Cost of Reasoning in non-English Languages: A Case Study on Japanese

Yuu Jinnai
CyberAgent / Tokyo, Japan
jinnai_yu@cyberagent.co.jp

Models: https://huggingface.co/collections/cyberagent/cat-thinking

Abstract

Reasoning Language Models (RLMs) achieve their strongest performance when they reason in English, the language for which reasoning-oriented training data is most abundant. However, reasoning trace is a clue for model interpretability and safety, and useful in practice for both the model users and for model developers. Thus, it is desirable to be able to develop a model that reasons in a language of the user's choice, while still maintaining strong reasoning performance. To this end, we study the feasibility of training a model that reasons in Japanese. We develop a Japanese-reasoning variant of Qwen-3-Swallow-8B, which is a Japanese LLM continually pretrained from Qwen-3-8B, with GRPO and evaluate it across coding, math, and science benchmarks. The study shows that reasoning-language control is feasible by training a Japanese continually pretrained model with GRPO. However, its performance is at best on par with strong English-reasoning baselines on several benchmarks. We also evaluate the trained model on Japanese cultural benchmarks and observe that the model's performance is worse than the baseline models, suggesting that the reasoning in Japanese does not immediately improve performance on culturally relevant tasks for free.

1 Introduction

English is the most resource-rich language by far, especially for reasoning-oriented training data. Thus, most of the Reasoning Language Models (RLMs) are trained to reason in English (Yang et al., 2025; Guo et al., 2025; Team et al., 2025; Agarwal et al., 2025; Muennighoff et al., 2025). Even multilingual LLMs tend to reason best in English: across many benchmarks, including multilingual ones, reasoning in English outperforms reasoning in the question's own language (Saji et al., 2026; Yong et al., 2025).

Yet, reasoning trace is a useful signal for the model users and for model developers that would ideally be available in the user's or developers language of choice. For example, reasoning trace is useful for model interpretability (Wei Jie et al., 2024) and safety (Jiang et al., 2025; Guan et al., 2025), and can be used to improve model performance (Wei et al., 2023; Seo et al., 2026). Such reasoning trace is also useful for model developers to understand the model's reasoning process and to debug the model's behavior. Users may also prefer to see reasoning trace in their own language for better accessibility.

The question is thus how to achieve a reasoning trace in a language of one's choice with minimum degradation of the model's performance. To answer this question anecdotally, we develop a Japanese-reasoning language model as a case study. We build upon Qwen-3-Swallow-8B (Fujii et al., 2024; Ma et al., 2025), a Japanese LLM continually pretrained from Qwen-3-8B (Yang et al., 2025), and train it with supervised warm-start and two-stage GRPO to produce reasoning traces in Japanese. The resulting model achieves 100% Japanese reasoning traces on all evaluated benchmarks, including those whose instructions are entirely in English. Its performance remains competitive with the English reasoning models on most benchmarks, but falls short on several tasks, suggesting that reasoning-language control is feasible but not free of cost without further improvements to the methodology.

We observe two ingredients that were necessary for the development of our Japanese-reasoning model. First, we find that **continual pretraining on Japanese data significantly helps the model to reason in Japanese.** Qwen-3-Swallow-8B learns to reason within hundreds of GRPO steps, while Qwen-3-8B (with SFT warmstart) shows no sign of generating Japanese even after 15k GRPO steps (Table 1). Second, warm-start training with a permissive reward model can be effective to improve the training speed by mitigating the sparse reward

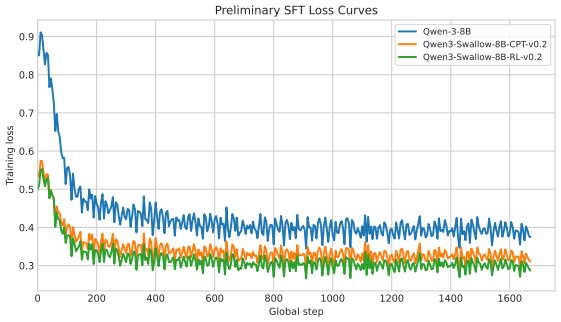

Figure 1: Loss curves of the three models during SFT on Japanese reasoning data. The loss curves show that Qwen3-Swallow-8B-RL-v0.2 have the lowest loss compared to Qwen-3-8B and Qwen3-Swallow-8B-CPT-v0.2, indicating that it is likely to be the best model for training a Japanese-reasoning model under low-resource conditions.

#Steps	#Completions	#Traces in Ja	Ratio
4,000	192	6	3.1%
15,000	192	6	3.1%

Table 1: Ratio of <think> blocks containing Japanese characters in Qwen-3-8B after 4,000 and 15,000 GRPO steps with a Japanese language reward. The model continues to reason in English throughout, with Japanese appearing only when explicitly requested by the prompt.

problem (Sutton and Barto, 2018). Because our model is much smaller and less capable than the frontier models, it achieves no reward using a strict reward model, making it impossible to further learn from the reward signal. In fact we observe that most batches have zero reward standard deviation using the strict reward model without a warm-start by a permissive reward model (Figure 2). To remedy this, we deploy a permissive reward that gives partial rewards for following the reasoning format, reasoning in Japanese, and answering correctly. The permissive reward serves as a reward shaping (Sutton and Barto, 2018) with a dense signal, providing learning signal to the model before the model fully meets all the conditions.

With these two ingredients, we are able to train a model that reasons in Japanese with competitive performance. As both continual pretraining and warm-start with a permissive reward are not specific to Japanese, we expect that they can be applied to other languages as well, yet further experiments and language resources are needed to confirm this.

2 Model Development

Our goal is to train a model that produces Japanese reasoning traces reliably. Prior work has trained models to reason multilingually by aligning the reasoning process in the target language with that of English, so that they reason in the same way (Zhu et al., 2024; Ranaldi and Pucci, 2025). While this approach is likely to be sample efficient and cost effective, our focus is different: we study how reliably reasoning-language behavior can be controlled in Japanese, what capability cost this induces, and how stable the behavior remains when reward pressure is removed.

We first conduct a preliminary experiment to evaluate which model is suitable for training a Japanese-reasoning model. We evaluate Qwen-3-8B, Qwen3-Swallow-8B-CPT-v0.2, and Qwen3-Swallow-8B-RL-v0.2. Qwen3-Swallow-8B-CPT-v0.2 is a continually pretrained model trained on a large amount of Japanese data on top of Qwen-3-8B. Qwen3-Swallow-8B-RL-v0.2 is a fine-tuned version of Qwen3-Swallow-8B-CPT-v0.2 with instruction-following and English reasoning ability. We run supervised fine-tuning (SFT) on the three models with a small amount of Japanese reasoning data and evaluate SFT loss as shown in Figure 1; Qwen3-Swallow-8B-RL-v0.2 achieves the lowest loss and is selected as the base model.

2.1 Warm Start with Supervised Learning

We start with supervised learning to train the model to follow the format of the reasoning. Concretely, we train the model to follow the following format:

```
<think>
[Reasoning in Japanese]
</think>
[Response in Japanese or English]
```

We use the math and coding subsets from tokyotech-llm/Swallow-Nemotron-Post-Training-Dataset-v1[1], one of the SFT datasets used to train Qwen3-Swallow-8B-RL-v0.2. We choose this dataset so that the model learns only the format and language of reasoning, allowing us to evaluate the effect of reinforcement learning in isolation. We translate the reasoning traces into Japanese using gpt-oss-120b (Agarwal et al., 2025) so that the model can learn Japanese reasoning-trace style and format. We deliberately halt training before

[1] https://huggingface.co/datasets/tokyotech-llm/Swallow-Nemotron-Post-Training-Dataset-v1

convergence, so that GRPO can shape the model's own reasoning style in subsequent stages rather than having it merely imitate translated English reasoning. The details of the training are described in Appendix B.

