Language Gateway to Tomorrow Module

【LGTM】みんなが使えるＡＩを。

LGTM.

Base Architecture by RWKV

Language Gateway to Tomorrow Module

About LGTM

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

LGTM開発統括(CTO)よりご挨拶

OpenMOSE @_m0se_ ・XX月XX日

LLMが数多く誕生していく中でLGTMはどこまでも
ローカルで動く賢い言語モデルを追求していきます。

昨今どんなスマートフォン / PCでもAIが当たり前に
機能の一つとして備わっているので、小さくて賢い
モデルが必要になると考えております。

「身近なAI」を目指し、LGTMを磨き続けます。
詳しくはX(@_m0se_)をご覧ください。

❤ 10100

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

よく聞くお悩みの声

AIを取り入れたいけどクラウドだと
情報漏洩が怖い、、

社内規則を社員に周知したい、、
社員同士の会話を増やしたい！

自分の代わりにコーディングしてくれる人
がいたら、、

活用シーンに合わせた会話ができる
AIチャットを作りたいけど予算が・・・

ChatGPTやLlaMAを使っても出来ない、
AIの活用方法が分からない、というお悩みの声をいただきます

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

LGTMがお悩みを解決

大規模言語モデルの一つ、RWKVを基にした言語モデルです。
「**みんなが使えるAI**」を実現するために、ローカルで学習/運用できる言語モデルと学習キットを開発し誰でも簡単にオリジナルのLLMを開発できるようにしました。

RWKV

弊社CTOがコアメンバーを
務めるLGTMのベースモデル
Windowsにも採用

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

チューニングが容易で、
特化型の自作AIを開発可能

製品例
デジタルヒューマン
受付や認知症防止に

日本初のRNN型LLMサービス

速い・安い・簡単が揃った唯一無二のLLM

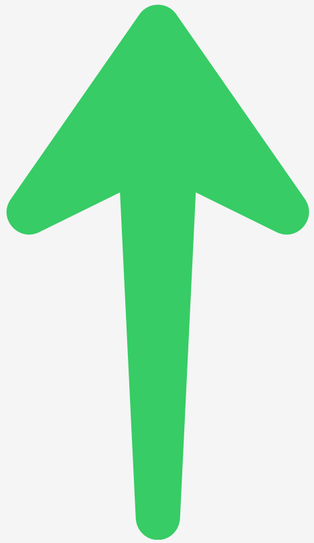
日本語性能検証結果

ベンチマークでGPT-4以外の他社LLMを超える結果

LGTM 7bの日本語性能(Rakuda,ELYZAベンチマーク)

検証日：2024年8月19日時点

gpt-4-turbo-preview	Rakuda	9.275	9.275	gpt-4-turbo-preview	ELYZA-tasks	4.47	8.94
gpt-4-turbo-2024-04-09	Rakuda	9.175	9.175	gpt-4-turbo-2024-04-09	ELYZA-tasks	4.39	8.78
CohereForAI__c4ai-command-r-plus	Rakuda	9.05128205	9.05128205	RWKV-x060-Jpn-7B-20240816-ctx4096 bancho	ELYZA-tasks	3.9	7.8
CohereForAI__c4ai-command-r-v01	Rakuda	8.625	8.625	Qwen__Qwen1.5-72B-Chat	ELYZA-tasks	3.8019802	7.6039604
RWKV-x060-Jpn-7B-20240816-ctx4096 State	Rakuda	8.175	8.175	CohereForAI__c4ai-command-r-plus	ELYZA-tasks	3.75	7.5
Qwen__Qwen1.5-72B-Chat	Rakuda	7.85	7.85	RWKV x060 14B ORPO Base State	ELYZA-tasks	3.73	7.46
karakuri-ai__karakuri-lm-70b-chat-v0.1	Rakuda	7.85	7.85	gpt-3.5-turbo-0125	ELYZA-tasks	3.62	7.24
gpt-3.5-turbo-0125	Rakuda	7.64102564	7.64102564	lightblue__ao-karasu-72B	ELYZA-tasks	3.5959596	7.19191919
Qwen__Qwen1.5-32B-Chat	Rakuda	7.51282051	7.51282051	Qwen__Qwen1.5-32B-Chat	ELYZA-tasks	3.54545455	7.09090909
lightblue__ao-karasu-72B	Rakuda	7.25	7.25	RWKV x060 7B ORPO Base bancho	ELYZA-tasks	3.45	6.9
RWKV x060 7B ORPO Base bancho	Rakuda	7.05263158	7.05263158	karakuri-ai__karakuri-lm-70b-chat-v0.1	ELYZA-tasks	3.43	6.86
RWKV x060 14B ORPO Base bancho	Rakuda	6.94871795	6.94871795	Qwen__Qwen1.5-14B-Chat	ELYZA-tasks	3.35	6.7
lightblue__20240422	Rakuda	6.675	6.675	lightblue__20240422	ELYZA-tasks	3.34	6.68
xverse__XVERSE-13B-Chat	Rakuda	6.65	6.65	xverse__XVERSE-13B-Chat	ELYZA-tasks	3.17	6.34
Rakuten__RakutenAI-7B-chat	Rakuda	6.575	6.575	RWKV x060 14B ORPO Base	ELYZA-tasks	3.06	6.12
Qwen__Qwen1.5-14B-Chat	Rakuda	6.545	6.545	CohereForAI__c4ai-command-r-v01	ELYZA-tasks	3.04	6.08
RWKV-x060-Jpn-7B-20240816-ctx4096	Rakuda	6.15789474	6.15789474	Rakuten__RakutenAI-7B-chat	ELYZA-tasks	2.96	5.92
cyberagent__calm2-7b-chat	Rakuda	5.75	5.75	openchat__openchat-3.5-0106	ELYZA-tasks	2.91	5.82
elyza__ELYZA-japanese-Llama-2-13b-instruct	Rakuda	5.625	5.625	mistralai__Mistral-7B-Instruct-v0.2	ELYZA-tasks	2.89	5.78
RWKV x060 14B ORPO Base	Rakuda	5.47222222	5.47222222	Nexusflow__Starling-LM-7B-beta	ELYZA-tasks	2.87	5.74
lightblue__qarasu-14B-chat-plus-unleashed	Rakuda	5.46153846	5.46153846	meta-llama__Llama-2-13b-chat-hf	ELYZA-tasks	2.82	5.64
RWKV x060 7B ORPO Base	Rakuda	5.35135135	5.35135135	elyza__ELYZA-japanese-Llama-2-13b-instruct	ELYZA-tasks	2.8	5.6
Qwen__Qwen1.5-7B-Chat	Rakuda	4.82051282	4.82051282	lightblue__qarasu-14B-chat-plus-unleashed	ELYZA-tasks	2.78787879	5.57575758
Nexusflow__Starling-LM-7B-beta	Rakuda	4.425	4.425	Qwen__Qwen1.5-7B-Chat	ELYZA-tasks	2.77	5.54
RWKV x060 14B Base	Rakuda	4.26315789	4.26315789	RWKV x060 7B ORPO Base	ELYZA-tasks	2.55	5.1
RWKV x060 3B ORPO Base	Rakuda	3.87179487	3.87179487	RWKV x060 14B Base	ELYZA-tasks	2.53	5.06
mistralai__Mistral-7B-Instruct-v0.2	Rakuda	3.8	3.8	RWKV-x060-Jpn-7B-20240816-ctx4096	ELYZA-tasks	2.52	5.04
openchat__openchat-3.5-0106	Rakuda	3.775	3.775	lightblue__starlingbeta_cult_dsir_megagon_fina	ELYZA-tasks	2.48	4.96
RWKV x060 3B Base	Rakuda	3.75675676	3.75675676	cyberagent__calm2-7b-chat	ELYZA-tasks	2.45	4.9
lightblue__starlingbeta_cult_dsir_megagon_final	Rakuda	3.15	3.15	meta-llama__Llama-2-7b-chat-hf	ELYZA-tasks	2.39	4.78
RWKV x060 7B Base	Rakuda	2.89473684	2.89473684	RWKV x060 7B Base	ELYZA-tasks	2.13	4.26
meta-llama__Llama-2-13b-chat-hf	Rakuda	2.275	2.275	RWKV x060 3B ORPO Base	ELYZA-tasks	2.01	4.02
augmxnt__shisa-7b-v1	Rakuda	2.225	2.225	augmxnt__shisa-7b-v1	ELYZA-tasks	1.86	3.72
meta-llama__Llama-2-7b-chat-hf	Rakuda	2.075	2.075	RWKV x060 3B Base	ELYZA-tasks	1.8	3.6

