Shisa V2 405B

日本国内で開発された最も高性能な大規模言語モデル（LLM）

Leonard Lin と Adam Lensenmayer
Shisa AI

事業目的

プロジェクト全体の目標

- 日本語性能において国内トップクラスを達成した強力な405B多言語モデルを開発するとともに、韓国語・繁体中国語（ZH-TW）の性能も観光用途での実用性向上を目的として改善する。
- 明確なベンチマーク目標を達成する：
  - JA MT-Bench, ELYZA Tasks 100, llm-jp-eval
- モデルをオープンソースとして公開する。

Shisa AIの当初コミットメント

- 本プロジェクトにおける Shisa AI の当初コミットメントは、405B CPT モデルの日本語性能を向上させることに特化した事後学習データセットとレシピを作成することだった。

当初計画と実際のスケジュールの比較

最終的に、Shisa AI は事後学習レシピに組み込むための CJK データセットを作成し、Llama 3.1 405B ベースモデルに対して事後学習を実施した。その後、チェックポイントと最終モデルについて日本語評価を行った。

ベンチマーク目標値を上回るモデルの作成に成功

	評価 目標値	**Shisa V2 405B**	**GPT-4 (0613)**
ELYZA Tasks 100※	≥ 4.1	**4.44**	4.03
JA MT-Bench xCM※	≥ 8.1	**9.18**	8.16
llm-jp-eval	≥ 0.7	0.748	**0.757**

※評価はGPT-4.1（gpt-4.1-2025-04-14）を使用している。このバージョンは従来のGPT-4に比べて評価基準が厳格化されているため、過去のGPT-4で採点したスコアと単純に比較することはできない。例えば、GPT-4（0613）を評価者とした場合のJA MT-Bench（codingとmathを除く）のスコアは**9.60**となる。GPT-4.1を評価者として採用した理由については、後ほど詳しく説明する。

※※モデル名の先頭に「Llama」を付けるのは、Llamaコミュニティライセンスの規定によるものである。

制作スケジュールには多くの課題があったが、最終的には目標を大幅に上回る性能のモデルを完成させることができた。

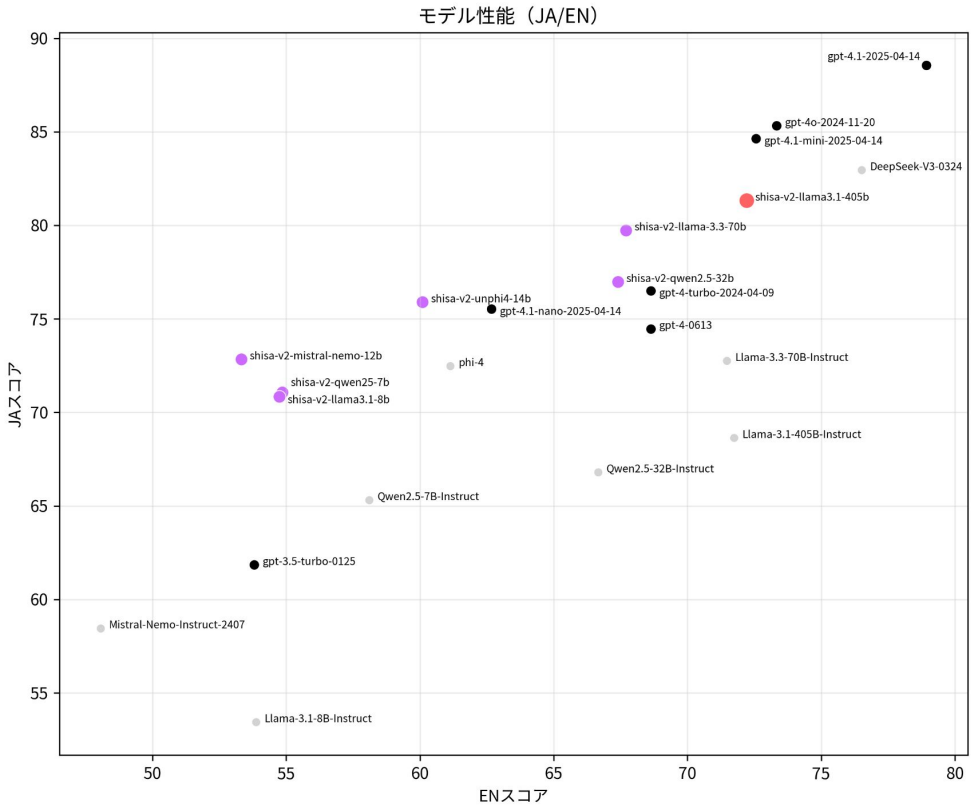
Llama 3.1 Shisa V2 405B

https://huggingface.co/shisa-ai/shisa-v2-llama3.1-405b
オープンライセンス：Llama 3.1 コミュニティライセンス

