Shisa V2 405B

Japan's Highest Performing LLM

Leonard Lin and Adam Lensenmayer
Shisa.AI

2025-06-02

Objectives

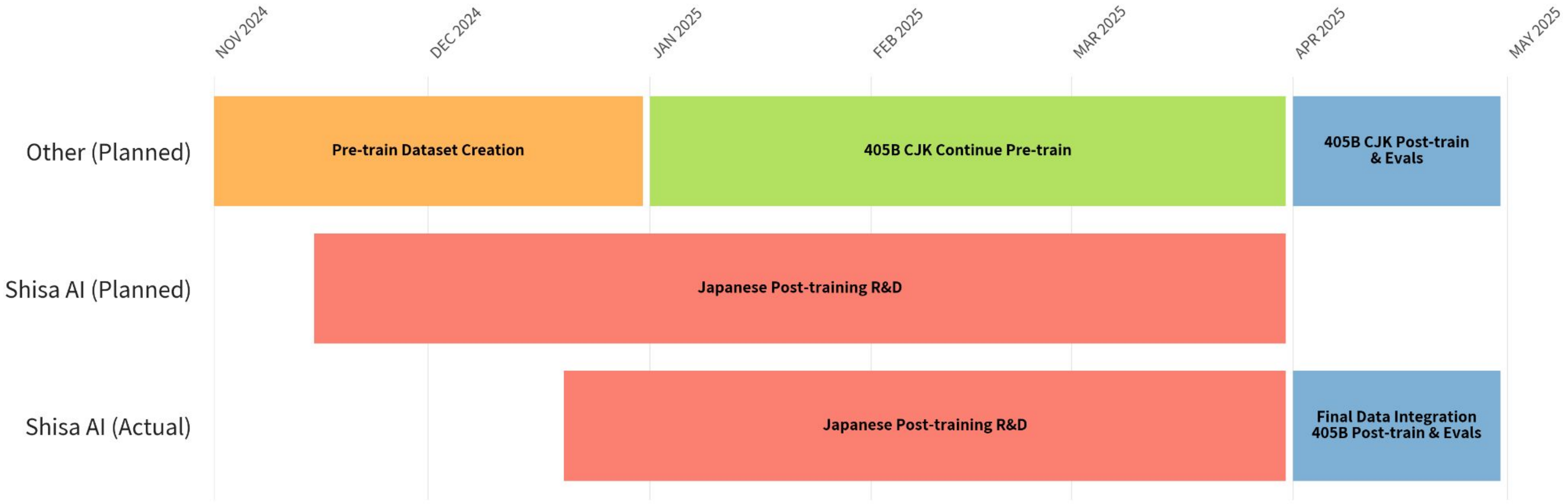
Overall Project Goals

- Develop a powerful 405B multilingual model achieving top-tier Japanese language performance domestically
- Improve Korean and Traditional Chinese (ZH-TW) capabilities, specifically enhancing practical usability for tourism-related applications
- Meet clearly defined benchmark targets:
  - JA MT-Bench, ELYZA Tasks 100, llm-jp-eval
- Release the model as open-source

Shisa.AI's Original Commitment

- Shisa.AI's initial commitment within this project was to create a specialized post-training dataset and recipe aimed at improving the Japanese language performance of the 405B CPT model

Initial Plan vs Actual Schedule



In addition to our original commitment, Shisa.AI agreed to create the final CJK dataset and train the final 405B model using **Llama 3.1 405B Instruct** as its base. We also ran all evaluations on both intermediate checkpoints and the final model.

Final Result: All Benchmark Targets Exceeded

	Target Scores	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**

※Judging was conducted using GPT-4.1 (gpt-4.1-2025-04-14). Since this version applies stricter evaluation standards than previous versions of GPT-4, scores obtained from earlier GPT-4 judging are not directly comparable. For example, the JA MT-Bench score (excluding coding and math) using GPT-4 (0613) as evaluator is 9.60. The rationale for using GPT-4.1 scores will be explained in detail later.

※※Per the Llama Community License, the official name of this model is prefixed with "Llama."

Despite facing challenges in the production schedule, we successfully developed a model whose performance surpassed the project targets.

