gitee 我的

9月17日，Gitee Xtreme 极智AI重磅发布，来Gitee直播间一起探索AI时代的软件研发新模式

# Tele-AI/TeleChat2

Watch 3 | Star 6 | Fork 1

代码 | Issues 0 | Pull Requests 0 | Wiki | 统计 | 流水线 | 服务

该仓库未声明开源许可证文件（LICENSE），使用请关注具体项目描述及其代码上游依赖。

main | 分支 5 | 标签 0 | 克隆/下载

姜卓 update README.md. 7a38abc 2个月前 108 次提交

| | | |
|---|---|---|
| MoE | update document | 6个月前 |
| autoAWQ | update awq and gptq | 6个月前 |
| autoGPTQ | update awq and gptq | 6个月前 |
| datas | Add files via upload | 10个月前 |
| deepspeed | remove unused import and refresh import | 9个月前 |
| images | 更新国产化运行文档，支持长文微调、LoRA微调、动态外推 | 7个月前 |
| inference_telechat | 添加README.md,部署+微调 | 9个月前 |
| langchain | Update README.md | 9个月前 |
| llama_factory_training | update document of llama-factory | 8个月前 |
| llama_index | Update README.md | 9个月前 |
| mindformers-telechat | 更新 mindformers 支持最新版本的moe | 6个月前 |
| model/115B | Update generation_config.json | 11个月前 |
| ollama | Merge branch 'add_ollama_patch' of https://github.com/luoy... | 6个月前 |
| text_generation_webui | add langchain & text_generation_webui | 10个月前 |
| tutorial | 更新 mindformers 支持最新版本的moe | 6个月前 |
| .gitignore | update awq and gptq | 6个月前 |
| Dockerfile_TeleChat_ms | 增加国产环境dockerfile | 10个月前 |
| README.md | update README.md. | 2个月前 |
| requirements.txt | Create requirements.txt | 11个月前 |

## README

AI 翻译由 translate.js 提供 Translate to English



# 星辰语义大模型-TeleChat2

🤗 Hugging Face • 🤖 ModelScope • 🏔 MindSpore • 🐾 gitee • 💬 WeChat

# 目录

## 简介

星辰语义大模型TeleChat2是由中国电信人工智能研究院研发训练的大语言模型，是首个完全国产算力训练并开源的千亿参数模型

暂无标签

README

6 Stars

3 Watching

1 Forks

## 发行版

暂无发行版

## 贡献者 (23) 全部

## 语言

Python 92.2%

C++ 7.2%

Shell 0.6%

## 近期动态

姜 2个月前推送了新[illegible]交到 main 分支，bd0...7a38abc

S 9个月前同步了仓库

S 9个月前同步了仓[illegible]

S 10个月前推送了[illegible]提交到 main 分支，767c436...1a936bb

S 10个月前创建了仓库

加载更多

# 最新动态

- 2025.03.14 开源MoE模型 TeleChat2-39B-A12B 模型。
- 2024.12.20 开源 **TeleChat2-35B-32K**。
- 2024.11.08 开源 **TeleChat2-3B**、**TeleChat2-7B**、**TeleChat2-35B**，该版本模型均具备 **Function Call** 功能。
- 2024.10.18 开源TeleChat2-35B模型。
- 2024.9.20 开源TeleChat2-115B模型，该模型是**首个完全国产算力训练并开源的千亿参数模型**。

# 模型介绍

## 星辰语义大模型-TeleChat2

- 星辰语义大模型**TeleChat2**是由中国电信人工智能研究院研发训练的大语言模型，该系列模型**完全基于国产算力**训练。
- 本次开源的 **TeleChat2-3B**、**TeleChat2-7B**、**TeleChat2-35B** 模型已支持**工具调用**功能。在 **Function Call** 方面，我们针对性进行了效果优化，在相关榜单评测上相比同尺寸模型均有较好表现。
- **TeleChat2-115B**模型采用10万亿 Tokens中英文高质量语料进行训练，同步开源对话模型**TeleChat2-115B**的多格式、多平台权重文件。
- **TeleChat2**在训练数据、训练方法等方面进行了改进，在通用问答和知识类、代码类、数学类榜单上相比**TeleChat1**均有大幅提升。
  - **TeleChat2**完全基于国产算力和国产深度学习框架进行训练，算力和算法框架更自主可控。优化MP、PP、SP实现方式提升模型性能，优化算子来提升训练速度。
  - 我们使用大量小模型实验来验证scaling law规律，在不同模型结构、不同数据配比和数据清洗方式中寻找最优设计。
  - 采用RingAttention及其他序列切分方式，实现长文训练性能提升；通过ntk-aware+attention-scaling的方式保证

