6502 words | 17 minutes

Introduction to Large Language Model (LLM) Theory

1. What is a Large Language Model (LLM)

1.1 Concept of Large Language Model (LLM)

A large language model (LLM) is an artificial intelligence model designed to understand and generate human language .

LLM usually refers to **language models containing tens of billions (or more) parameters** , which are trained on massive amounts of text data to gain a deep understanding of the language. At present, well-known LLMs abroad include GPT-3.5, GPT-4, PaLM, Claude and LLaMA, and domestic ones include Wenxin Yiyan, iFlytek Spark, Tongyi Qianwen, ChatGLM, Baichuan, etc.

In order to explore the limits of performance, many researchers began to train increasingly large language models, such as with `1750 亿` parameters `GPT-3` and `5400 亿` with parameters `PaLM` . Although these large language models use similar architectures and pre-training tasks as small language models (such as `3.3 亿` with parameters `BERT` and `15 亿` with parameters `GPT-2`), they exhibit completely different capabilities, especially showing amazing potential in solving complex tasks, which is called " **emergent ability** ". Taking GPT-3 and GPT-2 as examples, GPT-3 can solve few-sample tasks by learning context, while GPT-2 performs poorly in this regard. Therefore, the scientific research community has given these huge language models a name, calling them "large language models (LLM)". An outstanding application of LLM is **ChatGPT** , which is a bold attempt to use the GPT series LLM for conversational applications with humans, showing a very smooth and natural performance.

1.2 Development History of LLM

The study of language modeling can be traced back to `20 世纪 90 年代` the time when the research focused on using **statistical learning methods** to predict words, predicting the next word by analyzing the previous words. However, it has certain limitations in understanding complex language rules.

Subsequently, researchers continued to try to improve it. **Bengio** 2003 年 , a pioneer in deep learning , first incorporated the idea of deep learning into the language model in his classic paper . The powerful **neural network model** is equivalent to providing a powerful "brain" for computers to understand language, allowing the model to better capture and understand the complex relationships in language. 《A Neural Probabilistic Language Model》

2018 年 Around this time, **neural network models with Transformer architecture** began to emerge. These models were trained with large amounts of text data, enabling them to deeply understand language rules and patterns by reading large amounts of text, just like letting computers read the entire Internet. This gave them a deeper understanding of language and greatly improved the performance of the models on various natural language processing tasks.

At the same time, researchers found that as **the size of the language model increases (increasing the model size or using more data)** , the model shows some amazing capabilities and significantly improves performance in various tasks. This discovery marks the beginning of the era of large language models (LLMs).

1.3 Common LLM models

Although the development of large language models has only been less than five years, the development speed is quite amazing. As of June 2023, more than 100 large models have been released at home and abroad. The following figure shows the influential large language models with more than 10 billion model parameters from 2019 to June 2023 according to the timeline:

(This figure is from reference [1])

Next, we will mainly introduce several common large models at home and abroad (including open source and closed source)

1.3.1 Closed-source LLM (undisclosed source code)

1.3.1.1 GPT series

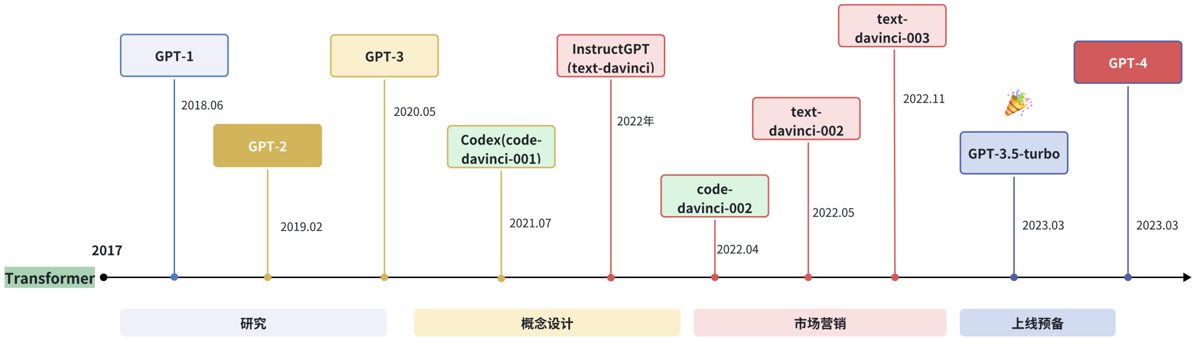
> OpenAI Model Introduction

The GPT (Generative Pre-Training) 2018 年 model proposed by **OpenAI** is a typical one. 生成式预训练语言模型

The basic principle of the GPT model is **to compress world knowledge into a decoder-only Transformer model through language modeling** , so that it can recover (or remember) the semantics of world knowledge and act as a general task solver. There are two key points to its success:

- Train a decoder-only Transformer language model that can accurately predict the next word
- Scaling the size of language models

OpenAI's research on LLM can be roughly divided into the following stages:



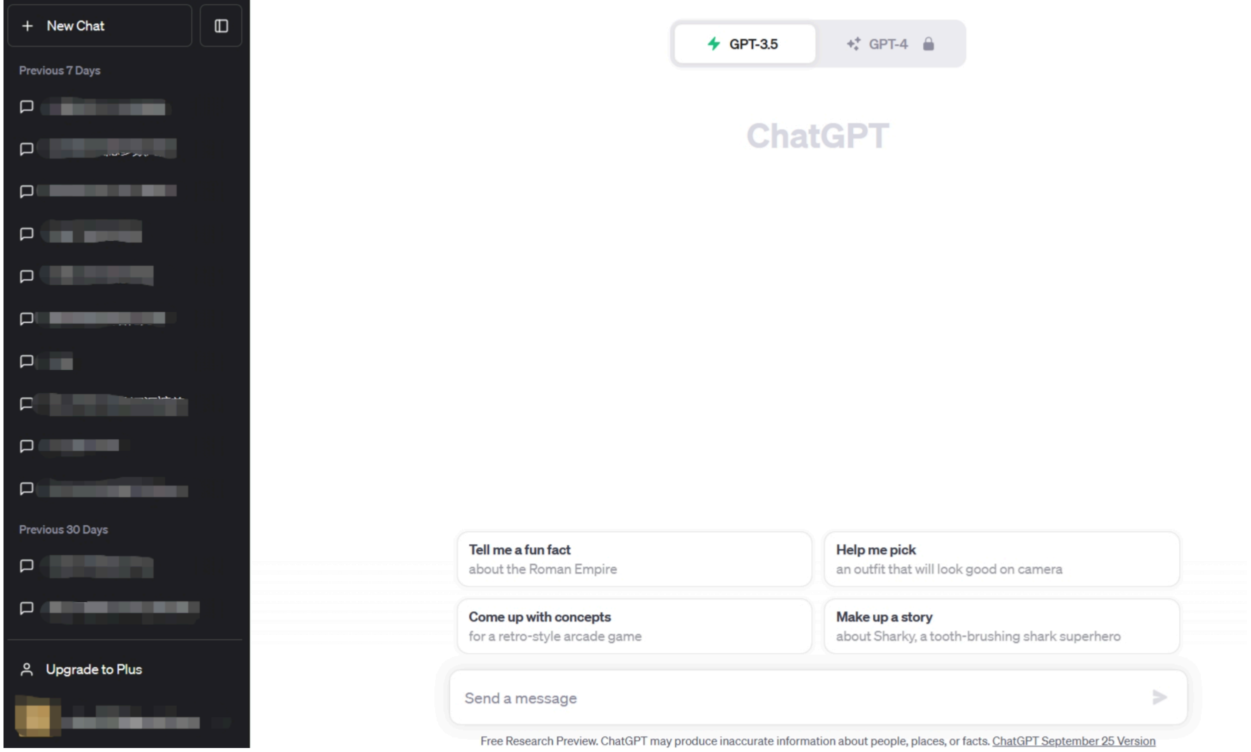
Next, we will introduce the well-known ChatGPT and GPT4 from the aspects of model scale and characteristics:

