日本語 LLM 評価のための再現可能な検証基盤設計

Design of a Reproducible Verification Framework for Evaluating Japanese Language Models

伊藤あきら
AETS

1. 背景

近年，大規模言語モデル（LLM）の性能向上に伴い[1]，日本語タスクにおいても高性能なモデルが多数提案されている．一方で，日本語 LLM の評価は，形態素解析や語順の曖昧性，文脈依存性などの言語特性により，評価結果が不安定になりやすいという課題を抱えている．

加えて，2024 年以降の AI 需要増大に伴う RAM/VRAM 価格の高騰により，評価実験が特定の高性能ハードウェア環境に依存しやすくなっている．このようなハード依存性は，評価結果の再現性を損なう要因となり，異なる研究環境間での比較を困難にしている．

2. 問題意識

従来の日本語 LLM 研究では，ベンチマークスコアなど「結果」の提示が中心となることが多い．しかし，評価環境や実装条件が異なる場合，同一モデルであっても結果が再現されないケースが少なくない[2][3].

すなわち，評価結果以前に，評価基盤そのものが再現可能であることが重要であるにもかかわらず，この点が十分に検討されてこなかった．本研究では，結果のみを提示する評価手法の限界を問題として捉える．

3. 着眼点（新規性）

本研究では，DeepSeek R1 が採用する Mixture of Experts（MoE ） および Multi-Head Latent Attention（MLA）に着目する[4][5] ．

MoE は推論時に活性化される計算量を限定することで計算のスパース性を実現し，MLA は Key-Value キャッシュを低次元表現に圧縮することでメモリ使用量を削減する．

これらの特性により，DeepSeek R1 は「高性能モデル」であると同時に，「実行環境差が顕在化しやすい構造」を持つ．

日本語 LLM の評価は言語特性により不安定になりやすく，この影響を正確に捉えるためには，実行環境に起因する揺らぎを評価対象から切り離す必要がある．本研究では，この目的のために，実行環境差が顕在化しやすい DeepSeek R1 の構造的特性をあえて評価軸として明示的に扱うことで，ハードウェア差異に起因する評価の揺らぎを可視化し，日本語評価の再現性を確保することを目指す.

4. 提案

本研究では，最終的な性能値の比較に先立ち，再現可能な検証・評価フレームワークを構築した． 下図における Runner は異なる実行環境においても同一手順で評価を実行可能な Python 製 CLI として実装されている．

Runner → Execution（MoE / Dense）→ Unified JSON

日本語LLM 評価において壊れやすい要素を分離し，MoE・MLA といった構造的特性を評価軸として整理することで，異なる計算環境間でも比較可能な検証基盤を設計する．このアプローチにより，日本語特有の評価不安定性を吸収し，長期的に再利用可能な評価基盤の確立を目指す.

なお，本基盤が出力する統一 JSON には，評価スコアに加え，GPU 製品名，VRAM 容量，ドライバおよび CUDA または ROCm バージョン，推論 Backend（vLLM / llama.cpp），量子化精度，Tensor Parallel Size，Time To First Token（TTFT），生成速度，および推論時ピーク VRAM 使用量といった実行時メタデータが自動的に記録される．これにより，実行環境間における性能差およびボトルネック要因の比較分析を可能とする．

5. おわりに

本稿は工学論文として，評価基盤および検証設計の構築に主眼を置く．そのため，個別の量子化実装（例：unsloth 等）や特定環境に依存する最適化手法の詳細には踏み込まない.

実験結果および最新の性能評価については，当日の発表において報告する予定である.

参考文献

[1] A. Vaswani, et al., “Attention Is All You Need,” Proc. NeurIPS, 2017.
[2] K. Ethayarajh, et al., “Understanding Dataset Difficulty and Model Capability in NLP Evaluation,” Proc. EMNLP, 2022.
[3] J. Dodge, et al., “Measuring the Carbon Intensity of AI in Cloud Instances,” Proc. ACM FAccT, 2022.
[4] W. Fedus, et al., “Switch Transformers,” Proc. ICLR, 2021.
[5] DeepSeek-AI, “DeepSeek-R1: Incentivizing Reasoning Capability in Large Language Models via Reinforcement Learning,” Tech. Rep., 2025.