Llama 3.1 Shisa V2 405B

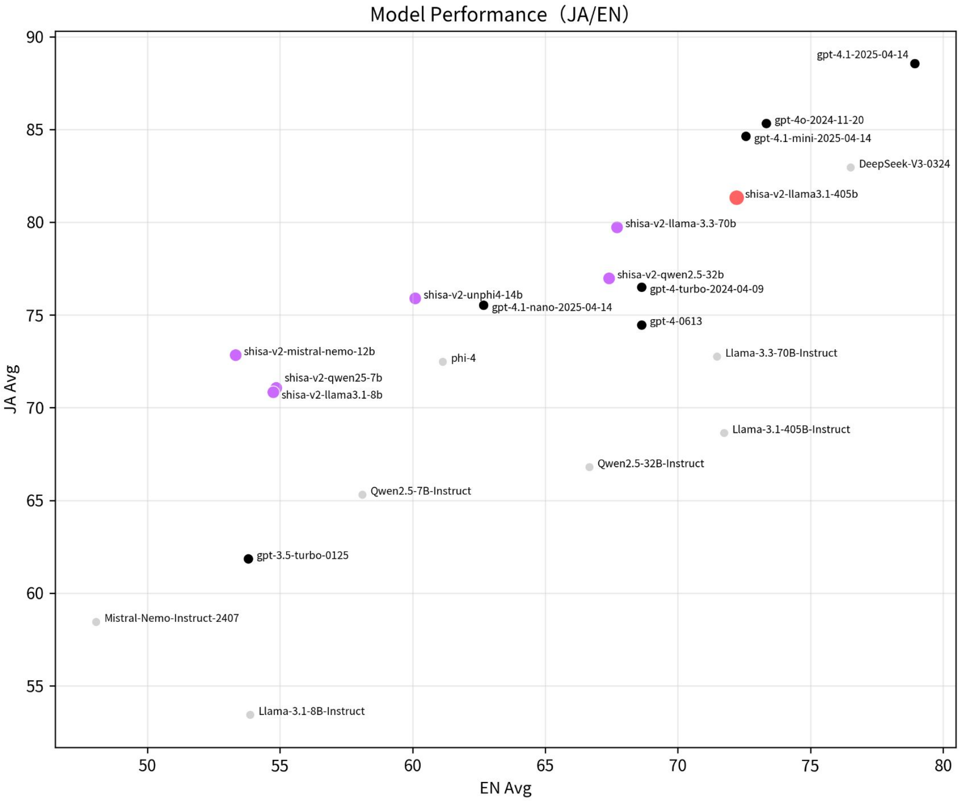
https://huggingface.co/shisa-ai/shisa-v2-llama3.1-405b

Open License: Llama 3.1 Community License

The highest-performing LLM ever developed in Japan.

- Exceeded all the project benchmark targets
- Achieved our stretch goal of matching or surpassing GPT-4 (0613) performance in Japanese.

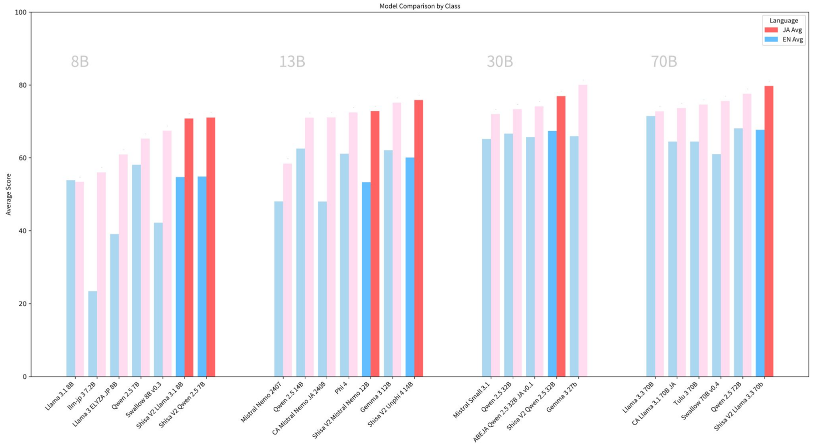
Japan's Highest Performing LLM

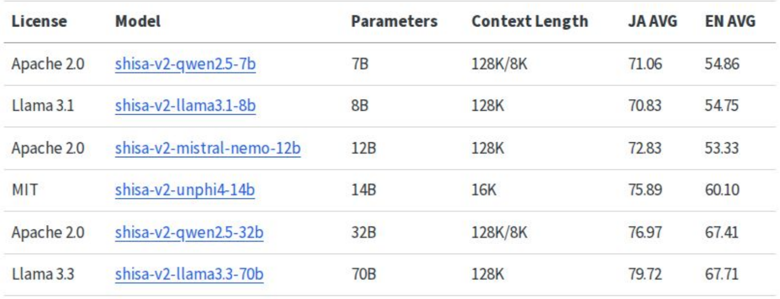


In addition to the primary target benchmarks, we evaluated the model's performance in Japanese and English using both standard and several newly developed benchmarks.