成高质量数据，并使用拒绝采样生成多样的推理路径；通过研究一择偏好对齐数据方案，基于适配数据最大限度提升模型效果。
  - 通用能力较TeleChat系列模型提升超过29%，在逻辑推理、总学计算上均有大幅提升。
- 同时，我们也开源了TeleChat2-MoE模型 TeleChat2-39B-A12B。

### 模型结构

我们采用标准的 `Decoder-only` 结构设计了 **TeleChat2** 模型，使用 Rot编码方法、使用 SwiGLU 激活函数来替代GELU激活函数、使用基于 RM Normalization进行层标准化操作。我们将**TeleChat2**的词嵌入层和输出助于增强训练稳定性和收敛性。我们选择了GQA以节约attention部分的训练和推理速度。

**TeleChat2**的模型结构配置如下表所示：

| | layer_num | hidden_size | ffn_hidden_size | head_nu |
|---|---|---|---|---|
| 3B | 24 | 3072 | 6144 | 24 |
| 7B | 30 | 4096 | 12288 | 32 |
| 35B | 64 | 6144 | 20480 | 48 |
| 115B | 96 | 8192 | 40960 | 64 |

我们开源的 **TeleChat2** 模型：

- 支持deepspeed微调，开源了基于deepspeed的训练代码，支持Ze集成了FlashAttention2
- 多轮能力支持。开源了多轮数据构建方式，针对多轮模型训练集成训练方式，更好的聚焦多轮答案，提升问答效果。

本次发布版本和下载链接见下表

| 模型版本 | 下载链接 |
|---|---|
| telechat2-3B | modelscope |
| telechat2-7B | modelscope |
| telechat2-35B | modelscope |
| telechat2-35B-32K | modelscope |
| telechat2-115B | modelscope |

# 效果评测

**TeleChat2** 模型相比同规模模型在评测效果方面也有较好的表现，我们MMLU、C-Eval、CMMLU、 GSM8K、MATH、HumanEval、BBH等数指令遵循、考试能力、数学计算和推理、代码生成等

## 评测集介绍

### 通用能力

- CEVAL 数据集是一个全面的中文评估测试集，包括初中、高中、大项选择题，涵盖了 52 个不同的学科领域。
- CMMLU 数据集同样是一个全面的中文评估测试集，涵盖了从基础67个主题。

### 推理和代码能力

- GSM8K 数据集包含了8.5K高质量的小学数学题，能够评估语言模型现。
- HumanEval 数据集是一个由openai提供的代码能力测试数据集，它成，要求根据给定的问题和代码模板，生成正确的代码片段。
  - BBH 数据集全名为BIG-Bench Hard（BBH），包含23个具有挑务，均为之前的语言模型评估中没有超过平均人类评审者表现
- MBPP 数据集包含大约1000个众包的Python编程问题，涵盖编程基等。每个问题包括任务描述、代码解决方案和3个自动化测试用例。

### 主观题能力

- AlignBench是一个多维度全面评估中文大模型对齐水平的评测基准评测题。
- MT-bench是一个用于评估聊天助手的具有挑战性的多轮开放式问题观评测题。

### 指令遵循能力

- IFEval旨在评估语言模型对指令的精确遵循能力，它包含了500条可pen LLM Leaderboard中使用的核心基准测试之一。

## 评测结果如下

| Dataset | Llama-3.1-70B | Qwen1.5-110B | Qwen2-72-instruct | DeepSeek-v2 | Te 11 |
|---|---|---|---|---|---|
| C-Eval | - | - | 83.8 | 78 | 8( |
| MMLU | **86** | 80.4 | 82.3 | 77.8 | 8( |
| CMMLU | 69.01 | 87.64 | 87.47 | 81.6 | 8 |
| BBH | - | 74.8 | - | 79.7 | 8 |
| GSM8K | **95.1** | 85.4 | 91.1 | 92.2 | 9 |
| HumanEval | 80.5 | 52.4 | **86** | 81.1 | 7 |
| MBPP | **86** | 58.1 | 80.2 | 72 | 7 |
| AlignBench | - | 7.86 | **8.27** | 7.91 | 8. |
| MT-bench | 8.79 | 8.88 | **9.12** | 8.97 | 8. |

| | 70B | 110B | instruct | v2 | 1T |
|---|---|---|---|---|---|
| IFEval | 87.5 | - | 77.6 | 63.8 | 8 |

