开源大模型调研 by kjn

开源大模型基本信息表

数据来源：list of models, LLMs

model	producer	price_input	price_output	download	paper	badcase
internlm2-chat-7b	上海人工智能实验室	0.3	0.3	link	link	link
internlm2_5-7b-chat	上海人工智能实验室	0.3	0.3	link	/	link
Yi-1.5-9B-Chat	零一万物	0.4	0.4	link	link	link
Llama-3-8B-Instruct	meta	0.4	0.4	link	link	link
Llama-3.1-8B-Instruct	meta	0.4	0.4	link	link	link
glm-4-9b-chat	智谱AI	0.6	0.6	link	link	link
gemma-2-9b-it	google	0.6	0.6	link	link	link
internlm2-chat-20b	上海人工智能实验室	1.0	1.0	link	link	link
deepseek-llm-67b-chat	深度求索	1.0	1.0	link	/	link
internlm2_5-20b-chat	上海人工智能实验室	1.0	1.0	link	/	link
Yi-1.5-34B-Chat	零一万物	1.3	1.3	link	link	link
deepseek-chat-v2	深度求索	1.0	2.0	/	link	link
qwen1.5-7b-chat	阿里巴巴	1.0	2.0	link	/	link
qwen2-7b-instruct	阿里巴巴	1.0	2.0	link	link	link
qwen2.5-7b-instruct	阿里巴巴	1.0	2.0	link	/	link
qwen1.5-14b-chat	阿里巴巴	2.0	4.0	link	/	link

model	producer	price_input	price_output	download	paper	badcase
Llama-3-70B-Instruct	meta	4.1	4.1	link	link	link
Llama-3.1-70B-Instruct	meta	4.1	4.1	link	link	link
qwen2.5-14b-instruct	阿里巴巴	2.0	6.0	link	/	link
qwen1.5-32b-chat	阿里巴巴	3.5	7.0	link	/	link
qwen2-57b-a14b-instruct	阿里巴巴	3.5	7.0	link	link	link
qwen2.5-32b-instruct	阿里巴巴	3.5	7.0	link	/	link
qwen1.5-72b-chat	阿里巴巴	5.0	10.0	link	/	link
qwen2-72b-instruct	阿里巴巴	5.0	10.0	link	link	link
qwen2.5-72b-instruct	阿里巴巴	4.0	12.0	link	/	link
aquilachat2-34b	智源研究院	/	/	link	/	link
AquilaChat2-70B-Expr	智源研究院	/	/	link	/	link
Phi-3-mini-128k-instruct	微软	/	/	link	link	link
openbuddy-deepseek-67b	openbuddy	/	/	link	/	link
openbuddy-mixtral-7bx8	openbuddy	/	/	link	/	link
openbuddy-llama3-8b	openbuddy	/	/	link	/	link
Baichuan2-13B-Chat	百川智能	/	/	link	link	link
Baichuan2-7B-Chat	百川智能	/	/	link	link	link
qwen1.5-0.5b-chat	阿里巴巴	/	/	link	/	link
qwen1.5-1.8b-chat	阿里巴巴	/	/	link	/	link
qwen1.5-4b-chat	阿里巴巴	/	/	link	/	link
gemma-7b-it	google	/	/	link	/	link
gemma-2b-it	google	/	/	link	/	link

model	producer	price_input	price_output	download	paper	badcase
MiniCPM-2B-dpo	面壁智能	/	/	link	/	link
qwen2-1.5b-instruct	阿里巴巴	/	/	link	link	link
qwen2-0.5b-instruct	阿里巴巴	/	/	link	link	link
internlm2-chat-1_8b	上海人工智能实验室	/	/	link	link	link
DeepSeek-V2-Lite-Chat	深度求索	/	/	link	link	link

Bencmark收集：

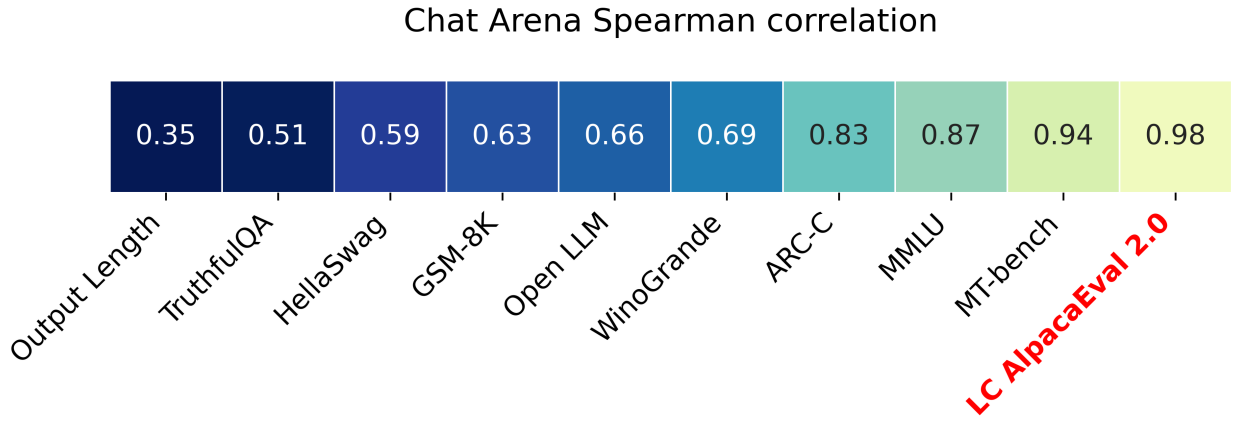
1. CLUEbenchmark/CLUE: 中文语言理解测评基准 Chinese Language Understanding Evaluation Benchmark: datasets, baselines, pre-trained models, corpus and leaderboard
   官网：CLUE中文语言理解基准测评