Necessity of Japanese pretraining. To test whether Japanese continual pretraining is a necessary condition for the language reward to work, we apply the same SFT warm-start and the first GRPO with permissive reward to Qwen-3-8B directly, bypassing the Swallow pretraining step. Inspection of model generations at approximately 4,000 and 15,000 GRPO training steps reveals that the model continues to reason in English throughout: Japanese text appeared in `<think>` blocks only when task instructions explicitly requested Japanese output (e.g., OR-Tools problems instructing the model to explain in Japanese), and no spontaneous Japanese reasoning was observed at either checkpoint (Table 1). This contrasts sharply with the Swallow-based model, which reaches 100% Japanese reasoning within a few hundred GRPO steps (Table 2). The result is consistent with the SFT loss comparison (Figure 1): without the Japanese substrate established by Swallow pretraining, the language reward finds no Japanese-capable region of the policy to reinforce. We conclude that Japanese continual pretraining and the GRPO language reward are jointly necessary.

2.2 Warm Start with GRPO with Permissive Dense Reward

We find that it is inefficient to train the model as-is with a strict reward that only rewards the model for following the format and answering correctly, as the model would not receive any reward and thus would not learn. To address this issue, we first use a permissive reward that gives partial rewards to the model. The reward is additive across three components: (1) a format component, awarded if the response contains `<think>` and `</think>` tags; (2) a language component, awarded if the reasoning trace is in Japanese; and (3) an answer component, awarded if the model's answer is correct.

Given that the model fails to follow the format when the task is not in math or coding, we expand the training to include several additional domains. Concretely, we use math, coding, STEM, and instruction-following tasks including rStar-Coder (Liu et al., 2026)[2], AReaL-boba-2-RL-Code (Fu et al., 2026)[3], Big-Math-RL-Verified (Albalak et al., 2025)[4], and DAPO-Math (Yu et al., 2026)[5]. In addition we synthesize Japanese counterparts of the tasks using gpt-oss-120b so that the model also learns with Japanese instructions.

Permissive reward improves learning signal. By using the permissive reward, we observe that most batches have some variation of rewards, which allows the model to learn and improve its generation. Figure 2(a) quantifies the necessity of the SFT warm-start: without it, the fraction of batches with zero reward standard deviation (*frac_reward_zero_std*) averages 0.58 and peaks at 0.78 from the very first steps, indicating that GRPO cannot extract any learning signal. With the SFT warm-start this rate stays below 0.07 throughout training, confirming that the warm-start is necessary for GRPO to function in this setting. Figure 2(b) shows the training reward across both GRPO stages (permissive then strict); the strict stage begins with negative reward because the harder reward function rejects outputs that the permissive reward had accepted, then recovers as the model adapts.

2.3 GRPO with Strict Sparse Reward

After a few iterations of training with the permissive reward, we find the model to be able to follow the format but not consistently answer the question correctly. We therefore apply a strict reward that requires both precise formatting (a `\boxed{}` answer) and correctness. In addition to the dataset used in the first stage of the GRPO, we also sample instructions from extraction-wiki-ja[6] and llm-jp/magpie[7] (LLM-jp et al., 2024) so that the model won't overfit to math and coding tasks.

3 Evaluation

We evaluate two models trained with the procedure of Section 2: **CAT-Thinking**, optimized for single-turn reasoning, and **CAT-Paws**, additionally trained on multi-turn coding, terminal interaction,

[2] https://huggingface.co/datasets/microsoft/rStar-Coder

[3] https://huggingface.co/datasets/inclusionAI/AReaL-boba-2-RL-Code

[4] https://huggingface.co/datasets/SynthLabsAI/Big-Math-RL-Verified

[5] https://huggingface.co/datasets/BytedTsinghua-SIA/DAPO-Math

[6] https://huggingface.co/datasets/llm-jp/extraction-wiki-ja

[7] https://huggingface.co/datasets/llm-jp/magpie-sft-v1.0

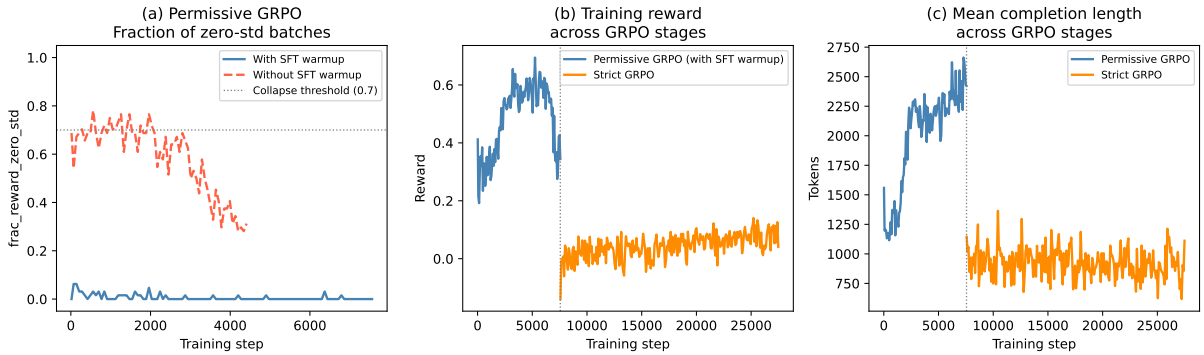

Figure 2: Training dynamics of the two GRPO stages. (a) Fraction of batches with zero reward standard deviation (*frac_reward_zero_std*) during permissive GRPO with and without SFT warm-start. Without warm-start the fraction exceeds 0.7 immediately and training yields no learning signal; with warm-start it remains below 0.07 throughout. (b) Training reward across both GRPO stages (permissive then strict, separated by the dashed line); the strict reward starts negative because its harder criterion rejects outputs the permissive reward had accepted. (c) Mean completion length across the two stages.

Benchmark	CAT-Thinking	Swallow	Qwen-3
	English		
PolyMath (En)	100	0	0
AIME 2026	100	0	0
HumanEval	100	25	0
MBPP	100	0	0
GPQA Main (En)	100	0	0
	Japanese		
PolyMath (Ja)	100	53	33
JHumanEval	100	61	43
GPQA Main (Ja)	100	73	34

Table 2: Fraction of reasoning traces in Japanese, measured as `<think>` blocks containing hiragana (%). CAT-Thinking reaches 100% on every benchmark including English-instruction tasks.

and tool-use tasks for agentic use. We focus here on single-turn math and coding ability for both models; agentic evaluation of CAT-Paws is reported in Appendix A. Throughout, we compare against the base model Qwen-3-8B reasoning in English (*Qwen-3*) and the continually pretrained Qwen-3-Swallow-8B-RL-v0.2 (*Swallow*). Unless noted otherwise, all runs use random sampling (temperature $= 0.8$, top-p $= 0.95$, maximum of 4096 new tokens). Some of the datasets we used for training were released after the benchmarks, so there may be some contamination. The absolute scores should be read as reference points rather than as held-out measurements.

3.1 Reasoning Languages

We first evaluate if the models reason in Japanese. Table 2 reports the fraction of `<think>` blocks with Japanese. We use a heuristic to measure Japanese reasoning as a text containing at least one hiragana character. CAT-Thinking achieves 100% Japanese reasoning on every benchmark, including those whose instructions are entirely in English (HumanEval, GPQA Main (En), MBPP). As shown in Appendix D, CAT-Thinking generates the reasoning trace fully in Japanese instead of partially in Japanese. By contrast, Swallow and Qwen-3 show mixed behavior: both produce little or no hiragana on English benchmarks, while on Japanese benchmarks they still fall short of consistent Japanese reasoning (e.g., GPQA Main (Ja): 73% for Swallow, 34% for Qwen-3), consistent with prior work documenting that multilingual models default to English (or Chinese) for intermediate reasoning steps (Wang et al., 2025).