# 模型推理和部署

## 模型推理

当前模型推理兼容了单卡和多卡推理，以及针对长文推理做了部分优化

### 模型推理方法示范

```
>>> import os
>>> import torch
>>> from transformers import AutoModelForCausalLM, AutoTokeniz
>>> tokenizer = AutoTokenizer.from_pretrained('TeleChat2/Telec
>>> model = AutoModelForCausalLM.from_pretrained('TeleChat2/Te
                                                  torch_dtype=
>>> prompt = "生抽与老抽的区别？"
>>> messages = [{"role": "user", "content": prompt}]
>>> text = tokenizer.apply_chat_template(messages,
>>>     tokenize=False,
>>>         add_generation_prompt=True
>>> )
>>> model_inputs = tokenizer([text], return_tensors="pt").to(m
>>> generated_ids = model.generate(
>>>     **model_inputs,
>>>     max_new_tokens=512
>>> )
>>> generated_ids = [
>>>     output_ids[len(input_ids):] for input_ids, output_ids
>>> ]

>>> response = tokenizer.batch_decode(generated_ids, skip_spec
生抽和老抽是两种不同的酱油，它们在风味、色泽和用途上都有所区别。

1.颜色：生抽的颜色比较淡，而老抽的颜色较深。生抽的颜色呈红褐色或棕红

2.味道：生抽具有鲜美的咸味和微甜的味浅，而老抽浓郁，颜色较深。根据个
```

## 模型部署

我们建议您在部署TeleChat时尝试使用vLLM。

**vllm推理**

## 模型工具调用能力

TeleChat2 目前已支持工具调用功能，具体使用方式参考文档TeleChat2

# 模型微调

TeleChat2 现已支持DeepSpeed微调方式，具体使用方式参考文档TeleC

# 国产化适配

## 昇腾Atlas 800T A2训练服务器实现训练、推理适配

gitee 我的

发环境。它支持多种硬件平台，并具有自动微分、模型优化等功能
务。

- MindSpore Transformers：该框架的目标是构建一个大模型训练、
署的全流程开发套件，提供业内主流的Transformer类预训练模型和
涵盖丰富的并行特性。期望帮助用户轻松的实现大模型训练和创新

**当前星辰语义大模型TeleChat2支持昇腾Atlas 800T A2训练服务器，可
架以及MindSpore Transformers框架进行模型训练和评测，详情请看te
对mindsformers相关特性有疑问，也可以查看mindformers。**

115B模型性能方面，具体对比如下：

| NAME | performance(samples/p/s) | Epochs | AMP_Type |
|---|---|---|---|
| 115B | 0.0192 | 1 | O1 |
| 115B | 0.0174 | 1 | O2 |

# 更多功能

## LLaMA-Factory

LLaMA-Factory 是一个专注于大语言模型（LLM）开发和优化的开源平
和部署的过程。该平台提供了多种工具和框架，支持用户根据特定需求
型。通过LLaMA-Factory，研究人员和开发者可以更高效地探索和实现
术，例如LoRA，QLoRA，Pre-Training，Supervised Fine-Tuning，DPO

TeleChat2 已支持使用LLaMA-Factory进行微调、权重合并、推理、部署
档TeleChat2-LLaMA-Factory微调文档。

## AWQ

TeleChat2已支持AWQ量化，能够快速实现int4精度的权重量化，降低推
性能，具体使用方式参考：TeleChat2-AutoAWQ文档。

## GPTQ

TeleChat2已支持GPTQ量化，能够快速实现int4和int8精度的权重量化，
高推理性能，具体使用方式参考：TeleChat2-AutoGPTQ文档

## Ollama

TeleChat2已支持Ollama推理框架，提供灵活高效的推理部署方案，具体
TeleChat2-Ollama文档

## text-generation-webui

text-generation-webui 是一个开源的Web用户界面，旨在简化大语言模
持多种预训练模型，使用户能够方便地进行文本生成、对话和其他自然
友好易用，适合研究人员和开发者快速构建和测试他们的应用程序。