- Notably, Shisa V2 405B achieved **superior performance to both GPT-4 and GPT-4 Turbo in both Japanese and English**

Shisa V2: A Family of Bilingual Models





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

As part of the development of the 405B model we created SOTA bilingual (JA/EN) open models across all class sizes from 7B to 70B parameters. All Shisa V2 models have been released under open licenses allowing commercial use.

Shisa.AI is already using these models in real-world production environments.

For more about the Shisa V2 family, including download links: https://shisa.ai/posts/shisa-v2/

How Did We Do It?

The single most important factor for model performance was **data quality**.

- We extended our synthetic data generation pipelines to produce an exceptionally rigorously curated dataset
- Extensive optimization and validation:
    - Performed over 200 ablation runs to validate combinations of training techniques, hyperparameter tuning, and dataset composition
    - Implemented a comprehensive evaluations suite, including new benchmarks reflecting common real-world downstream use-cases
    - Continually assessed whether each experiment improved model performance

No benchmark-specific training (eg, rephrased benchmark data) was conducted.

Post-Training Recipe

We experimented with numerous new post-training methods, but the most effective recipes turned out to be relatively standard:

- **SFT**: Single stage supervised fine-tuning of mixed JA/EN corpus
  - 420M tokens (362K samples)
- **DPO**: Preference tuning via Direct Preference Optimization
  - 115M tokens (113K samples)

All Shisa V2 models utilize the same JA/EN dataset, but for the 405B model, the following additional datasets were included in the SFT stage:

- Korean dataset: 133M tokens (511K samples) provided by Ubitus
- Taiwanese Mandarin (ZH-TW) dataset: 3.7M tokens (23K samples) provided by Ubitus

While other supplementary datasets were provided, they were not used in the final model as they negatively impacted performance due to poor quality.

It cannot be emphasized enough that **data quality was the single most important factor** in determining the final quality of the model.

※All token counts are with the Llama 3 tokenizer

Japanese Datasets

All datasets were created using Shisa.AI's custom synthetic data pipelines:

SFT

shisa-ai/shisa-v2-sharegpt
A high-performance JA/EN dataset filtered and regenerated from the original Shisa V1 dataset

shisa-ai/rewild-set-deepseek-subset
JA translations of Rewild (WildChat) prompts with filtered responses from DeepSeek-V3-0324

shisa-ai/magpie-ultra-set
JA dataset generated based on argilla/magpie-ultra-v1.0

shisa-ai/magpie-advanced-questions-set
Advanced university-level question set generated using Magpie

shisa-ai/japan-magpie-set
Magpie-generated set covering Japanese economics, history, culture, and business practices

shisa-ai/shisa-v2-roleplaying-sft
Role-playing dataset featuring diverse characters and scenarios

shisa-ai/translation_expanded_master_set_filtered
Translation dataset encompassing various tasks, including essays, conversations, and novels

shisa-ai/shisa-v2-instruction-following-sft
Instruction-following dataset based on Aratako/Magpie-Tanuki prompts

DPO

shisa-ai/deepseekv3-ultrafeedback-armorm-dpo
princeton-nlp/gemma2-ultrafeedback-armorm DPO prompts with regenerated responses. Surprisingly, this relatively small, English-only DPO dataset outperformed our tested JA/EN mixed DPO sets and even larger ones like the Tulu 3 preference mixture.

shisa-ai/shisa-v2-roleplaying-dpo
Role-playing DPO dataset evaluated in UltraFeedback style

shisa-ai/translation-no-extra-text-dpo-dataset
DPO dataset specifically designed to control unnecessary explanatory text during translation tasks

shisa-ai/shisa-v2-instruction-following-dpo
A DPO dataset focusing on enhancing instruction-following capabilities

shisa-ai/politeness-dpo-set
Dataset created for fine-grained control over Japanese politeness and speaking styles

Dataset Quality Challenges

Sample Unused Data

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。 Ек 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。

Although Shisa.AI’s Japanese datasets were carefully curated and filtered over several months, additional language datasets and certain tourism-related datasets were contributed in the final stages of development.

When added to the mix, these new datasets had a significant negative impact on model performance and eval scores.

Upon reviewing the provided datasets, we found that much of it lacked relevance or meaningful context, and also contained numerous errors, corrupted output, and inaccuracies/hallucinations.

This highlights the difficulties in creating suitable synthetic datasets. Particularly during the post-training phase (where the model's behavior is concretely defined), it is essential to rigorously filter and meticulously review all training data.