3.2 Math and Coding Tasks

The question is if the reasoning in Japanese comes at a cost to performance, and if so, how large that cost is. We evaluate on coding with MBPP (Austin et al., 2021), HumanEval (Chen et al., 2021), its Japanese translation JHumanEval[8], and LiveCodeBench v6 (Jain et al., 2025), and on math and science using GPQA Main (Rein et al., 2024) and PolyMath (Wang et al., 2026), each in English and Japanese, and AIME 2026 (Dekoninck et al., 2026). Figure 3 reports the results.

The results suggest that it is feasible to enforce Japanese reasoning traces while largely retaining performance on reasoning tasks. CAT-Thinking is competitive with the other models in English tasks,

[8] https://github.com/KuramitsuLab/jhuman-eval

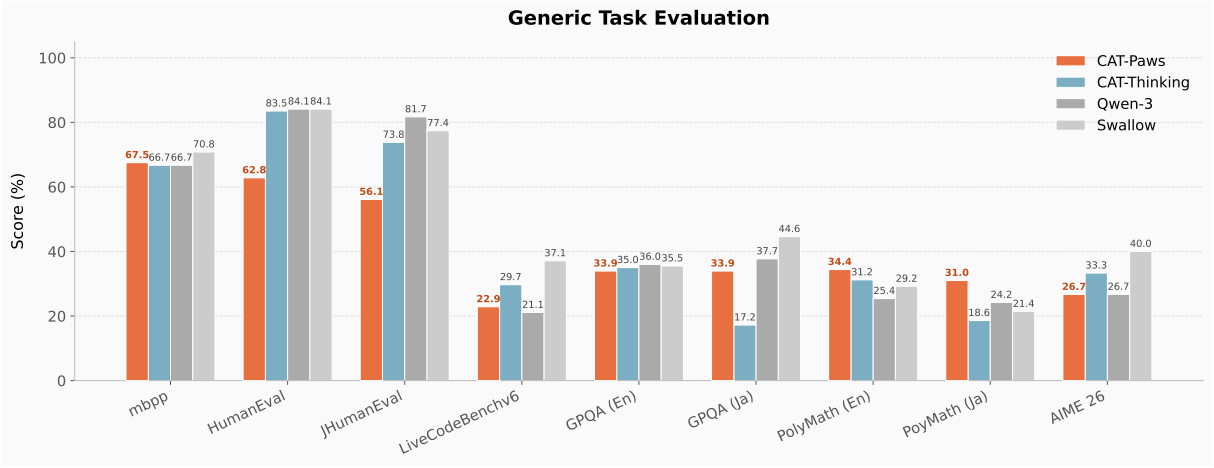

Figure 3: Evaluation on single-turn coding and math benchmarks. We compare our Japanese-reasoning models, CAT-Thinking and CAT-Paws, against the base models Qwen-3-8B (reasoning in English) and Qwen-3-Swallow-8B-RL. For every instance, CAT-Thinking and CAT-Paws generate their reasoning trace in Japanese, whereas the comparison models reason in English. CAT-Thinking remains competitive with English reasoning on most English benchmarks but trails on the Japanese benchmarks (GPQA (Ja), PolyMath (Ja)); CAT-Paws, optimized for multi-turn agentic use, is weaker across single-turn tasks.

Benchmark	CAT-Thinking	Swallow	Qwen-3-8B
JCM	86.7	**88.3**	70.8
ECM (JA translated)	72.2	**75.9**	72.9
Jubaku	85.8	**97.9**	92.9

Table 3: Cross-benchmark cultural evaluation summary (%). The scores of CAT-Thinking are worse than Swallow in all three benchmarks.

but falls short on several Japanese tasks. We speculate this gap is due to the training mixture. Most reasoning tasks used during SFT and GRPO are in English, even though the reasoning traces are in Japanese (Section 2). This is due to the lack of Japanese resources in the math and coding domains, which is a known limitation for developing multilingual LLMs (Yong et al., 2025). This motivates future work to expand Japanese training data for reasoning tasks, and to explore whether the gap can be closed by training on more Japanese reasoning tasks.

3.3 Cultural Benchmarks

Intuitively, one would expect that reasoning in Japanese would improve performance on tasks that are culturally specific to Japan. Prior work has shown that multilingual models can exhibit cultural bias, and that reasoning in a target language can improve performance on culturally specific tasks (Yong et al., 2025). On the other hand, Ri et al. (2026) show that reasoning tokens do not matter in terms of safety judgments, which suggests that reasoning-language control may not improve performance on culturally specific safety tasks. CAT-Thinking is trained for math, coding, and logical reasoning tasks to improve the model's reasoning ability, but not for cultural or social judgment tasks. We evaluate the effect of reasoning-language control to the model's performance on cultural benchmarks, specifically on the commonsense morality and cultural-bias tasks.

We use the Japanese Commonsense Morality benchmark (JCM; Takeshita et al. (2023))[9], a true/false dataset of 3,992 moral scenarios drawn from everyday situations in Japan. To single out the effect of Japanese language and Japanese culture, we also evaluate on commonsense morality subset of ETHICS dataset (Hendrycks et al., 2021) translated to Japanese using gpt-oss-120b, which we call ECM-JA. ETHICS is collected from English speakers from United States, Canada, and Great Britain.

Table 3 summarizes the cross-benchmark comparison. CAT-Thinking scores lower than Swallow in both JCM and ECM-JA. Because degradation appears on both Japanese and translated-English cultural evaluations, it is more likely that the performance drop is due to a general degradation of social/cultural judgment rather than a Japanese-specific effect. In fact the training dataset for CAT-Thinking is heavily skewed toward math, coding,

[9] https://github.com/Language-Media-Lab/commonsense-moral-ja

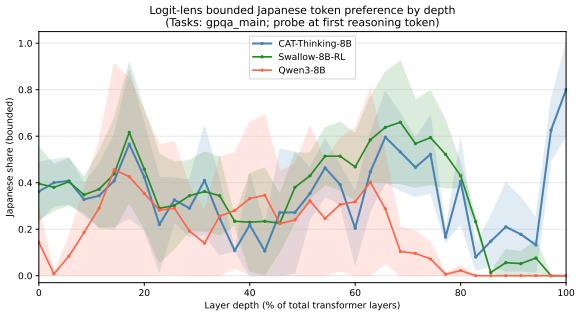

Figure 4: Layerwise logit-lens Japanese-token preference on GPQA Main. CAT-Thinking and Swallow have similar global area, but CAT is substantially higher in the final layers.

and logical reasoning tasks, which may have caused the model to forget its social/cultural judgment ability. The result suggests that it would be safe to include cultural and safety datasets in the reasoning training because the model's performance on cultural and safety tasks may degrade by not including them enough in the training dataset.

3.4 Layerwise Visualization by Logit-Lens

Is the reasoning language control we observe a surface-level effect, or does it reflect deeper changes in the model's internal representations? To study this question, we use the logit-lens technique (nostalgebraist, 2020), which projects hidden states at each layer through the model's unembedding matrix to produce a distribution over the vocabulary. Using logit-lens, we can measure if the model's representations at each layer are leaning toward English or *non-English* tokens. In this way we may visualize what change makes the reasoning model to reason in Japanese. Note that logit-lens is a probing tool and should be read as suggestive rather than causal evidence. Also note that we cannot distinguish between Japanese and Chinese at the token level, because both languages share many characters in CJK scripts; we therefore report non-English vs. English tokens.

