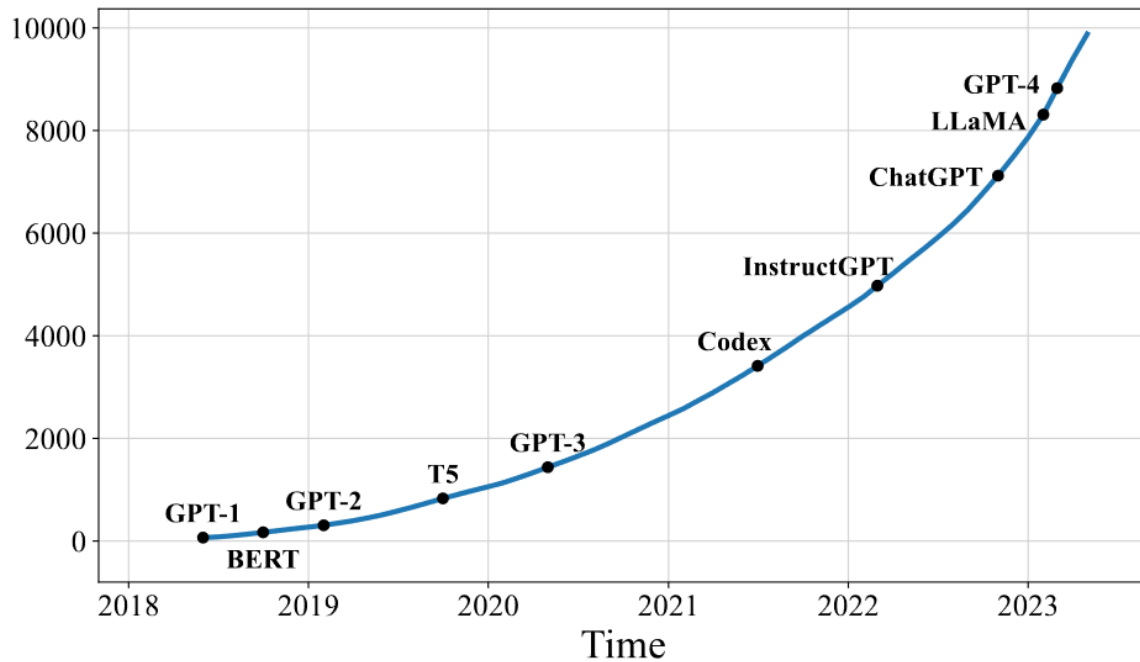
## Questions

- What is Large Language Models(LLMs)?
- Why are LLMs important?
- How can we utilize LLMs?

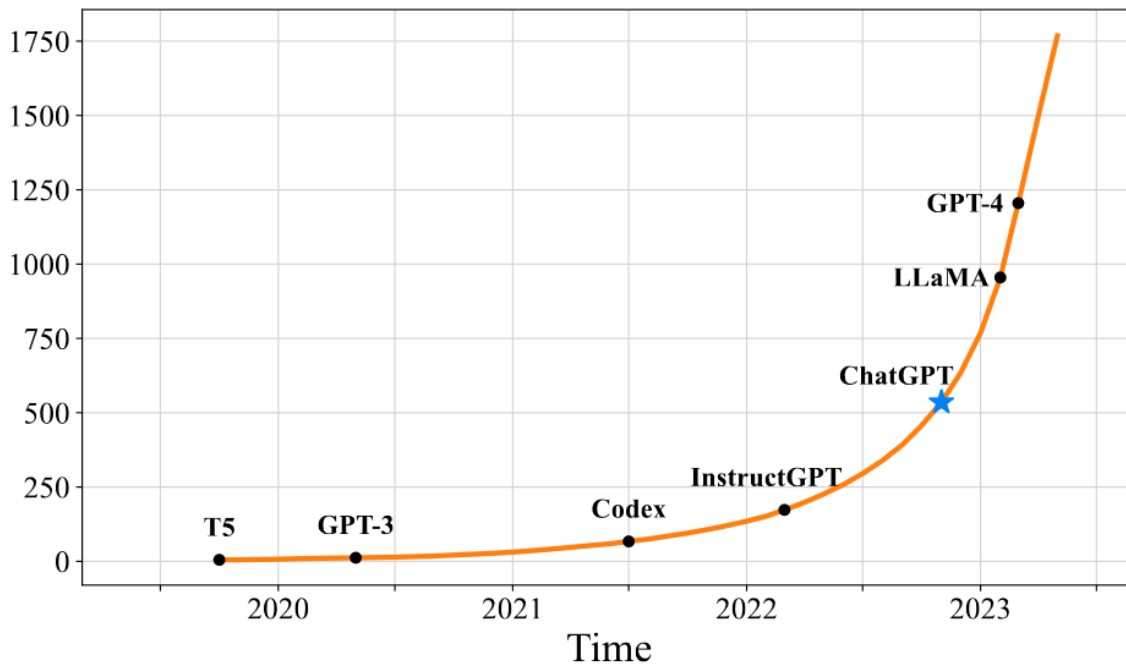
# Language Model (LM)