1.3.1.1.1 ChatGPT

> ChatGPT usage address

`2022 年 11 月` , **OpenAI** released **ChatGPT, a conversational application** based on the GPT model (GPT-3.5 and GPT-4) . ChatGPT has sparked excitement in the AI community since its release due to its outstanding ability to communicate with humans. ChatGPT is developed based on the powerful GPT model with specially optimized conversational capabilities.

ChatGPT is essentially an LLM application, which is developed based on the base model and is fundamentally different from the base model. It supports two versions: GPT-3.5 and GPT-4.



Tell me a fun fact
about the Roman Empire

Help me pick
an outfit that will look good on camera

Come up with concepts
for a retro-style arcade game

Make up a story
about Sharky, a tooth-brushing shark superhero

Send a message

Free Research Preview. ChatGPT may produce inaccurate information about people, places, or facts. ChatGPT September 25 Version

Today's ChatGPT supports up to 32,000 characters, with a knowledge deadline of September 2021. It can perform a variety of tasks, including **code writing, math problem solving, writing suggestions** , and more. ChatGPT has demonstrated outstanding ability to communicate with humans: it has a rich knowledge reserve, the skill to reason about math problems, accurately tracks context in multi-turn conversations, and is very consistent with the values of human safety use. Later, ChatGPT supported a plug-in mechanism, which further expanded the capabilities of ChatGPT with existing tools or applications. So far, it seems to be the most powerful chatbot in the history of artificial intelligence. The launch of ChatGPT has a significant impact on future artificial intelligence research, and it provides inspiration for exploring human-like artificial intelligence systems.

1.3.1.1.2 GPT-4

2023年3月 GPT-4 was released, which **expands text input to multimodal signals** . GPT3.5 has 175 billion parameters, and the number of parameters of GPT4 has not been officially announced, but relevant personnel have speculated that GPT-4 contains a total of 1.8 trillion parameters in 120 layers, which means that the scale of GPT-4 is more than 10 times that of GPT-3. Therefore, GPT-4 **is more capable of solving complex tasks than GPT-3.5, and has shown significant performance improvements on many evaluation tasks** .

A recent study investigated the capabilities of GPT-4 by qualitatively testing it on artificially generated questions covering a wide variety of difficult tasks and showed that GPT-4 can achieve superior performance than previous GPT models such as GPT3.5. In addition, thanks to six months of iterative calibration (with additional safety reward signals in RLHF training), GPT-4 responds more safely to malicious or provocative queries and applies some intervention strategies to mitigate issues that may arise with LLM, such as hallucinations, privacy, and over-reliance.

> **Note: On November 7, 2023, OpenAI held its first developer conference, where it launched its latest large language model, GPT-4 Turbo, which is equivalent to an advanced version. It extends the context length to 128k, equivalent to 300 pages of text, and updates the training knowledge to April 2023.**

GPT3.5 is free, but GPT-4 is paid. You need to subscribe to the plus membership for $20/month.

`2024 年 5 月 14 日` , the new generation flagship generative model **GPT-4o** was officially released. GPT-4o has the ability to deeply understand the three modalities of text, voice, and image, and is quick to respond, emotional, and extremely human. Moreover, GPT-4o is completely free, although the number of free uses per day is limited.

Usually we can call the model API to develop our own applications. The comparison of mainstream model APIs is as follows:

Language model name	Context length	Features	Input Fee ($/million tokens)	Output Fee ($/ 1M tokens)	Knowledge Deadline
GPT-3.5-turbo-0125	16k	Economy, specialized dialogue	0.5	1.5	September 2021
GPT-3.5-turbo-instruct	4k	Instruction Model	1.5	2	September 2021
GPT-4	8k	Better performance	30	60	September 2021
GPT-4-32k	32k	Strong performance, long context	60	120	September 2021
GPT-4-turbo	128k	Better performance	10	30	December 2023
GPT-4o	128k	Highest performance, faster speed	5	15	October 2023

Embedding model name	Dimensions	Features	Fees ($/ 1M tokens)
text-embedding-3-small	512/1536	Smaller	0.02
text-embedding-3-large	256/1024/3072	Larger	0.13
ada v2	1536	Tradition	0.1

1.3.1.2 Claude Series

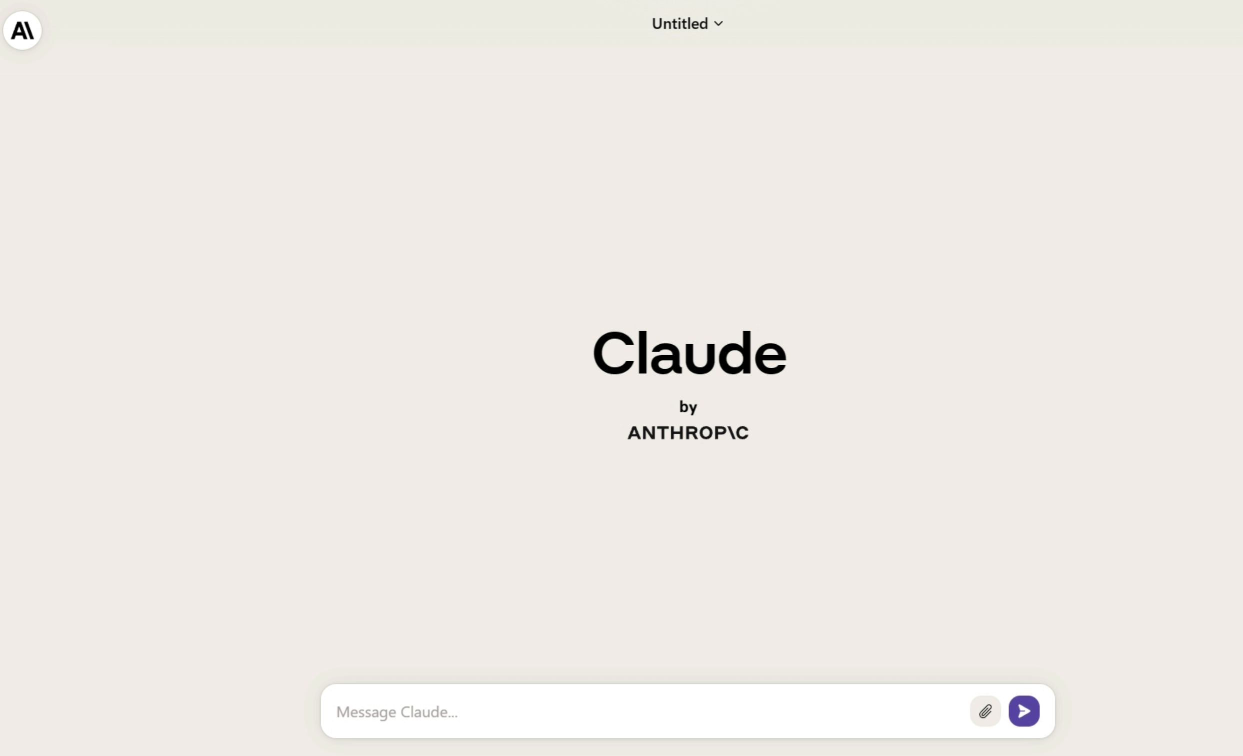
The Claude series of models are large closed-source language models developed by **Anthropic,** a company founded by former OpenAI employees .