We run logit-lens across four datasets (GPQA Main, GPQA Ja, AIME 2026, HumanEval), $k = 5$, with 50 prompts per task (30 for AIME) for Qwen-3, Swallow, and CAT-Thinking. We visualize the logits of the first token of the reasoning trace, which is the first token after `<think>`. Figure 4 shows the result on GPQA Main (En). See Appendix C for the visualization of the other datasets. Qwen-3 shifts to English earlier in the model compared to Swallow and CAT-Thinking. Swallow and CAT-Thinking

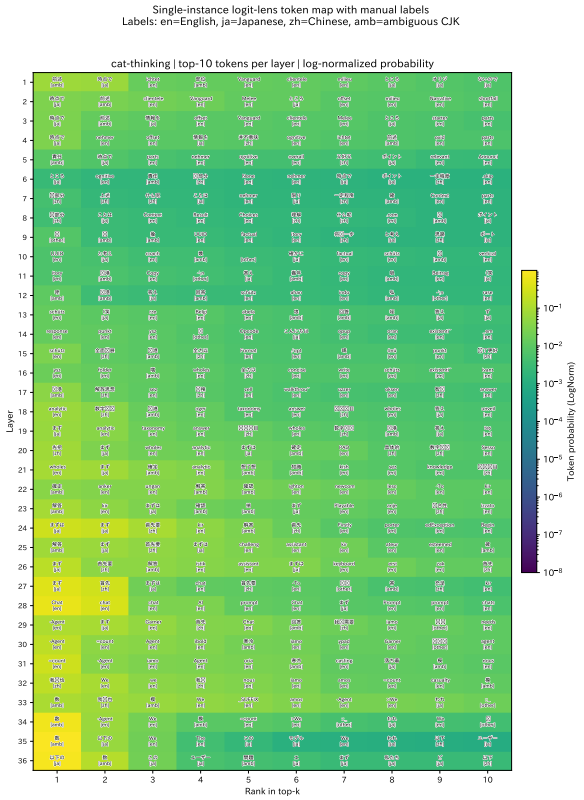

Figure 5: Visualization of the top-k tokens for CAT-Thinking on the first instance of the GPQA Main (En). Japanese tokens appear to have high probability from the early layers.

have similar curves in most of the layers except the last few layers, where CAT-Thinking shifts to non-English tokens whereas Swallow shifts to English tokens. This suggests that the Japanese reasoning traces are not a surface-level effect, but happens relatively shallow layers in the model's computation. We manually inspect the tokens and find that the tokens by CAT-Thinking seem to be mostly Japanese from the early layers (Figure 5) unlike Qwen-3 and Swallow which show a lot of Chinese tokens (Figure 9 in Appendix C).

More fine grained study would be needed to understand the mechanism of reasoning-language control. Especially for CJK scripts, it is difficult to distinguish between Japanese and Chinese at the token level, so we cannot quantitatively conclude that the model is reasoning in Japanese rather than Chinese.

4 Qualitative Analysis

The following analysis characterises CAT-Thinking specifically and *should not be interpreted as a general account of Japanese reasoning or of non-*

Representative opening sentences
済みの問題を解くことは許可されていますか？
済みの数学問題かを確認する必要がありますか？
済みの数学的問題かを確認する必要がありますか？
済みの問題を解くことは許可されているか確認します。
済みの問題を解くことは許可されたカテゴリに含まれていますか？

Table 4: Selected variants of the permission-phrase artifact. The model inserts one of these sentences (or a close variant) at the start of 84.8% of its reasoning traces.

English reasoning models. As one of the few available models explicitly optimized for Japanese reasoning traces, we believe documenting its behaviours provides a useful reference for future work.

4.1 An Emergent RL Artifact: The Permission Phrase

In 84.8% of non-empty reasoning traces (643 of 758 across GPQA, AIME, HumanEval, and MBPP), CAT-Thinking opens its ⟨think⟩ block with a permission-seeking sentence of the form *"[Question type]*を解くことは許可されていますか？*"* ("Am I permitted to solve this [question type]?"). The rate is near-universal on reasoning-heavy benchmarks (GPQA 99.6%, AIME 100%, MBPP 100%) but falls to 31% on HumanEval, which is consistent with the phrase having been reinforced primarily on math and STEM training examples.

Table 4 shows representative variants. The phrase is not fixed: the model inserts a description of the problem type ("数学的問題", "計算問題", "数学解答問題", etc.) and occasionally frames the question as a category check ("許可されたカテゴリに含まれていますか？"). Despite the surface variation, all instances follow the same schema.

We attribute this to the format reward used during permissive GRPO: a minimum reasoning-trace length of 10 tokens was required for a non-zero reward. We hypothesise that the model discovered the permission-question as a low-cost preamble that reliably satisfies the length threshold, and converged on it. The phenomenon shows how RL can produce stylistically distinctive artifacts that have no counterpart in English-trained models.

4.2 Generation Example

Table 5 shows the reasoning traces of CAT-Thinking and Qwen-3 on a matched GPQA biology question (both models answered correctly). CAT-Thinking's trace is markedly shorter (1,646 vs. 11,963 characters) and reasons by posing a sequence of verification questions to itself, reaching the conclusion without explicit step numbering. We speculate this is because we set the maximum number of output tokens to be 4,096 and penalized the model for exceeding this limit during training, which encourages the model to reason more concisely. Qwen-3 opens with "Okay, let's try to figure out…", proceeds through explicit enumeration of options, and engages in overt backtracking. See Appendix D for additional examples.

5 Related Work

Redirecting the language of reasoning. The line of work closest to ours trains models to reason in a chosen non-English language. Garcia et al. (2026) fine-tune a model to reason in Portuguese via supervised learning, and Gurgurov et al. (2026) combine supervised fine-tuning with reinforcement learning to shift reasoning into German, French, Italian, and Spanish, reporting little performance loss. Reasoning-focused systems such as Magistral (Mistral-AI et al., 2025) and MAI-Thinking-1 (Team, 2026) also use a language-consistency reward, though there the goal is to suppress code-switching rather than to install a specific reasoning language.

Language-agnostic reasoning. A complementary view holds that reasoning should be disentangled from any particular language: Zhao et al. (2025) improve multilingual reasoning by removing language-specific components, and Ranaldi and Pucci (2025) train language-agnostic reasoning trajectories. Our goal is complementary: we study the practical cost and stability of enforcing a specific reasoning language at inference time, which makes these approaches a useful point of contrast.

Does the language of reasoning matter? A growing body of analysis examines how the choice of reasoning language affects models. Sprague et al. (2025) find that explicit reasoning helps mainly on math and symbolic tasks, and work on language mixing documents that RLMs spontaneously fall back to English (or Chinese) during intermediate steps (Wang et al., 2025), a phenomenon consistent with the reversion we observe.

CAT-Thinking (1,646 chars, correct)	**Qwen-3** (11,963 chars, correct)
合成されたDNAはクロマチン構造を保持するための固定剤としての作用が不十分である 可能性がありますか？質問は、PFA単独の固定化サンプルでは見つかったが、PFA＋DSG 固定化サンプルでは見つからなくなったChIPピークがどこに最も多く存在するかというものです。考えられる理由としては、DSGがタンパク質間相互作用を切断したり、クロマチン構造を変化させることが考えられますが、特に活性化されたプロモーター・エンハンサー領域はタンパク質結合が頻繁で、固定が不十分だと消失しやすいのかもしれません。しかし質問は「最も多く見つかる場所」を尋ねているようです。[...]	Okay, let's try to figure out this ChIP-seq question. So, the question is about why some ChIP peaks disappear when using PFA+DSG fixation compared to just PFA fixation. The antibody is for IKAROS, a transcription factor in human B cells. The options are about where these disappearing peaks are located. First, I need to recall what PFA and DSG fixation do. PFA (paraformaldehyde) is a common fixative that crosslinks proteins to DNA. DSG [...]