Model Development

The setup and provisioning of the H100 Slurm cluster used for this training project were handled by our compute sponsor Ubitus.

- Primary development for all the Shisa V2 models (except the 405B) was conducted on a small-scale cluster consisting of 4 H100 nodes (Each node was an AWS P5 with 8 x NVIDIA H100 GPUs)
- While this configuration was adequate for most of our development phase, resource constraints created training bottlenecks as half the nodes were continuously occupied for (dual-purposed) evaluation and synthetic generation models
    - Unfortunately, during primary development, additional nodes were unavailable/allocated elsewhere
- When Shisa.AI ultimately took responsibility for training the 405B model, the cluster was first expanded to 16 nodes, then to 32 nodes to meet the project deadline

Our open-source development stack included:

- Linux, Python, Slurm, PyTorch, HuggingFace TRL, Axolotl, DeepSpeed, OpenRLHF, Magpie, and more

Our model development would not have been possible without the open-source software we used.

Training logs, source code, and datasets, and our development models are shared for the benefit of the research community:

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

Model Training Time

Computational resources grow exponentially as model sizes increase (8B → 70B → 405B).

	8B	70B	405B
SFT (H100 hours)	30	1,000	65,000+
DPO (H100 hours)	30	200	4,000
Eval (H100 hours)	4	20	100
SFT (min nodes)	1	4	16
DPO (min nodes)	1	8	32

※For our largest-scale models, we switched from Axolotl to OpenRLHF for stable and reliable sequence parallelism (ring attention) support, ensuring models could fit within GPU memory.

Training Challenges

Training the 405B model was extremely difficult. Only three other groups that we know of: Nous Research, Bllossom, and AI2 have published Llama 405B full fine-tunes. During our training, we corresponded with Nous Research and AI2.

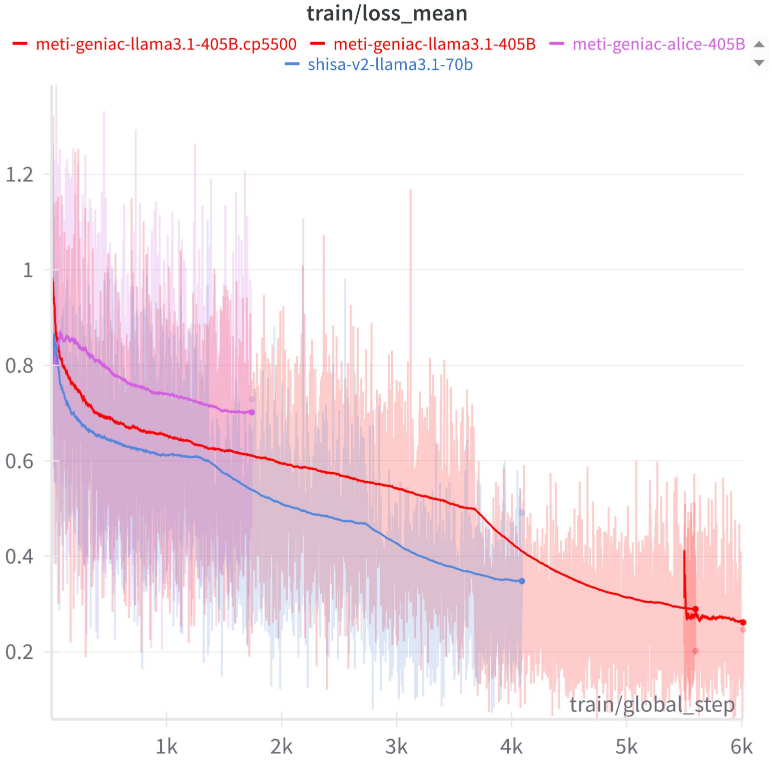
- We implemented every optimization at our disposal including: DeepSpeed ZeRO-3 parameter and activation offloading, gradient accumulation, 8-bit paged optimizer, and sequence parallelism
    - Even so, the 405B model still barely fit within the H100's memory limits
- Due to memory constraints and Axolotl limitations with sequence parallelism, we switched to OpenRLHF for our largest models (this required additional data reprocessing, but there were no other viable options for the 405B model training)
- The enormous size of the 405B model meant that model loading took several hours each time settings were changed or the run was restarted, even with our high-speed FSX network storage
- We had empirically checked the optimal learning rate scaling up to the 70B model, but for the 405B, we had to rely purely on theoretical formulas for scaling the configuration
- To conduct continuous evals on checkpoints, we allocated 2 of the 32 nodes for running evals, leaving 30 nodes (240 H100s) for training during the SFT phase

Training Challenges (continued)

Throughout the training process, we encountered numerous bugs, some specific to the 405B runs, which required significant efforts to resolve.

