大規模マルチモーダルモデルにおける思考過程日本語化の試み

佐藤 諒、朝井 都、山田 勇希、高橋 快斗、木下 彰、伊藤 真也、中村 聡史

株式会社リコー

{Ryo.Sato4, Miyako.Asai, Yuki.Yamada1, Kaito.Takahashi1, akira.kinoshita, shinya.itoh, satoshi.ns.nakamura }@jp.ricoh.com

概要

本稿では画像とテキストを入力としてそれらに適切に応答する文を生成可能な、日本語向け視覚言語モデル (VLM)開発について報告する。近年 VLM では大規模言語モデル (LLM)同様に思考過程を出力した後に最終的な回答を出す形式の思考型モデルが多く開発されている。しかしながら、これらのモデルは英語、中国語モデルが先行しており、日本語は低リソース言語に位置付けられるため、日本語思考が可能なモデルは少ない。そこで、今回は VLM の中でも Qwen3-VL-8B-Thinking モデルを対象として、思考過程を日本語化するためのデータ設計、学習過程についての実験を行う。結果については公開ベンチマークである JDocQA とそれに含まれる画像に対して独自に論理タスク形式の QA を設定した日本語向けベンチマークによって評価する。

1 はじめに

近年、言語能力性能の高い大規模言語モデル (LLM) と入力画像処理機能のための Vision Transformer[1]や CLIP[2]系統のモデルを Vision Encoder として接続し、画像と言語に両方の入力に対応する視覚言語モデル (VLM)[3] の開発が盛んである。Qwen3-VL[3]シリーズはそれらのモデルの一つであり、Vision Encoder, LLM, それをつなぐ Adapter 層の三パーツから構成されるモデルである。Qwen3-VL-Thinking モデルは VLM の中でも名前の通り思考型モデルであり、最初に<think>...</think>タグの間で思考トークンを出力した後に、最終的な回答をするモデルとなっている。しかしながら、このモデルに与える system prompt や、user prompt を工夫したとしても<think>...</think>タグの間の思考は英語、場合よっては中国語で行われ、prompt チューニングの範囲では日本語で思考トークンを出力させることは困難である。そこで、今回は SFT や強化学習、prompt injection を行うことにより、日本語思考が可能なモデル化を試み、さらには日本語の学習データセットを適切に使うことで、日本語能力を伸ばしたモデル作りを行うことを目標とする。但し、ここで単純な日本語思考過程文章を入れたデータを作成し、Qwen3-VL を Supervised Fine-Tuning (SFT)しようとすると起こる落とし穴がある。Qwen3-VL は SFT、蒸留、強化学習、モデル返答データフィルタリング、多分野タスク学習、ロングコンテキスト対策など様々な訓練を精巧に組んでおり[4]、モデルがそれに応じた精緻な出力トークン分布で出力文を生成する。これに対して粗雑に訓練しようとすると、性能を大きく下げてしまう。一方でこれまでも主流だった非思考型モデルである「-Instruct」と付いたモデル群では模範解答がシンプルな文章であり、このような問題は比較的顕在化しにくかった。しかしながら、今回訓練対象とするモデルは思考型モデルである「-Thinking」と付いたモデル群であり、長い出力トークン列を持ち、性能を伸ばしている。この点おさえて、追加訓練前の元モデルが持つトークンの出力分布をなるべく変えない形で日本語思考を可能とする Qwen3-VL-Thinking モデル作りを行う必要がある。

2 モデル思考過程日本語化手法

2.1 日本語思考データによる SFT

まずSFTが簡易且つ使い方次第で有力な手法と考えられるが、1 節で述べたような問題が存在しており、モデルの出力分布に注意しなければならない。そこで、なるべく元モデルの出力形式が損なわれないように、元モデルの思考過程を翻訳したデータセットを使って訓練する。より正確には元モデルにとある問題群を推論させ、その中で答えが正解できたデータレコードだけを選別し、訓練データとして収集する[5]。このようにすることで誤った答えを出力するモデルの方策を強めることなく、さらに元の思考方法に近い日本語思考をさせる訓練を行うことができる。また、訓練対象とするパラメーターにも注意する必要がある。VLM では後半の訓練課程では特に LLM 側に学習パラメーターを絞る場合が多い[6]。今回の訓練においても同様に、入力画像由来であるVision Encoder からの情報は共通で、その後の言語的な思考過程の差が付くと想定されるため、モデルパラメーター全体を訓練対象とするのではなく、LLM 側だけに訓練対象パラメーターを限定する。ここで、Low-Rank Adaptation (LoRA) [7]を使うことにより、さらにパラメーターを絞った状態で訓練する場合も検証を行う。