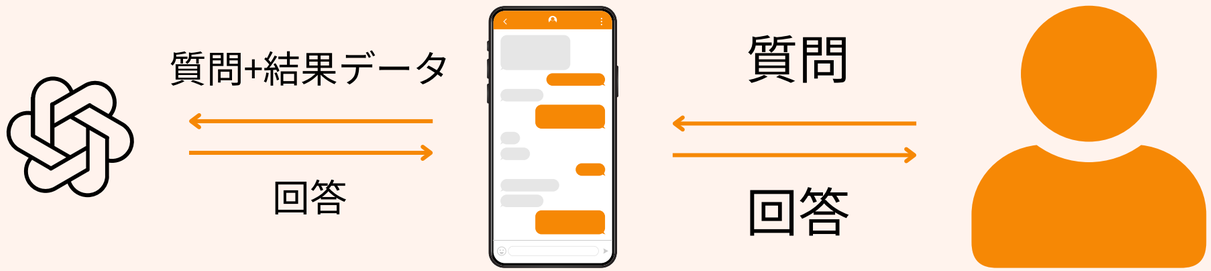


賢さ

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

ChatGPTとLGTMの違い

ChatGPTの場合

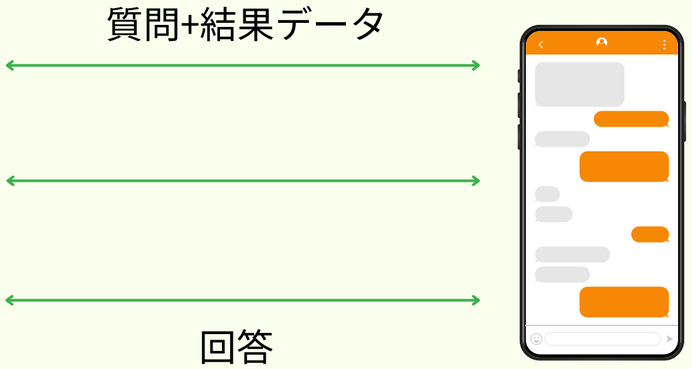


LGTMの場合

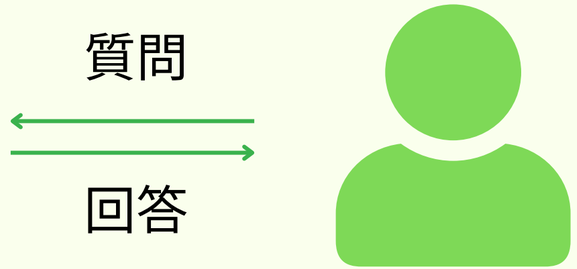
特化型言語モデル開発

特化型言語モデル

内容に応じてAIを切り替え



自社/分野に特化した言語モデルを
シーンによって切り替えて使用



LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

情報漏えい/通信環境への依存を無くす

LGTMなら・・・

クラウドとの接続不要

自社サーバー

デスクトップ

LGTMの場合

外部通信の必要がない

機密保持、機密情報の取り扱いが可能

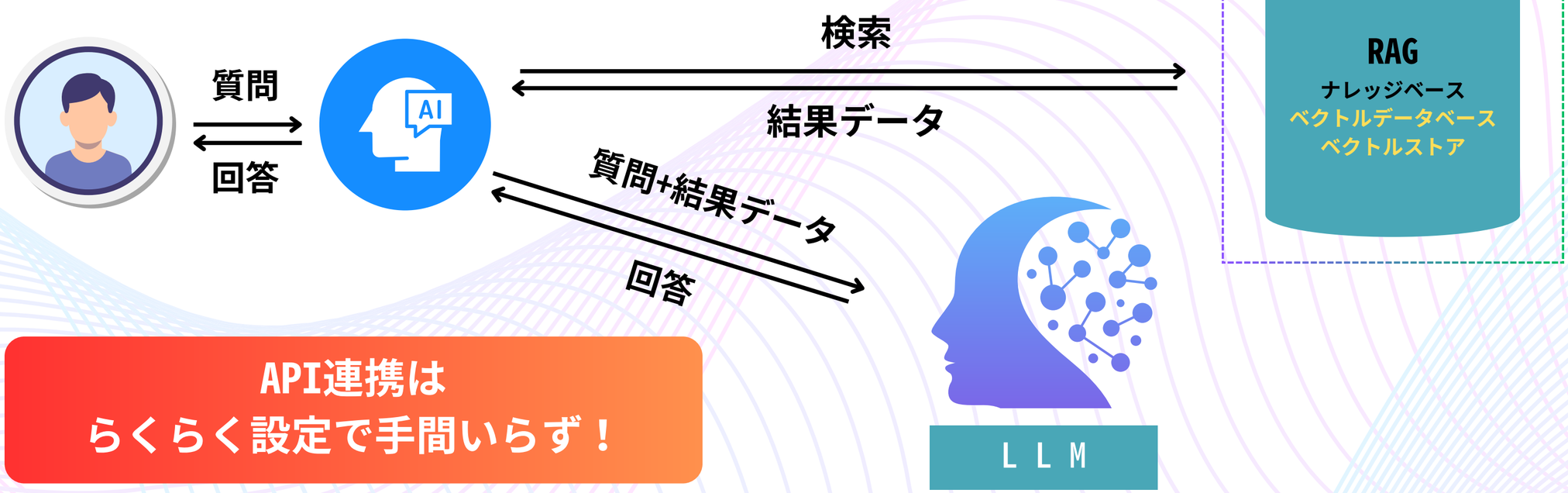
ChatGPTの場合

社内用チャットボットを作る場合でも
社外にデータをアップロードする必要がある。
社内機密や個人情報を先に削除しておく等の
データセットの精査が必要。

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

ローカルでRAGを定額で実装可能

裏側のイメージ図



API連携は
らくらく設定で手間いらず！

RAGを活用して間違いを減らし
LLMに考えて会話してもらう
ことがLGTMでも実現可能

誤回答をLGTMで大幅ダウン

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

開発しやすく、質の高いLLM

文字の生成が1秒以内

Chatgptなどのトランスフォーマーと比較して質問した際の回答スピードが圧倒的に速く、約20倍のスピードになります。

コンテキスト長が理論上無限

例えばChatGPTに質問する際コンテキスト長を超えた場合古い内容を忘れていきます。LGTMはこれが無限なので、事前情報を長文で入力したときでも文脈を理解できます。

短期記憶力に優れている

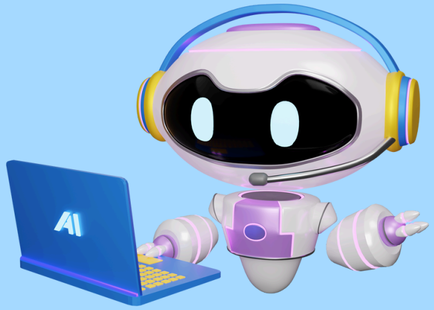
ユーザーに対し回答した際に続けて関連のある質問をされた場合でも記憶力が高いので通常回答が難しい質問でも前の質問から推測して正確な回答を出すことが可能です。

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

LGTMの活用シーン

社内の情報屋 wikiボット

社内のあらゆる情報を学習させることで社内ルールや、社員の趣味や嬉しいことなどをいつでも聞いて相談ができ、コミュニケーションの促進だけでなく新卒社員の退職率軽減にも役立ちます。



コーディングAI

プログラマー、足りてますか？
御社のプログラミング規則、独自ライブラリ、これまでに制作したコードを学習させることで、御社独自のコーディングAIを作成することが可能です。

デジタルヒューマン

自由に口調や性格をカスタマイズ可能なキャラクター性の高いデジタルヒューマンを作成し、AI Tuberの制作や、サイトに設置するチャットボットなど様々な活用をご支援します。

LGTMは他にも様々な課題を解決することができます

LGTM.