- Even the relatively modest 32-node cluster experienced multiple node failures, necessitating hardware replacements
- We also encountered critical bugs across almost every layer of the software stack（NVIDIA NCCL, PyTorch, TRL, DeepSpeed, Axolotl, OpenRLHF, LightEval, LiteLLM, Bespoke Curator, etc.), many of which required implementing our own fixes
- Considerable time was spent ensuring checkpoint-based restarts worked properly
    - This was crucial not only as a safeguard against hardware failures but also to mitigate persistent low-level software issues that gradually reduced training speeds for extended continuous runs
    - As this issue remained unsolved, we built a workaround to dynamically calculate optimal timing for restarts and checkpointing (accounting for slowdowns and restart times)
- DPO training required all 32 nodes (256 H100s) as attempts with only 30 nodes consistently resulted in failures. At present, without further software optimization, we do not recommend full fine-tuning of the Llama 3 405B model with fewer nodes

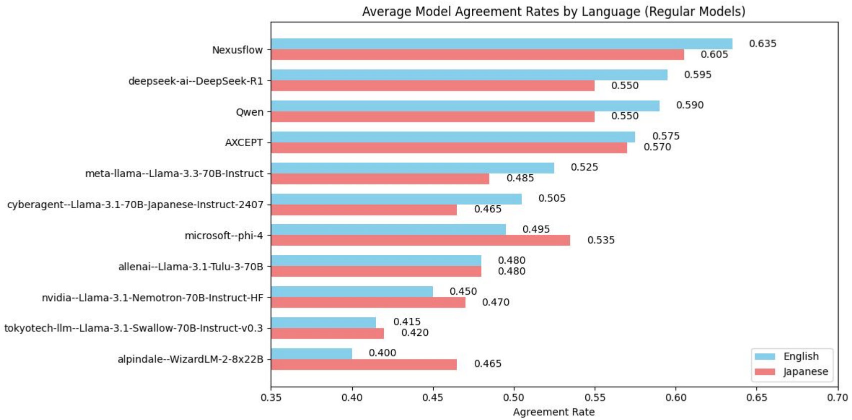
Supplementary Note on CPT Post-Training



The initial plan was to use Llama 3.1 405B Instruct as a placeholder to verify training configuration for a 405B model, after which we intended to replace it with a Japanese CPT 405B model provided by Deepreneur. However, that model was ultimately not delivered to us.

We look forward to seeing the results of Deepreneur's development.

Model Evaluation (Evals)



Because we anticipated needing to run hundreds, if not thousands of model evaluations for this project, our first priority was to implement a robust eval framework.

We first validated an LLM-as-a-Jury approach, comparing various models against both GPT-4 and human gold-standard judgements.

Ultimately, we selected three diverse, high-quality open models as our judges:

- Nexusflow Athene V2
  A Qwen 2.5 72B-based model that showed the highest correlation with GPT-4 judgements and surprisingly strong Japanese capabilities
- Tulu 3 405B (FP8)
  A fully open model based on Llama 3.1 405B (Base), with SOTA multilingual, reasoning, and alignment performance
- Llama 3.3 70B Instruct
  Meta’s most advanced fine-tuned model, achieving benchmark performance comparable to Llama 3.1 405B Instruct

※Evaluations performed with this panel maintained stable model rankings and showed less than 5% average deviation compared to our standard GPT-4 Turbo judged Shaberi results.

評価 (Evals)

We built a multieval evaluation framework to automatically test JA/EN performance after each training run.

Japanese Evals

ELYZA Tasks 100
Instruction-following tasks (writing, reasoning, summarization, etc.)

JA MT-Bench (Turn 1)
Japanese version of MT-Bench covering a wide range of conversational tasks

Rakuda
Open-ended questions on Japanese culture and history

Tengu-Bench
Diverse downstream tasks including function calling, math, politeness, ethics, etc

llm-jp-eval (v1.4.1)
Ever-growing set of Japanese NLP benchmarks

shisa-jp-ifeval
Our Japanese adaptation of IFEval for grammar and linguistic instruction following

shisa-jp-rp-bench
Multi-turn role-playing and narrative evaluation

shisa-jp-tl-bench
English-to-Japanese translation quality evaluation

English Evals

MixEval
A mix of real-world prompts and closed-form testing with (Arena correlation of 0.96)

LiveBench
Regularly-updated contamination-free tasks (math, coding, reasoning, etc)

IFEval
English-language instruction-following (keywords, formatting, word counts, etc)