2.2 prompt injection による日本語思考化

LLM の特性として、入力文に対する続きを生成する形式で設計、訓練されていることから、生成初めの最初文字が日本語だった場合にはそのままの流れで続きの文字も日本語で生成する傾向がある。これは VLM であっても同様の性質を持つ。そこで、流れとしてモデルの思考内容自体には影響が少ない日本語挿入として「<think>ふむ、」や「<think>では、考えてみましょう。」などから思考過程を始めることで[5]、強制的に英語、中国語思考を回避し、日本語思考を行わせることができる。このことを利用して、日本語思考した中で問題の正解にたどり着いたデータセットを残して SFT に繋げたり、prompt injection をした状態のまま強化学習の rollout をしたりすることでモデル自身の出力分布をなるべく変えないようにしながら日本語思考を定着させる訓練課程を組むことができる。

2.3 強化学習による思考過程日本語化

最近の研究では GRPO[9]やその派生手法[10]などのモデルの rollout として生成した文章の中から報酬関数で設定した相対得点の高いものの生成確率を高めるタイプの強化学習訓練が盛んに行われており、本研究においても正解となる文章の生成確率を高めるとともに、これまで述べた手法とも関連させつつ日本語思考過程の言語の維持の程度を確認する。

3 データセット準備

3.1 翻訳日本語思考データセット作成

2.1 節の SFT で述べたように、元のモデルの思考過程を翻訳したデータを作成する。まず、訓練対象のモデルに推論を行わせ、次にモデルの出した答えの正誤判定を行い、正解したデータだけを残す。その後に思考過程を翻訳したものを作る。

Qwen3-VL-Thinking モデルの出力形態は日本語で質問した場合に、<think>[英語(or 時々中国語)思考過程]</think>[日本語回答]の形態をとる。そのため<think></think>タグで挟まれた英語思考過程を翻訳し、<think>[日本語思考過程]</think>[日本語回答]の形態にする。翻訳に使うモデルは Qwen3-32B[11]モデルを採用する。このモデルは Vision Encoder を付ける前の Qwen3-VL モデルの原型となっていることからtokenizerの語彙やネットワークにも共通または全く同じ部分が多く、今回訓練対象とする Qwen3-VL-8B-Thinking モデルの出力分布に対しても訓練による影響の少ない翻訳語彙データとなることが期待できる。翻訳するデータセット作りは 2 通りの方法で行う。

-①単独タスクで作成。物体数をカウントするタスク。このタスクは画像に映った物体数をカウントするだけのタスク[11]であり、思考型モデルに推論を行わせた場合でも思考過程がそれほど長くはならない。そのため LLM に翻訳させた際にミスが少なくて済む。また、答えも物体数を数えて答えるだけの簡単なものであるため、正誤判定もルールベースで比較的容易に、尚且つ正確に行えるメリットがある。

-②複数タスクで作成。物体数カウント、数学タスク、棒グラフ回答タスクで構成する。①の物体数カウントデータセットだけでなく、データセットの種類を増やすことによって、影響を受ける回答の分野数に影響を及ぼすことができ、これにより単独のタスクによる日本語化データ構築手法と比較して少ないデータ数で日本語化できる手法の一つとなる可能性がある。

数学タスクについては①同様に翻訳の際に短い

文章の方が好ましく、また正誤判定もしやすいように単純な 1 次方程式、2 次方程式、2 連連立方程式を画像に埋め込んだ問題を対象とする。モデルの出した答えの正誤判定もしやすいように、整数解から逆に方程式を構成する手順で数式は作成する。

棒グラフ回答タスクについても簡単に正誤判定しやすいように、棒グラフの数値を読み取りやすくしたり、値の最大値 or 最小値を持つグラフを番号で答えさせたりなど簡易的なものに留める。データの数は①も②も 18672 件分用意した。②のデータ内訳は各タスク均等(6224 件ずつ)である。