   CLUE定位：为更好的服务中文语言理解、任务和产业界，做为通用语言模型测评的补充，通过搜集整理发布中文任务及标准化测评等方式完善基础设施，最终促进中文NLP的发展。中文语言理解测评基准，包括代表性的数据集、基准(预训练)模型、语料库、排行榜。

2. AI4LIFE-GROUP/OpenXAI: OpenXAI : Towards a Transparent Evaluation of Model Explanations
   OpenXAI是第一个通用的轻量级库，它提供了一个全面的函数列表，用于系统地评估由基于属性的解释方法生成的解释的质量。OpenXAI支持开发新的数据集（包括合成的和真实的）和解释方法，并强烈倾向于促进对解释方法的系统的、可重复的和透明的评估。

   OpenXAI是一个开源计划，它包含一系列精心策划的高风险数据集、模型和评估指标，并提供了一个简单且易于使用的工具

3. tatsu-lab/alpaca_eval: An automatic evaluator for instruction-following language models. Human-validated, high-quality, cheap, and fast.
   AlpacaEval 2.0 with length-controlled win-rates与ChatBot Arena的speraman相关性为0.98，同时花费不到10美元的OpenAI积分，运行时间不到3分钟。我们的目标是为聊天llm建立一个基准：快速（< 5分钟），便宜（< 10美元），并且与人类高度相关（0.98）。以下是与其他基准测试的比较：



4. Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots
   Chatbot Arena使用人肉众包、随机进行评测。 使用 50 万以上的用户投票来计算 Elo 评分。

   人类用户在Chatbot Arena提问，会并排显示两个不同的模型的响应，但是不知道哪个模型生成了哪个响应。 然后人类用户投票决定他们更喜欢哪种回答。

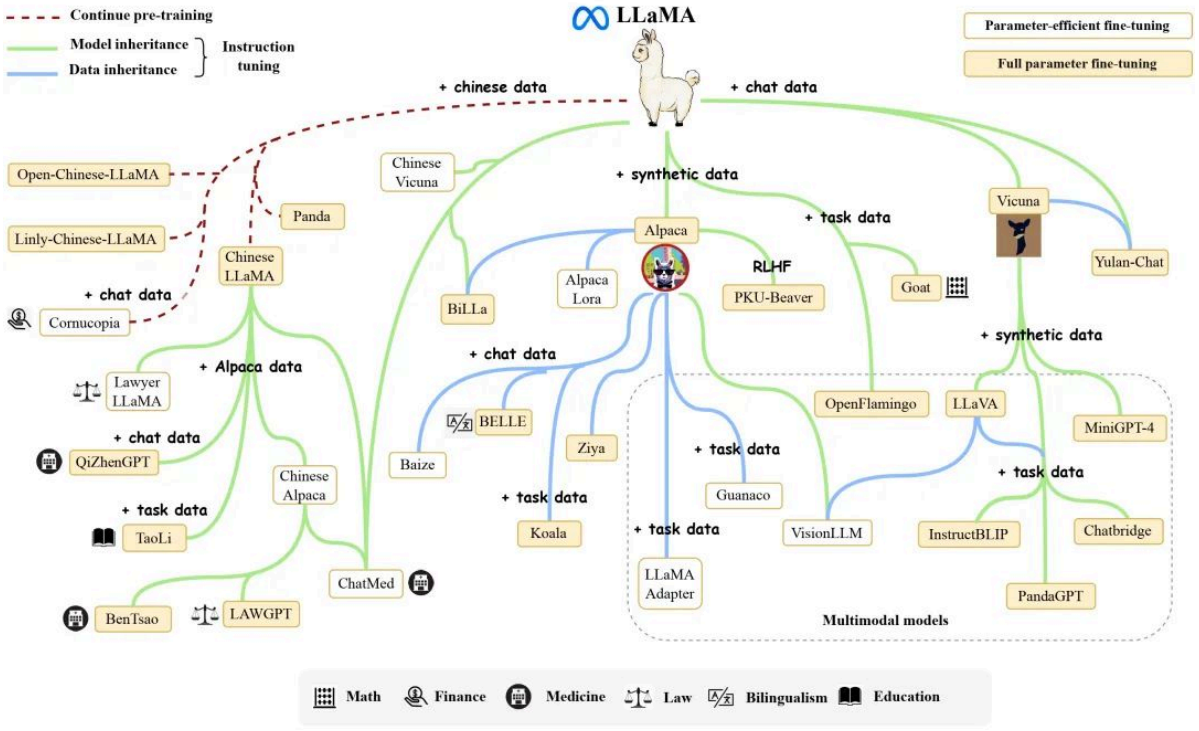
To be continued...