EvalPlus
Rigorous code-gen evaluation (HumanEval+ and MBPP+)

New Japanese Benchmarks

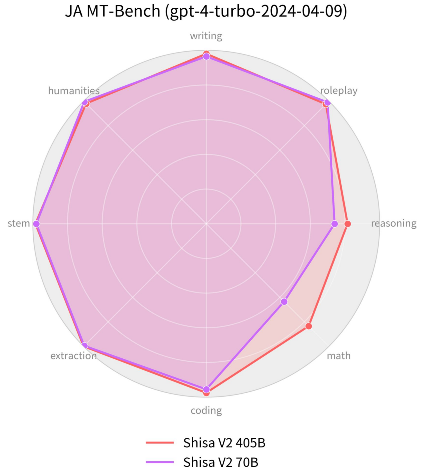
In the course of model development, we created the following new benchmarks to more accurately evaluate performance on important downstream tasks in Japanese:

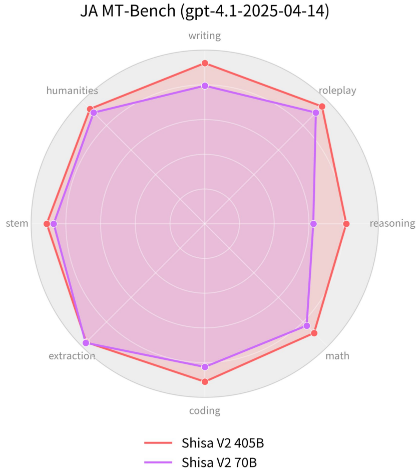
- shisa-jp-ifeval
  Inspired by IFEval, this benchmark specifically evaluates instruction-following abilities in Japanese grammar and linguistic expression (automated verification).
- shisa-jp-rp-bench
  Based on Aratako's Japanese-RP-Bench, this benchmark assesses multi-turn conversation and character adherence (LLM Judge)
- shisa-jp-tl-bench
  Evaluates translation quality between a variety of Japanese and English scenarios (uses LLM Judge for pairwise comparison with logistic transform scoring)

We believe that these benchmarks will broadly benefit the Japanese LLM research community and plan to release these as open-source in the near future.

Why use GPT-4.1 rather than GPT-4 as a Judge?

Our Shisa V2 70B and 405B models both exceed GPT-4 in Japanese performance, and we found that GPT-4 could not accurately distinguish performance differences among models stronger than itself. This phenomenon has previously been reported in the literature, but it was interesting to see it for ourselves:

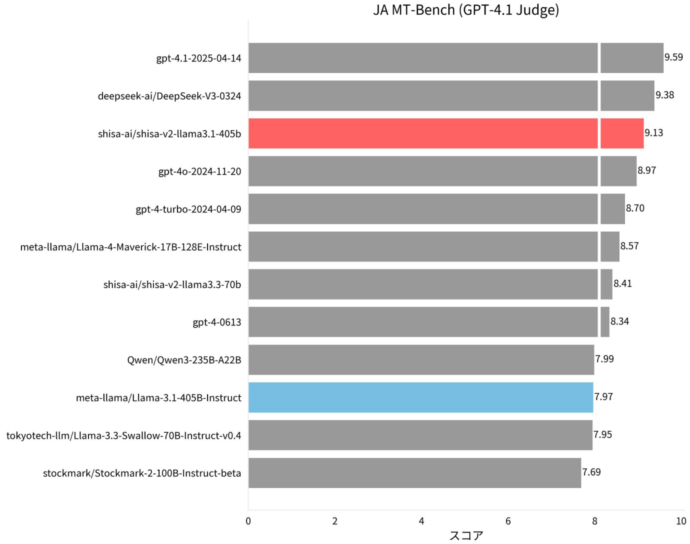

JA MT-Bench (gpt-4-turbo-2024-04-09)

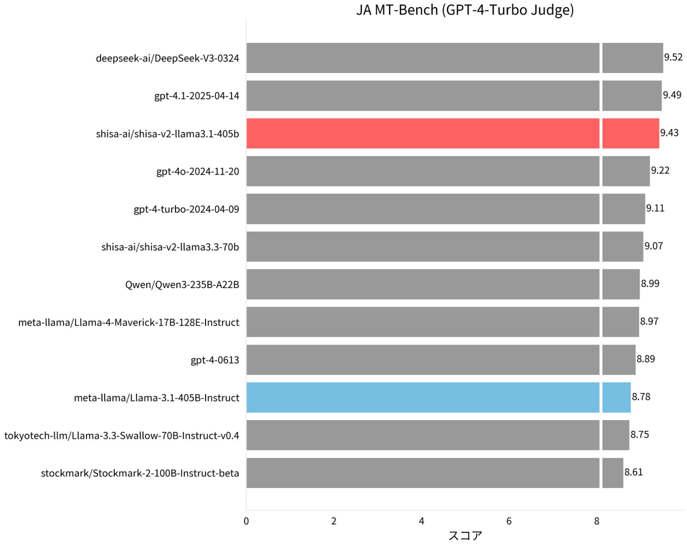

JA MT-Bench (gpt-4.1-2025-04-14)