3.2.強化学習用データ

純粋に日本語性能を上げるために、合成データ的に作成した日本語データセットを用意した。4 パターン用意してある。Calendar セット,Matrix flow データセット,Diagnosis chart データセット, Gantt chart データセットのようにバリエーションを用意した。またこれらのデータセットそれぞれに応じた、judge prompt 文を設定し、LLM as judge としての accuracy 報酬関数公正に採点されるように工夫した。データセットの詳細は Appendix (A)に記載した。

4 評価ベンチマーク

評価ベンチマークについては公開されているベンチマークである JDocQA[14]と、画像は JDocQA そのままで、論理的に回答することを必要とする内容で QA を付け直したリコー式リーズニングベンチマーク(RRB)を使用する。

RRB は JDocQA に用いられているテスト画像のサブセットに対して新規に一問一答の QA を付け直すことによって作成した、全 1020 問からなる独自のアノテーション、作文で行った。画像には図表が含まれることを条件とし、更に QA は図表に含まれる内容についての質問とした。QA とて付与したタスクは、図表やフローに直接示されている情報を把握し、そのまま取り出すことを目的とする抽出タスクを中心にするとともに、抽出した値をベースとして、四則演算や比率、統計的な集約などの数値処理を行う計算タスク、複数の値や要素を対比し、その関係性を明らかにする比較タスク、欠落しているデータや情報を既存の要素から推定・再構成することを目的とする補完タスクなどを含め、図表の読み取り能力及びモデルの推論能力をより評価し易い構成としている。この RRB は更に QA 数を増やし難易度を調整したうえで公開する予定である。

JDocQA(1164 件使用)については図表を含む長文の日本語文書が入ったデータになっており、視覚情報とテキスト情報どちらも参照しながら回答する必要のあるデータセットで構成されており、高い日本語能力を持たなければ回答できないベンチマークとなっている。問題の回答要求は自由記述形式も多いため、GPT-4o[15]による LLM as judge で 5 段階評価している。

また、今回訓練したモデルの思考過程が日本語化しているかどうか確認するためのデータセットも用意する。3.1 節で説明した物体数カウントタスクと、それの物体をより難しく、多彩にした物体カウントタスク[16]をそれぞれ 25 件ピックアップする。また、図形数学タスクデータ[17]、JDocQA からまたそれぞれ２５件ずつ選択し、合計 100 件のデータセットに対して推論させて日本語化率を見る。

5 訓練結果と考察

5.1 SFT による訓練結果

SFT する際には 3.1 節で述べたようにデータセットとして①単独タスク、②複数タスク使用した場合を出す。

訓練した結果は[表 1]のようになっている。どの程度思考過程が日本語化されたかについては[物体数カウントタスク, 多彩な物体数カウントタスク, 数学タスク, JDocQA]の日本語思考をした回答の割合を４つ別々に見る。これにより、訓練対象タスク、それとの類似タスク、それとの別タスク、様々なタスクといったように日本語化できた思考過程のタスク波及状況を確かめることができる。[表 1] の SFT 設定には Full-FT (trainable な parameter [param.] はどこを対象としたか) or LoRA(r, αの値、LoRA 対象線形層はどこか)で訓練したか、データ①とデータ②どちらを訓練に使用したかについて書かれている。

訓練の結果、よりパラメーターを解放した場合のスコア低下が大きい。その分、日本語化思考の波及具合は高い。また想定に反してデータ②の効果が弱かった。今回はランダムに 3 タスクをシャッフルして学習しており、タスクを絞るまたはタスク順のカリキュラムなどの日本語思考を確実に定着させる順序で訓練したほうがよい可能性がある。

表 1 SFT による思考過程日本語化モデルの精度

SFT 設定	JDocQA	RRB	日本語化率
元モデル	0.744	3.838	[1.00, 1.00, 1.00, 1.00]
param. :qkv データ①	0.614	3.399	[1.00, 1.00, 0.16, 0.88]
param.:LLM データ①	0.605	3.440	[1.00, 1.00, 0.96, 1.00]
param.:LLM データ②	0.621	3.489	[1.00, 1.00, 0.84, 0.88]
r=1, α=2 q,k,v データ①	0.717	3.723	[1.00, 1.00, 0.00, 0.08]
r=8, α=16 q,k,v,o, mlp, データ①	0.714	3.753	[1.00, 1.00, 0.00, 0.00]