开源大模型调研

1. **Llama**

   - **LLaMA 1** 于2023年2月发布，是一系列从 7 B到 65B 参数的基础语言模型。Meta 训练这些模型使用了数万亿个 token，并且**证明了完全可以只使用公开可得的数据集来训练最先进的模型，而无需使用专有和不可获取的数据集**。特别是，LLaMA-13B 在大多数基准测试中表现优于GPT-3（175B），而 LLaMA-65B 在竞争中与最佳模型

Chinchilla70B 和PaLM-540B 持平。因为开源协议问题，Llama-1不可免费商用。



- [Llama 2](#) 于2023年7月发布，参数规模从70亿到700亿不等。Meta有不少微调LLMs，比如为对话场景而优化的Llama 2-Chat，专注于代码生成的Code-Llama。Llama 2模型在大多数基准上都比开源的对话模型表现得更好，并且根据人类评估的有用性和安全性，可能是闭源模型的合适替代品。Meta提供了他们对Llama 2-Chat进行微调和安全改进的方法的详细描述。相比于 Llama 1 ，Llama 2 的训练数据多了 40%，上下文长度也翻倍，并采用了分组查询注意力机制。具体来说，Llama 2预训练模型是在2 万亿的token上训练的，精调 Chat 模型是在100 万人类标记数据上训练的。

- Llama 2 模型优于 Llama 1 模型。特别是，与 Llama 1 65B 相比，Llama 2 70B 在MMLU 和 BBH上的结果分别提高了≈5和≈8个点。Llama 2 7B 和 30B 模型在除代码基准外的所有类别上都优于相应大小的 MPT模型。对于 Falcon 模型，Llama 2 7B 和34B 在所有类别的基准上都优于 Falcon 7B 和 40B 模型。

- **Llaama 3** 于2024年4月18日发布，是Meta最先进开源大型语言模型的下一代，包括具有80亿和700亿参数的预训练和指令微调的语言模型，能够支持广泛的应用场景。

仓库：meta-llama/llama3: The official Meta Llama 3 GitHub site
官网：Introducing Meta Llama 3: The most capable openly available LLM to date
Huggingface：meta-llama (Meta Llama)
参考文章：
LLMs之Llama3：Llama-3的简介、安装和使用方法、案例应用之详细攻略-CSDN博客
【首发】万字长文解读：Llama系列的全面考察！-CSDN博客
一文读懂Llama 2（从原理到实战）
仓库：Meta Llama

2. **Phi**

- **Phi** 是由微软 AI 研究院开发的一款开源「小型语言模型」，具有商用属性，其主要特点是体积小巧，所需资源较少。

- **Phi-3-Mini** 模型拥有 3.8B 的参数，并经过 3.3T token 的训练数据。在多项学术基准测试中，Phi-3-Mini 的性能与市场上的大型模型相接近或等同。例如，在 MMLU 测试中，其得分为 69%；在 MT-bench 测试中，得分为 8.38 分。

- **Phi-3-Small** 模型具备 7B 参数，并使用了 4.8T token 的训练数据。在 MMLU 测试中，其得分为 75%，在 MT-bench 测试中，得分为 8.7 分。

- **Phi-3-Medium** 模型则拥有 14B 参数，同样使用了 4.8T token 的训练数据。在 MMLU 测试中，其得分为 78%，在 MT-bench 测试中，得分为 8.9 分。

- **核心优势**：在于其小巧的体积（Phi-3-Mini 的内存占用约为 1.8GB），尤其适用于移动设备。Phi-3 模型的小巧之处体现在，它能够在手机上运行，例如在 iPhone 上，每秒能生成 16 个 token 的信息，相当于大约 12 个单词。这种便携性和高效性使得 Phi-3 成为移动端应用和实时交互的理想选择。可以实现离线部署，加强隐私保护。

- **一些缺点**：由于模型体积较小，其存储的知识量相对有限。在处理需要广泛事实知识支持的任务（如 TriviaQA 测试）时，这一局限性尤为明显，可能导致性能下降和产生不准确的输出。然而，这种问题可以通过与搜索引擎集成来部分解决，利用搜索引擎提供额外的信息支持，从而增强模型的知识库和应对复杂任务的能力。只能处理英语。

