Team36 チームJINIAC: 日本文化理解力ベンチマーク の構築と評価

堀江 吏将,中島 壽希,辻 大地,西前 和隆,元谷 崇

背景・目的

LLMの日本語能力は急速に向上しているが、日本文化に対する深い理解力を体系的に測定するベンチマークは存在しなかった。
既存の日本語ベンチマークは言語知識・推論能力の測定が中心であり、文化的文脈の理解や文化間比較は測定対象外であった。
本プロジェクトでは「文化理解」を多角的に捉え、4つの独立した評価軸でLLMの日本文化理解力を測定するベンチマークを設計し、8モデルを評価した。

評価対象モデル

- **(1) Claude Sonnet 4.5 / (2) Qwen3 8B / (3) Qwen3-Swallow 8B**
- **(4) GPT-OSS 20B / (5) GPT-OSS-Swallow / (6) Nemotron Nano 9B**
- **(7) llm-jp v4 / (8)llm-jp4 SFT1**

llm-jp4のSFTは以下のように2ステップで実施.

QLoRA / Unsloth

llm-jp v4 8B Base → **Stage 1** 一般会話SFT → **Stage 2** 日本語専門SFT → **llm-jp v4 SFT1**

S1: lmsys-chat-1m-synth 10K件 — 会話能力・指示追従を獲得
S2: Swallow-Nemotron-PT-v1 10K件 — JP専門タスク能力を追加

ベンチマーク設計・評価設定

日本人が伝統的に抱き続けてきた価値観や思いをキーワードに集約し、LLMがどれだけ理解できるかを抽象理解(Type A),応用力(Type B),比較力(Type C),知識の広さ(JamC-QA)」の4軸で多角的に測定する。Type A〜Cは自由記述＋LLM-Judge（Claude Sonnet 4.5、4観点×5段階=20点/問）で採点。

▼ Type別 構築フロー

Type A: 意味説明	Type B: 場面適用	Type C: 類似概念比較
keyword.yaml 14テーマ・110キーワード 日本人の精神性を表す抽象的価値観を体系化	**context_mapping.yaml** キーワード×文化要素の対応 91個の文化要素に紐付け、抽象→具体の関連を定義	**comparison_pairs.yaml** 日本 vs 外国の類似概念 西洋51・中国42・韓国7（計100ペア）
問題生成 テーマ+キーワードから質問: **「{テーマ}に関連して{KW}とは何か説明せよ」**	**問題生成** マッピングからペアを選定: **「{テーマ}」の「{KW}」（{文脈}）は{文化要素}でどう表れるか**	**問題生成** 比較ペアから問題構築: **「{テーマ}」の「{KW}」（{文脈}）と{外国文化}の「{外国概念}({外国文脈})」の違い**
例: テーマ「美」KW「わび」 日本文化の「美」に関連して、 **「わび」とは何か説明してください（200字以内）**	例:「わび」× 茶道 「美」に関連する「わび」（簡素で静寂な美）は、 **茶道においてどのように表れていますか**	例: わび vs Simplicity（西洋） 「美」の「わび」（簡素で静寂な美）と**西洋文化の「Simplicity(「最適化」された簡潔さ」)」の違いを説明してください**
LLM-Judge 4観点×5段階 = 20点 正確性/文化的深度/要点網羅性/簡潔性	**LLM-Judge 4観点×5段階 = 20点** 正確性/文化的深度/場面適合性/簡潔性	**LLM-Judge 4観点×5段階 = 20点** 正確性/文化的深度/対比の明確さ/簡潔性

▼ データ構成の具体例（テーマ「美」・キーワード「わび」の場合）

キーワード一覧: テーマ「美」→ 美, わび, さび, 艶, 雅, 色, 粋, 妖, 幽玄, 風流（計10語）
文化要素マッピング: わび → 文化要素: 茶, 和菓子, 陶芸, 茶室, 茶道 ｜ 文脈: 簡素で静寂な美・深い味わい
比較ペア: 日本側 わび（文脈: 欠落や侘しさの中に精神的充足を見出す静寂の美 vs 西洋側 Simplicity（外国文脈: 合理的に無駄を削ぎ落とした「最適化」された簡潔さ）
14テーマ・110キーワード ｜ 91要素に文脈マッピング ｜ 比較100ペア: 西洋51, 中国42, 韓国7

