



ZIKUU Distillery は LLM を 4 倍速く推論させるためのプラットフォームです。

Description

Inference Broker は、LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

- LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。
- LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。
- LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

LLM の推論を高速化するためのプラットフォームです。LLM の推論を 4 倍速く推論させるためのプラットフォームです。

ZIKUU

DISTILLERY

WORKBENCH

01 教材を作る

入力と正解を決める

02 学習する

正解を小さなモデルに覚えさせる

03 別の入力で試す

学習に使っていない文章で確かめる

大きなモデルの能力を、
小さな部品へ。

distillery

文脈抽出

▶ モデル：Qwen/Qwen3-0.6B / 学習済みLoRAあり

STEP 01

教材を作る

入力と正解を決める

教材の入力

工場の音声認識はサーバーで処理することに決めた。
端末にはマイクだけを入れたい。
通信は有線に限定する。
マイクの設置場所はまだ決めていない。

確定した正解

****決定事項****

- 音声認識はサーバーで処理することにした。
- 端末にはマイクだけを入れる。

****制約****

- 通信は有線に限定する。

****未解決の問い****

- マイクの設置場所はまだ決めていない。

▶ 指示と教師の元の回答

[教材を作り直す](#)

▶ 作業履歴 / 22件

Step2ã•-å'ç¿?ã??

Seedã?¢ã??ã?«ã•«ã!ç¿?ã•?ã•?ã??ã??
ã!ç¿?çµ•œ??ã•LoRAã?¢ã??ã??ã?¿ã?¼ã•ã•?ã•!ä¿•ã?ã•?ã??ã??

ZIKUU

DISTILLERY

WORKBENCH

01 教材を作る

入力と正解を決める

02 学習する

正解を小さなモデルに覚えさせる

03 別の入力で試す

学習に使っていない文章で確かめる

distillery

文脈抽出

▶ モデル：Qwen/Qwen3-0.6B / 学習済みLoRAあり

STEP 02

学習する

正解を小さなモデルに覚えさせる

学習が完了しました

学習済みモデルを保存しました。次は別の文章で出力を確かめます。

1件の教材で学習しています。未知の入力での品質は未判定です。

別の入力で試す

▼ 学習の結果と条件

損失 0.508828 → 0.000548 / 3.209 秒

{

"run_id": "76ac07c3301f4fa1a646ecaa916b64d4",

"adapter_path": "/model/adapters/76ac07c3301f4fa1a646ecaa916b64d4",

"model_id": "Qwen/Qwen3-0.6B",

"model_revision": "c1899de289a04d12100db370d81485cdf75e47ca",

"definition_revision": 5,

"dataset_sha256": "8311ff1c60c8f6a8d1f3a0371632f561ced60be8133e7f833b9ce65bf5e5cf73",

"examples": 1,

"steps": 20,

"learning_rate": 0.0002,

"rank": 8,

"alpha": 16,

"dropout": 0,

"target_modules": [

"q_proj",

"k_proj",

"v_proj",

Step3ã•æ?è¼ã??

Step2ã•Sä¿•å?ã•?ã•?ã?¢ã??ã??ã?¿ã?¼ã??ã½¿ã•£ã•!ã••å?¥ã•®å ¥å??ã•Sæ?`è«?çµ•æ??ã??ç¢°èª•ã•
?ã??ã??

default watermark

ZIKUU

DISTILLERY

WORKBENCH

01 教材を作る

入力と正解を決める

02 学習する

正解を小さなモデルに覚えさせる

03 別の入力で試す

学習に使っていない文章で確かめる

distillery

文脈抽出

▶ モデル：Qwen/Qwen3-0.6B / 学習済みLoRAあり

STEP 03

別の入力で試す

学習に使っていない文章で確かめる

学習に使っていない文章を入力してください。学習前と学習後の出力を並べます。

新しい入力

工場の温度記録はサーバーに保存することに決めた。
端末には温度センサーだけを入れたい。
通信はWi-Fiに限定する。
設置場所はまだ決めていない。

この操作で追加学習はしません。

出力を比較する

比較した入力

工場の温度記録はサーバーに保存することに決めた。
端末には温度センサーだけを入れたい。
通信はWi-Fiに限定する。
設置場所はまだ決めていない。

学習例とは異なる入力です。可否は未判定です。

学習前

- 温度記録はサーバーに保存することに決めた。
- 端末には温度センサーだけを入れたい。

学習後

- **決定事項**
- 食品の温度記録はサーバーに保存すること

ã•?ã•@3ã•ã•@ã•¹ã??ã??ã??ã??ç¹°ã??è¿?ã•?ã•¾ã•?ã??