## Introduction to Language Model

Language modeling (LM) is one of the major approaches to advancing language intelligence of machines. In general, LM aims to model the generative likelihood of word sequences, so as to predict the probabilities of future (or missing) tokens.



(a) Query="Language Model"



(b) Query="Large Language Model"

Figure. The trends of the cumulative numbers of arXiv papers that contain the keyphrases “language model” (since June 2018) and “large language model” (since October 2019), respectively. The statistics are calculated using exact match by querying the keyphrases in title or abstract by months. We set different x-axis ranges for the two keyphrases, because “language models” have been explored at an earlier time. We label the points corresponding to important landmarks in the research progress of LLMs. A sharp increase occurs after the release of ChatGPT: the average number of published arXiv papers that contain “large language model” in title or abstract goes from 0.40 per day to 8.58 per day (Figure (b)).

## A Story from LM to Large LM

### Timeline of Language Model Development

Statistical language models (SLM)→  
 Neural language models (NLM)→  
 Pre-trained Language Models (PLM)→  
 Large Language Models(LLM,denoted as PLM<sup>+</sup>)

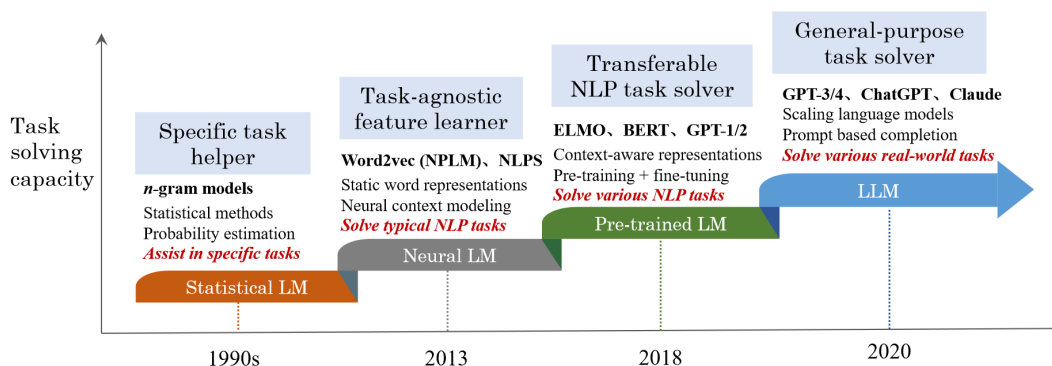


Figure. An evolution process of the four generations of language models (LM) from the perspective of task solving capacity. Note that the time period for each stage may not be very accurate, and we set the time mainly according to the publish date of the most representative studies at each stage. For neural language models, we abbreviate the paper titles of two representative studies to name the two approaches: NPLM [1] (“A neural probabilistic language model”) and NLPS [2] (“Natural language processing (almost) from scratch”). Due to the space limitation, we don’t list all representative studies in this figure.