- **技术性能**：从分数上来看，即使是 Phi-3 系列中最小的版本——Phi-3-Mini，也已经全面超越了最近发布的 Llama 3 模型。Phi-3-Mini 在大规模多任务语言理解（MMLU）基准测试中取得了 69% 的得分，在 MT-bench 测试中得分为 8.38。这一成绩表明，即使是 Phi-3 系列中最小的模型，也具备了与大型模型如 Mixtral 8x7B 和 GPT-3.5 相匹敌的竞争力。

- Phi-3-Small（7B 参数）和 Phi-3-Medium（14B 参数）的表现更为出色。在 MMLU 测试中，Phi-3-Small 的得分为 75%，而 Phi-3-Medium 的得分则进一步提升至 78%。在 MT-bench 测试中，Phi-3-Small 和 Phi-3-Medium 的得分分别为 8.7 和 8.9。

Phi-3官网：Phi-3 - a microsoft Collection
仓库：microsoft/Phi-3CookBook
Huggingface：Phi-3 - a microsoft Collection
参考文章：Phi-3：微软小模型今日发布，手机上超越 Llama3 - 知乎

3. **Qwen**

- **Qwen2.5**于2024年9月19日发布，旗舰模型Qwen2.5-72B性能超越Llama405B。Qwen2.5全系列模型都在18T tokens数据上进行预训练，相比Qwen2，整体性能提升18%以上，拥有更多的知识、更强的编程和数学能力。Qwen2.5-72B模型在MMLU-rudex基准（考察通用知识）、MBPP基准（考察代码能力）和MATH基准（考察数学能力）的得分高达86.8、88.2、83.1。

- Qwen2.5支持高达128K的上下文长度，可生成最多8K内容。模型拥有强大的多语言能力，支持中文、英文、法文、西班牙文、俄文、日文、越南文、阿拉伯文等29种以上语言。模型能够丝滑响应多样化的系统提示，实现角色扮演和聊天机器人等任务。在指令跟随、理解结构化数据（如表格）、生成结构化输出（尤其是JSON）等方面Qwen2.5都进步明显。

- 语言模型方面，Qwen2.5开源了7个尺寸，0.5B、1.5B、3B、7B、14B、32B、72B（旗舰模型），它们在同等参数赛道都创造了业界最佳成绩，型号设定充分考虑下游场景的不同需求，3B是适配手机等端侧设备的黄金尺寸；32B是最受开发者期待的“性价比之王”，可在性能和功耗之间获得最佳平衡，Qwen2.5-32B的整体表现超越了Qwen2-72B。

官网：通义大模型_企业拥抱_AI_时代首选-阿里云
Huggingface：Qwen (Qwen)
参考文章：通义千问重磅开源Qwen2.5，性能超越Llama-阿里云开发者社区

4. **Bert系列**

- **BERT**（Bidirectional Encoder Representations from Transformers）全称是“双向编码器表征法”或简单地称为“双向变换器模型”，是一种基于Transformer架构的预训练语言模型，由Google在2018年推出，代码已开源。BERT在自然语言处理（NLP）领域具有广泛的应用和出色的性能，为多种语言理解任务提供了强大的预训练模型基础。

仓库：google-research/bert: TensorFlow code and pre-trained models for BERT
Huggingface：google-bert (BERT community)
参考文章：AI大模型介绍-BERT - 知乎

5. **零一万物**

- 2023年11月5日上午，零一万物正式发布首款研发的最强开源人工智能（AI）大模型系列，名为“Yi”。此次 Yi 系列基础模型的首个公开版本包括两款：Yi-6B（数据参数量为60亿）、Yi-34B（340亿），均是双语（英文/中文）、支持开源。

- **Yi系列模型**以双语能力领先领域，利用3T多语言语料库训练，具备卓越的语言理解、常识推理和阅读理解等能力。2024年1月数据显示，Yi-34B-Chat在AlpacaEval排名第二，仅次于GPT-4 Turbo，超越其他LLM如GPT-4、Mixtral、Claude。在各种基准测试中，Yi-34B排名第一，超越Falcon-180B、Llama-70B、Claude等开源模型。其中，基于超强Infra下模型训练成本实测下降40%，模拟千亿规模训练成本可下降多达50%，并以更小模型尺寸的基准结果超过LLaMA2-34B/70B、Falcon-180B等大尺寸开源模型，以及百川智能（王小川创立）的Baichuan2-13B。

- **LiveCodeBench**：Yi-Coder-9B-Chat 在 LiveCodeBench 评测平台上的通过率高达 23.4%，成为唯一一个在 10B 参数以下的模型中通过率超过 20% 的产品。它甚至超越了 CodeGeex4 和 DeepSeek-Coder 等高参数模型。

- 在 **HumanEval 和 MBPP** 等流行的代码生成和推理任务中，Yi-Coder-9B-Chat 也表现突出，分别达到了 85.4% 和 73.8% 的通过率，并且是首个在 CRUXEval-O 基准测试中通过率超过 50% 的开源模型。