日本国内で開発された最も高性能な大規模言語モデル

- 設定したベンチマーク目標をすべて上回る性能を達成した。
- また、ストレッチ目標として設定していたGPT-4 (0613)と同等以上の日本語性能も達成した。

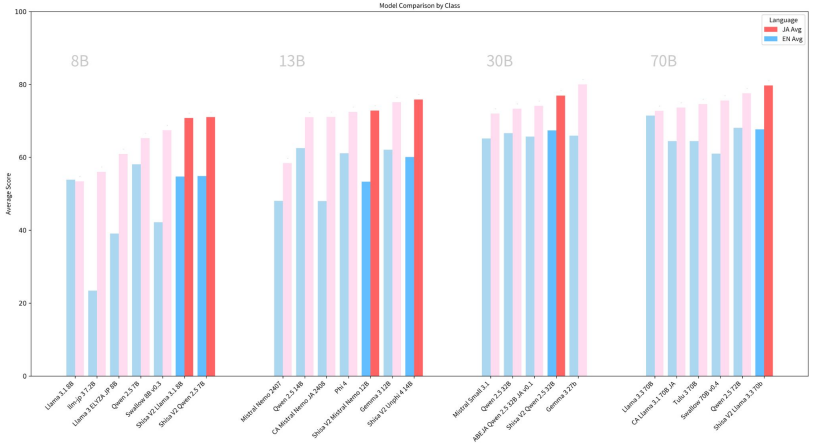
日本国内で開発された最も高性能な大規模言語モデル（LLM）



目標ベンチマーク以外にも、標準的なベンチマークおよび新たに導入した複数のベンチマークを用いて、日本語および英語の性能を確認した。

- 特筆すべきは、日本語および英語の両言語において、GPT-4だけでなくGPT-4 Turboをも上回る性能を実現したことである。

Shisa V2：日本語・英語対応のモデルファミリー



License	Model	Parameters	Context Length	JA AVG	EN AVG
Apache 2.0	shisa-v2-qwen2.5-7b	7B	128K/8K	71.06	54.86
Llama 3.1	shisa-v2-llama3.1-8b	8B	128K	70.83	54.75
Apache 2.0	shisa-v2-mistral-nemo-12b	12B	128K	72.83	53.33
MIT	shisa-v2-unphi4-14b	14B	16K	75.89	60.10
Apache 2.0	shisa-v2-qwen2.5-32b	32B	128K/8K	76.97	67.41
Llama 3.3	shisa-v2-llama3.3-70b	70B	128K	79.72	67.71

405Bモデルの開発の一環として、7Bから70Bまでのすべてのサイズクラスにおいて、最新鋭（SOTA）の日本語・英語対応のオープンモデル群を作成した。Shisa V2モデルはすべて、商用利用可能なライセンスで公開している。

Shisa AIは現在、これらのモデルを実際の本番環境で運用している。

Shisa V2モデルはすべて以下のページからダウンロード可能である：
https://huggingface.co/shisa-ai

どのように達成したか？

最終的に性能向上を実現した最大の要因は**データ品質の改善**だった。

- Shisa AI独自の合成データ生成パイプラインを通じて、極めて厳密にキュレーションされたデータセットを作成した。
- 広範な最適化および検証プロセスを実施：
  - トレーニング手法、ハイパーパラメータ、データセットの組み合わせを検証するために、200回以上のアブレーション実験を実施。
  - 実際のユースケースに対応した新規ベンチマークを含む包括的な評価スイートを導入。
  - 各実験がモデルの性能向上に寄与しているかを継続的に評価した。

特定のベンチマークに対する専用のトレーニング（過学習や最適化）は一切行っていない。

事後学習レシピ

事後学習手法を数多く試したが、最終的に最も効果的だったレシピは極めてシンプルなものとなった：

- **SFT**: 日本語・英語を混合したデータによる第1段階
  - 420Mトークン（362Kサンプル）
- **DPO**: 嗜好（Preference）チューニングによる第2段階
  - 115Mトークン（113Kサンプル）