## LLM by Definition

### LLM: Large-sized PLM

Researchers find that scaling PLM (e.g., scaling model size or data size) often leads to an improved model capacity on downstream tasks (i.e., following the scaling law [3]). A number of studies have explored the performance limit by training an ever larger PLM (e.g., the 175B-parameter GPT-3 and the 540B parameter PaLM). Although scaling is mainly conducted in model size (with similar architectures and pre-training tasks), these large-sized PLMs display different behaviors from smaller PLMs (e.g., 330M-parameter BERT and 1.5B parameter GPT-2) and show *surprising* abilities (called **emergent abilities** [4]) in solving a series of complex tasks. For example, GPT-3 can solve few-shot tasks through in-context learning, whereas GPT-2 cannot do well. Thus, the research community coins the term **large language models (LLM)**<sup>1</sup> for these large-sized PLMs [5–8], which attract increasing research attention. A remarkable application of LLMs is ChatGPT2 that adapts the LLMs from the GPT series for dialogue, which presents an amazing conversation ability with humans.

1. Note that a LLM is not necessarily more capable than a small PLM, and emergent abilities may not occur in some LLMs.

## An Overview on LLM Market

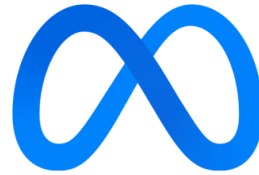
Common LLM Products



OpenAI

ChatGPT

gpt-3.5-turbo/gpt-4



Meta AI



LLaMA



Google



PALM



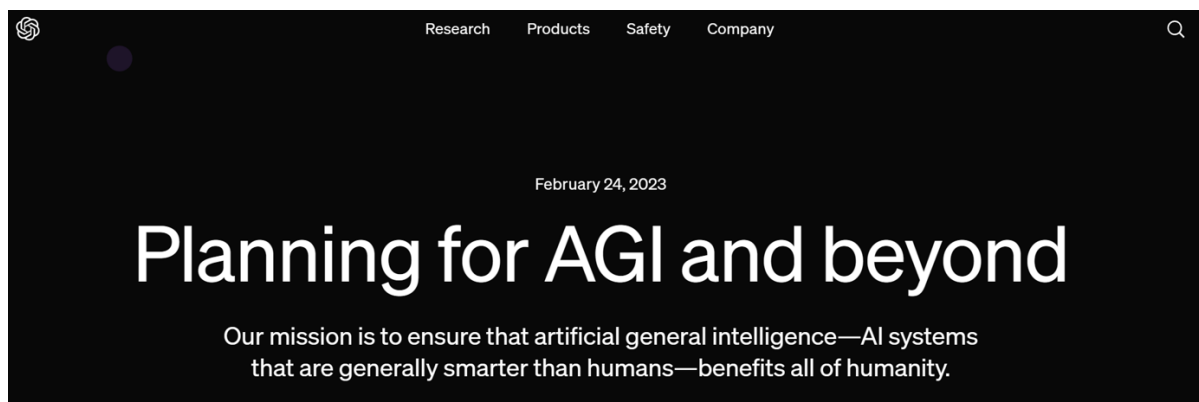
Zhipu AI



GLM

Mainstream LLM with their AI companies

## LLM: A Path for AGI



- We can imagine a world in which humanity flourishes to a degree that is probably impossible for any of us to fully visualize yet.
- We hope to contribute to the world an AGI aligned with such flourishing.<sup>1</sup>

1. <https://openai.com/index/planning-for-agi-and-beyond?ref=hir.harvard.edu>

## Guidance for AI Green-hands/Researchers

Choose Your Favorite LLM Projects in [GitHub](#)