- **代码编辑和跨文件代码补全**：CodeEditorBench 涵盖了四个关键领域：代码调试、代码翻译、代码优化和代码需求转换。结果表明，在开源代码大语言模型中，Yi-Coder-9B-Chat 取得了优异的成绩，在 Primary 和 Plus 两个子集中始终优于 DeepSeek-Coder-33B-Instruct 和 CodeQwen1.5-7B-Chat。

- **CrossCodeEval** 基准测试中，Yi-Coder 在有检索和无检索上下文的情况下都表现优异，展现了强大的代码补全能力，尤其适合处理跨文件依赖的代码库。

- **数学推理能力**：Yi-Coder 还可以通过编程解决复杂的数学推理问题。在代码辅助解题的测试中，Yi-Coder-9B 的准确率达到 70.3%，远超同类模型。

官网：零一万物-大模型开放平台
Huggingface：01-ai (01-ai)
仓库：01-ai/Yi
参考文章：小而强大！零一万物 Yi-Coder 模型震撼发布！-CSDN博客

6. **智源BGE**

- **BGE**，全称BAAI General Embedding，是北京智源人工智能研究院研发的开源通用向量模型，该系列模型专为各类信息检索及大语言模型检索增强应用而打造，于2023年2月推出，至今在huggingface上下载量破亿。BGE 是当前**中文任务下最强语义向量模型**，各项语义表征能力全面超越同类开源模型。
- **BGE-M3**是BGE的进阶版本，主要优化为：
  1. 多功能：具备嵌入式模型的三种常用检索方法 — 稠密检索、稀疏检索和多向量检索。
  2. 跨语言：它可以支持100多种语言。
  3. 长文本：最多能够处理8192个Tokens。

Huggingface：BGE - a BAAI Collection
仓库：FlagOpen/FlagEmbedding: Retrieval and Retrieval-augmented LLMs
参考文章：BGE（BAAI General Embedding）解读 - 知乎

7. **InternLM**

- **InternLM**是2023 年上海人工智能实验室和商汤联合研发的大模型，104B 模型经过 1.6T tokens 数据训练，在多个榜单上取得了仅次于 GPT4 的成绩
- **InternLM2** ，即书生·浦语大模型第二代，开源了面向实用场景的70亿参数基础模型与对话模型 （InternLM2-Chat-7B）。模型具有以下特点：
  1. 有效支持20万字超长上下文：模型在20万字长输入中几乎完美地实现长文“大海捞针”，而且在 LongBench 和 L-Eval 等长文任务中的表现也达到开源模型中的领先水平。 可以通过 LMDeploy 尝试20万字超长上下文推理。
  2. 综合性能全面提升：各能力维度相比上一代模型全面进步，在推理、数学、代码、对话体验、指令遵循和创意写作等方面的能力提升尤为显著，综合性能达到同量级开源模型的领先水平，在重点能力评测上 InternLM2-Chat-20B 能比肩甚至超越 ChatGPT （GPT-3.5）。
  3. 代码解释器与数据分析：在配合代码解释器（code-interpreter）的条件下，InternLM2-Chat-20B 在 GSM8K 和 MATH 上可以达到和 GPT-4 相仿的水平。基于在数理和工具方面强大的基础能力，InternLM2-Chat 提供了实用的数据分析能力。
  4. 工具调用能力整体升级：基于更强和更具有泛化性的指令理解、工具筛选与结果反思等能力，新版模型可以更可靠地支持复杂智能体的搭建，支持对工具进行有效的多轮调用，完成较复杂的任务。

- **InterLM2.5**于2024年7月3日发布，相比上一代模型，InternLM2.5 有三项突出亮点：
  1. 推理能力大幅提升，领先于国内外同量级开源模型，在部分维度上甚至超越十倍量级的 Llama3-70B；
  2. 支持 1M tokens 上下文，能够处理百万字长文；
  3. 具有强大的自主规划和工具调用能力，比如可以针对复杂问题，搜索上百个网页并进行整合分析。

  官网：InternLM(书生·浦语)
  仓库：InternLM
  Huggingface：internlm (InternLM)
  参考文章：
  InternLM：商汤研发的书生大模型 - 知乎
  [大模型]InternLM2-7B-chat FastAPI 部署 internlm2-chat-7b-CSDN博客