Table 5: Reasoning traces for a GPQA biology question (both models correct). CAT-Thinking reasons via a sequence of questions to itself; Qwen-3 uses explicit step enumeration and overt backtracking. CAT-Thinking's trace is 7× shorter.

6 Conclusions

We presented an empirical case study of reasoning-language control in Japanese with an 8B model. We find that a Japanese reasoning model is feasible, and two key ingredients make it work in our experiment: (1) Japanese continual pretraining to establish a Japanese substrate, and (2) GRPO with a language reward to steer the model's reasoning traces into Japanese.

Still the question is if it is worth the cost to enforce Japanese reasoning traces. On raw capability, the answer is currently no: a model that reasons in English remains stronger on our benchmarks. The value of Japanese reasoning therefore lies elsewhere in potentially reducing English-anchoring for sovereignty-oriented development (Takase and Honda, 2026), and in producing reasoning that the Japanese-speaking community can read and audit directly. We caution that reasoning trace does not reflect the computation that produced the answer, and its interpretability is still an open question (Lanham et al., 2023).

More broadly, our results temper the expectation that reasoning-language redirection is uniformly near-free beyond closely related European languages. For typologically distant languages, language control appears to require persistent optimization pressure and careful objective design. We also highlight two evaluation lessons for the community: parser choices can change benchmark conclusions, and CJK token-level language attribution is often ambiguous and should be reported with explicit uncertainty (e.g., bounds or non-English vs. English splits). Our cross-benchmark cultural results (JCM, ECM-JA, Jubaku) further indicate that reasoning-language control alone does not improve cultural alignment and may degrade it without explicit alignment/culture supervision during post-training.

We hope this work provides a practical baseline for future multilingual reasoning research at larger scales and across more languages.

7 Limitations

Our study has several limitations.

First, we study only one target language (Japanese) and one model family at 8B scale. Although Japanese is typologically distant from English, it is still relatively high-resource; the same training recipe may fail or require substantially more data for genuinely low-resource languages.

Second, our conclusions are based on one training pipeline and a limited set of ablations. We do not isolate which ingredients are most responsible for the observed capability gap and reversion dynamics (e.g., reward design, instruction language mix, continual pretraining data, or model scale).

Third, our evaluation suite is broad but not exhaustive. Today there are many applications for LLMs beyond what we evaluate here, and the performance gap may be larger or smaller on other tasks.

References

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. 2024. On-policy distillation of language models: Learning from self-generated mistakes. In *The Twelfth International Conference on Learning Representations*.

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K.

Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, Kai Chen, and 106 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Alon Albalak, Duy Phung, Nathan Lile, Rafael Rafailov, Kanishk Gandhi, Louis Castricato, Anikait Singh, Chase Blagden, Violet Xiang, Dakota Mahan, and Nick Haber. 2025. Big-math: A large-scale, high-quality math dataset for reinforcement learning in language models. *arXiv preprint arXiv:2502.17387*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik R Narasimhan. 2026. τ^2-bench: Evaluating conversational agents in a dual-control environment. In *Forty-third International Conference on Machine Learning*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Jasper Dekoninck, Nikola Jovanović, Tim Gehrunger, Kári Rögnvaldsson, Ivo Petrov, Chenhao Sun, and Martin Vechev. 2026. Beyond benchmarks: Matharena as an evaluation platform for mathematics with llms. *arXiv preprint arXiv:2605.00674*.

Wei Fu, Jiaxuan Gao, Xujie Shen, Chen Zhu, Zhiyu Mei, Chuyi He, Shusheng Xu, Guo Wei, Jun Mei, WANG JIASHU, Tongkai Yang, Binhang Yuan, and Yi Wu. 2026. AREAL: A large-scale asynchronous reinforcement learning system for language reasoning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Kazuki Fujii, Taishi Nakamura, Mengsay Loem, Hiroki Iida, Masanari Ohi, Kakeru Hattori, Hirai Shota, Sakae Mizuki, Rio Yokota, and Naoaki Okazaki. 2024. Continual pre-training for cross-lingual LLM adaptation: Enhancing japanese language capabilities. In *First Conference on Language Modeling*.

Gabriel Lino Garcia, André da F. Schuck, João R. R. Manesco, Pedro Henrique Paiola, Leandro A. Passos, and João Paulo Papa. 2026. Think Portuguese with bode reasoning. In *Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026) - Vol. 1*, pages 953–958, Salvador, Brazil. Association for Computational Linguistics.

Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. 2025. Deliberative alignment: Reasoning enables safer language models. *arXiv preprint arXiv:2412.16339*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning. *arXiv preprint arXiv:2501.12948*.

Daniil Gurgurov, Tom Röhr, Sebastian von Rohrscheidt, Josef van Genabith, Alexander Löser, and Simon Ostermann. 2026. ReasonXL: Shifting LLM reasoning language without sacrificing performance. *arXiv preprint arXiv:2604.12378*.

Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. 2021. Aligning {ai} with shared human values. In *International Conference on Learning Representations*.

Jonas Hübotter, Frederike Lübeck, Lejs Deen Behric, Anton Baumann, Marco Bagatella, Daniel Marta, Ido Hakimi, Idan Shenfeld, Thomas Kleine Buening, Carlos Guestrin, and Andreas Krause. 2026. Reinforcement learning via self-distillation. In *Forty-third International Conference on Machine Learning*.

Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2025. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *The Thirteenth International Conference on Learning Representations*.

Fengqing Jiang, Zhangchen Xu, Yuetai Li, Luyao Niu, Zhen Xiang, Bo Li, Bill Yuchen Lin, and Radha Poovendran. 2025. SafeChain: Safety of language models with long chain-of-thought reasoning capabilities. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 23303–23320, Vienna, Austria. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton

Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, and 11 others. 2023. Measuring faithfulness in chain-of-thought reasoning. *arXiv preprint arXiv:2307.13702*.

Yifei Liu, Li Lyna Zhang, Yi Zhu, bingcheng dong, Xudong Zhou, Ning Shang, Fan Yang, Cheng Li, and Mao Yang. 2026. rstar-coder: Scaling competitive code reasoning with a large-scale verified dataset. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

LLM-jp, :, Akiko Aizawa, Eiji Aramaki, Bowen Chen, Fei Cheng, Hiroyuki Deguchi, Rintaro Enomoto, Kazuki Fujii, Kensuke Fukumoto, Takuya Fukushima, Namgi Han, Yuto Harada, Chikara Hashimoto, Tatsuya Hiraoka, Shohei Hisada, Sosuke Hosokawa, Lu Jie, Keisuke Kamata, and 64 others. 2024. Llm-jp: A cross-organizational project for the research and development of fully open japanese llms. *arXiv preprint arXiv:2407.03963*.

Youmi Ma, Sakae Mizuki, Kazuki Fujii, Taishi Nakamura, Masanari Ohi, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Koki Maeda, Kakeru Hattori, Takumi Okamoto, Shigeki Ishida, Rio Yokota, Hiroya Takamura, and Naoaki Okazaki. 2025. Building instruction-tuning datasets from human-written instructions with open-weight large language models. In *Second Conference on Language Modeling*.

Mistral-AI, :, Abhinav Rastogi, Albert Q. Jiang, Andy Lo, Gabrielle Berrada, Guillaume Lample, Jason Rute, Joep Barmentlo, Karmesh Yadav, Kartik Khandelwal, Khyathi Raghavi Chandu, Léonard Blier, Lucile Saulnier, Matthieu Dinot, Maxime Darrin, Neha Gupta, Roman Soletskyi, Sagar Vaze, and 82 others. 2025. Magistral. *arXiv preprint arXiv:2506.10910*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 20275–20321, Suzhou, China. Association for Computational Linguistics.

nostalgebraist. 2020. interpreting gpt: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`.