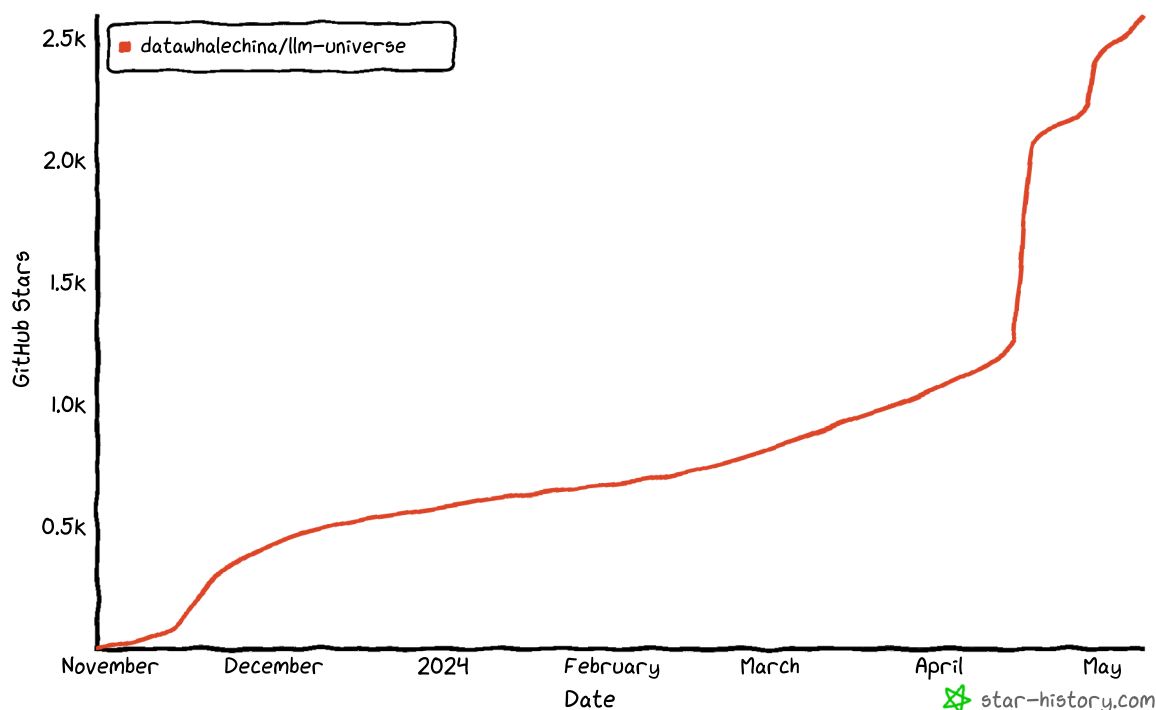
# 动手学大模型 应用开发

— LLM-UNIVERSE —

奇想星球

Datawhale

Star History



## ► 项目简介

本项目是一个面向小白开发者的大模型应用开发教程，旨在基于阿里云服务器，结合个人知识库助手项目，通过一个课程完成大模型开发的重点入门，主要包括：

1. 大模型简介，何为大模型、大模型特点是什么、LangChain 是什么，如何开发一个 LLM 应用，针对小白开发者的简单介绍；
2. 如何调用大模型 API，本节介绍了国内外知名大模型产品 API 的多种调用方式，包括调用原生 API、封装为 LangChain LLM、封装为 Fastapi 等调用方式，同时将包括百度文心、讯飞星火、智谱AI等多种大模型 API 进行了统一形式封装；
3. 知识库搭建，不同类型知识库文档的加载、处理，向量数据库的搭建；

4. 构建 RAG 应用，包括将 LLM 接入到 LangChain 构建检索问答链，使用 Streamlit 进行应用部署
5. 验证迭代，大模型开发如何实现验证迭代，一般的评估方法有什么；