> Claude uses the address

The earliest **Claude** was released on `2023 年 3 月 15 日` , and on July 11, 2023, it was updated to **Claude-2** , and then `2024 年 3 月 4 日` to **Claude-3** .

The Claude 3 series includes three different models, namely Claude 3 Haiku, Claude 3 Sonnet and Claude 3 Opus, with increasing capabilities to meet the needs of different users and application scenarios.

Model Name	Context length	Features	Input Fee ($/1M tokens)	Output Fee ($/1M tokens)
Claude 3 Haiku	200k	Fastest	0.25	1.25
Claude 3 Sonnet	200k	balance	3	15
Claude 3 Opus	200k	Highest performance	15	75



Untitled

Claude

by

ANTHROP\C

1.3.1.1.3 PaLM/Gemini Series

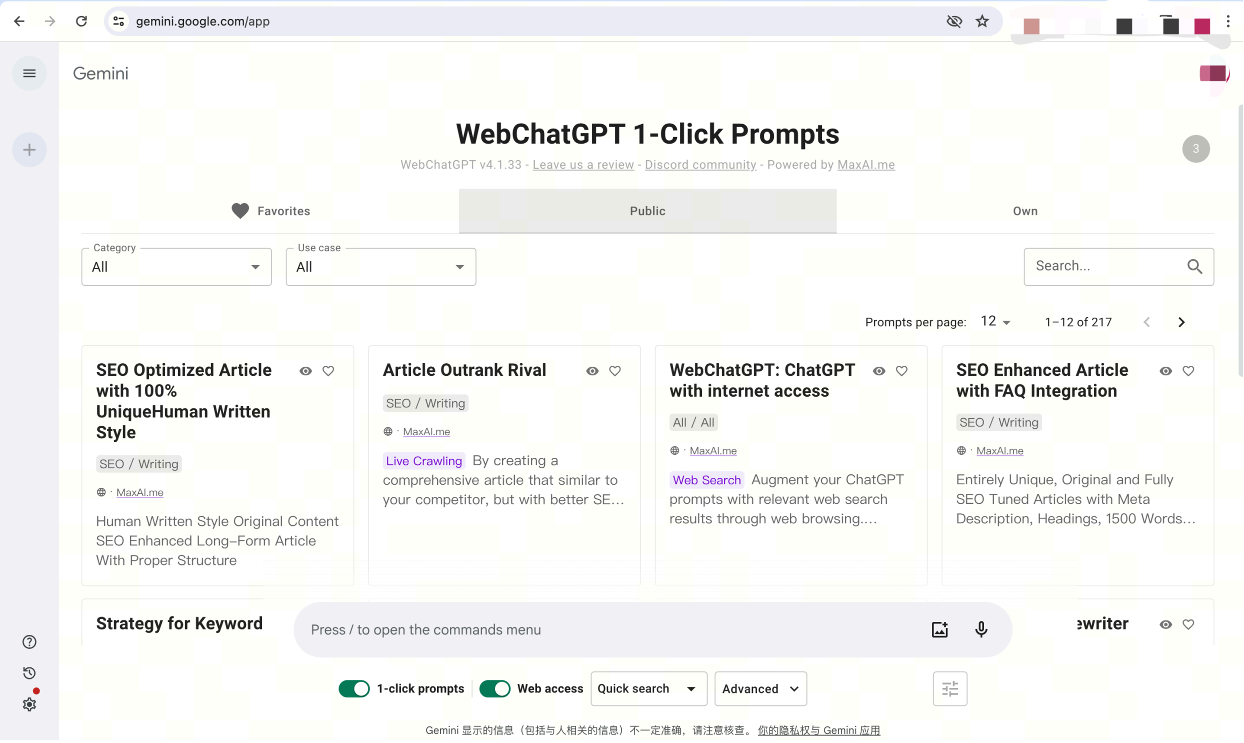
The PaLM series of language models was developed by **Google** . Its initial version was 2022 年 4 月 released in 2020, and its API was made public in March 2023. In May 2023, Google released **PaLM 2.** Google 2024 年 2 月 1 日 changed the underlying model driver of Bard (a previously released conversational application) from PaLM2 to **Gemini** , and also renamed the original Bard to **Gemini** .

> PaLM official website

> Gemini usage address

The current Gemini is the first version, Gemini 1.0, which is divided into three versions: Ultra, Pro and Nano according to different parameter quantities.

The following window is the Gemini interface:



WebChatGPT 1-Click Prompts

WebChatGPT v4.1.33 - Leave us a review - Discord community - Powered by MaxAI.me

Category: All
Use case: All
Search...

Prompts per page: 12 1–12 of 217

SEO Optimized Article with 100% UniqueHuman Written Style
SEO / Writing
MaxAI.me
Human Written Style Original Content SEO Enhanced Long-Form Article With Proper Structure

Article Outrank Rival
SEO / Writing
MaxAI.me
Live Crawling By creating a comprehensive article that similar to your competitor, but with better SE...

WebChatGPT: ChatGPT with internet access
All / All
MaxAI.me
Web Search Augment your ChatGPT prompts with relevant web search results through web browsing....

SEO Enhanced Article with FAQ Integration
SEO / Writing
MaxAI.me
Entirely Unique, Original and Fully SEO Tuned Articles with Meta Description, Headings, 1500 Words...