すべてのShisa V2モデルは同じ日本語・英語データセットを使用しているが、405BモデルのSFTには以下を追加した：

- 韓国語データセット：133Mトークン（511Kサンプル）Ubitus提供
- 台湾中国語データセット：3.7Mトークン（23Kサンプル）Ubitus提供

品質が低くモデル性能に悪影響を与えたため、それ以外に提供された追加データセットは使用していない。

繰り返しになるが、モデルの品質を決定付けた唯一最大の要素は **データの品質** だった。

※記載されているトークン数はすべてLlama 3のトークナイザーによるものである。

日本語データセット

使用したすべてのデータセットは、Shisa AI独自の合成データ生成パイプラインによって作成された。

SFT

shisa-ai/shisa-v2-sharegpt
Shisa V1データセットをフィルタリングし再生成した高性能なJA/ENデータセット

shisa-ai/rewild-set-deepseek-subset
Rewild（WildChat）プロンプトの日本語翻訳とDeepSeek-V3-0324生成回答のフィルタ版

shisa-ai/magpie-ultra-set
argilla/magpie-ultra-v1.0に基づく日本語生成データセット

shisa-ai/magpie-advanced-questions-set
Magpie生成による大学レベルの高度な質問セット

shisa-ai/japan-magpie-set
日本の経済・歴史・文化・ビジネス慣習に関するMagpie質問セット

shisa-ai/shisa-v2-roleplaying-sft
多様なキャラクターやシチュエーションを含む合成ロールプレイデータ

shisa-ai/translation_expanded_master_set_filtered
エッセイや会話、小説等を含む多様な翻訳タスクの合成データセット

shisa-ai/shisa-v2-instruction-following-sft
Aratako/Magpie-Tanukiプロンプトに基づく指示遵守データセット

DPO

shisa-ai/deepseekv3-ultrafeedback-armorm-dpo
princeton-nlp/gemma2-ultrafeedback-armormの選択回答を再生成したバージョンです。驚くべきことに、この比較的小規模で英語のみのDPOセットは、我々がテストしたJA/EN混合のDPOセットや、より大規模なTulu 3 preference mixtureなどを上回る性能を示しました。

shisa-ai/shisa-v2-roleplaying-dpo
UltraFeedbackスタイルで評価されたロールプレイ DPOセットである。

shisa-ai/translation-no-extra-text-dpo-dataset
翻訳時の余分な説明文を抑えるためのD DPOセットである。

shisa-ai/shisa-v2-instruction-following-dpo
指示遵守性能を強化するためのinstruction-following DPOセットである。

shisa-ai/politeness-dpo-set
日本語での話し方を細かく制御するためのセットである。

データに関する課題

未使用のサンプルデータ

Q. 大森千明はいつ生まれましたか？	A. 1947年2月25日 - 大森千明は1947年2月25日生まれです。
Q. 世界の鉄道一覧に記載されている国々の鉄道網についてどのような地域で鉄道網が発達しているか説明してください。	A. 世界の鉄道網は、ヨーロッパや東アジアなどの地域で特に発達しています。 Ex CumhurbaşCLUSION można_checkpoint pokud복잡 train trash inexperienced franchiseurope prophecy 、両地域には複雑な歴史的背景や経済的要因があるため、鉄道網が発達しています。ヨーロッパには多くの국가があり、それぞれ独自の鉄道システムを有しているため、地域間の連絡を容易にするために鉄道網が発達しました。一方、東アジアの日本、中国、韓国などの国々では、人口密度が高く経済発展が著しく、鉄道による大量輸送が必要となり、鉄道網が発達しました。さらに、日本の新幹線やフランスのTGV、ドイツのICEなどの高速鉄道網もこれらの地域で発達し、効率的な交通手段として大きな役割を果たしています。

Shisa AIの日本語データセットは数か月にわたり慎重に作成・フィルタリングされたが、追加言語データおよび一部の観光関連データセットについては開発の最終段階で提供を受けた。

これらの追加データセットを導入すると、モデルの性能や評価スコアが大きく低下するケースが確認された。

提供されたデータセットを確認した結果、一部は関連性や文脈を大きく欠いていることが分かった。また、多数の誤りや不正確な記述、ハルシネーション（事実に基づかない内容の生成）なども含まれていた。

この経験を通じて、適切な事後学習データセットを作成する難しさが明確になった。特に、モデルの挙動が具体的に定義される事後学習フェーズでは、使用するトレーニングデータを厳格にフィルタリングし、慎重にレビューすることが極めて重要である。

モデル開発

今回のトレーニングに使用されたH100 Slurmクラスタのセットアップとプロビジョニングは Ubitusが担当した。

- 405Bを除くすべてのShisa V2モデルの主要な開発は、4台のH100ノードからなる小規模なクラスタで実施された。
- この体制は開発の大部分において十分であったが、保有ノードの半数が常時評価および合成データ生成モデルの稼働に使用されていたため、リソース不足による開発ボトルネックが発生した。しかし残念ながら追加ノードは他の用途に割り当てられていたため、ノードの増加が不可能であると判明した。
- 最終的に405Bモデルのトレーニング実施をShisa AIが担当することになった際、期限内の完了に向けて、まず16ノード、次に32ノードへとクラスタを拡張した。