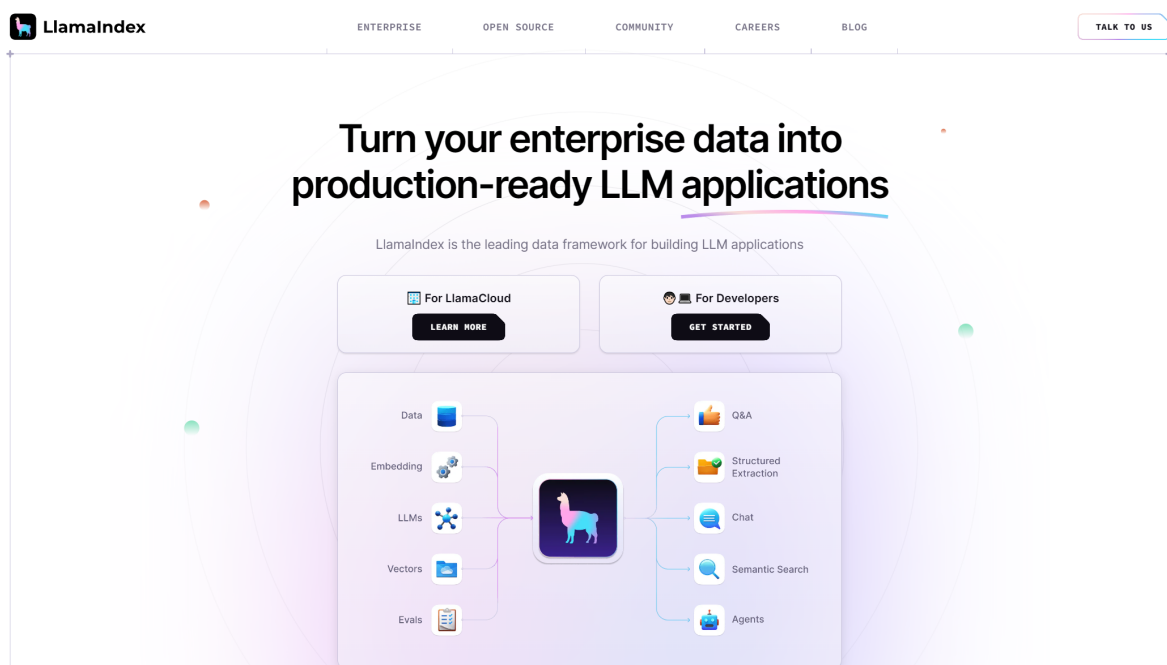
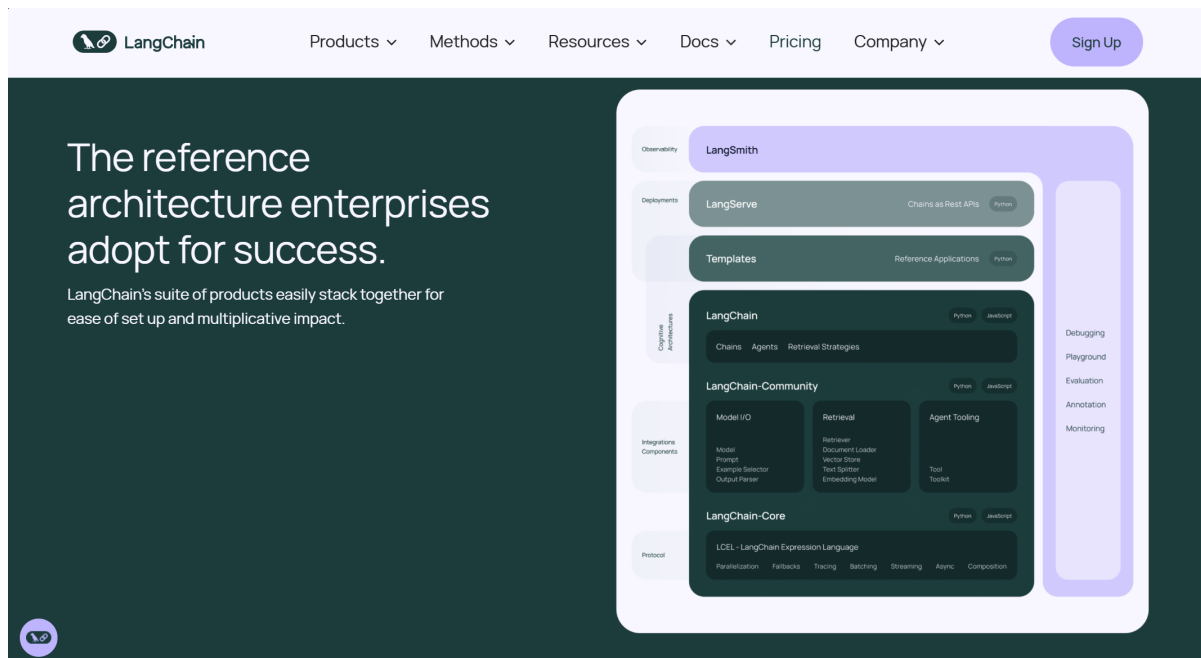
本项目主要包括三部分内容：