8. **GLM**

- **GLM-130B**由清华智谱AI于2022年8月开源发布。它是GLM系列模型中最大的模型，拥有1300亿参数，支持中英文双语，其目标是能够训练出开源开放的高精度千亿中英双语语言模型，让每个人都能用的上的千亿模型。GLM-130B大模型第一次将千亿模型量化到int4层次，并且在没有量化感知训练的条件下，性能损失也很少，这让模型能够在4块3090(24G)或8块2080Ti(11G)就可以推理GLM-130B模型。且GLM-130B从预训练到评估都是可复现的，所有评估代码也是开源的。这对于不具备大语言业务深耕经验的开发者而言带来了本地快速部署的可能，为行业大语言模型提供了重要的基石支撑。
- **GLM-4-9B** 是智谱 AI 推出的最新一代预训练模型 GLM-4 系列中的开源版本。 在语义、数学、推理、代码和知识等多方面的数据集测评中，GLM-4-9B 及其人类偏好对齐的版本 **GLM-4-9B-Chat** 均表现出较高的性能。 除了能进行多轮对话，GLM-4-9B-Chat 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）等高级功能。 本代模型增加了多语言支持，支持包括日语，韩语，德语在内的 26 种语言。我们还推出了支持 1M 上下文长度（约 200 万中文字符）的模型。
  官网：智谱AI开放平台
  仓库：THUKEG&THUDM
  Huggingface：THUDM (Knowledge Engineering Group (KEG) & Data Mining at Tsinghua University)
  参考文章：
  【GLM-4部署实战】GLM-4-9B-Chat模型本地部署实践指南-CSDN博客
  预训练大模型解析：GLM - 知乎

9. **Gemma**)

- **Gemma**模型是在2024.2.21号Google新发布的大语言模型，Gemma复用了Gemini相同的技术(Gemini也是Google发布的多模态模型)，Gemma这次发布了了2B和7B两个版本的参数，不仅提供了预训练的checkpoints，还提供了用于对话、指令跟随等fine-tune的checkpoints。所有版本均可在各类消费级硬件上运行，无需数据量化处理，拥有高达 8K tokens 的处理能力
- Gemma模型非常适合执行各种文本生成任务，包括问答、摘要和推理。它们相对较小的尺寸使得可以在资源有限的环境中部署，例如笔记本电脑、桌面电脑或您自己的云基础设施，使每个人都能获得最先进的AI模型，促进创新。
  官网：Google AI Gemma 开放模型 |Google for Developers
  仓库：google-deepmind/gemma: Open weights LLM from Google DeepMind.

Huggingface：google (Google)
参考文章：
[转]Gemma模型论文详解(附源码) - 知乎
gemma 大模型（gemma 2B,gemma 7B）微调及基本使用_gemma-2b-CSDN博客

10. **Deepseek**

- **DeepSeek LLM 67B**由DeepSeek的AI团队发布，在代码、数学和推理任务中均超越了Llam-2-70B，而 **DeepSeek LLM 67B Chat** 在开放性评估中更是超越了 GPT-3.5。这一系列的表现为开源 LLM 的未来发展奠定了一定基础。
- **Deepseek-V2**由私募基金幻方于2024年5月6日发布，每百万Tokens仅需1元-2元，被誉为“价格屠夫”，引发了大模型价格站。DeepSeek-V2沿袭了1月发布的 Deepseek-MoE（混合专家模型）的技术路线，采用大量的小参数专家进行建模，同时在训练和推理上加入了更多的优化。沿袭了一贯的作风，Deepseek对模型（基座和对话对齐版本）进行了完全的mit协议开源，可以商用。
- **DeepSeek-Coder-V2**是一个开源的混合专家 (MoE) 代码语言模型，在代码特定任务中实现了与 GPT4-Turbo 相当的性能。具体来说，DeepSeek-Coder-V2 从 DeepSeek-V2 的中间检查点进一步预训练，增加了 6 万亿个 token。通过这种持续的预训练，DeepSeek-Coder-V2 大幅增强了 DeepSeek-V2 的编码和数学推理能力，同时在一般语言任务中保持了相当的性能。
  官网：DeepSeek | 深度求索
  仓库：deepseek-ai/DeepSeek-Coder-V2:
  Huggingface：deepseek-ai (DeepSeek)
  参考文章：
  大模型·DeepSeek(2)：DeepSeek-Code-V2 - 知乎
  Deepseek-V2技术报告解读！全网最细！-CSDN博客

11. **Mistral AI**

- **Mistral Large 2**，参数123B，于2024年7月24日发布，参数123B，用不到三分之一的参数量性能比肩Llama 3.1 405B，也不逊于GPT-4o、Claude 3 Opus等闭源模型。代码能力方面，Mistral Large 2支持包括Python、Java、C、C++、JavaScript和Bash在内的**80多种编程语言**，吸取Codestral 、Codestral Mamba经验，表现远超之前的Mistral Large。Human Eval、MBPP基准上，Mistral Large 2代码生成能力可与GPT-4o、Claude 3 Opus和Llama 3.1 405B等最强模型相媲美。
- **Mistral-Small-Instruct-2409 (22B)**，是 Mistral AI 最新的企业级小型模型，是 Mistral Small v24.02 的升级版。
- **Pixtral 12B** 是一个多模态LLM。具有自然场景理解，代码生成，图像转代码，图像理解，多图指令跟随，图表理解与分析以及复杂图形推理等多项能力。从效果演示来看模型的能力很强，其中对中文能力的理解也很好。主要特点：
  - Mistral Nemo 12B 的直接替代品。
  - 从头开始训练的新型 400M 参数视觉编码器；
  - 搭配基于 Mistral Nemo 的 12B 多模态解码器；
  - 能处理可变的图像尺寸和纵横比；
  - 支持128k上下文窗口中的多个图像。