Base Architecture by RWKV

Language Gateway to Tomorrow Module

ライセンス

LGTM Easy Trainer

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

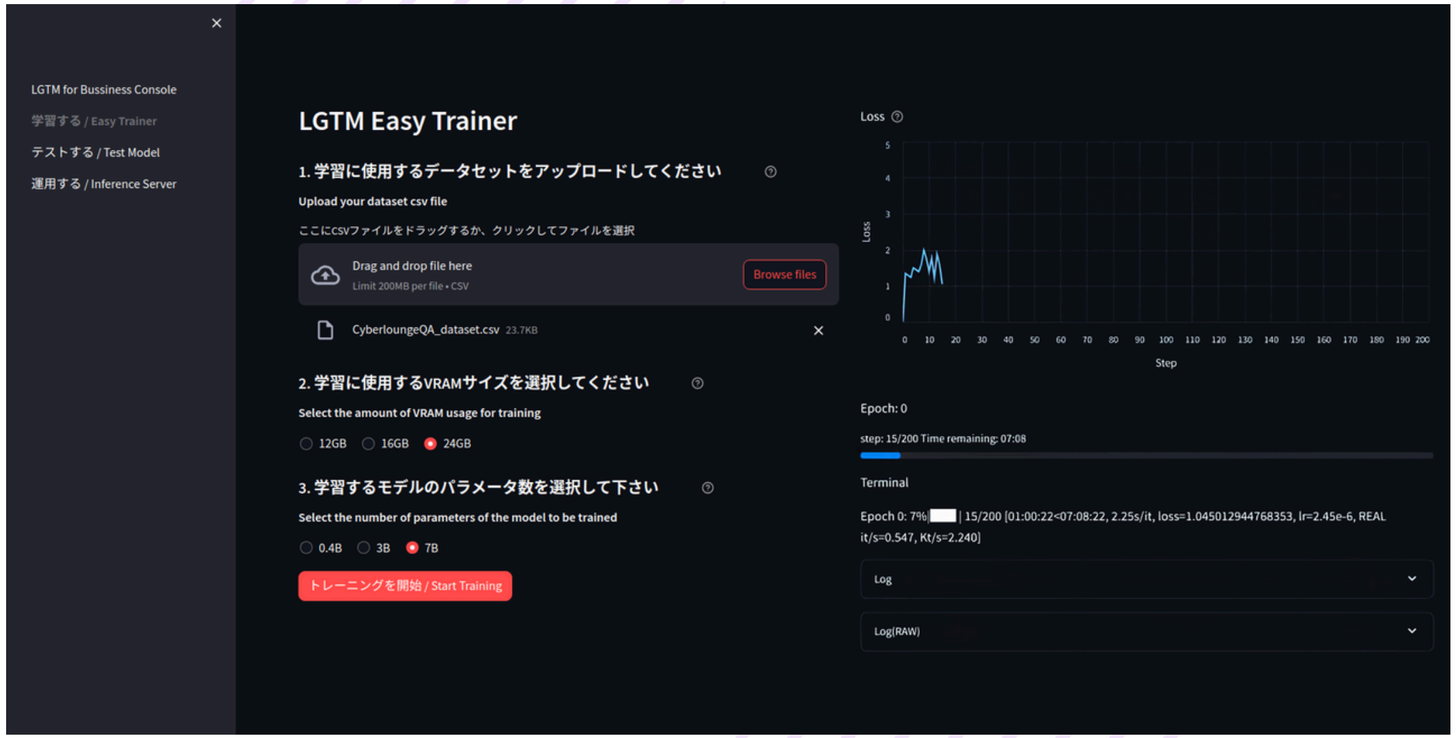
直感的に操作できる学習環境
LGTM Easy Trainer

「LGTM Easy Trainer」では、所定の形式のCSVファイルを読み込ませて、画面の指示に従っていくだけで、あなただけの特化言語モデルを作成できます。

直感的な操作

簡単なデータセット

圧倒的な学習スピード



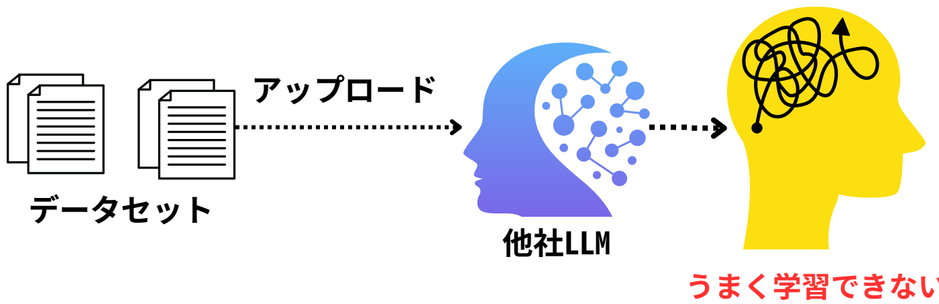
LGTM Easy Trainerライセンス

LETライセンスの特徴

～言語生成AIを簡単に作成可能に～

オリジナル言語モデルを作成できる
基盤を定額で提供！

従来



- **コスト**が膨らむ
- **専門性の高い技術者**が必要
- 個性がなく特化させることが難しい

他社サービスではRAGを
実装しただけのモデルがほとんど

LETライセンス

- **定額で学習し放題**
- IT未経験者でも操作可能
- **個性**を持った**特化型言語モデル**が開発可能

誰でも簡単に操作できて
特定型のオリジナル言語モデルを開発可能

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

ファインチューニング(学習)工程

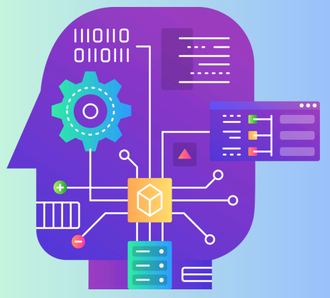
LGTM

データセット

RAG

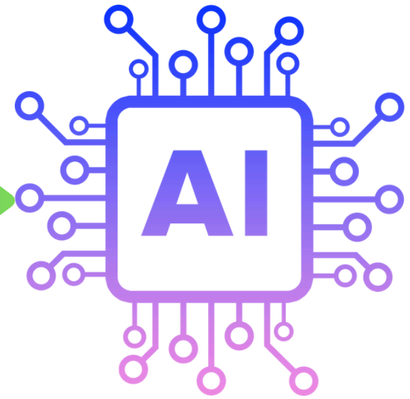
- 学習内容をCSV形式にて用意
- 誤回答を防ぐためRAGを用意

- 面倒な工程もノンストレス
- 定額でファインチューンし放題!
- 直感的な操作で作業可能



LGTM Easy Trainer

ファインチューニング



オリジナル言語モデル

活用シーンに特化した
オリジナルLLMが完成

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

シンプルで使いやすいチャットツール
LGTM Web UI

何でも気になったことを聞けるモードや、あなたが開発した「特化言語モデル」をシーンに合わせて利用することを意識して使いやすい「チャットツール」をご用意しました。

シンプルなUI

簡単モード切替

RAGも実装可能

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

比較表

	LGTM	GPT	LLaMA