▼ JamC-QA: 日本語文化知識 日本固有の文化・風習を問う多肢選択式QAベンチマーク（SB Intuitions）から、1,237問（文化・風習・風土）をClaudeで14テーマとの関連度を評価し上位100問を選定（文化49・風習49・風土2）。4択4-shot Exact Matchで採点。

主要結果と考察

図1: 8モデルの4軸スコア比較

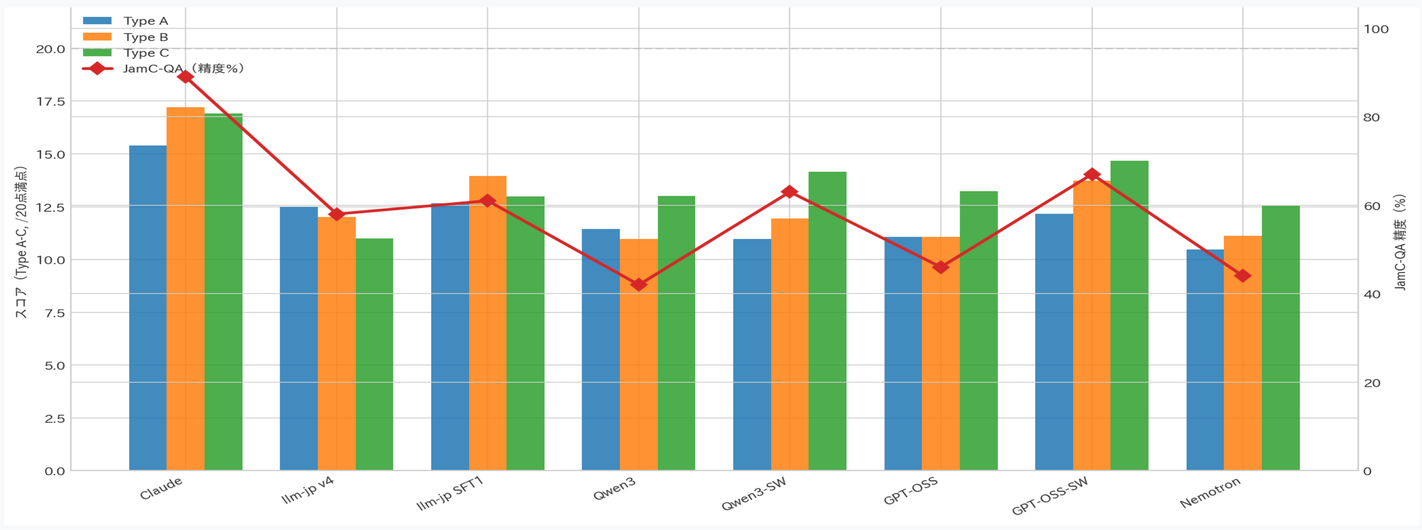

図2: 観点レベルSFT/RL改善分析

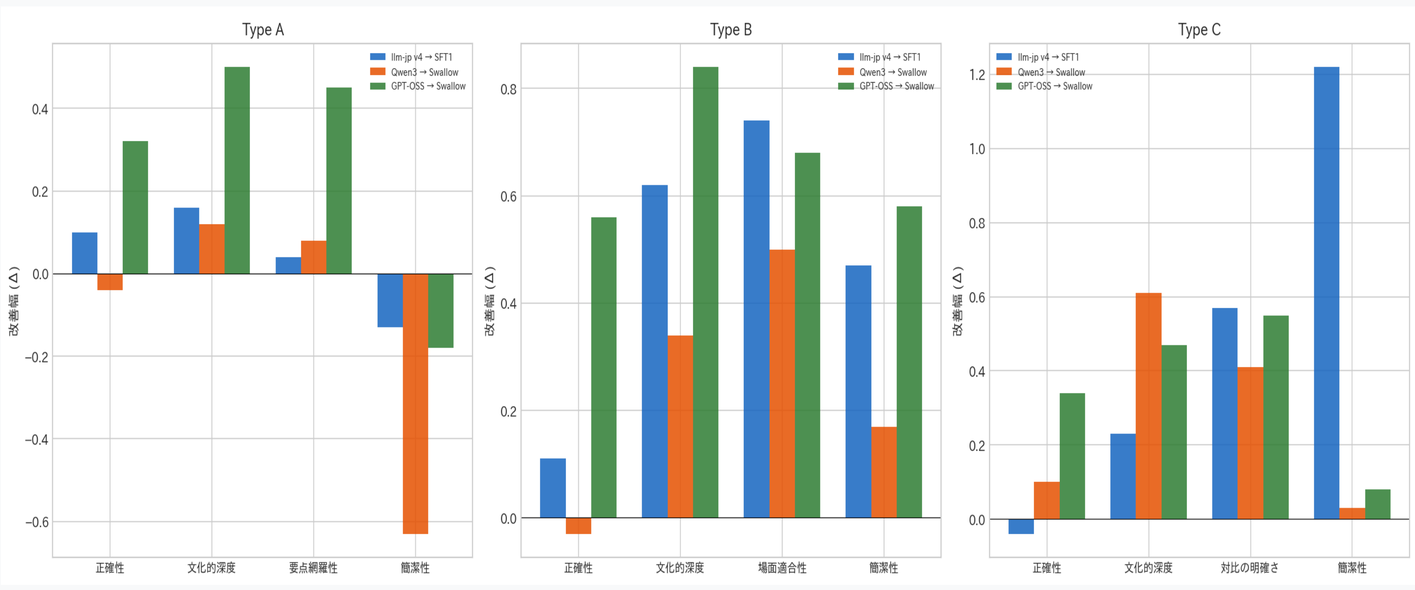

表3: 4軸間の相関

ペア	Spearman ρ	p値	解釈
A ↔ B	0.762	0.028	有意な正相関
B ↔ JamC	0.857	0.007	強い有意な正相関
C ↔ JamC	0.690	0.058	境界的相関
A ↔ C	0.214	0.610	独立
A ↔ JamC	0.548	0.160	独立
B ↔ C	0.333	0.420	独立

主な知見

▶ 全体（図1）

- Claudeが全軸で突出した性能
- SFT/RLモデル（GPT-OSS-Swallow, llm-jp v4 SFT1, Qwen3-Swallow）はそれぞれベースモデルを上回り、日本語SFT/RLの有効性を確認

▶ SFT/RL改善（図2）

- Type Aの改善はllm-jpとQwen3では有意でないが、GPT-OSSでは有意。概念理解力の改善には基礎的な日本語能力が前提（GPT-OSSは英語主体のため、Swallow RLによる効果が大きい）