※GPT-4.1 applies stricter evaluation criteria compared to GPT-4, resulting in somewhat lower scores. However, we believe that GPT-4.1 provides a more accurate and practically useful assessment. Nonetheless, in certain cases (such as JA MT-Bench), we also conducted evaluations using GPT-4 to allow for 1:1 comparisons with previously evaluated models.

JA MT-Bench

Our project benchmark targets were for JA MT-Bench scores excluding coding and math (9.18), but in 2025, we consider these tasks effectively solved by large language models (our 405B model shows only a 0.05 difference when coding and math is included), so here the JA MT-Bench score averages include all categories:

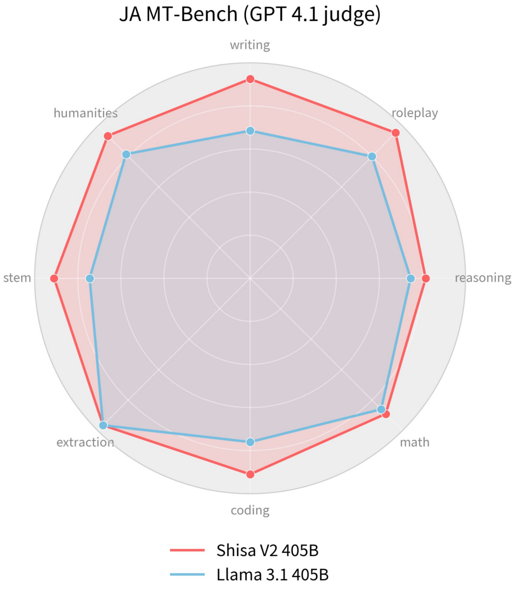


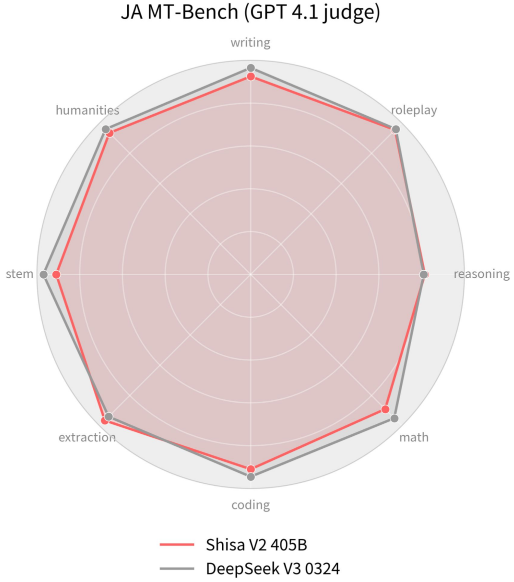


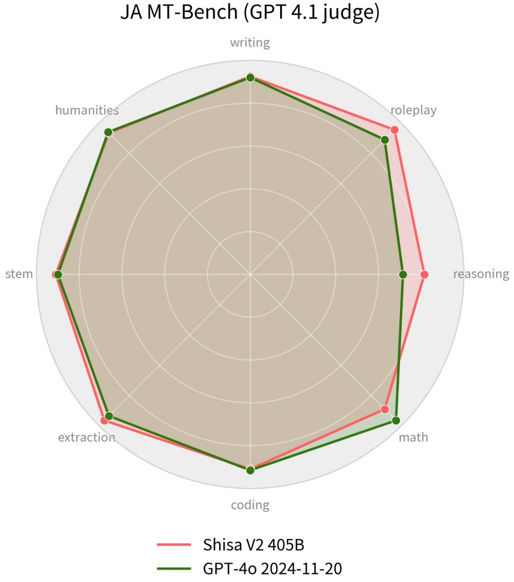
※We believe the GPT-4.1 scores to be more accurate, but also include GPT-4-Turbo evaluation results to facilitate comparisons with previously evaluated models.

