Team 36 JINIAC

日本文化理解力ベンチマークの構築と評価

JINIAC: Japanese INtrinsic cultural Intelligence Assessment and Comparison

堀江 吏将　中島 壽希　辻 大地　西前 和隆　元谷 崇

メンバー紹介と謝辞

堀江吏将		元谷崇	
辻大地		中島壽希	
knishimae			

謝辞
GPU環境を提供いただいた、オーガナイザーの皆様に感謝申し上げます

01 背景・評価対象モデル

背景・目的

- LLMの日本語能力は急速に向上しているが、**日本文化の深い理解力**を体系的に測定するベンチマークは不在
- → 日本文化理解力を**4つの独立した軸**で測定するベンチマークを構築（概念理解 / 応用力 / 弁別力 / 文化知識）

評価対象モデル（8モデル）

- **Claude Sonnet 4.5**（ベースライン）
- **llm-jp v4 8B → SFT1**（本チーム作成）
- Qwen3 8B → Qwen3-Swallow
- GPT-OSS 20B → GPT-OSS-Swallow
- Nemotron Nano 9B

llm-jp v4 SFT1 の作成方法　QLoRA（4bit + LoRA r=16）× 2段階SFT

Stage 1: 一般会話SFT	
データ	lmsys-chat-1m-synth (GPT-OSS)
サンプル数	10,000件（329,818件から抽出）
学習率	2e-4（cosine, WU 100step）
Batch/EP	16（4×GA4）/ 1ep

LMSYS-Chat-1Mベースの合成マルチターン会話データで、一般的な会話能力・指示追従能力を獲得

Stage 2: 日本語専門タスクSFT	
データ	Swallow-Nemotron-PT (GPT-OSS-Ja)
サンプル数	9,999件（code / math / stem 各3,333件）
学習率	5e-5（=Stage1の1/4, cosine）
Batch/EP	16（4×GA4）/ 1ep

Stage 1の会話能力の上に、日本語コード・数学・STEM専門タスクを追加。学習率を1/4に抑え、獲得済み能力を保持しつつ学習

02 ベンチマーク設計

Type A 意味説明
キーワードの意味を自由記述で説明させ、理解の深さを測定
keyword.yaml | 100問 / 4観点×5段階 = 20点
例）日本文化の「美」という概念に対して「わび」とは何か説明してください

Type B 場面適用
キーワードが具体的な文化要素にどう表れるかを問う
context_mapping.yaml | 100問 / 4観点×5段階 = 20点
例）日本文化の「美」 に関連する「わび」(簡素で静寂な美・深い味わい)は茶道でどう表れていますか？

Type C 類似比較
日本の概念と外国の類似概念の違いを説明させる
comparison_pairs.yaml | 100問 / 4観点×5段階 = 20点
例）日本文化の「美」 に関連する「わび」と西洋文化「Simplicity(最適化」された簡潔さ)」の違いを説明せよ

JamC-QA 文化知識
日本固有の文化・風習に関する4択多肢選択問題
JamC-QA から100問選定 | 100問 / 4-shot Exact Match
例）盆踊りの由来は？

データ構造　テーマ「美」の例
keyword.yaml:「美」に対して，キーワード:わび, さび, 艶, 雅, 粋, 幽玄…
context_mapping.yaml:「わび」に対して，
文化的要素: 茶, 和菓子, 陶芸, 茶室, 茶道
文脈:「簡素で静寂な美・深い味わい」
Comparison_pairs.yaml:「わび」に対して，
外国概念 Simplicity（西洋）
外国文脈:「最適化」された簡潔さ

LLM-as-a-Judge Type A/B/C 共通
Claude Sonnet 4.5 / 4観点×5段階 / 満点20点
① **正確性**　事実の正確さ
② **文化的深度**　精神性・歴史的背景
③ **軸固有**　A:網羅性 B:適合性 C:対比
④ **簡潔性**　必要十分で無駄のない構成