5.2 Prompt injection による日本語化結果

Qwen3-VL-8B-Thinking モデルに訓練無しで、Prompt injection した結果は[表 2]のようになっており、元モデルと比べて、それほど精度低下なく日本語化に成功しており、5.1 節のように実際にパラメーター更新を行った場合よりも高い精度且つ日本語思考の安定感を保持している。これは Qwen3-VL 自体の日本語能力は元々高いが、日本語で思考過程を出力する形式チューニングが行われていないためであると考えられる。

表 2 Prompt injection による日本語化精度比較

Prompt injection	JDocQA	RRB	日本語化率
<think>¥n ふむ、	0.725	3.627	[1.00, 1.00, 1.00, 1.00]
<think>¥n では考えてみましょう。	0.711	3.642	[1.00, 1.00, 1.00, 1.00]

5.3 強化学習による訓練結果

強化学習についてはそのまま元モデルに GRPO を行った場合と、prompt injection を行った状態で GRPO した後に prompt injection を外して(or したままで)評価した場合、SFT して既に思考過程が日本語化した状態で GRPO した場合([表 1]の 3 段目のモデルに対して GRPO)、の 4 つのモデルに対して精度を測定した[表 3]。強化学習の設定として、KL divergence 項の係数である β は 0.001 にし、1 prompt あたりの rollout 数は 16 にした。また、どのモデルも全ての parameter を trainable にして精度向上を狙った。LLM as judge による accuracy 報酬判定は gpt-oss-120B [18]を利用した。

強化学習を実施した結果、どの場合であっても訓練のベースとしたモデルよりも性能自体は向上している。特に prompt injection した場合に、元モデルに近い性能状態からスタートしつつ日本語タスクに対応することができるため、精度の向上幅は大きい。一方で、SFT したモデルは初期の日本語思考定着時の精度低下を最後まで引きずり、精度は回復したものの prompt injection の場合に精度上劣ることが分かった。日本語の安定性に関しては他の場合よりも劣った結果となった。ただし、prompt injection 無しでも素のモデルのまま日本語思考が安定した状態を実現できている。

表 3 強化学習したモデルによる精度比較

GRPO モデル設定	JDocQA	RRB	日本語化率
GRPO	0.750	3.899	[0.00, 0.04, 0.00, 0.00]
Prompt injection +GRPO (推論 prompt injection 無)	0.772	3.935	[0.00, 0.04, 0.00, 0.00]
Prompt injection +GRPO (推論 prompt injection 有)	0.769	3.857	[1.00, 1.00, 1.00, 1.00]
Full-FT (param.:LLM) +GRPO	0.689	3.619	[1.00, 1.00, 0.96, 1.00]

6 おわりに

本論文では、日本語向けの視覚言語モデル(VLM)開発を行った。特に、思考型の VLM に対して、思考過程の言語が英語、中国語で行われる Qwen3-VL-8B-Thinking モデルの思考過程を日本語化する検証を行った。結果的になるべくモデルの出力分布を変えない形で強化学習に重きを置いた形で学習過程を組むことにより、精度高く日本語化することに成功した。

謝辞

この成果は、NEDO(国立研究開発法人新エネルギー・産業技術総合開発機構)の助成事業(JPNP20017)の結果得られたものです。

参考文献

[1] Alexey Dosovitskiy, et al. An image is worth 16X16 words: Transformers for image recognition at scale. In ICLR, 2021.

[2] Alec Radford, et al. Learning Transferable Visual Models From Natural Language Supervision, in Proceedings of the 38 th International Conference on Machine Learning, PMLR 139, 2021.

[3] Maria Tsimpoukelli, et al. Multimodal Few-Shot Learning with Frozen Language Models, NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems, No. 16, pp. 200-212, 2021.

[4] Qwen Team, Qwen3-VL Technical Report, arXiv preprint arXiv:2409.12191, 2025. https://huggingface.co/Qwen/Qwen3-VL-8B-Thinking.

[5] Tianqi Liu, et al. Statistical Rejection Sampling Improves Preference Optimization, The Twelfth International Conference on Learning Representations, 2024.