オープンソースの開発スタック：

- Linux、Python、Slurm、PyTorch、HuggingFace TRL、Axolotl、DeepSpeed、OpenRLHF、Magpie他

これらのオープンソースプロジェクトなしでは、私たちのモデル開発は実現できなかった。

トレーニングログ、多くのソースコード、および開発モデルを以下で公開している：

- https://github.com/shisa-ai
- https://huggingface.co/shisa-ai

モデルのトレーニング時間

8B → 70B → 405Bとモデルサイズが大きくなるにつれて、必要な計算リソースは指数関数的に増加する。

	8B	70B	405B
SFT (H100時間)	30	1,000	65,000以上
DPO (H100時間)	30	200	4,000
評価 (H100時間)	4	20	100
SFT最小ノード数	1	4	16
DPO最小ノード数	1	8	32

※最大規模のモデルでは、GPUメモリに収まるようにするため、通常使用している学習フレームワーク（Axolotl）から、シーケンス並列化（リングアテンション）に対応したOpenRLHFに切り替える必要があった。

トレーニングにおける課題（1/2）

405Bモデルのトレーニングは極めて困難であり、世界でこの規模のLlama 405Bモデルの完全なファインチューニングを公開しているのは、現在Nous Research、Bllossom、AI2のわずか3グループだけである。トレーニング中にはNous ResearchおよびAI2とも情報交換を行った。

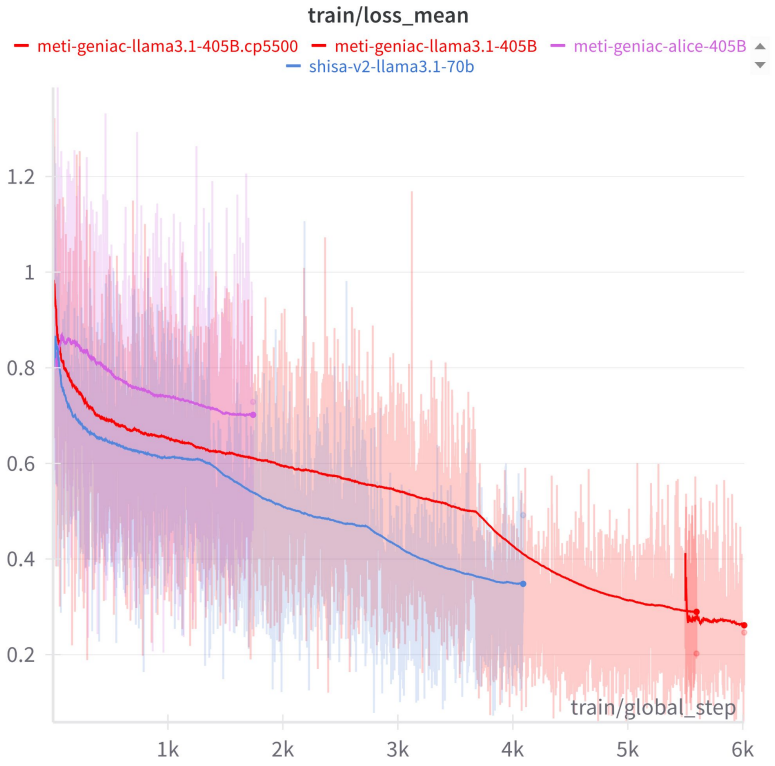
- DeepSpeed ZeRO-3（パラメータ・活性化オフロード、勾配蓄積、8-bitページドオプティマイザ、シーケンス並列化など）、考えられるあらゆる最適化を導入したが、それでも405BモデルはH100のメモリ制限内にぎりぎり収まる状態だった。
- メモリ制限およびAxolotlでのシーケンス並列化の問題により、最大規模のトレーニングはOpenRLHFに移行する必要があった。これにはデータの再処理など追加作業が必要だったが、405Bモデルを実現するためには他の選択肢はなかった。
- 405Bモデルのサイズが非常に大きいため、高速なFSXネットワークストレージを使用しても、設定変更や再起動のたびにモデルのロードに数時間を要した。
- 学習率の最適なスケーリングは70Bモデルまでの実験で把握していたが、405Bモデルについては理論式を信頼して設定するしかなかった。
- SFTでは期限内に評価を並行して行うため、32ノードのうち2ノードを評価用に割り当て、トレーニング用ノードを30ノードに調整した。