Qwen Team. 2026. Qwen3.6-27B: Flagship-level coding in a 27B dense model.

Leonardo Ranaldi and Giulia Pucci. 2025. Multilingual reasoning via self-training. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11566–11582, Albuquerque, New Mexico. Association for Computational Linguistics.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2024. GPQA: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*.

Narutatsu Ri, Abhishek Panigrahi, and Sanjeev Arora. 2026. Do thinking tokens help with safety? *arXiv preprint arXiv:2606.25013*.

Alan Saji, Raj Dabre, Anoop Kunchukuttan, and Ratish Puduppully. 2026. The reasoning lingua franca: A double-edged sword for multilingual AI. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 329–344, Rabat, Morocco. Association for Computational Linguistics.

Wonduk Seo, Wonseok Choi, Junseo Koh, Juhyeon Lee, Hyunjin An, Minhyeong Yu, Jian Park, Qingshan Zhou, Seunghyun Lee, and Yi Bu. 2026. Toward culturally aligned llms through ontology-guided multi-agent reasoning. *arXiv preprint arXiv:2601.21700*.

Idan Shenfeld, Mehul Damani, Jonas Hübotter, and Pulkit Agrawal. 2026. Self-distillation enables continual learning. In *Forty-third International Conference on Machine Learning*.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

Sho Takase and Ukyo Honda. 2026. Toward LLMs beyond English-centric development. *arXiv preprint arXiv:2506.10910*.

Masashi Takeshita, Rafal Rzepka, and Kenji Araki. 2023. Jcommonsensemorality: Japanese dataset for evaluating commonsense morality understanding. In *In Proceedings of The Twenty Nineth Annual Meeting of The Association for Natural Language Processing (NLP2023)*, pages 357–362. In Japanese.

GLM Team, Aohan Zeng, Xin Lv, Qinkai Zheng, Zhenyu Hou, Bin Chen, Chengxing Xie, Cunxiang Wang, Da Yin, Hao Zeng, Jiajie Zhang, Kedong Wang, Lucen Zhong, Mingdao Liu, Rui Lu, Shulin Cao, Xiaohan Zhang, Xuancheng Huang, Yao Wei, and 152 others. 2025. Glm-4.5: Agentic, reasoning,

and coding (arc) foundation models. *arXiv preprint arXiv:2508.06471*.

The Microsoft AI Team. 2026. Mai-thinking-1: Building a hill-climbing machine. Technical report, Microsoft AI.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, Nathan Lambert, Shengyi Huang, Kashif Rasul, and Quentin Gallouédec. 2020. TRL: Transformers Reinforcement Learning.

Weixuan Wang, Minghao Wu, Barry Haddow, and Alexandra Birch. 2025. Demystifying multilingual reasoning in process reward modeling. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 9775–9788, Suzhou, China. Association for Computational Linguistics.

Yiming Wang, Pei Zhang, Jialong Tang, Hao-Ran Wei, Baosong Yang, Rui Wang, Chenshu Sun, Feitong Sun, Jiran Zhang, Junxuan Wu, Qiqian Cang, Yichang Zhang, Fei Huang, Junyang Lin, Fei Huang, and Jingren Zhou. 2026. Polymath: Evaluating mathematical reasoning in multilingual contexts. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Yeo Wei Jie, Ranjan Satapathy, Rick Goh, and Erik Cambria. 2024. How interpretable are reasoning explanations from prompting large language models? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2148–2164, Mexico City, Mexico. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zheng-Xin Yong, M. Farid Adilazuarda, Jonibek Mansurov, Ruochen Zhang, Niklas Muennighoff, Carsten Eickhoff, Genta Indra Winata, Julia Kreutzer, Stephen H. Bach, and Alham Fikri Aji. 2025. Crosslingual reasoning through test-time scaling. *arXiv preprint arXiv:2505.05408*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, YuYue, Weinan Dai, Tiantian Fan, Gaohong Liu, Juncai Liu, LingJun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, and 17 others. 2026. DAPO: An open-source LLM reinforcement learning system at scale. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Weixiang Zhao, Jiahe Guo, Yang Deng, Tongtong Wu, Wenxuan Zhang, Yulin Hu, Xingyu Sui, Yanyan Zhao, Wanxiang Che, Bing Qin, Tat-Seng Chua, and Ting Liu. 2025. When less language is more: Language-reasoning disentanglement makes LLMs better multilingual reasoners. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Yanli Zhao, Andrew Gu, Rohan Varma, Liang Luo, Chien-Chin Huang, Min Xu, Less Wright, Hamid Shojanazeri, Myle Ott, Sam Shleifer, Alban Desmaison, Can Balioglu, Pritam Damania, Bernard Nguyen, Geeta Chauhan, Yuchen Hao, Ajit Mathews, and Shen Li. 2023. Pytorch fsdp: Experiences on scaling fully sharded data parallel. *arXiv preprint arXiv:2304.11277*.

Wenhao Zhu, Shujian Huang, Fei Yuan, Cheng Chen, Jiajun Chen, and Alexandra Birch. 2024. The power of question alignment in multilingual reasoning: Broadened scope and deepened insights. *arXiv preprint arXiv:2405.01345*.

A Agentic Evaluation

In addition to the math and coding datasets used for CAT-Thinking, we also use tool usage datasets and multi-turn datasets to improve the model's agentic capability.

We find the model to struggle to follow the format of the tool call in addition to reason in Japanese at the same time even with the warm-start pipeline used for CAT-Thinking. We speculate that because there is little to no tool usage dataset in Japanese at the time of writing, it is difficult for the model to generate a sequence of tokens to reason in Japanese in a way it leads to using a tool correctly. Concretely, we want a tool calling SFT dataset that contains a sequence of reasoning in Japanese and also somehow on-policy. Inspired by the self-distillation approaches (Agarwal et al.,

2024; Hübotter et al., 2026; Shenfeld et al., 2026), we synthesize a SFT dataset by prompting the developing model but with a system prompt that gives the correct answer of the task. The following is an example of the prompt used to synthesize the SFT dataset for tool usage:[10]

```
CONTEXT: ...

USER: Fetch news from the Guardian about
↪  technology in the UK.

SYSTEM: The answer is [{"name": "news",
↪  "arguments": {"category": "technology",
↪  "country": "gb", "domain":
↪  "theguardian.com"}}]. Please take into
↪  account of the given context and
↪  summarize it yourself. Then, respond
↪  based on your own judgment.
```

In this way, the model generates a Japanese reasoning trace that leads to the correct tool call with higher probability. Interestingly, we find that the model does not mention the fact that it is given the answer in the system prompt, and instead it generates a reasoning trace that looks like it is reasoning to the answer. This unfaithful reasoning is a known phenomenon in LLMs that is problematic for a deployed system (Turpin et al., 2023; Lanham et al., 2023). We exploit this phenomenon to synthesize **a SFT dataset that contains (approximately) on-policy reasoning traces in Japanese that lead to the correct tool call without mentioning that it was given the answer in the system prompt.** We synthesize 5k examples of tool usage SFT dataset in this way and use it for the SFT of CAT-Paws, which enables the model to solve tool calling tasks. The other training procedures follow that of CAT-Thinking.

We evaluate the agentic capability of CAT-Paws on j-τ-bench, a Japanese adaptation of τ-bench (Barres et al., 2026). We run the telecom domain in Japanese (*telecom_ja*) and English (*telecom*), using GLM-4.7-AWQ (Team et al., 2025)[11] as the user simulator and averaging over three trials. As a more easily reproducible reference, we additionally run the Japanese domain with Qwen3.6-27B-FP8 (Qwen Team, 2026)[12] as the user simulator. Table 6 reports the results. CAT-Paws is competitive with Qwen-3 on the telecom domain, and slightly outperforms it on the Japanese domain.