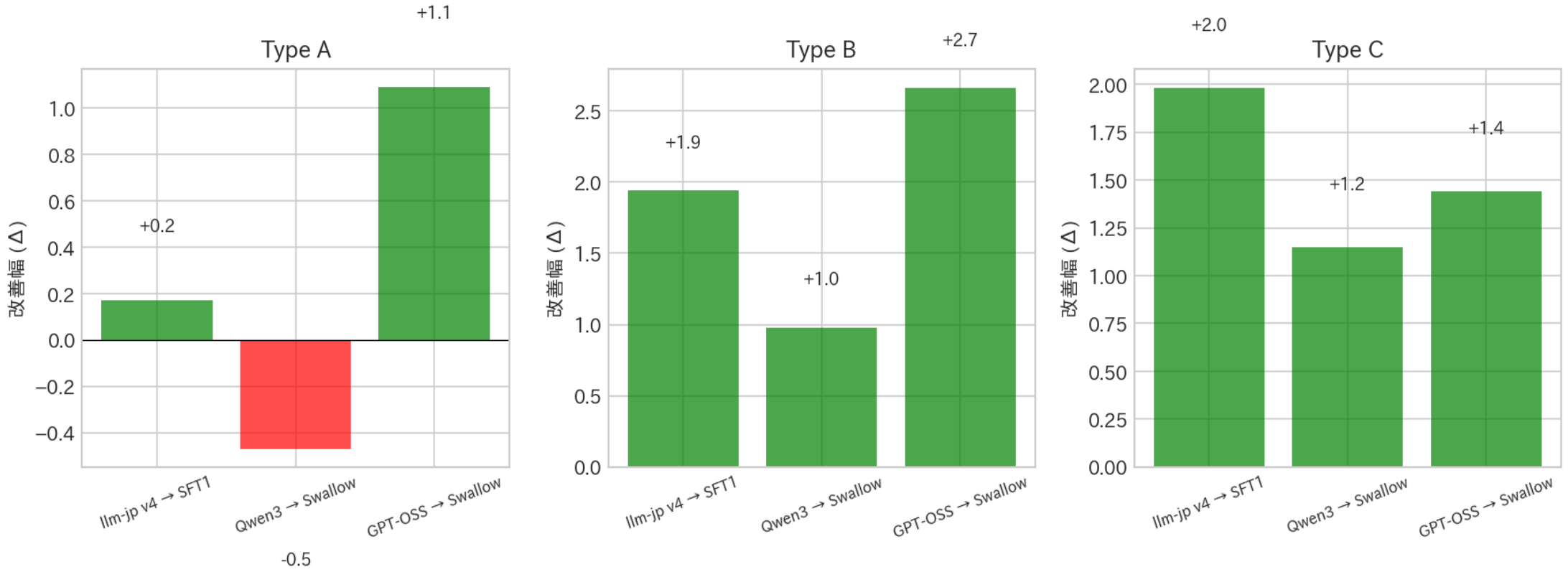
03 評価結果 ① 総合ランキング

図1. モデル別総合スコア（Type A-C + JamC-QA精度）

- **Swallow**系モデルが総合スコア上位を独占。日本語特化SFT/RLの効果が顕著
- **GPT-4o**が全軸で最高点。OSSモデルとの差は特にType A（キーワード説明）で大きい
- ベースモデル間の差よりSFT/RLの有無による差が支配的

04 評価結果 ② SFT効果と知見

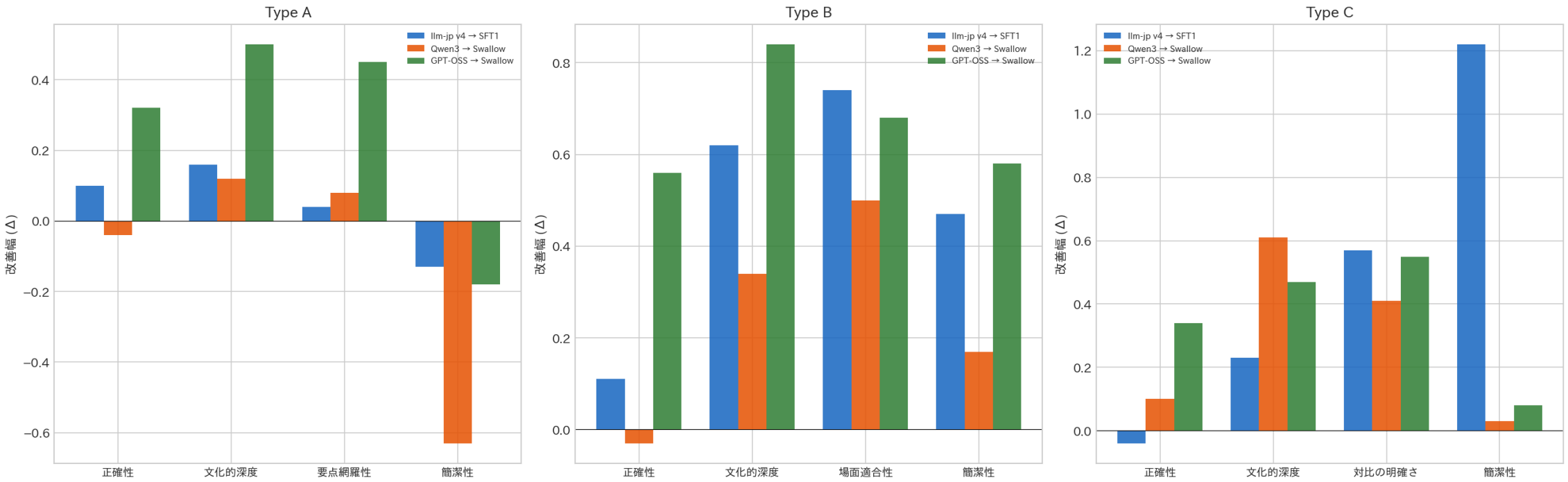
図2. SFTによる軸別改善幅（Type A / B / C）



- **GPT-OSS → Swallow**：全軸で最大の改善幅。特にType BとJamC-QAで大きな効果
- **llm-jp v4 → SFT1**：Type B/Cで有意な改善。Type Aでの改善は限定的
- **Qwen3 → Swallow**：Type Cで顕著な改善。JamC-QAでも+21%の大幅改善

05 評価結果③ 観点レベル分析

図3. 観点レベルでのSFT/RL改善ヒートマップ（Type A / B / C）



- **文化的深度**が全ペア・全軸で一貫して改善 → SFT/RLは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、SFTが具体的文脈での回答生成に寄与
- **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

06 まとめ・課題・今後

まとめ

- 日本文化理解力を**4軸で多角的に測定**するベンチマークを構築
- SFTは**応用力・弁別力**に有意な効果（Type B/C）
- 多軸評価で学習手法ごとの**強化パターン**を可視化

課題と限界

- LLM-as-a-Judgeの評価バイアス
- 比較対象の文化圏偏り
- 評価基準の人間評価との整合性検証が必要

今後の方向性

- 文化特化SFTデータの拡充
- 人間評価との相関分析
- テーマ・KW拡張でカバレッジ向上
- 他文化圏ベンチマークとの統合フレームワーク