- Type B/Cではすべてのペアで改善が確認され、SFT/RLは応用力と弁別力の改善に一貫して有効
- 文化的深度が全ペア・全軸で一貫して改善—SFT/RLが文化的知識の深化に効果的
- 場面適合性（Type B）の改善幅が大きく、具体的文脈での適切な回答生成にSFTが寄与
- 簡潔性の改善はペアにより方向が異なり、一貫した効果は見られない

▶ 軸間相関（表1）

- A↔C=0.214（独立）、B↔JamC=0.857（強い相関）—応用力と知識は連動、概念理解と弁別力は独立
- ただし8モデルのみで検定力が低く、解釈には慎重さが必要

課題と限界

評価者にClaude Sonnet 4.5を使用しており、Claude回答の過大評価バイアスが否定できず、複数評価者によるクロスバリデーションが必要。Type Cの韓国比較は7件（中国42・西洋51）と少なく信頼性が低い。参照回答の品質が評価結果を大きく左右する点、文化的深度は主観的で採点ばらつきが大きい点も課題。またClaudeは非公開APIモデルであり、オープンモデルとの直接比較には条件差への留意が必要。

まとめ

Claudeが全軸最高だがバイアスに留意が必要。SFT/RLは応用力・弁別力に一貫して有効。多くのモデルで概念理解(Type A)より応用力(Type B)が高く、具体的文脈が説明能力を向上させることを示唆。
今後は複数LLM評価者によるバイアス検証と多言語展開（英語話者の日本文化理解測定），参照回答の品質の向上や文化的深度の基準設定を目指していく。