1. LLM 开发入门。V1 版本的简化版，旨在帮助初学者最快、最便捷地入门 LLM 开发，理解 LLM 开发的一般流程，可以搭建出一个简单的 Demo。
2. LLM 开发技巧。LLM 开发更进阶的技巧，包括但不限于：Prompt Engineering、多类型源数据的处理、优化检索、召回精排、Agent 框架等
3. LLM 应用实例。引入一些成功的开源案例，从本课程的角度出发，解析这些应用范例的 Idea、核心思路、实现框架，帮助初学者明白其可以通过 LLM 开发什么样的应用。

目前，第一部分已经完稿，欢迎大家阅读学习；第二、三部分正在创作中。

How to Deploy a Personalized LLM

Python Packages: [LangChain](#) | [LlamaIndex](#)



Software Application: [Ollama](#)



## Get up and running with large language models.


Run [Llama 3](#), [Phi 3](#), [Mistral](#), [Gemma](#), and other models. Customize and create your own.

[Download ↓](#)

Available for macOS, Linux, and Windows (preview)

### User-friendly LLM-based Chatbot Project

×



当前版本: v0.2.6-preview

对话

知识库管理

请选择对话模式:

LLM 对话

选择LLM模型:

Qwen-14B-Chat

Temperature:

0.10

0.001.00

历史对话轮数:

3

- +

导出记录

清空对话

🔄 RUNNING... Stop ⋮

🗨️ 给我讲一个童话故事

🗨️ 从前有一只小猪，它住在一个小村庄里。它非常喜欢吃甜食，每天都会去村里的糖果店买糖果。有一天，小猪发现糖果店的老板娘正在偷糖果，于是它决定帮助老板娘。它把糖果店的糖果都藏了起来，然后把它们还给了老板娘。老板娘非常感激小猪，于是她给小猪一些糖果作为回报。小猪很高兴，它把糖果带回家，和它的朋友们一起分享。从此以后，小猪和老板娘成为了好朋友，它们一起度过了许多快乐的时光。

🗨️ 好的，你能概述刚才的故事吗

🗨️ 故事讲述了一只小猪发现糖果店的老板娘正在偷糖果，于是它决定帮助老板娘。它把糖果店的糖果都藏了起来，然后把它们还给了老板娘。老板娘非常感激小猪，于是她给小猪一些糖果作为回报。小猪很高兴，它把糖果带回家，和它的朋友们一起分享。从此以后，小猪和老板娘成为了好朋友，它们一起度过了许多快乐的时光。

请输入对话内容，换行请使用Shift+Enter

➤



LangChain-Chatchat

当前版本: v0.2.6-preview

对话

知识库管理

请选择或新建知识库:

samples (faiss @ bge-large-zh-v1.5)

上传知识文件:

Drag and drop files here

Limit 200MB per file • HTML, MD, JSON, CSV, PDF, PNG, JPG, JPEG, BMP, EML, MSG, RST, RTF, TXT, XML, DOCX, EPUB, ODT, PPT, PPTX, TSV, HTM

Browse files

文件处理配置

单段文本最大长度: 250 相邻文本重合长度: 50 ☐ 开启中文标题加强

添加文件到知识库

知识库 samples 中已有文件:

知识库中包含源文件与向量库, 请从下表中选择文件后操作

	分词器	文档数量	源文件	向量库
redFileLoader	RecursiveCharacterTextSplitter	299	✓	✓

下载选中文档

添加至向量库

从向量库删除

从知识库中删除

依据源文件重建向量库

删除知识库

## Summary

Recall the three Questions

- What is Large Language Models?
- Why are LLMs important?
- How can we utilize LLMs?