導入レベル	初心者～上級者	中級～	上級者
生成スピード	◎（RTX3090にて7B 3000Token/sec）	○	○
学習	オンプレミス/クラウドで簡単 最短２分で学習	限定的なファインチューン/ 基本はプロンプトチューニング	難しい
コスト	世界最安・月額固定でファインチューニング可能	APIによる従量課金	高額な機材が必要
ビジネス面	製品用特化モデルの作成にかかる時間が削減できるため、エンド製品を早くリリースできる	OpenAI社に依存したシステムを構築する必要がある	ファインチューニングに時間がかかるため、製品実用化までに時間がかかる

パラメーター数

他言語モデル比較

LGTM	LLaMA-3	Phi-3	Mistral	Mixtral
1.6B、3B、7B、14B	8B	3.8B、14B(4bit)	7B	7B×8台、22B×8台

LGTM

14Bはチャットgpt3.5相当

2026年12月までに、
56B(ローカル)を目指して研究開発を進めております。

ローカルでも動かせるように
小さいパラメーター数も豊富

特化型言語モデルを作るのに
適したパラメーター数

学習最低環境

他言語モデル比較

LGTM	LLaMA-3	Phi-3	Mistral	Mixtral
20GB GPU	24GB GPU	24GB GPU(48GB推奨)	24GB GPU	80GB GPU×8台

LGTM

20GBで14Bを学習可能

LGTMは20GBでも14Bの学習を問題なく学習させることができます。

学習最低環境は最低限必要なGPUメモリのことを指します

他言語モデルは上記環境下では限定的な学習しかできない

学習手法

他言語モデル比較

LGTM	LLaMA-3	Phi-3	Mistral	Mixtral
LoRA(全層)、State、フルパラメーター	PEFT-LoRA(限定的)	LoRA	PEFT-LoRA(限定的)	PEFT-LoRA(限定的)

LGTM

20GBで14Bを学習可能

※State
RNN型アーキテクチャ独自のファインチューンニング手法。
通常のファインチューニングの上位互換として誕生した最新技術。
Transformer型だとモデル一つの運用につき24GB以上のVRAMが必要。
※PEFT-LoRA(限定的)
モデルの一部の層のみをピックアップしLoRA学習させる手法。
24GB GPUでも学習が可能であるが、学習効果が薄い。

State学習ができるのは
LGTMだけ

1台のGPUに無限に学習後モデルが
導入でき、瞬時に切り替え可能

推論環境

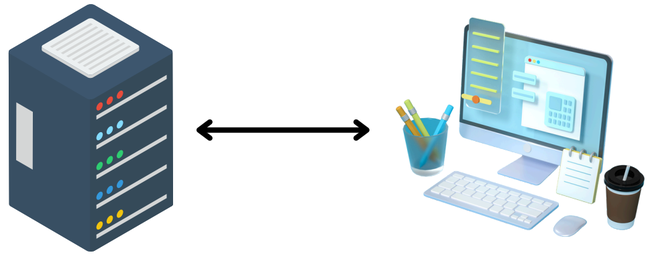
LGTM	LLaMA-3	Phi-3	Mistral	Mixtral
20GB 多重推論可能	24GB シングル推論	24GB シングル推論	24GB シングル推論	80GB シングル理論