  官网：Mistral AI | Frontier AI in your hands
  仓库：Mistral AI
  Huggingface：mistralai (Mistral AI)
  参考文章：

开源大模型杀疯了！Mistral新模型三分之一参数卷爆Llama 3.1
Mistral AI 又又又开源了闭源企业级模型——Mistral-Small-Instruct-2409_mistral-small 2409-CSDN博客
Mistral AI 开源 Pixtral 12B 多模态 LLM，多场景能力理解，支持中文指令遵循

12. Baichuan

- **Baichuan-7B** （发布时间：2023-09-06；模型文件大小：14GB）是百川智能匠心打造的一款开源大规模预训练语言模型，采用先进的Transformer架构。该模型配置有70亿参数，在超过1.2万亿的tokens上进行训练，支持中文和英文双语环境。其上下文窗口长度达到了4096，特别值得关注的是，它在C-EVAL和MMLU等权威的中文和英文基准测试中展现出了同类模型中的顶级性能。
- **Baichuan-13B-Chat** （发布时间：2023-07-08）是由百川智能继 Baichuan-7B 之后开发的包含 130 亿参数的开源可商用的大规模语言模型，在权威的中文和英文 benchmark 上均取得同尺寸最好的效果。本次发布包含有预训练 (Baichuan-13B-Base) 和对齐 (Baichuan-13B-Chat) 两个版本。
- **Baichuan2**（发布时间：2023-12-29）是百川智能推出的新一代开源大语言模型，采用2.6万亿Tokens的高质量语料训练。在多个权威的中文、英文和多语言的通用、领域 benchmark 上取得同尺寸最佳的效果。包含有7B、13B的Base和Chat版本，并提供了Chat版本的4bits量化。
- 2023年 10 月 30 日，百川智能正式发布 Baichuan2-192K 长窗口大模型，将大语言模型（LLM）上下文窗口的长度一举提升到了 192K token。这相当于让大模型一次处理约 35 万个汉字，长度达到了 GPT-4（32K token，约 2.5 万字）的 14 倍，Claude 2.0（100K token，约 8 万字） 的 4.4 倍。
  官网：百川大模型-汇聚世界知识 创作妙笔生花-百川智能
  仓库：Baichuan Intelligent Technology
  Huggingface：baichuan-inc (Baichuan Intelligent Technology)
  参考文章：
  百川智能大模型：Baichuan-7B 开源项目实战指南-CSDN博客
  百川智能RAG方案总结：搜索出生的百川智能大模型RAG爬坑之路-CSDN博客

13. 腾讯混元

- 11 月 5 日,腾讯混元一天内正式开源 2 大核心模型：MoE 模型“混元 Large”以及 3D 生成模型。
- **腾讯混元 Large** 的模型总参数量 389B，激活参数量 52B ，上下文长度高达256K，是当前业界参数规模最大、效果最好的 MoE 模型，同时通过技术的优化，也更适配开源框架的精调和部署，具有较强的实用性。MoE(Mixture of Experts)结构是目前主流的大模型结构，MoE 模型的每一层都包含多个并行的同构专家，一次 token 的前向计算只会激活部分专家，推理成本远低于同等参数的稠密模型。
- **混元 3D 生成大模型**是业界首个同时支持文字、图像生成 3D 的开源模型，具有强大泛化能力和可控性，可重建各类尺度物体，大到建筑，小到工具花草。
  官网：https://hunyuan.tencent.com/
  仓库：https://github.com/Tencent/Tencent-Hunyuan-Large
  Huggingface：https://huggingface.co/tencent
  参考文章：
  刚刚，腾讯混元开源两大核心模型

To be continued...

参数-模型对应表

参数大小	模型名称
8b	Llama 3.1 8b
12b	Nemo 12b
22b	Mistral Small
27b	Gemma-2 27b
35b	Command-R 35b 08-2024
70b	Llama 3.1 70b
103b	Command-R+ 103b
123b	Mistral Large 2
141b	WizardLM-2 8x22b
230b	Deepseek V2/2.5
405b	Llama 3.1 405b

To be continued...