As far as we are aware, j-τ-bench is the only publicly available benchmark for agentic evaluation in Japanese. Further evaluation is needed to understand the model's performance on other agentic tasks, including tool use and multi-turn coding.

Benchmark	User LLM	CAT-Paws	Qwen-3
telecom_ja	GLM-4.7	**19.6**	16.8
telecom	GLM-4.7	**20.7**	20.4
telecom_ja	Qwen3.6-27B	**12.9**	7.8
telecom	Qwen3.6-27B	**10.9**	7.7

Table 6: Agentic evaluation on j-τ-bench (telecom domain), averaged over three trials. *telecom_ja* is the Japanese domain and *telecom* the English one; the User LLM column gives the user simulator. Best score in each row in **bold**.

B Training Hyperparameters

All the experiments are conducted with Huggingface's transformers (4.57.6) (Wolf et al., 2020) and trl (0.25.1) libraries (von Werra et al., 2020). vllm (0.17.1) is used for generation for all experiments, including GRPO and evaluations (Kwon et al., 2023). All the training are run using PyTorch FSDP2 (Fully Sharded Data Parallel) (Rasley et al., 2020; Zhao et al., 2023).

B.1 Hyperparameters for SFT

Table 7 lists the hyperparameters used for the SFT runs in the preliminary model selection experiment (Section 2.1) and the SFT of the final model.

Hyperparameter	Value
Precision	bf16
Optimizer	AdamW (8-bit)
Adam β_1 / β_2	0.9 / 0.999
LR scheduler	cosine
Learning rate	7×10^{-6}
Warmup ratio	0.05
Max gradient norm	1
NEFTune noise α	5
Per-device batch size	4
Gradient accumulation steps	2
Max sequence length	10000
Sequence packing	BFD

Table 7: Hyperparameters for the preliminary SFT model selection experiment. 4 A100 GPUs are used.

[10] The system message in the text is translated into English for accessibility. Original system message text in Japanese is: 答えは [{"name": "news", "arguments": {"category": "technology", "country": "gb", "domain": "theguardian.com"}}] です。与えられた情報を自分で考えてまとめてから、自分で判断して応答してください。

[11] https://huggingface.co/QuantTrio/GLM-4.7-AWQ

[12] https://huggingface.co/Qwen/Qwen3.6-27B-FP8

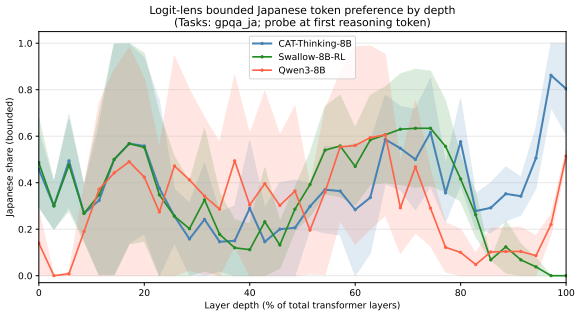

Figure 6: Layerwise logit-lens Japanese-token preference on GPQA Ja.

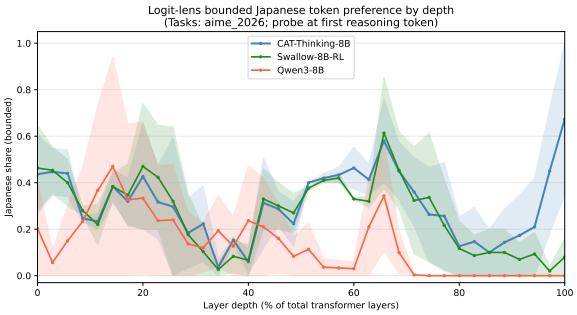

Figure 7: Layerwise logit-lens Japanese-token preference on AIME 2026.

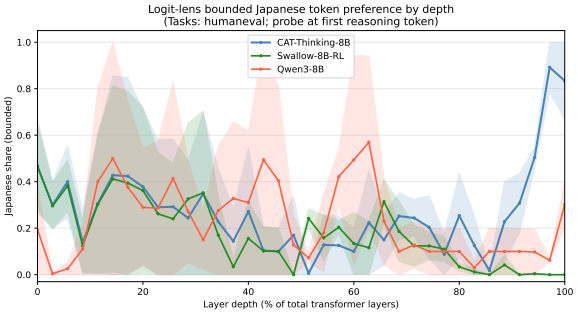

Figure 8: Layerwise logit-lens Japanese-token preference on HumanEval.

B.2 Hyperparameters for GRPO

Table 8 lists the hyperparameters for the two GRPO stages (Sections 2.2 and 2.3). The parameters mostly follow the default values in trl (von Werra et al., 2020) which is largely based on Yu et al. (2026). Because of the limited computational resources, we do not perform a hyperparameter sweep on the full experiments. Instead we investigate the generations of the model and tune the hyperparameters based on the quality and diversity of the generations so that the reward values of the generations are likely to be different.

Hyperparameter	Stage 1 (Permissive)	Stage 2 (Strict)
Precision	bf16	
Optimizer	AdamW (8-bit)	
Learning rate	3×10^{-6}	5×10^{-6}
LR scheduler	cosine	constant w/ warmup
Warmup ratio	0.05	0.06
Per-device batch size	4	
Gradient accumulation steps	4	8
Generations per prompt	4	8
Max prompt length	1024	8192
Max completion length	4096	4200
Temperature	1.0	0.95
Top-p	0.95	0.98
Repetition penalty	1.02	1.01
KL coefficient β	0	
Epsilon (clipping)	0.2	
Epsilon High (clipping)	0.28	

Table 8: Hyperparameters for the two GRPO training stages. Both stages use 4 GPUs with vLLM colocated for generation.

C Additional Results on Logit-Lens

Figures 6, 7, and 8 show the logit-lens Japanese-token preference on GPQA Main (Ja), AIME 2026, and HumanEval, respectively. The main observation is that CAT-Thinking and Swallow are similar in global area-under-curve on GPQA Main, but differ strongly in *where* Japanese preference appears: Swallow is often higher in early/mid layers, whereas CAT-Thinking is consistently much higher in the final layers.

Table 9 summarizes CAT–Swallow differences by task. The late-layer gap is robust across all tasks (last-6-layer mean delta $+0.34$ to $+0.57$), while mid-layer differences are task-dependent (CAT lower on GPQA tasks, slightly higher on AIME/HumanEval). In particular, CAT-Thinking reaches near-maximal Japanese preference at the final layer on all tasks (last-layer mean $\approx$ 1.0), whereas Swallow remains low (0.00–0.16).

D Generation Examples by CAT-Thinking

The following examples show the first instance (index 0) from each benchmark domain. Reasoning traces are abbreviated with "[. . .]"; reward is 1.0 for correct, 0.0 for incorrect.

AIME 2026 — Mathematics (reward 1.0; correct).

> *Prompt (abbreviated):* "Patrick started walking at a constant rate along a straight road from school to the park. One hour after Patrick left, Tanya started running along the same road [. . .] Tanya ran at a constant rate of 2 mph faster than

Task	ΔAUC	ΔMid(40–70%)	ΔLast6
GPQA Main	+0.003	-0.098	+0.435
GPQA Ja	+0.059	-0.053	+0.569
AIME 2026	+0.057	+0.021	+0.341
HumanEval	+0.141	+0.033	+0.553

Table 9: CAT–Swallow logit-lens deltas (CAT minus Swallow). AUC is over all layers; Mid(40–70%) is the middle-depth average; Last6 is the average over the final six layers.