Strategy for Keyword

ewriter

Press / to open the commands menu

1-click prompts Web access Quick search Advanced

Gemini 显示的信息（包括与人相关的信息）不一定准确，请注意核查。你的隐私权与 Gemini 应用

1.3.1.1.4 Wen Xin Yi Yan

> Wenxinyiyan usage address

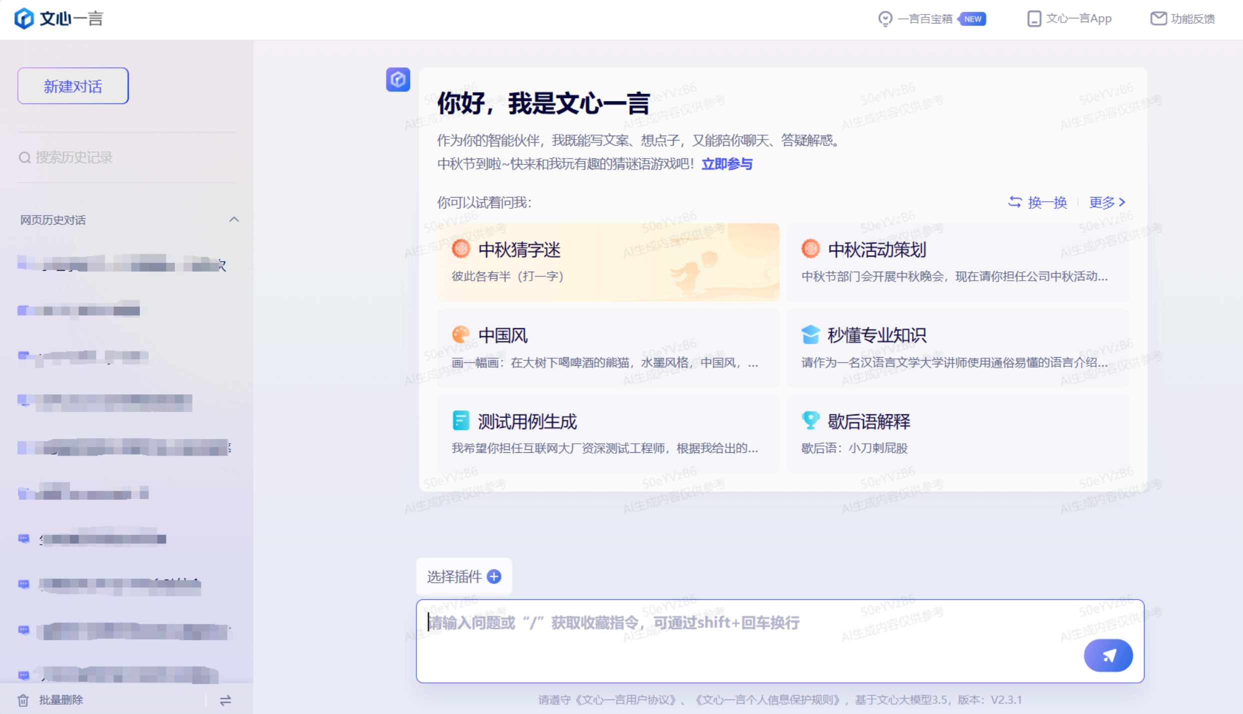
Wenxinyiyan is a knowledge-enhanced language big model based on Baidu's Wenxin big model . It 2023 年 3 月 was the first to be launched in China. Wenxinyiyan's basic model, Wenxin big model, released version 1.0 in 2019 and has now been updated to version **4.0** . Further classification, Wenxin big model includes NLP big model, CV big model, cross-modal big model, biocomputing big model, and industry big model. It is a closed-source model with relatively good Chinese capabilities.

The web version of Wenxin Yiyan is divided into **free version** and **professional version** .

- The free version uses Wenxin 3.5, which can already meet most of the needs of individual users or small businesses.
- The professional version uses Wenxin 4.0. The price is 59.9 yuan/month, and the monthly discount price is 49.9 yuan/month.

You can also use the API to make calls (billing details).

The following is the user interface of Wenxinyiyan:



[redacted]

你好，我是文心一言

作为你的智能伙伴，我既能写文案、想点子，又能陪你聊天、答疑解惑。

中秋节到啦~快来和我玩有趣的猜谜语游戏吧！**立即参与**

你可以试着问我：

中秋猜字迷
彼此各有半（打一字）

中秋活动策划
中秋节部门会开展中秋晚会，现在请你担任公司中秋活动...

中国风
画一幅画：在大树下喝啤酒的熊猫，水墨风格，中国风，...

秒懂专业知识
请作为一名汉语言文学大学讲师使用通俗易懂的语言介绍...

测试用例生成
我希望你担任互联网大厂资深测试工程师，根据我给出的...

歇后语解释
歇后语：小刀刺屁股

请输入问题或“/”获取收藏指令，可通过shift+回车换行

请遵守《文心一言用户协议》、《文心一言个人信息保护规则》，基于文心大模型3.5，版本：V2.3.1

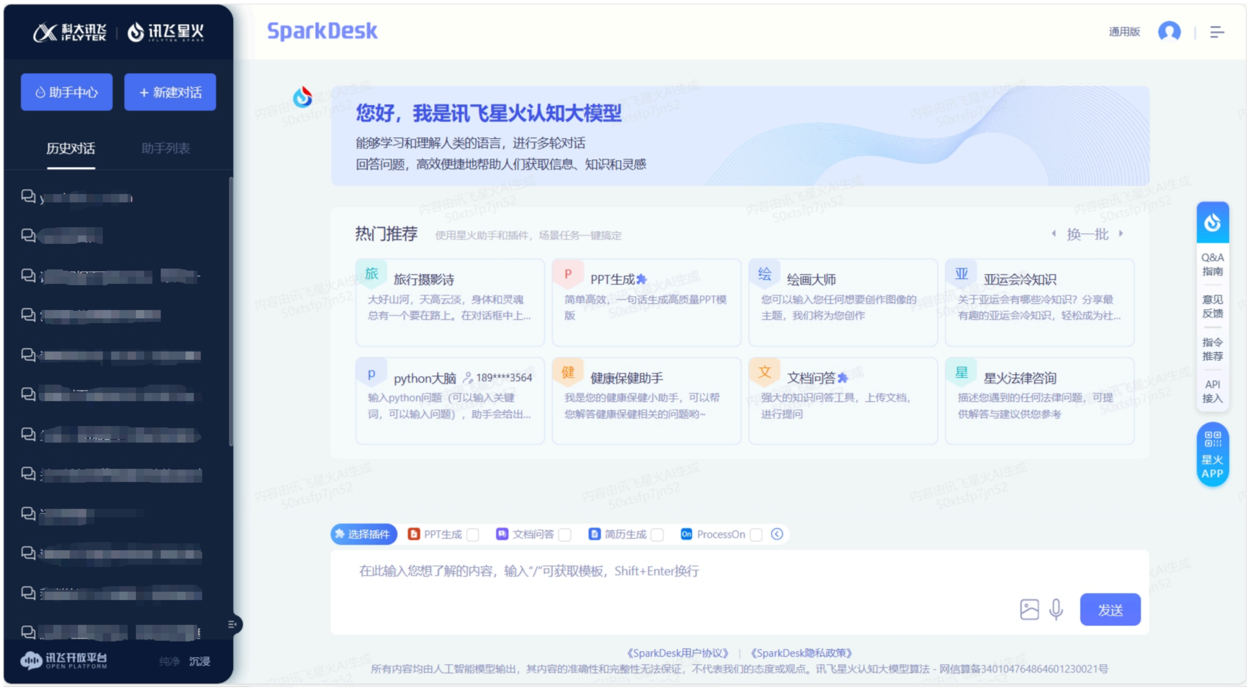
1.3.1.1.5 Spark Large Model

> Spark Large Model Usage Address

iFlytek Spark Cognitive Big Model is a language big model released by **iFlytek** 2023 年 5 月 , which supports a variety of natural language processing tasks. The model was first released and has been upgraded many times. 2023 年 10 月 iFlytek released **iFlytek Spark Cognitive Big Model V3.0** . 2024 年 1 月 iFlytek released **iFlytek Spark Cognitive Big Model V3.5** , which has been upgraded in seven aspects including language understanding, text generation, knowledge question and answer, and supports multiple functions such as system instructions and plug-in calls.