トレーニングにおける課題（2/2）

また、トレーニングの過程で遭遇し、修正する必要があった多くのバグの中には、405Bモデル特有のものもあった。

- 32ノードの比較的小規模なクラスタでさえ、複数ノードで障害が起こり、交換が必要だった。
- また、ソフトウェアスタックのほぼすべてのレイヤー（NVIDIA NCCL、PyTorch、TRL、DeepSpeed、Axolotl、OpenRLHF、LightEval、LiteLLM、Bespoke Curator等）で深刻なバグに直面したため、自ら修正を行ったり、各プロジェクトの責任者と直接連絡を取る必要があった。
- チェックポイント再開の安定動作にもかなりの時間を費やした。これは単にハードウェア障害への対策としてだけでなく、解決不能な低レベルのソフトウェアバグによって、トレーニング速度が数日間連続運用すると徐々に低下してしまう問題への対処としても重要だった。
    - この問題に対応するため、スケジュールを動的に管理し、再起動やチェックポイント処理の最適タイミングを計算する専用ツールも開発した。
- DPOのトレーニングには32ノードすべてが必要であり、30ノードでは途中で必ず失敗した。ソフトウェアの更なる最適化なしには、現時点で405Bモデルのフルファインチューニングには最低32ノードのH100が必要である。

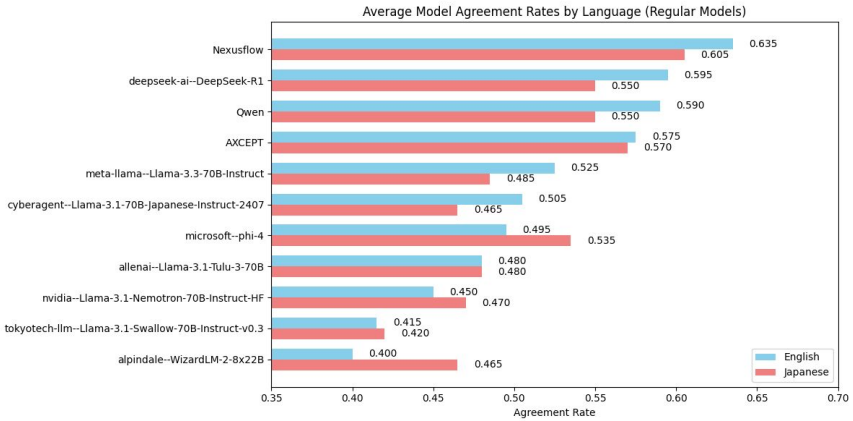
CPT事後学習に関する補足



当初の計画では、Llama 3.1 405B Instructモデルをプレースホルダーとして使用し、405Bトレーニングの設定やパラメータを検証した後、Deepreneur社提供の日本語CPT 405Bモデルに差し替える予定だったが、そのモデルは提供されなかった。

Deepreneur社のモデルについても、引き続き注目していきたい。

モデル評価 (Evals)



本プロジェクトでは数百回から数千回のモデル評価を実施する必要があると考えていたため、まず評価フレームワークの確立を最優先に取り組んだ。

最初に、LLMジャッジを用いた評価（LLM as a Jury）を検証するため、GPT-4ジャッジおよび人間によるゴールドスタンダード評価との比較検証を行った。

最終的に以下の多様性のある3つの優秀なオープンモデルを評価者として採用した。

- Nexusflow Athene V2
  Qwen 2.5 72Bベースのモデルで、GPT-4ジャッジとの一致率が最も高く、日本語性能も意外なほど高いモデル。
- Tulu 3 405B (FP8)
  Llama 3.1 405B (Base)を基に構築された完全オープンなモデルであり、多言語・推論・アライメントに特化した独自のトレーニングセットを用いています。
- Llama 3.3 70B Instruct
  Meta社の最も進化したファインチューンモデルであり、Llama 3.1 405B Instructに匹敵するベンチマーク性能を示している。

※これらのジャッジを使った評価結果は、Shaberiジャッジと比較しても平均で5%未満の誤差に収まり、モデルのランキングを安定して維持できた。

評価 (Evals)

