. | 1 Branch | 0 Tags | Go to file | Go to file | Code

HanxiaoZhang update README eb1c3c7 · last week

docs	init repo	last week
examples/sft/llama_fact...	init repo	last week
figures	init repo	last week
models	init repo	last week
LICENSE	init repo	last week
README.md	update README	last week
requiremetns.txt	init repo	last week

About

No description, website, or topics provided.

Readme
MIT license
Activity
Custom properties
4 stars
0 watching
1 fork
Report repository

Releases

README | MIT license

Packages

No packages published

Languages

Python 100.0%

Ling-V2.5

Hugging Face | ModelScope

Introduction

Ling-2.5-1T, Inclusive Intelligence, Instant Impact.

Thinking models raise the ceiling of intelligence, while instant models expand its reach by balancing efficiency and performance—making AGI not only more powerful, but also more accessible. As the latest flagship instant model in the Ling family, Ling-2.5-1T delivers comprehensive upgrades across model architecture, token efficiency, and preference alignment, designed to bring universally accessible AI to a new level of quality.

- Ling-2.5-1T features 1T total parameters (with 63B active parameters). Its pre-training corpus has expanded from 20T to 29T tokens compared to the previous generation. Leveraging an efficient hybrid linear attention architecture and refined data strategy, the model delivers exceptionally high throughput while processing context lengths of up to 1M tokens.
- By introducing a composite reward mechanism combining "Correctness" and "Process Redundancy", Ling-2.5-1T further pushes the frontier of efficiency-performance balance in instant models. At comparable token efficiency levels, Ling-2.5-1T's reasoning capabilities significantly outperform its predecessor, approaching the level of frontier "thinking models" that typically consume ~4x the output tokens.
- Through refined alignment strategies—such as bidirectional RL feedback and Agent-based instruction constraint verification—Ling-2.5-1T achieves substantial improvements over the previous generation in preference alignment tasks, including creative writing and instruction following.
- Trained with Agentic RL in large-scale high-fidelity interactive environments, Ling-2.5-1T is compatible with mainstream agent platforms such as Claude Code, OpenCode, and OpenClaw. It achieves leading open-source performance on the general tool-calling benchmark, BFCL-V4.

Model Downloads

Model	Context Length	Download
Ling-2.5-1T	256K -> 1M (YaRN)	🤗 HuggingFace 🤖 ModelScope

Note: If you are interested in previous version, please visit the past model collections in Huggingface or ModelScope.

Deployment

SGLang

Environment Preparation

We will later submit our model to SGLang official release, now we can prepare the environment following steps:

```
git clone -b ling_2_5 git@github.com:antgroup/sglang.git
cd sglang

# Install the python packages
pip install --upgrade pip
pip install -e "python"
```

Run Inference

Both BF16 and FP8 models are supported by SGLang now. It depends on the dtype of the model in ${MODEL_PATH}.

Here is the example to run Ling-1T with multiple GPU nodes, where the master node IP is ${MASTER_IP} and server port is ${PORT}:

- Start server:

```
# Node 0:
python -m sglang.launch_server --model-path $MODEL_PATH --tp-size 8 --pp-size 4 --dp-size 1 --trust-remote-code --di
# Node 1:
python -m sglang.launch_server --model-path $MODEL_PATH --tp-size 8 --pp-size 4 --dp-size 1 --trust-remote-code --di
# Node 2:
python -m sglang.launch_server --model-path $MODEL_PATH --tp-size 8 --pp-size 4 --dp-size 1 --trust-remote-code --di
# Node 3:
python -m sglang.launch_server --model-path $MODEL_PATH --tp-size 8 --pp-size 4 --dp-size 1 --trust-remote-code --di

# This is only an example. Please adjust arguments according to your actual environment.
```

- Client:

```
curl -s http://${MASTER_IP}:${PORT}/v1/chat/completions \
  -H "Content-Type: application/json" \
  -d '{"model": "auto", "messages": [{"role": "user", "content": "What is the capital of France?"}]}'
```

More usage can be found here

Finetuning

We recommend you to use Llama-Factory to finetune Ling-1T.