JA MT-Bench (Radar Charts)

Our post-training has improved the Japanese performance of the Llama 3.1 405B Instruct base model across all evaluation categories. Shisa V2 405B’s JA MT-Bench scores are competitive with the flagship models published by both leading US and Chinese Frontier Labs.



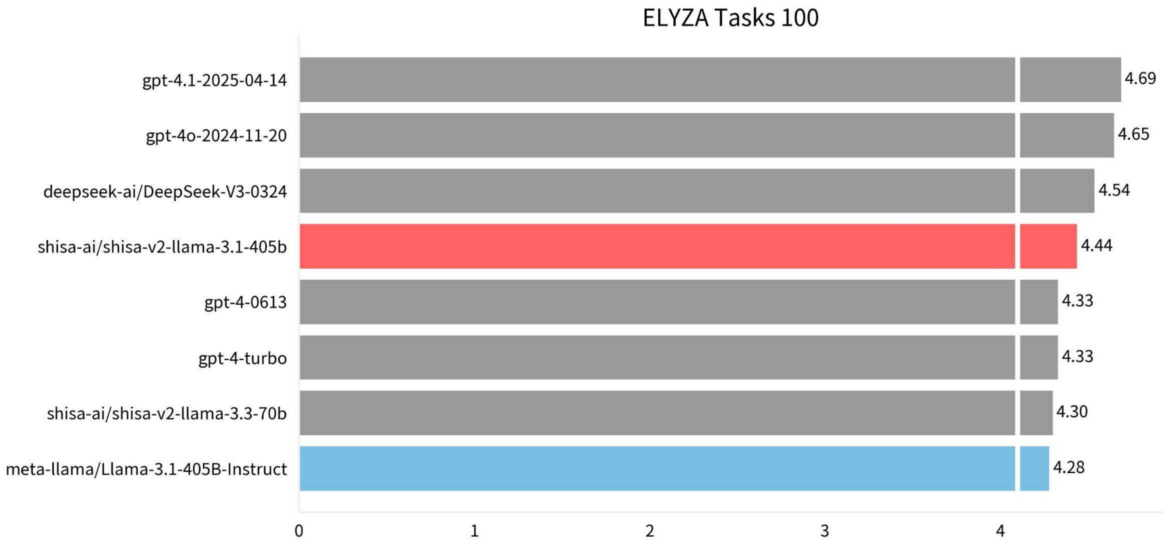




※GPT-4.1 (gpt-4.1-2025-04-14) Judge

ELYZA Tasks 100

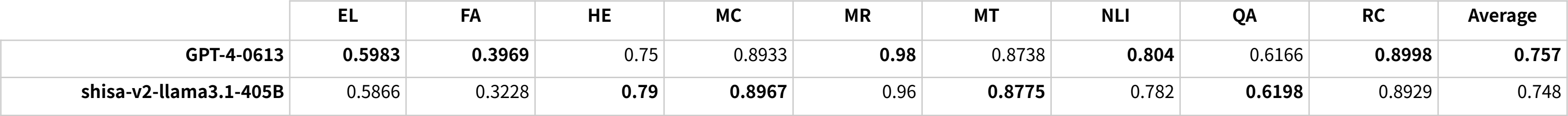
Since the base Llama 3.1 405B Instruct model had already achieved a score above 4.1 on the ELYZA Tasks 100, our 405B model easily met this target.



※GPT-4.1 (gpt-4.1-2025-04-14) Judge via Shaberi benchmark suite

llm-jp-eval

For testing, we used llm-jp-eval v1.4.1 with default settings. Given the large number of configurable options and frequent updates to the evaluation framework, it was not entirely clear which specific settings should be targeted. Here, we present a comparison with GPT-4 (0613).



	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

The llm-jp-eval README explicitly notes that its results are not necessarily a reliable indicator of model performance:

> It has become clear that models instruction-tuned using jaster can achieve very high llm-jp-eval evaluation scores, even if test data was not used for instruction tuning. Therefore, please note that obtaining a high evaluation score does not necessarily mean better performance than other LLMs.

For a prior analysis of llm-jp-eval scoring issues, see our article: https://shisa.ai/posts/llm-jp-eval-eval/

Final Deliverables

Llama 3.1 Shisa V2 405B

The highest-performing large language model developed in Japan

- Download: https://huggingface.co/shisa-ai/shisa-v2-llama3.1-405b

Core Dataset

A best-in-class synthetic dataset, freely available for use to improve the Japanese capabilities of any model
Licensed under Apache 2.0

- Download: https://huggingface.co/datasets/shisa-ai/shisa-v2-sharegpt

New Benchmarks

Created to contribute to Japan's LLM research community and scheduled for open-source release in Q3 2025