科大讯飞 iFLYTEK | 1024
第六届世界声博会暨2023科大讯飞全球1024开发者节
星火认知大模型V3.0正式发布 七大能力均持续提升
+7% 文本生成
+6% 语言理解
+7% 知识问答
+8% 逻辑推理
+3% 数学能力
+9% 代码能力
+7% 多模态能力
讯飞星火 IFLYTEK SPARK
*测试结果来源：按照《通用认知智能大模型评测体系》，基于真实用户数据，构建的全新测试集
世界声博会 WORLD VOICE EXPO
第六届世界声博会暨2023科大讯飞全球1024开发者节

The following is the user interface of iFlytek Spark:



[redacted]

您好，我是讯飞星火认知大模型

能够学习和理解人类的语言，进行多轮对话
回答问题，高效便捷地帮助人们获取信息、知识和灵感

热门推荐 使用星火助手和插件，场景任务一键搞定 换一批

旅行摄影诗
大好山河，天高云淡，身体和灵魂总有一个要在路上。在对话框中上...

PPT生成
简单高效，一句话生成高质量PPT模版

绘画大师
您可以输入您任何想要创作图像的主题，我们将为您创作

亚运会冷知识
关于亚运会有哪些冷知识？分享最有趣的亚运会冷知识，轻松成为社...

python大脑 189****3564
输入python问题（可以输入关键词，可以输入问题），助手会给出...

健康保健助手
我是您的健康保健小助手，可以帮您解答健康保健相关的问题哟~

文档问答
强大的知识问答工具，上传文档，进行提问

星火法律咨询
描述您遇到的任何法律问题，可提供解答与建议供您参考

在此输入您想了解的内容，输入"/"可获取模板，Shift+Enter换行

《SparkDesk用户协议》 | 《SparkDesk隐私政策》
所有内容均由人工智能模型输出，其内容的准确性和完整性无法保证，不代表我们的态度或观点。讯飞星火认知大模型算法 - 网信算备340104764864601230021号

1.3.2. Open Source LLM

1.3.2.1 LLaMA series

> LLaMA official website

> LLaMA open source address

LLaMA 系列模型是 Meta 开源的一组参数规模 **从 7B 到 70B** 的基础语言模型。LLaMA 于 2023 年 2 月 发布，2023 年 7 月发布了 LLaMA2 模型，并于 2024 年 4 月 18 日 发布了 **LLaMA3** 模型。它们都是在数万亿个字符上训练的，展示了如何**仅使用公开可用的数据集来训练最先进的模型**，而不需要依赖专有或不可访问的数据集。这些数据集包括 Common Crawl、Wikipedia、OpenWebText2、RealNews、Books 等。LLaMA 模型使用了**大规模的数据过滤和清洗技术**，以提高数据质量和多样性，减少噪声和偏见。LLaMA 模型还使用了高效的**数据并行**和**流水线并行**技术，以加速模型的训练和扩展。特别地，LLaMA 13B 在 CommonsenseQA 等 9 个基准测试中超过了 GPT-3 (175B)，而 **LLaMA 65B 与最优秀的模型 Chinchilla-70B 和 PaLM-540B 相媲美**。LLaMA 通过使用更少的字符来达到最佳性能，从而在各种推理预算下具有优势。

与 GPT 系列相同，LLaMA 模型也采用了 **decoder-only** 架构，同时结合了一些前人工作的改进：

- `Pre-normalization 正则化` ：为了提高训练稳定性，LLaMA 对每个 Transformer 子层的输入进行了 RMSNorm 归一化，这种归一化方法可以避免梯度爆炸和消失的问题，提高模型的收敛速度和性能；
- `SwiGLU 激活函数` ：将 ReLU 非线性替换为 SwiGLU 激活函数，增加网络的表达能力和非线性，同时减少参数量和计算量；
- `旋转位置编码（RoPE，Rotary Position Embedding）` ：模型的输入不再使用位置编码，而是在网络的每一层添加了位置编码，RoPE 位置编码可以有效地捕捉输入序列中的相对位置信息，并且具有更好的泛化能力。

LLaMA3 在 LLaMA 系列模型的基础上进行了改进，提高了模型的性能和效率：

- `更多的训练数据量` ：LLaMA3 在 15 万亿个 token 的数据上进行预训练，相比 LLaMA2 的训练数据量增加了 7 倍，且代码数据增加了 4 倍。LLaMA3 能够接触到更多的文本信息，从而提高了其理解和生成文本的能力。

- `更长的上下文长度` ：LLaMA3 的上下文长度增加了一倍，从 LLaMA2 的 4096 个 token 增加到了 8192。这使得 LLaMA3 能够处理更长的文本序列，改善了对长文本的理解和生成能力。

- `分组查询注意力（GQA，Grouped-Query Attention）` ：通过将查询（query）分组并在组内共享键（key）和值（value），减少了计算量，同时保持了模型性能，提高了大型模型的推理效率（LLaMA2 只有 70B 采用）。

- `更大的词表` ：LLaMA3 升级为了 128K 的 tokenizer，是前两代 32K 的 4 倍，这使得其语义编码能力得到了极大的增强，从而显著提升了模型的性能。