LGTM

同時に5アカウント
利用可能

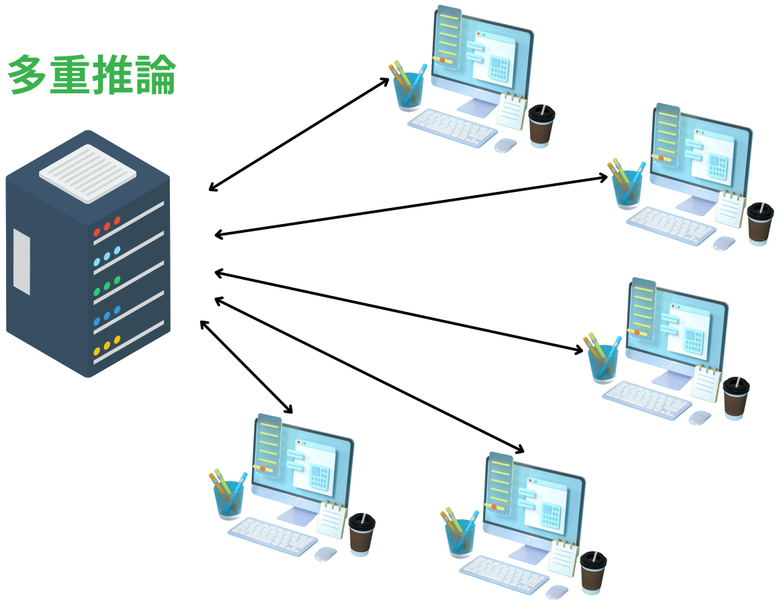
コスト削減に
大きく貢献

シングル推論



1つのGPUに対し１アカウント

多重推論



LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

ライセンス

LGTM	LLaMA-3	Phi-3	Mistral	Mixtral
商用可能	Meta ライセンス 限定的商用可	商用可能	商用可能	商用可能

LGTM

商用利用可能

定額でコスパ最強

他の言語モデルで「開発キット」と「言語モデル」を**定額で提供**をできるところはありません。

定額提供

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

製品化までの流れ

ファインチューニング

1. 少量のターゲットデータで
   高度に最適化されたモデル調整
2. ユーザー固有の用途や業界に特化
3. 短時間でのデプロイメントが現実に

RAG

1. 関連情報の検索と内容生成が
   一体で行われる
2. データベースからの情報取得と
   自然で正確な文の生成

製品例:チャットボット

1. 自然な日本語で流暢に対話し、
   ユーザーからの質問に対して適切な回答
2. 幅広いトピックに対応した
   満足度の高い対話体験

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

料金プラン

LGTMトライアル	LETライセンス契約	データセットコンサル
5万円/週	30万円/月	＋ 30万円/月
1週間お試しプラン	初回1週間 技術サポートあり	オンラインMTG （1時間×月4回）
• ファインチューニングし放題 • ローカルで学習可能 • 時間制限あり • 商品化不可(API連携不可) ※クラウドでの学習/運用も可能	• ファインチューニング可能なツール使い放題 • 開発、API連携レクチャ • Ver.アップデート対応	• データセットについて ◦ 作成方法 ◦ 内製化サポート

バックエンド開発、フロント開発、マーケティングのご相談も承ります。
詳しいアーキテクトや技術的な内容に関しては弊社SEにご相談ください。

LGTM.

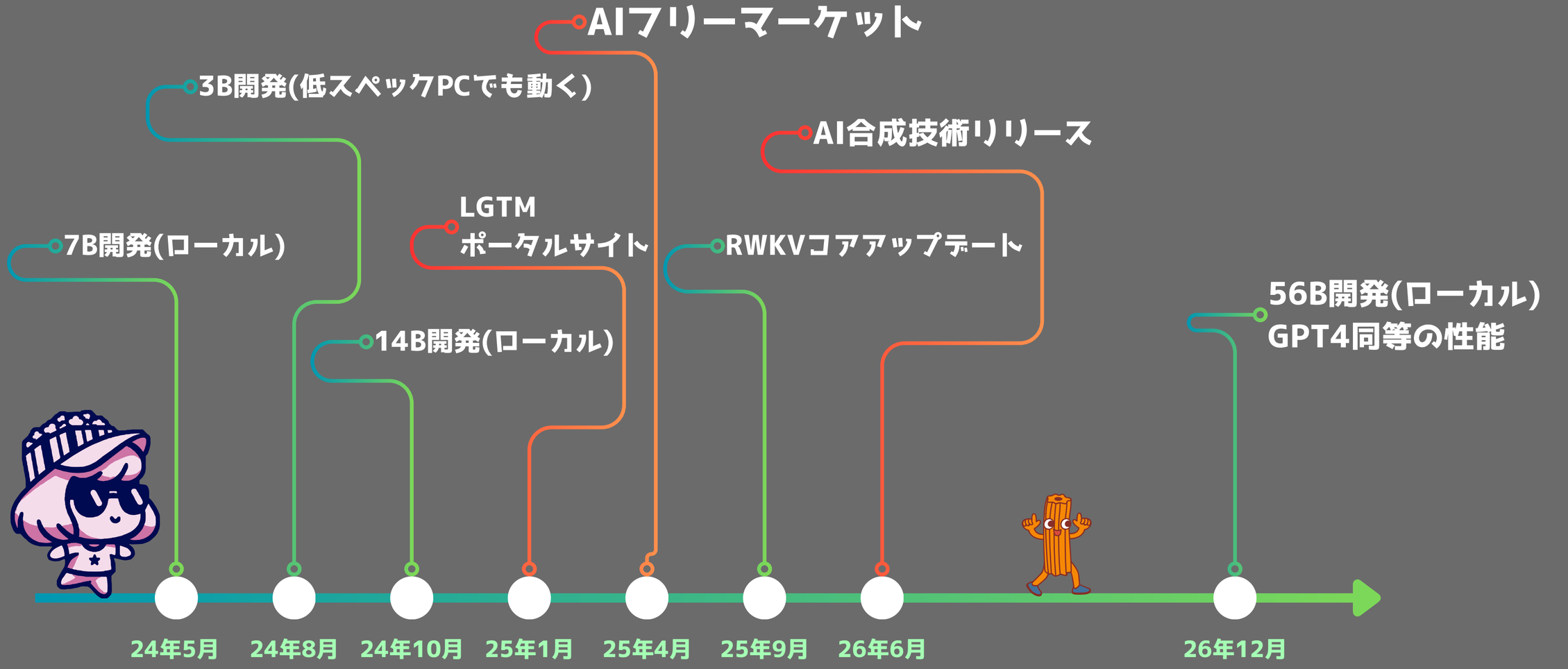
Base Architecture by RWKV

Language Gateway to Tomorrow Module

ロードマップ

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

LGTM　ロードマップ



LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

技術資料（LGTM for Business）

システム稼働要件(推奨)