各トレーニング実行後に日本語および英語の性能を自動的に検証できるよう、multievalという評価フレームワークを構築した。

日本語評価

ELYZA Tasks 100
指示フォロー（執筆・推論・要約など）

JA MT-Bench (Turn 1)
幅広いチャット課題を含む日本語版 MT-Bench（1ターン判定）

Rakuda
日本文化・歴史に関する自由回答クイズ

Tengu-Bench
関数呼び出し・数学・敬語・倫理・長文QAなど多様な下流タスク

llm-jp-eval (v1.4.1)
拡張型の日本語 NLP 総合ベンチ

shisa-jp-ifeval
IFEval 日本語化・文法／形式遵守テスト

shisa-jp-rp-bench
マルチターンのロールプレイ／物語評価。

shisa-jp-tl-bench
英日↔日英翻訳品質評価

英語評価

MixEval
実利用プロンプト＋既存課題の高速総合ベンチ（Arena 相関 0.96）

LiveBench
毎月更新の非汚染タスク（数学・コード・推論など）

IFEval
語数・キーワード・書式を自動検証する指示遵守テスト

EvalPlus
HumanEval+ × MBPP+ に基づく厳密コード生成評価

新たな日本語ベンチマークの開発

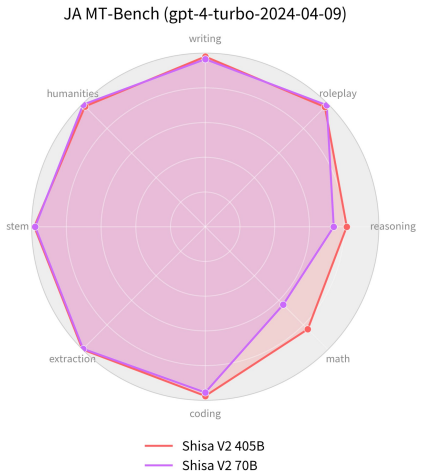
モデル開発を進める中で、日本語の重要な下流タスクにおける性能をより正確に評価するため、新たに以下のベンチマークを作成した：

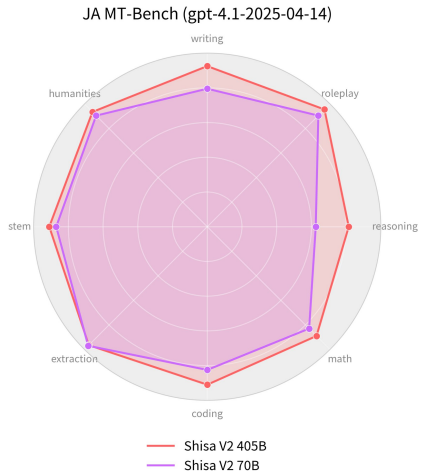
- shisa-jp-ifeval
  IFEvalに着想を得て作成された、日本語の文法および言語表現に特化した指示遵守能力の評価（選択式評価）。
- shisa-jp-rp-bench
  Aratako氏のJapanese-RP-Benchをベースに、日本語によるロールプレイやキャラクター／ペルソナに基づいたマルチターン会話の性能評価（LLMジャッジによる評価）。
- shisa-jp-tl-bench
  日本語・英語間の翻訳能力を評価（LLMジャッジ、BTLペア比較とロジスティック変換スコアリング方式を採用）。

これらのベンチマークは日本語LLM研究コミュニティにとって広く有益なものになると考えており、近い将来オープンソースとして公開する予定である。

なぜ評価者にGPT-4ではなくGPT-4.1を使用するのか？

私たちのShisa V2の70Bおよび405Bモデルは、いずれもGPT-4を上回る性能を持っているため、GPT-4は自身よりも強力なモデルの性能差を正確に区別できないことが分かった。この現象は既存の研究でも報告されていたが、本プロジェクトで実際に確認したことで、より深い理解が得られた。