1.3.2.2 通义千问

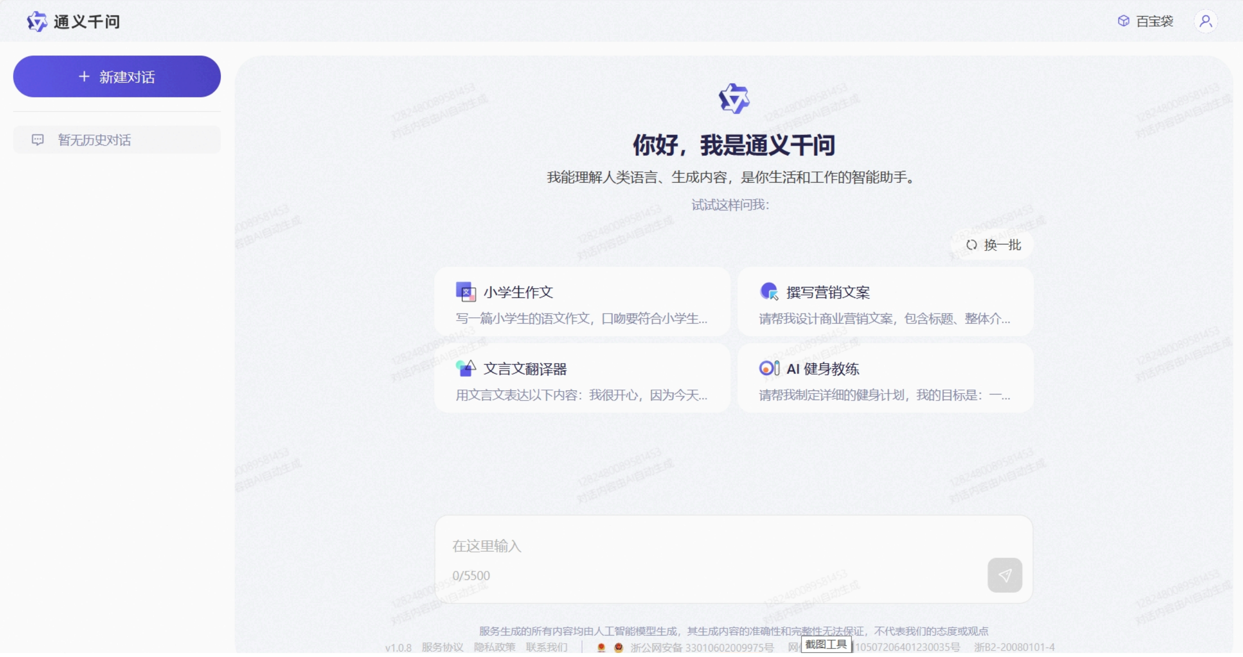
> **通义千问使用地址**

> **通义千问开源地址**

通义千问由阿里巴巴基于“通义”大模型研发，于 2023 年 4 月 正式发布。2023 年 9 月，阿里云开源了 Qwen（通义千问）系列工作。2024 年 2 月 5 日，开源了 **Qwen1.5**（Qwen2 的测试版）。并于 2024 年 6 月 6 日 正式开源了 **Qwen2**。 Qwen2 是一个 **decoder-Only** 的模型，采用 SwiGLU 激活 、 RoPE 、 GQA 的架构。中文能力相对来说非常不错的开源模型。

目前，已经开源了 5 种模型大小：**0.5B、1.5B、7B、72B 的 Dense 模型和 57B (A14B)的 MoE 模型**；所有模型均支持长度为 **32768 token** 的上下文。并将 Qwen2-7B-Instruct 和 Qwen2-72B-Instruct 的上下文长度扩展至 **128K token**。

以下是通义千问的使用界面：



你好，我是通义千问

我能理解人类语言、生成内容，是你生活和工作的智能助手。

试试这样问我：

小学生作文
写一篇小学生的语文作文，口吻要符合小学生...

撰写营销文案
请帮我设计商业营销文案，包含标题、整体介...

文言文翻译器
用文言文表达以下内容：我很开心，因为今天...

AI 健身教练
请帮我制定详细的健身计划，我的目标是：一...

在这里输入
0/5500

服务生成的所有内容均由人工智能模型生成，其生成内容的准确性和完整性无法保证，不代表我们的态度或观点

v1.0.8 服务协议 隐私政策 联系我们 | 浙公网安备 33010602009975号 网 10507206401230035号 浙B2-20080101-4

1.3.2.3 GLM 系列

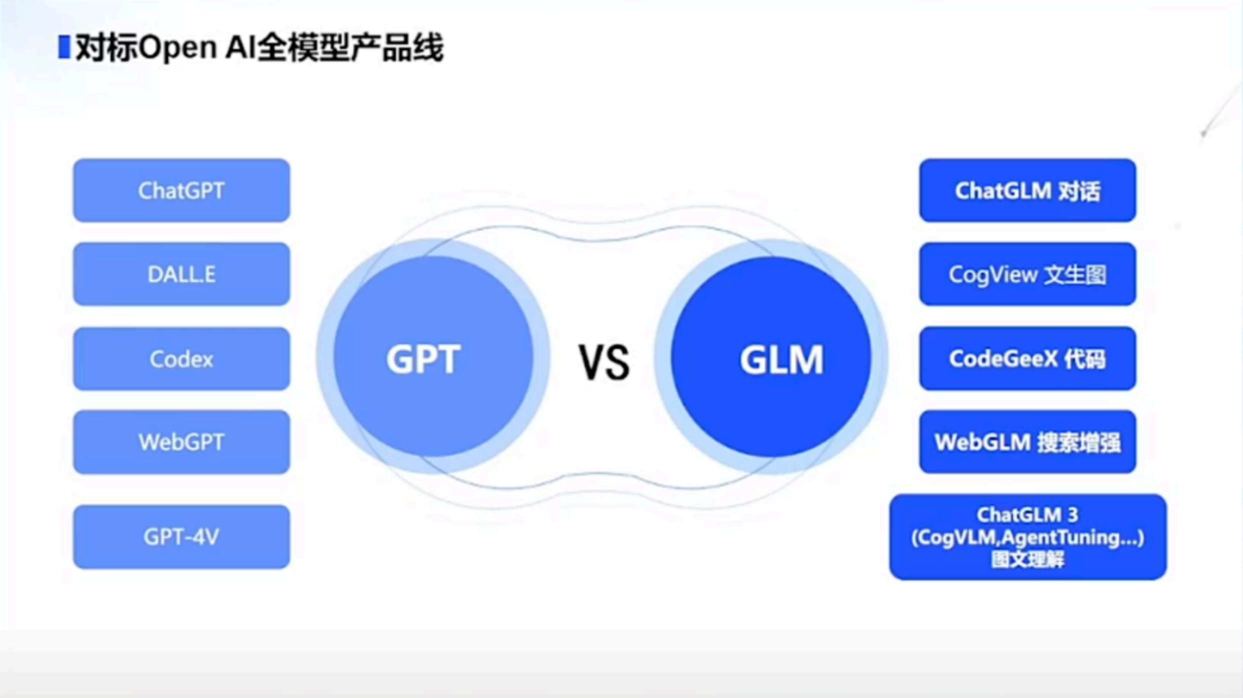
> **ChatGLM 使用地址**

> **ChatGLM 开源地址**

GLM 系列模型是清华大学和智谱 AI 等合作研发的语言大模型。2023 年 3 月 发布了 **ChatGLM**。6 月发布了 **ChatGLM 2**。10 月推出了 **ChatGLM3**。2024 年 1 月 16 日 发布了 **GLM4**，并于 2024 年 6 月 6 日 正式开源。

GLM-4-9B-Chat 支持多轮对话的同时，还具备网页浏览、代码执行、自定义工具调用（Function Call）和长文本推理（支持最大 **128K** 上下文）等功能。

开源了 对话模型 **GLM-4-9B-Chat**、 基础模型 **GLM-4-9B**、 长文本对话模型 **GLM-4-9B-Chat-1M**（支持 1M 上下文长度）、 多模态模型 **GLM-4V-9B** 等全面对标 OpenAI：