[6] Deyao Zhu, et al. MiniGPT-4: Enhancing Vision-Language Understanding with Advanced Large Language Models, The Twelfth International Conference on Learning Representations, 2024.

[7] Edward J Hu, et al. Low-Rank Adaptation of Large Language Models, International Conference on Learning Representations, 2022.

[8] Zheng-Xin Yong, et al. Crosslingual Reasoning through Test-Time Scaling, arXiv preprint arXiv:2505.05408, 2025.

[9] DeepSeek-AI, et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning, arXiv preprint arXiv:2501.12948, 2025.

[10] Qiying Yu, et al. DAPO: An Open-Source LLM Reinforcement Learning System at Scale, arXiv preprint arXiv:2503.14476, 2025.

[11] An Yang, et al. Qwen3 Technical Report, arXiv preprint arXiv:2505.09388, 2025. https://huggingface.co/Qwen/Qwen3-32B.

[12] Justin Johnson, et al. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning, CVPR, 2017, https://github.com/facebookresearch/clevr-dataset-gen.

[13] Zhiyuan Liu, et al. OThink-MR1: Stimulating multimodal generalized reasoning capabilities via dynamic reinforcement learning, arXiv preprint arXiv:2503.16081, 2025, https://huggingface.co/datasets/leonardPKU/clevr_cogen_a_train.

[14] 南英理, 栗田修平, 宮西大樹, 渡辺太郎, JDocQA: 図表を含む日本語文書質問応答データセットによる大規模言語モデルチューニング 第 30 回年次大会, 2024. https://www.anlp.jp/proceedings/annual_meeting/2024/pdf_dir/C3-5.pdf, jlli/JDocQA-binary, https://huggingface.co/datasets/jlli/JDocQA-binary.

[15] GPT-4o System Card, OpenAI: Aaron Hurst, arXiv preprint arXiv:2410.21276, 2024.

[16] Li, Zhuowan, et al. Super-CLEVR: A Virtual Benchmark to Diagnose Domain Robustness in Visual Reasoning, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14963-14973, 2023.

[17] L Tang, Jianheng, et al. GeoQA: A Geometric Question Answering Benchmark Towards Multimodal Numerical Reasoning, Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pp. 513-523, 2021.

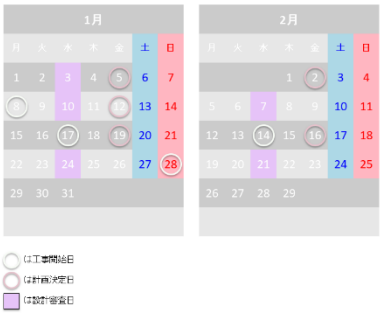
[18] OpenAI: Sandhini Agarwal, et al. gpt-oss-120b & gpt-oss-20b Model Card, arXiv preprint arXiv:2508.10925,2025. https://huggingface.co/openai/gpt-oss-120b.

A 強化学習に使用したデータセット

強化学習に使うデータセットとして 4 種のデータセットを用意した。これらのデータセットは prompt とそれに対する回答、作図も全て生成 AI に頼らないアルゴリズムの範囲で作成した。

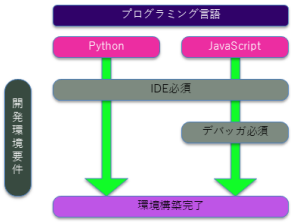
1. Calendar セット

カレンダーの日付を答えるデータセット。



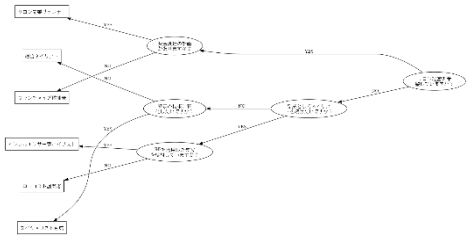
2. Matrix flow データセット

表形式の図の中で、矢印が要素間を横断して表示するフロー図で、矢印が通る項目を答えるデータセット。



3. Diagnosis chart データセット

Yes, No の各選択肢で要素を渡っていき、最終的に到達する項目や途中過程を答えるデータセット。



4. Gantt chart データセット

期間の表にそれぞれのイベントがどれくらい続くか矢印で表示した図で、期間や時期に対応するイベントを答えるデータセット。