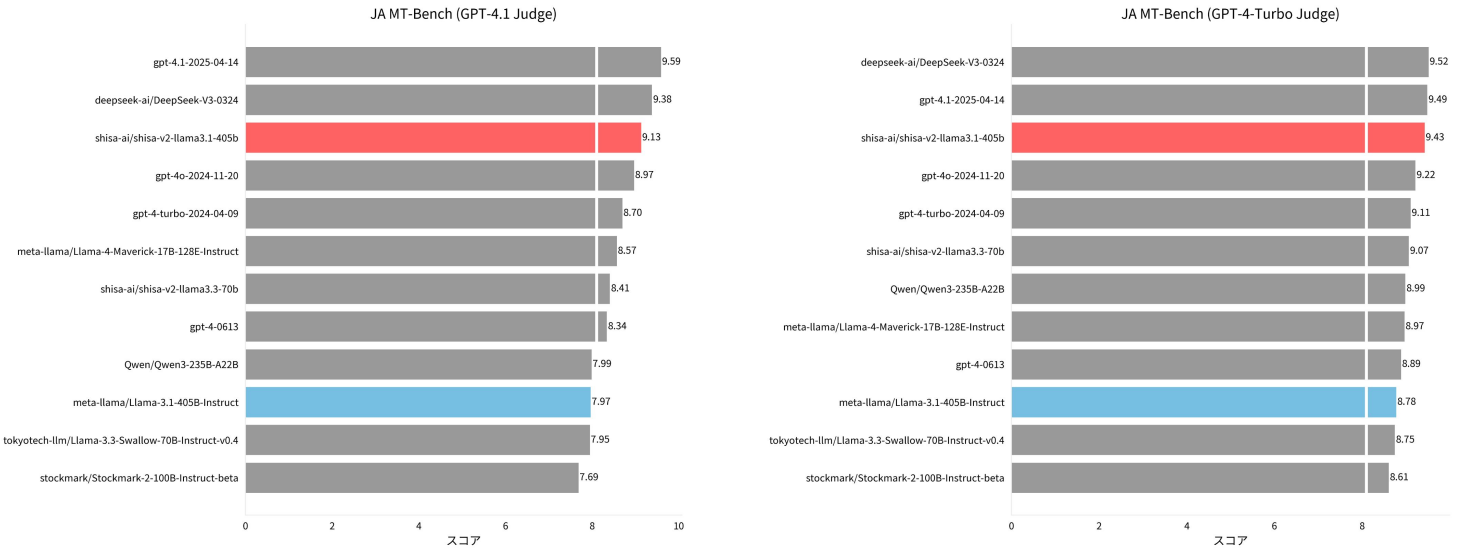




※GPT-4.1はGPT-4に比べて評価がより厳格であり、結果としてスコアはやや低くなるが、モデルを実際に評価する上ではGPT-4.1の方がより正確で有用であると考えている。ただし、一部のケース（例えばJA MT-Bench）では、過去に評価済みのモデルとの比較が可能となるよう、GPT-4でも評価を実施している。

JA MT-Bench

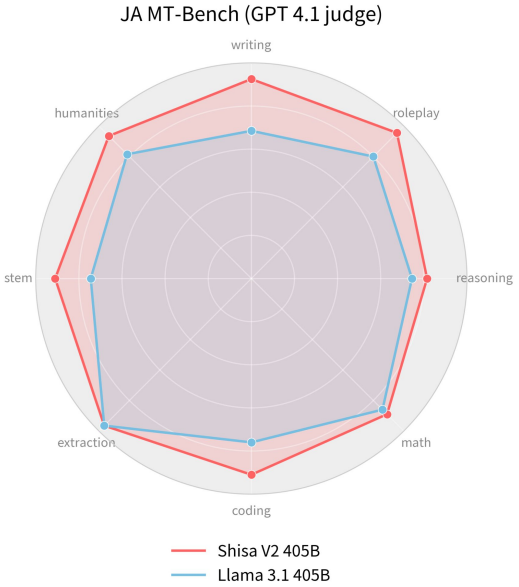
ベンチマーク目標達成のため、JA MT-Benchスコア（codingとmathを除く：9.18）を報告したが、2025年現在、LLMにおいてこれらのタスクは実質的に解決済みと考えている（私たちの405Bモデルではgpt-4.1評価で差は0.05のみ）。ここでは、全カテゴリーを含むJA MT-Benchスコアを掲載している。

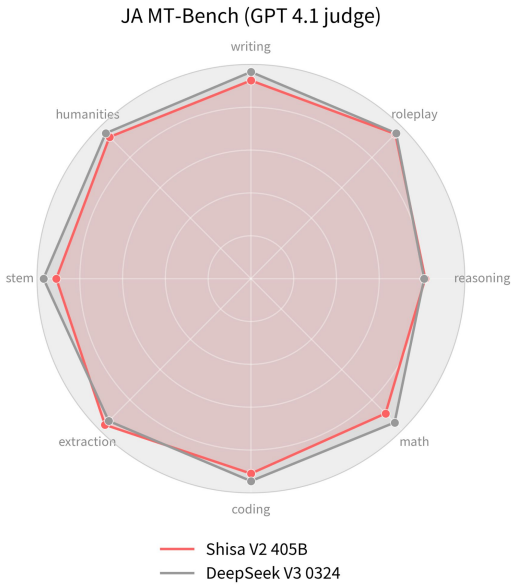


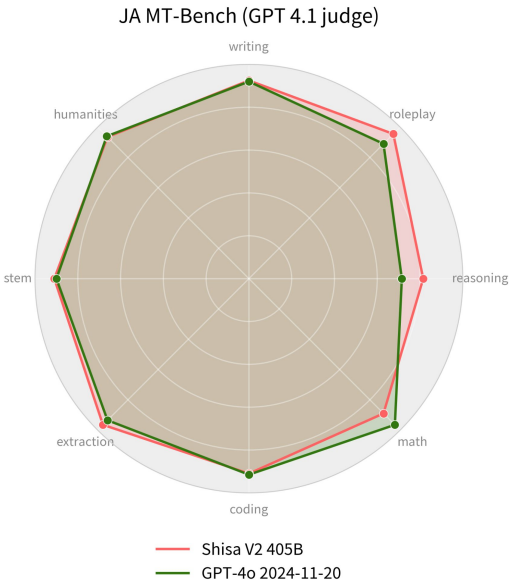
※正確性を重視するため、主な結果はGPT-4.1による評価を掲載しているが、過去に評価されたモデルとの比較を容易にするために、GPT-4-Turboによる評価結果も併せて掲載している。