TeleChat2 已支持使用text-generation-webui实现界面应用，具体使用方
text-generation-webui部署文档。

## LangChain

行交互。通过LangChain，用户可以快速创建复杂的对话系统、智能助
应用。

TeleChat2 已支持使用LangChain进行高效向量知识库检索问答，具体
TeleChat2-LangChain文档。

### LlamaIndex

LlamaIndex 是一个用于构建和管理与大型语言模型（LLM）交互的数据
息检索的效率。它允许用户将结构化和非结构化数据转化为可供语言模
升模型的响应准确性和相关性。LlamaIndex适用于各种应用场景，包括
档检索等。

TeleChat2 已支持使用LlamaIndex进行高效向量知识库检索问答，具体
TeleChat2-LlamaIndex文档。

## MoE模型

### MoE模型介绍

TeleChat2-39B-A12B模型采用MoE架构，总16路由专家，激活4个专家
活参数为12B。

### 技术创新-训练方式

采用课程学习的方式，首先聚焦低难度、高质量教育知识以及多语言数
得较好的模型初始性能；然后引入复杂数据，增大数学、逻辑推理、代
型逻辑推理能力；最后，使用高质量数据进行退火，持续提升模型效果

### 技术创新-国产算力优化

在MoE模块将Tensor并行域转换成专家并行域，从而将MoE的AllToAll
通讯效率;把MoE输入切成多个副本依次下发，将dispatch通信/FFN计算
连成流水线，实现MoE的计算通信掩盖;基于对内存和计算的开销建模，
能最优的流水线并行的负载配置，实现流水线负载均衡。

### 效果评测

综合评测数据集上，TeleChat2-39B-A12B模型以12B激活参数量接近Te

| Dataset | TeleChat2-35B | TeleChat2-39B-A12B | TeleChat2 |
|---|---|---|---|
| C-Eval | 85 | 89 | 82 |
| MMLU | 82 | 83 | 79.6 |
| CMMLU | 90.18 | 90 | 84.6 |
| GSM8K | 91 | 83.5 | 86.8 |
| HumanEval | 73 | 68 | 56 |
| MBPP | 75 | 67 | 62.6 |
| AlignBench | 7.88 | 7.56 | 6.96 |
| IFEval | 79.63 | 76.48 | 73.1 |

## 声明

我们在此声明，不要使用TeleChat模型及其衍生模型进行任何危害国家动。同时，我们也要求使用者不要将TeleChat模型用于没有安全审查和们希望所有使用者遵守上述原则，确保科技发展在合法合规的环境下进

我们已经尽我们所能，来确保模型训练过程中使用的数据的合规性。然了巨大的努力，但由于模型和数据的复杂性，仍有可能存在一些无法预由于使用TeleChat开源模型而导致的任何问题，包括但不限于数据安全或模型被误导、滥用、传播或不当利用所带来的任何风险和问题，我们

## 协议

社区使用 TeleChat 模型需要遵循《TeleChat模型社区许可协议》。Tele途，如果您计划将 TeleChat 模型或其衍生品用于商业目的，您需要通过tele_ai@chinatelecom.cn，提交《TeleChat模型社区许可协议》要求的后，将特此授予您一个非排他性、全球性、不可转让、不可再许可、可

## 引用

如需引用我们的工作，请使用如下 reference:

```
@misc{wang2025technicalreporttelechat2telechat25,
      title={Technical Report of TeleChat2, TeleChat2.5 and T1
      author={Zihan Wang and Xinzhang Liu and Yitong Yao and C
      year={2025},
      eprint={2507.18013},
      archivePrefix={arXiv},
      primaryClass={cs.CL},
      url={https://arxiv.org/abs/2507.18013},
}
```

gitee

北京奥思研工智能科技有限公司版权所有

Git 大全
Git 命令学习
CopyCat 代码克隆检测
APP与插件下载
Gitee 封面人物
GVP 项目
Gitee 博客
Gitee 公益计划
Gitee 持续集成
OpenAPI
MCP Server
帮助文档
在线自助服务
更新日志
关于我们
加入我们
使用条款
意见建议
合作伙伴

client@oschina.cn
企业版在线使用：400-606-0201
专业版私有部署：13670252304
13352947997

技术交流QQ群
微信服务号

开放原子开源基金会 合作代码托管平台
违法和不良信息举报中心
京ICP备2025119063号
简 体 / 繁 體 / English