对标Open AI全模型产品线

ChatGPT

DALLE

Codex

WebGPT

GPT-4V

GPT VS GLM

ChatGLM 对话

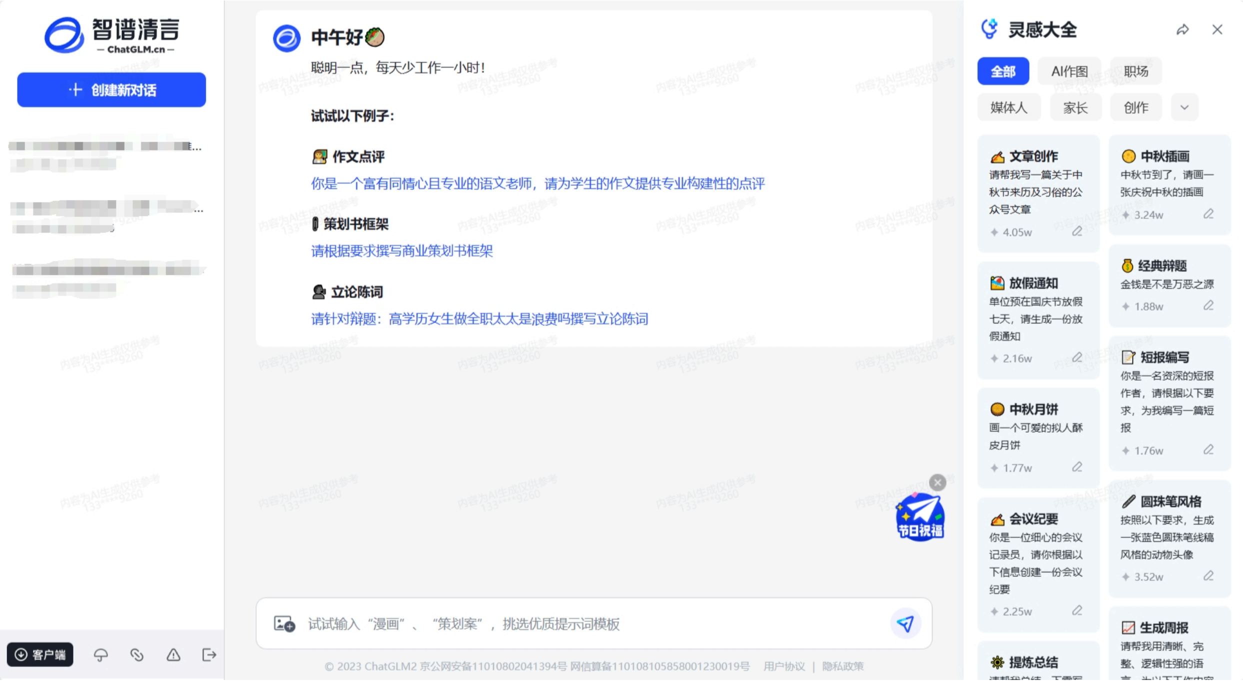
CogView 文生图

CodeGeeX 代码

WebGLM 搜索增强

ChatGLM 3
(CogVLM,AgentTuning...)
图文理解

以下是智谱清言的使用界面：



[redacted]

中午好

聪明一点，每天少工作一小时！

试试以下例子：

作文点评
你是一个富有同情心且专业的语文老师，请为学生的作文提供专业构建性的点评

策划书框架
请根据要求撰写商业策划书框架

立论陈词
请针对辩题：高学历女生做全职太太是浪费吗撰写立论陈词

试试输入“漫画”、“策划案”，挑选优质提示词模板

© 2023 ChatGLM2 京公网安备11010802041394号 网信算备110108105858001230019号 用户协议 | 隐私政策

灵感大全

文章创作
请帮我写一篇关于中秋节来历及习俗的公众号文章
✦ 4.05w

中秋插画
中秋节到了，请画一张庆祝中秋的插画
✦ 3.24w

放假通知
单位预在国庆节放假七天，请生成一份放假通知
✦ 2.16w

经典辩题
金钱是不是万恶之源
✦ 1.88w

中秋月饼
画一个可爱的拟人酥皮月饼
✦ 1.77w

短报编写
你是一名资深的短报作者，请根据以下要求，为我编写一篇短报
✦ 1.76w

会议纪要
你是一位细心的会议记录员，请你根据以下信息创建一份会议纪要
✦ 2.25w

圆珠笔风格
按照以下要求，生成一张蓝色圆珠笔线稿风格的动物头像
✦ 3.52w

提炼总结

生成周报
请帮我用清晰、完整、逻辑性强的语

1.3.2.4 Baichuan 系列

> **百川使用地址**

> **百川开源地址**

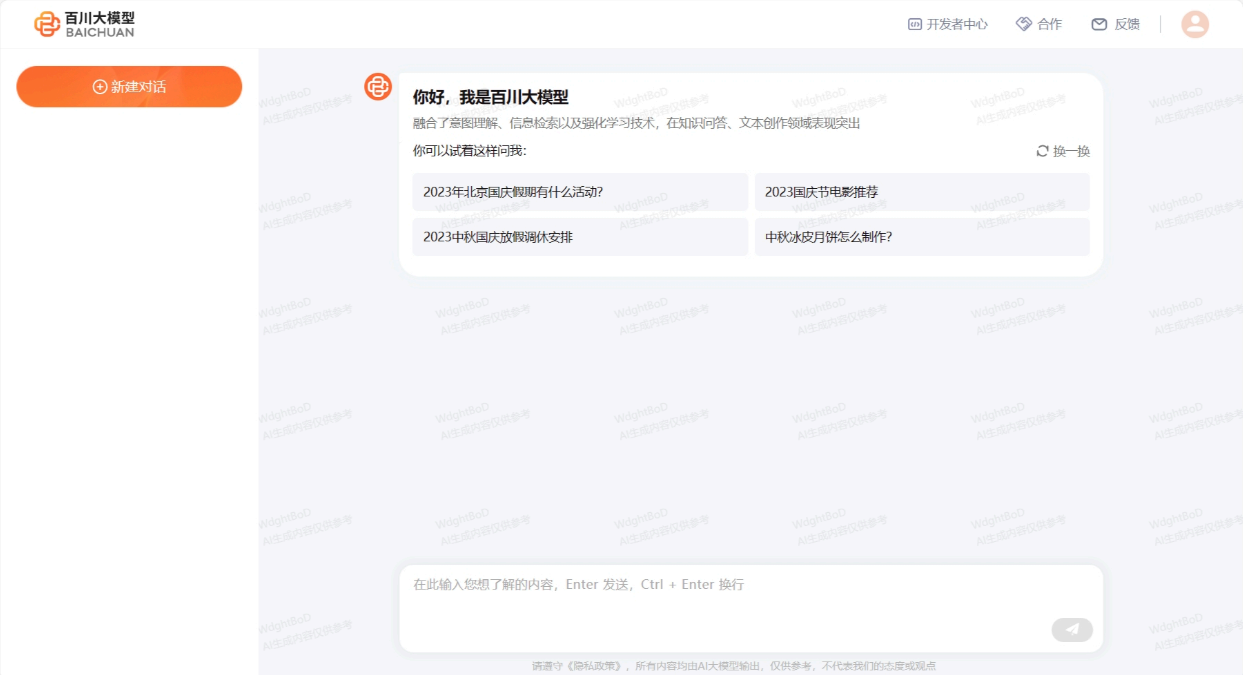
Baichuan 是由**百川智能**开发的**开源可商用**的语言大模型。其基于Transformer **解码器架构**（decoder-only）。

2023 年 6 月 15 日发布了 **Baichuan-7B** 和 **Baichuan-13B**。百川同时开源了**预训练**和**对齐**模型，预训练模型是面向开发者的“基座”，而 对齐模型则面向广大需要对话功能的普通用户 。

Baichuan2 was launched in 2011. **7B and 13B** Base **and** Chat **versions** 2023年 9 月 6 日 were released , and **4bits quantization** was provided for the Chat version .