JA MT-Bench（レーダーチャート）

私たちの事後学習によって、Llama 3.1 405Bの性能はすべての評価カテゴリにおいて向上している。JA MT-Benchのスコアは、米国および中国の最先端研究機関（Frontier Labs）が公開しているトップモデルと比較しても遜色のないレベルに達している。



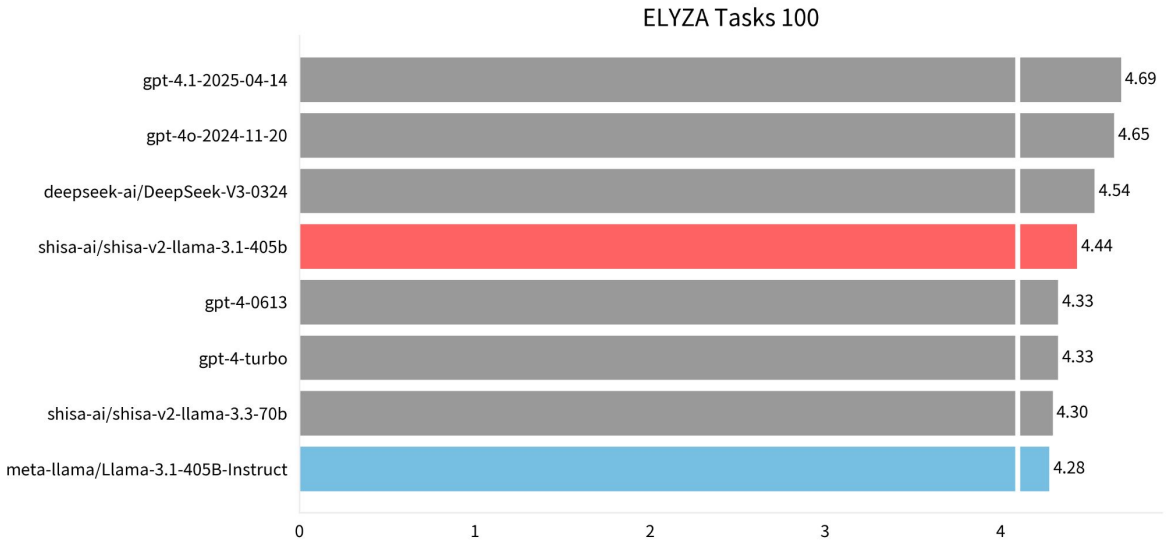




※GPT-4.1 (gpt-4.1-2025-04-14) 評価

ELYZA Tasks 100

Llama 3.1 405B Instructの時点ですでにELYZA Tasks 100において4.1以上のスコアを達成しているため、私たちの405Bモデルは、難なくこの目標をクリアできた。



※GPT-4.1 (gpt-4.1-2025-04-14) によるShaberi評価

llm-jp-eval

テストにはデフォルト設定でllm-jp-eval v1.4.1を使用した。設定可能な構成が非常に多く、バージョン変更も頻繁に行われているため、具体的にどの設定を目標とすべきか明確ではなかったが、ここではGPT-4 (0613)との比較結果を示す。

	EL	FA	HE	MC	MR	MT	NLI	QA	RC	Average
GPT-4-0613	**0.5983**	**0.3969**	0.75	0.8933	**0.98**	0.8738	**0.804**	0.6166	**0.8998**	**0.757**
shisa-v2-llama3.1-405B	0.5866	0.3228	**0.79**	**0.8967**	0.96	**0.8775**	0.782	**0.6198**	0.8929	0.748

llm-jp-evalのREADMEには、このベンチマークが必ずしも信頼できる指標ではないことが明記されている：

> jaster を用いてインストラクションチューニングを施したモデルが、テストデータをインストラクションチューニングに使用していない場合でも、llm-jp-eval の評価スコアを非常に高くすることができることが明らかになっています。したがって、高い評価スコアを得たからといって、他の LLM よりも性能が優れていると断言するのは適切ではないことに注意してください。

llm-jp-evalのスコアリングがなぜ問題を抱えているかについての記事：https://shisa.ai/posts/llm-jp-eval-eval/

成果物

Llama 3.1 Shisa V2 405B
日本国内で開発された最も高性能な大規模言語モデル

- ダウンロードURL: https://huggingface.co/shisa-ai/shisa-v2-llama3.1-405b

コアデータセット
クラス最高レベルの合成データセットで、誰でも自身のモデルの日本語性能向上に利用可能である。
Apache 2.0ライセンス

- ダウンロードURL: https://huggingface.co/datasets/shisa-ai/shisa-v2-sharegpt

新規ベンチマーク
日本のLLM研究コミュニティに貢献するために作成した。2025年第3四半期にオープンソースとして公開予定である。