## References

First Survey on LLM

arXiv > cs > arXiv:2303.18223

Search... All fields Search

Help | Advanced Search

Computer Science > Computation and Language

[Submitted on 31 Mar 2023 (v1), last revised 24 Nov 2023 (this version, v13)]

A Survey of Large Language Models

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, Yifan Du, Chen Yang, Yushuo Chen, Zhipeng Chen, Jinhao Jiang, Ruiyang Ren, Yifan Li, Xinyu Tang, Zikang Liu, Peiyu Liu, Jian-Yun Nie, Ji-Rong Wen

Language is essentially a complex, intricate system of human expressions governed by grammatical rules. It poses a significant challenge to develop capable AI algorithms for comprehending and grasping a language. As a major approach, language modeling has been widely studied for language understanding and generation in the past two decades, evolving from statistical language models to neural language models. Recently, pre-trained language models (PLMs) have been proposed by pre-training Transformer models over large-scale corpora, showing strong capabilities in solving various NLP tasks. Since researchers have found that model scaling can lead to performance improvement, they further study the scaling effect by increasing the model size to an even larger size. Interestingly, when the parameter scale exceeds a certain level, these enlarged language models not only achieve a significant performance improvement but also show some special abilities that are not present in small-scale language models. To discriminate the difference in parameter scale, the research community has coined the term large language models (LLM) for the PLMs of significant size. Recently, the research on LLMs has been largely advanced by both academia and industry, and a remarkable progress is the launch of ChatGPT, which has attracted widespread attention from society. The technical evolution of LLMs has been making an important impact on the entire AI community, which would revolutionize the way how we develop and use AI algorithms. In this survey, we review the recent advances of LLMs by introducing the background, key findings, and mainstream techniques. In particular, we focus on four major aspects of LLMs, namely pre-training, adaptation tuning, utilization, and capacity evaluation. Besides, we also summarize the available resources for developing LLMs and discuss the remaining issues for future directions.

Comments: ongoing work; 124 pages, 946 citations

Subjects: Computation and Language (cs.CL); Artificial Intelligence (cs.AI)

Cite as: arXiv:2303.18223 [cs.CL] (or arXiv:2303.18223v13 [cs.CL] for this version) https://doi.org/10.48550/arXiv.2303.18223

Access Paper:

- View PDF
- TeX Source
- Other Formats
- 双语版本(沉浸式翻译)

view license

Current browse context: cs.CL < prev | next > new | recent | 2303

Change to browse by: cs cs.AI

References & Citations

- NASA ADS
- Google Scholar
- Semantic Scholar

1 blog link (what is this?)

Export BibTeX Citation

Bookmark

[1] Y. Bengio, R. Ducharme, P. Vincent, and C. Janvin, "A neural probabilistic language model," J. Mach. Learn. Res., vol. 3, pp. 1137–1155, 2003.

[2] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. P. Kuksa, "Natural language processing (almost) from scratch," J. Mach. Learn. Res., vol. 12, pp. 2493–2537, 2011.

- [3] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," CoRR, vol. abs/2001.08361, 2020.
- [4] J. Wei, Y. Tay, R. Bommasani, C. Raffel, B. Zoph, S. Borgeaud, D. Yogatama, M. Bosma, D. Zhou, D. Metzler, E. H. Chi, T. Hashimoto, O. Vinyals, P. Liang, J. Dean, and W. Fedus, "Emergent abilities of large language models," CoRR, vol. abs/2206.07682, 2022.
- [5] M. Shanahan, "Talking about large language models," CoRR, vol. abs/2212.03551, 2022.
- [6] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. H. Chi, Q. Le, and D. Zhou, "Chain of thought prompting elicits reasoning in large language models," CoRR, vol. abs/2201.11903, 2022.
- [7] J. Hoffmann, S. Borgeaud, A. Mensch, E. Buchatskaya, T. Cai, E. Rutherford, D. de Las Casas, L. A. Hendricks, J. Welbl, A. Clark, T. Hennigan, E. Noland, K. Millican, G. van den Driessche, B. Damoc, A. Guy, S. Osindero, K. Simonyan, E. Elsen, J. W. Rae, O. Vinyals, and L. Sifre, "Training compute-optimal large language models," vol. abs/2203.15556, 2022.
- [8] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," CoRR, vol. abs/2211.09085, 2022.