2024 年 1 月 29 日 **Baichuan 3** has been released . However, **it is not open source yet** .

The following is the user interface of Baichuan model:



你好，我是百川大模型

融合了意图理解、信息检索以及强化学习技术，在知识问答、文本创作领域表现突出

你可以试着这样问我:

2023年北京国庆假期有什么活动?

2023国庆节电影推荐

2023中秋国庆放假调休安排

中秋冰皮月饼怎么制作?

在此输入您想了解的内容，Enter 发送，Ctrl + Enter 换行

请遵守《隐私政策》，所有内容均由AI大模型输出，仅供参考，不代表我们的态度或观点

II. LLM’s Capabilities and Characteristics

2.1 LLM Capabilities

2.1.1 Emergent abilities

One of the most notable features that distinguish large language models (LLMs) from previous pre-trained language models (PLMs) is their 涌现能力 emergence ability, a surprising ability that is not obvious in small models but is particularly prominent in large models. Similar to the phase transition phenomenon in physics, the emergence ability is like the model performance rapidly improving as the scale increases, exceeding the random level, which is what we often call **quantitative change leading to qualitative change** .

Emergent capabilities can be related to certain complex tasks, but we are more concerned with its general capabilities. Next, we briefly introduce three typical emergent capabilities of LLM:

1. **Contextual learning** : The contextual learning capability was first introduced by GPT-3. This capability allows the language model to perform tasks by understanding the context and generating corresponding outputs when provided with natural language instructions or multiple task examples, without the need for additional training or parameter updates.

2. **Instruction Following** : Fine-tuned on multi-task data using natural language descriptions, so-called 指令微调 LLMs are shown to perform well on unseen tasks formally described using instructions. This means that LLMs are able to perform tasks based on task instructions without having seen specific examples beforehand, demonstrating their strong generalization capabilities.

3. **Step-by-step reasoning** : Small language models often have difficulty solving complex tasks that involve multiple reasoning steps, such as math problems. However, LLMs `思维链（CoT, Chain of Thought)` solve these tasks by adopting a reasoning strategy that uses a hint mechanism that includes intermediate reasoning steps to arrive at the final answer. It is speculated that this ability may be acquired through training on the code.

These emergent capabilities allow LLMs to excel in a variety of tasks, making them powerful tools for solving complex problems and applying them in multiple fields.

2.1.2 Ability to support multiple applications as a base model

In 2021, researchers from Stanford University and other universities proposed the concept of foundation model, clarifying the role of pre-trained models. This is a new AI technology paradigm that uses training on massive amounts of unlabeled data to obtain large models (single or multimodal) that can be applied to a large number of downstream tasks. In this way, **multiple applications can rely on only one or a few large models for unified construction** .

The large language model is a typical example of this new model. Using a unified large model can greatly improve R&D efficiency. Compared with the way of developing a single model each time, this is an essential improvement. Large models can not only shorten the development cycle of each specific application and reduce the required manpower investment, but also achieve better application results based on the reasoning, common sense and writing ability of large models. Therefore, large models can become a unified base model for AI application development. This is a new paradigm that achieves multiple goals at one stroke and deserves to be vigorously promoted.

2.1.3 Supporting the ability to use dialogue as a unified entry point

The opportunity that made the large language model really popular was **ChatGPT** , which is based on conversational chat. The industry has long discovered users' special preference for conversational interaction. When Lu Qi was at Microsoft, he promoted the strategy of "conversation as a platform" in 2016. In addition, products based on voice conversations such as Apple Siri and Amazon Echo are also very popular, reflecting the preference of Internet users for chat and conversation as interaction modes. Although there were various problems with previous chatbots, the emergence of large language models has once again allowed chatbots as an interaction mode to re-emerge. Users are increasingly looking forward to artificial intelligence like "Jarvis" in Iron Man, who is omnipotent and omniscient. This has triggered our 智能体 `(Agent)` thinking about the prospects of type applications. Projects such as Auto-GPT and Microsoft Jarvis have already appeared and received attention. I believe that many similar projects will emerge in the future to allow assistants to complete various specific tasks in the form of conversations.

2.2 Characteristics of LLM

Large language models have several notable features that have made them a popular area of interest and research in natural language processing and other fields. Here are some of the main features of large language models:

1. **Huge scale:** LLMs usually have huge parameter scale, which can reach billions or even hundreds of billions of parameters. This allows them to capture more linguistic knowledge and complex grammatical structures.

2. **Pre-training and fine-tuning:** LLM adopts a pre-training and fine-tuning learning method. First, it is pre-trained on large-scale text data (unlabeled data) to learn general language representation and knowledge. Then it is adapted to specific tasks through fine-tuning (labeled data), so that it performs well in various NLP tasks.

3. **Context-aware:** LLMs have strong context-awareness when processing text, and are able to understand and generate text content that depends on previous text. This makes them excellent in conversation, article generation, and situational understanding.

4. **Multi-language support:** LLMs can be used in multiple languages, not just English. Their multi-lingual capabilities make cross-cultural and cross-linguistic applications easier.

5. **Multimodal support:** Some LLMs have been extended to support multimodal data, including text, images, and sound, allowing them to understand and generate content of different media types and achieve more diverse applications.

6. **Ethical and risk issues:** Although LLMs have excellent capabilities, they also raise ethical and risk issues, including the generation of harmful content, privacy issues, cognitive biases, etc. Therefore, research and application of LLMs require caution.

7. **High computing resource requirements:** LLM parameters are large in scale and require a lot of computing resources for training and reasoning. Usually, high-performance GPU or TPU clusters are required to implement it.

Large language models are a technology with powerful language processing capabilities that have demonstrated potential in many fields. They provide powerful tools for natural language understanding and generation tasks, but also raise concerns about their ethical and risk issues. These characteristics make LLM an important research and application direction in computer science and artificial intelligence today.

III. Application and Impact of LLM

LLM has had a profound impact in many fields. In the field of **natural language processing** , it can help computers better understand and generate text, including writing articles, answering questions, translating languages, etc. In the field of **information retrieval** , it can improve search engines and make it easier for us to find the information we need. In the field of **computer vision** , researchers are also working to make computers understand images and text to improve multimedia interactions.

Most importantly, the emergence of LLM has made people rethink the possibility of **artificial general intelligence (AGI)** . AGI is artificial intelligence that thinks and learns like humans. LLM is considered an early form of AGI, which has triggered many thoughts and plans for the future development of artificial intelligence.

In summary, LLM is an exciting technology that allows computers to better understand and use language, is changing the way we interact with technology, and is also triggering endless exploration of the future of artificial intelligence.

> **In the next chapter we will introduce RAG, an important technology in the LLM period.**

【Reference content】：

1. A Survey of Large Language Models
2. Zhou Feng: When we talk about big models, what new capabilities should we focus on?

Next Chapter >

2. Introduction to Retrieval Enhancement Generation RAG