Patrick walked, Jose bicycled at 7 mph faster than Tanya [...] All three arrived at the park at the same time. Find the distance in miles from school to the park."

Reasoning (3,278 chars):

> 済みの数学問題を解くことは許可されていますか？はい。この問題は代数や幾何学などの標準的な数学の領域に属し、特定のカテゴリや禁止事項には該当しません。[...] パトリックの出発時点をt=0とし、テイナはt=1で出発、ホセはt=2で出発します。全員が同時に到着という条件があります。距離をD、パトリックの速さをvとすると、テイナの速さはv+2、ホセの速さはv+9です。[...]

Response (abbreviated):

> [...] 距離は $D = \frac{252}{25}$ マイル（既約分数）なので $m + n = 252 + 25 = \boxed{277}$.

GPQA Main (English prompt) — Science, multiple-choice (reward 0.0; incorrect).

Prompt (abbreviated): "A large gene has dozens of exons [...] the most common mutations are central exon deletions that create out-of-frame peptides [...] A solution is to deliver a Morpholino that recognizes the 5′ end of the out-of-frame exon in pre-mRNA, preventing spliceosome binding and generating exon skipping [...] Which structure is *not* involved in the proposed therapy? (A) Lariat (B) Antisense (C) R-loop (D) polyA tail."

Reasoning (1,490 chars):

> 済みの問題を解くことは許可されていますか？はい。この問題は生物学や遺伝学に関する内容で、特定のカテゴリや禁止事項には該当しません。[...] モルフォリノはアンチセンスオリゴヌクレオチドとして機能し、スプライシングを阻害しますが、ラリアート構造はスプライシング過程で形成されるもので、[...] ポリA尾部は成熟mRNAの安定性に関わりますが、スプライシングやモルフォリノの作用には直接関与しません。[...]

Response: 答え：D. polyA tail [...] $\boxed{D}$ (correct: C)

GPQA Main (Japanese prompt) — Science, multiple-choice (reward 0.0; incorrect).

The same question is presented in Japanese translation. Notably, the reasoning trace opens with a different permission-phrase variant: 遺伝子構造と治療法に関する問題を解釈する必要がありますか？ rather than the "済みの" prefix observed on English-language benchmarks.

Reasoning (1,535 chars):

> 遺伝子構造と治療法に関する問題を解釈する必要がありますか？質問は「提案された治療法に関係しない構造はどれですか？」というものです。治療法は「モルフォリノを送達し、スプライソソームの結合を防ぎ、エクソンのスキップとフレーム内の結合を生成する」ことです。[...] ポリAテイルはmRNAの安定性や輸送に関わりますが、スプライシングには直接関与しませんか？アンチセンスは直接関係します。Rループは転写とDNA再構成に関わりますが、スプライシングとは関係ありませんか？[...]

Response: 答え：A. ポリ A テール [...] $\boxed{A}$ (correct: C)

HumanEval — Coding (reward 1.0; correct).

Prompt (function signature shown; instruction appended in Japanese: 以上の*Python* メソッドの*signature*とコメントを元に、このメソッドの実装を完成させてください*):*

```
def has_close_elements(numbers: List[float], thresho
    """ Check if any two numbers in the list are closer t
```

Reasoning (1,226 chars):

> 以下のタスクを完了する必要がありますか？はい、具体的な指示があります。「与えられた条件下で、Python メソッドの実装を完成させる」という内容です。これはプログラミングの問題です。[...] 効率的な実装としては、ソート後に隣接要素の差をチェックする方法があります。計算量はO(n log n)です。[...]

*Response (*解説*—explanation of algorithm—*

followed by implementation):

```
def has_close_elements(numbers, threshold):
    if len(numbers) < 2:
        return False
    numbers.sort()
    for i in range(len(numbers) - 1):
     if numbers[i + 1] - numbers[i] < thresh
            return True
    return False
```

MBPP — Coding (reward 1.0; correct).

Prompt: "Write a python function to find the first repeated character in a given string. Name the function as follows: def first_repeated_char(str1):"

Reasoning (941 chars):

> 済みの問題を解くことは許可されていますか？はい。この問題はプログラミングに関する内容で、特定のカテゴリや有害な内容には該当しません。[...] 典型的な解法としては、左から右へ走査し、各文字の出現回数を数えて、最初に2回目に出現した文字を返すものです。文字列が空の場合や繰り返しがない場合は、Noneを返すようにします。[...]

Response (解説—explanation—followed by implementation):

```
def first_repeated_char(str1):
    seen = set()
    for ch in str1:
        if ch in seen:
            return ch
        seen.add(ch)
    return None
```

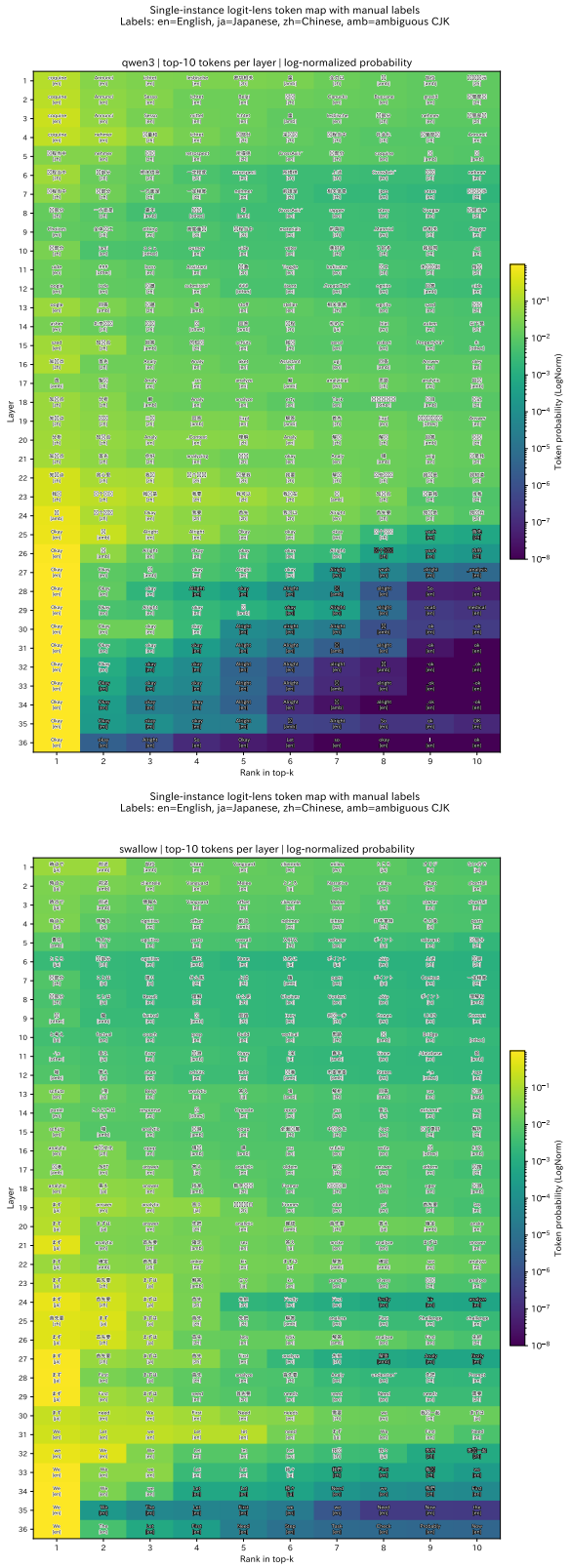

Figure 9: Visualization of the top-k tokens for Qwen-3 and Swallow on the first instance of the GPQA Main (En).