OS：Ubuntu 22.04 LTS
CPU：x86-64互換CPU
RAM：128GB以上
GPU：90TFlops(BF16)，VRAM 24GB以上
　　CUDA 12.3以上 もしくは ROCm 6.0以上
　　が動作するもの

配布形態

実行環境はDockerイメージで配布
（導入後は自動更新されます）
※ インターネットに接続できない環境への導入につきましてはお問い合わせください。

サーバ要件

・SSHでアクセスできる状態であること
・Webサーバーが動作し、クライアントとの送受信ができる環境であること

推論結果（テキスト生成）の利用方法

LGTM for Businessを導入したサーバから発行されるAPIエンドポイントとAPIキーをクライアント側で設定してください。
（詳細はオンラインダッシュボードからご確認ください。）

学習方法

オンラインダッシュボードの「LGTM Easy Trainer」から学習（ファインチューニング）が可能です。

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

技術資料（LGTMのモデル構造・特徴）

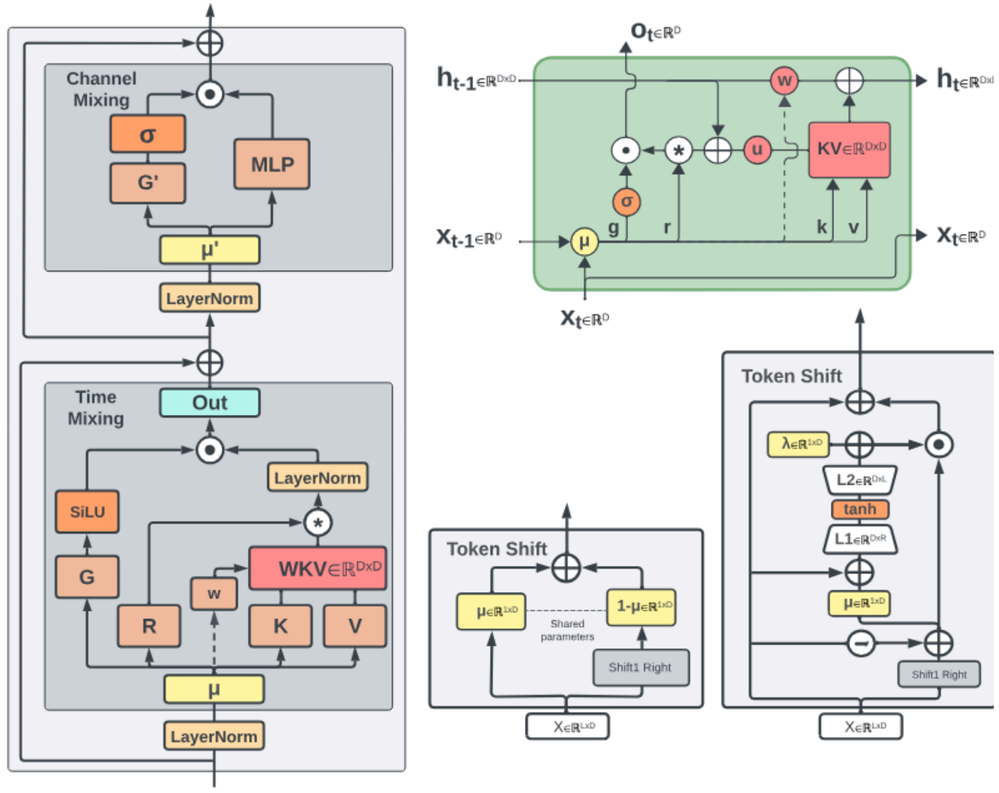
モデル構造

LGTMはRNN(回帰型ニューラルネットワーク)を使用したLLM(大規模言語モデル)です。

基本モデルアーキテクチャとして、RWKVという言語モデルを採用し、それを元に自社開発の学習手法とデータセットを用いてチューニングを行っています。

モデルの特徴

・実行およびトレーニング時の計算が少なく、大きなコンテキストサイズを持つTransformer系モデルと比べて、計算能力の要求が10分の1以下。
・計算負荷が任意のコンテキスト長に線形でスケールする。（Transformerは二乗にスケールする）
・ベースモデルが多言語でトレーニングされており、日本語だけでなく英語、中国語にも対応可。



図．モデル構造
Eagle and Finch: RWKV with Matrix-Valued States and Dynamic Recurrence (https://arxiv.org/abs/2404.05892)

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

技術資料（学習手法）

学習手法(Fine-Tuning)

「LGTM Easy Trainer」では State-Tuning という手法とRLHF（人間からのフィードバックによる強化学習）を組み合わせたものを採用しています。

その他の手法を使用してLGTMをファインチューニングする方法につきましてはお問い合わせください。

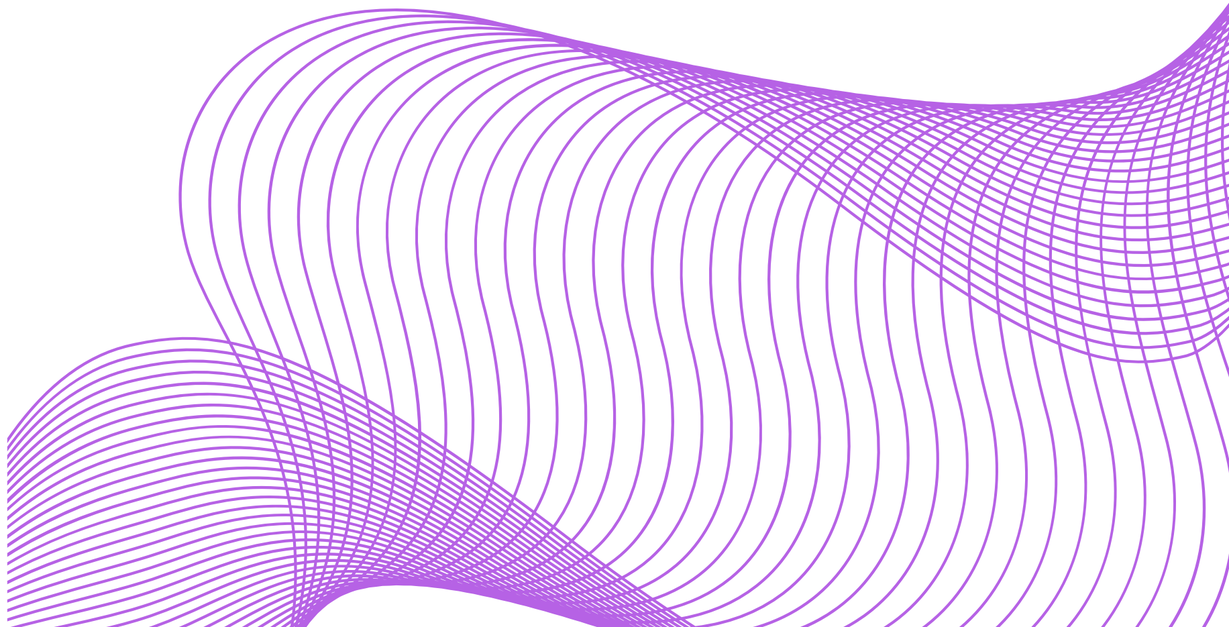
State-Tuning

State-Tuningとは、入力データセットに対して適切なpromptベクトルをバックプロパゲーションし、その時点のモデルのニューロン発火状態をニューラルネットワーク内の各層に埋め込む手法です。