大模型评测榜单：

1. **chinese-llm-benchmark**,中文大模型能力评测榜单,囊括128个大模型，不仅有商用大模型排行榜，还有开源大模型排行榜，同时支持多维度能力评测，包括分类能力、信息抽取能力、阅读理解能力、数据分析能力、中文编码效率、中文指令遵从、算术能力。
2. **Deepseek官网发布的参数榜单**（旨在宣传他们做的DeepSeek-V2.5）

	是否开源	中文综合	英文综合	知识	基础算数	数学解题	逻辑推理	编程
		AlignBench	MT-Bench	MMLU	GSM8K	MATH	BBH	HumanEval
DeepSeek-V2.5	开源	8.04	9.02	80.4	95.1	74.7	84.3	89.0
DeepSeek-V2	开源	7.89	8.85	80.6	94.8	71.0	83.4	84.8
GPT-4-Turbo-1106	-	8.01	9.32	84.6	93.0	64.1	-	82.2
GPT-4-0613	-	7.53	8.96	86.4	92.0	52.9	83.1	84.1
GPT-3.5	-	6.08	8.21	70.0	57.1	34.1	66.6	48.1
Gemini1.5 Pro	-	7.33	8.93	81.9	91.7	58.5	84.0	71.9
Claude3 Opus	-	7.62	9.00	86.8	95.0	61.0	86.8	84.9
Claude3 Sonnet	-	6.70	8.47	79.0	92.3	40.5	82.9	73.0

	是否开源	中文综合	英文综合	知识	基础算数	数学解题	逻辑推理	编程
Claude3 Haiku	-	6.42	8.39	75.2	88.9	40.9	73.7	75.9
abab-6.5（MiniMax）	-	7.97	8.82	79.5	91.7	51.4	82.0	78.0
abab-6.5s（MiniMax）	-	7.34	8.69	74.6	87.3	42.0	76.8	68.3
ERNIE-4.0（文心一言）	-	7.89	7.69	-	91.3	52.2	-	72.0
GLM-4（智谱清言）	-	7.88	8.60	81.5	87.6	47.9	82.3	72.0
Moonshot-v1（月之暗面）	-	7.22	8.59	-	89.5	44.2	-	82.9
Baichuan 3（百川）	-	-	8.70	81.7	88.2	49.2	84.5	70.1
Qwen1.5 72B（通义千问）	开源	7.19	8.61	76.2	81.9	40.6	65.9	68.9
LLaMA 3 70B	开源	7.42	8.95	80.3	93.2	48.5	80.1	76.2
Mixtral 8x22B	开源	6.49	8.66	77.8	87.9	49.8	78.4	75.0

3. Huggingface 上提供的**Open Chinese LLM Leaderboard**，旨在跟踪、排名和评估开放式中文大语言模型（LLM），地址：Open Chinese LLM Leaderboard - a Hugging Face Space by BAAI
4. Huggingface上提供的 **Open LLM Leaderboard**，地址：Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots
5. **Chatbot Arena LLM Leaderboard**: Community-driven Evaluation for Best LLM and AI chatbots，地址：Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots，以下为截取部分。

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
1	1	ChatGPT-4o-latest (2024-09-03)	1340	+4/-3	33743	OpenAI	Proprietar
1	1	o1-preview	1335	+4/-4	21071	OpenAI	Proprietar
3	6	o1-mini	1308	+4/-4	23128	OpenAI	Proprietar
3	4	Gemini-1.5-Pro-002	1303	+4/-4	15736	Google	Proprietar
4	4	Gemini-1.5-Pro-Exp-0827	1299	+4/-3	32385	Google	Proprietar

Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organization	License
6	9	Grok-2-08-13	1290	+3/-3	40873	xAI	Proprietar
6	3	Claude 3.5 Sonnet (20241022)	1286	+6/-6	7284	Anthropic	Proprietar
6	11	Yi-Lightning	1285	+4/-4	20973	01 AI	Proprietar
6	4	GPT-4o-2024-05-13	1285	+3/-3	102960	OpenAI	Proprietar
10	19	Gemini-1.5-Flash-002	1272	+5/-6	12379	Google	Proprietar
10	26	Llama-3.1-Nemotron-70b-Instruct	1271	+5/-7	6228	Nvidia	Llama 3.1
10	14	Gemini-1.5-Flash-Exp-0827	1269	+4/-4	25503	Google	Proprietar
11	6	Claude 3.5 Sonnet (20240620)	1268	+3/-3	81086	Anthropic	Proprietar
11	25	Grok-2-Mini-08-13	1267	+4/-3	34105	xAI	Proprietar
11	8	Meta-Llama-3.1-405b-Instruct-fp8	1267	+4/-3	43099	Meta	Llama 3.1 Community
11	7	Gemini Advanced App (2024-05-14)	1266	+3/-3	52235	Google	Proprietar
11	7	Meta-Llama-3.1-405b-Instruct-bf16	1266	+5/-6	14607	Meta	Llama 3.1 Community
12	14	Yi-Lightning-lite	1265	+3/-5	17271	01 AI	Proprietar
12	9	GPT-4o-2024-08-06	1264	+3/-4	34765	OpenAI	Proprietar
12	19	Qwen-Max-0919	1263	+5/-5	15384	Alibaba	Qwen

To be continued...

阅读资料

1. 人工智能 - Mistral AI vs. Meta：两大 Top 开源模型的对比 - IDP技术干货 - SegmentFault 思否
2. AI大模型【基础 01】智能AI开源模型与大模型接口整理（8个开源模型+7个大模型接口）
3. Langgpt大模型性能全景图(非常全面)