RLHF

RLHFとは、機械学習モデル学習時に人間の価値基準を元に、趣向に沿う回答が出るように調整する強化学習の手法です。

本製品では弊社独自実装の学習手法によりこれを実現しています。



LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

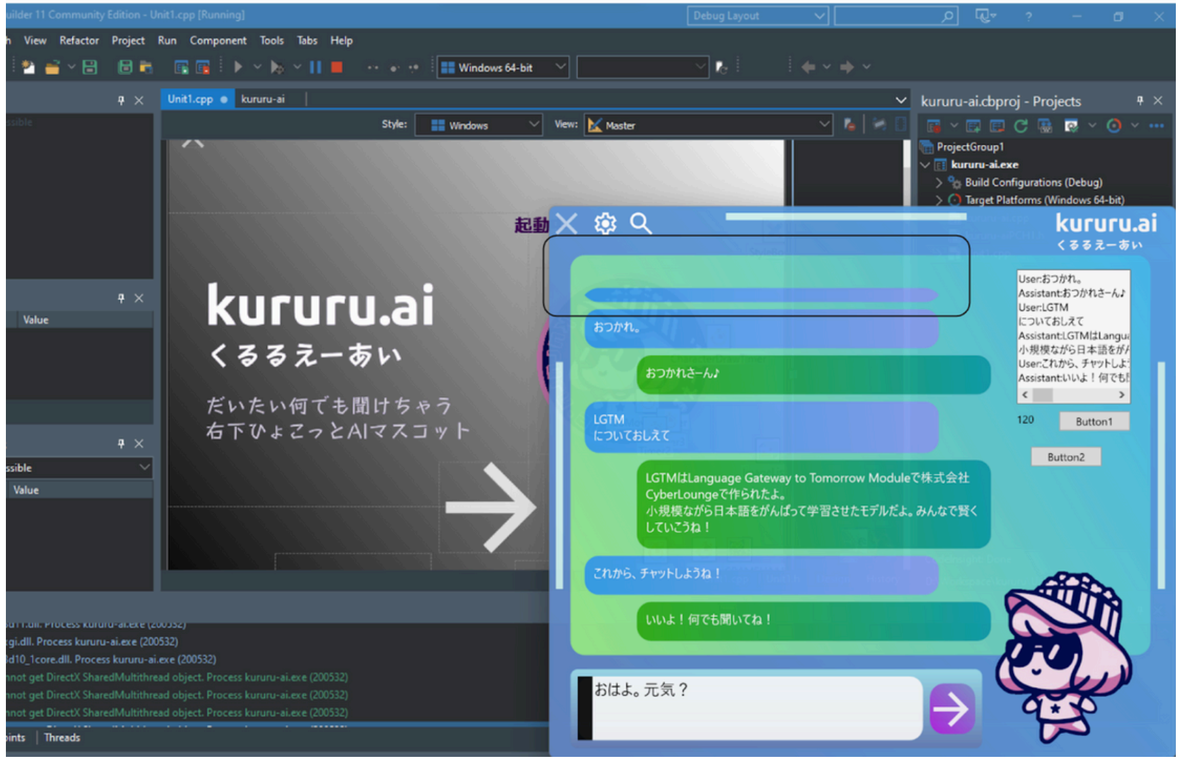
kururu.ai

AIデスクトップマスコット

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

kururu.aiは、LGTMを使った無料AIチャットボットです。

ネット環境不要の Windowsアプリ



1 PCにインストールするだけオフラインで利用可能

2 音声合成や天気予報などプラグインの追加が可能

3 プロンプトチューニングでキャラ設定ができる

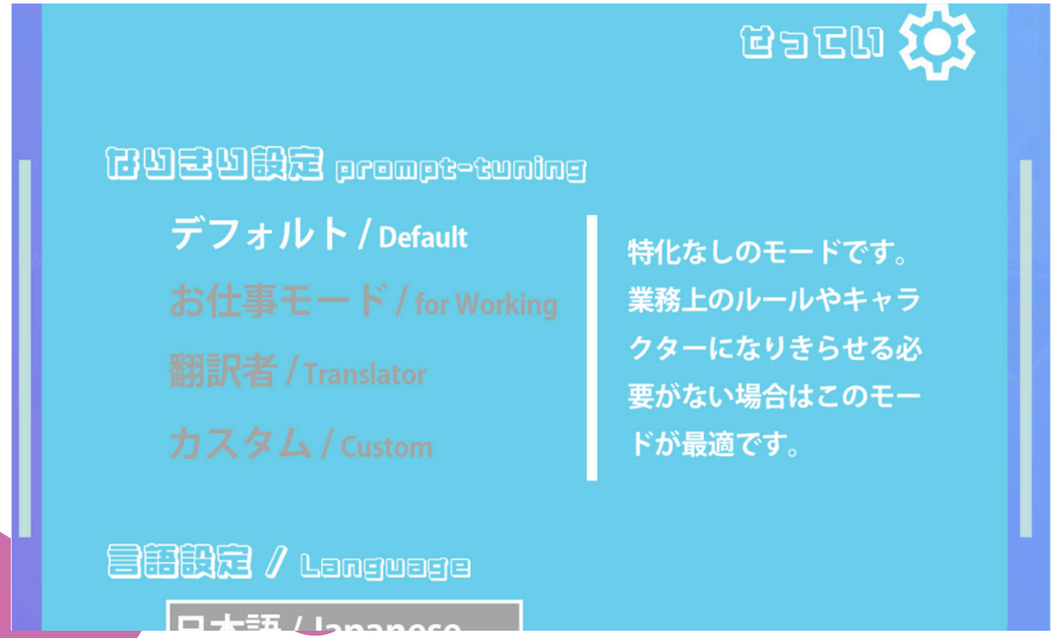
4 クオリティを上げるならファインチューンで簡単学習

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

簡単にキャラ設定

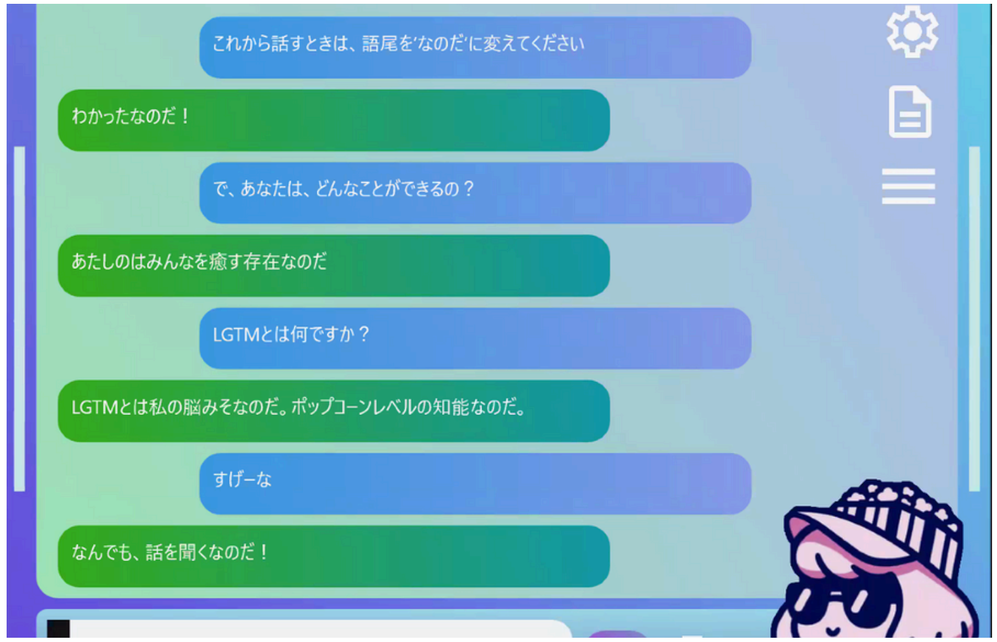
初期設定でカスタマイズ

初期画面で設定するだけの簡単になりきり設定！



プロンプト(チャット)で指示

カスタマイズで足りないならチャットで直接指示してみよう！

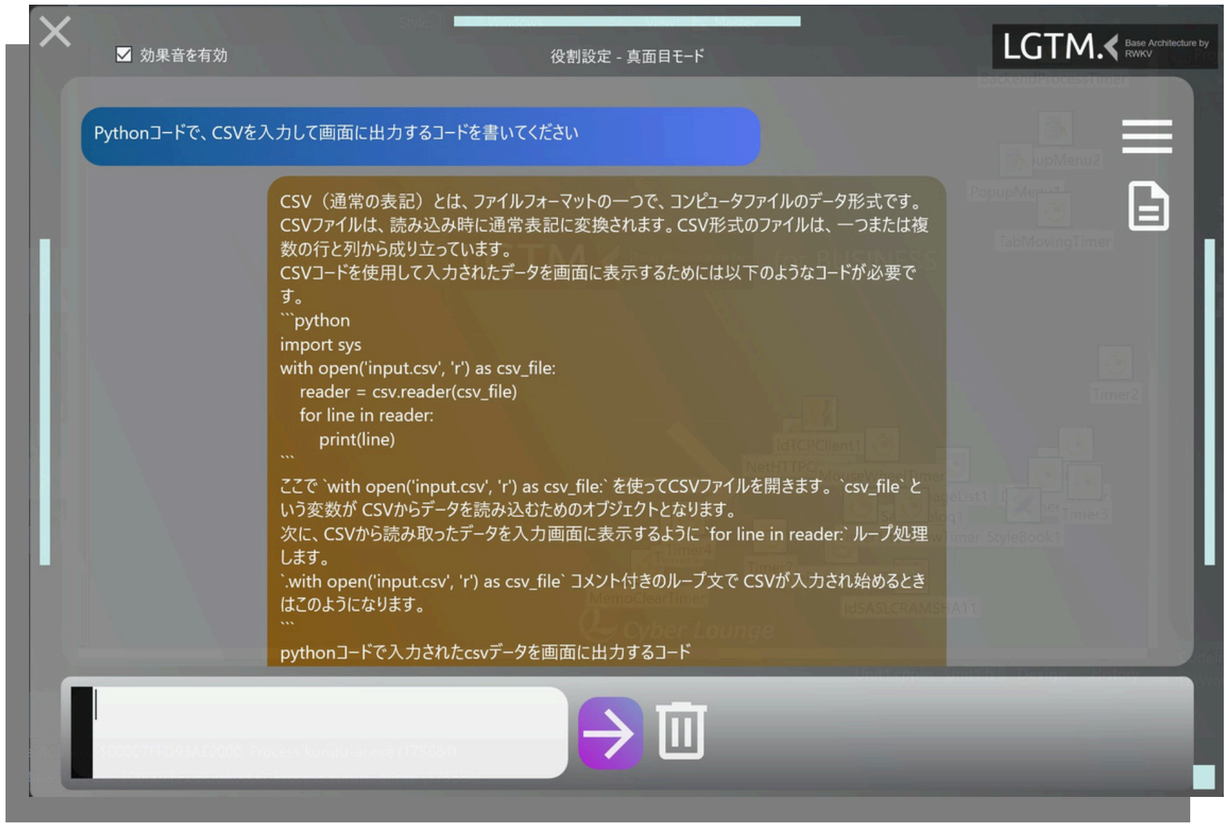


LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

For Business

エンド向けサービス開発をご検討されている企業様
社内でのAIチャット活用をご検討されている企業様

例)コーディングAI



ビジネス版UIで多様な活用方法

社内情報の共有や研修、教育など多岐わたるシーンでAIチャットは形を変えて活用が可能です。
例にあるAIチャットはプログラマーのコードを学習させコーディングAIとして活用している事例でございます。

OEMのご相談可能

機能をカスタマイズしたOEMも可能です。
様々なカスタマイズに対応いたしますのでOEMに興味/関心のある企業様はご連絡くださいませ。

LGTM. Base Architecture by RWKV
Language Gateway to Tomorrow Module

動作環境

- OS Windows 11
- CPU Intel Core i5 以上もしくは同等の互換CPU
- RAM 16GB以上
- SSD 20GB以上
- **GPU 不要（生成）**

PCのメモリで動くから
wi-fiががなくても動いちゃうぞ

アプリダウンロード後、初回起動時に
言語モデルファイルがダウンロードされます。

LGTM.

Base Architecture by RWKV

Language Gateway to Tomorrow Module

ご拝読ありがとうございました

TEL：03-5795-0067
MAIL：info@cyberlounge.co.jp

Cyber Lounge