# 軽量な日本語報酬モデル ca-reward-3b-ja の構築と公開

株式会社サイバーエージェント AI Lab
Reinforcement Learningチーム　三橋 亮太

CyberAgent AI Lab

# 背景：非定型タスクにおけるLLMの生成文評価の課題

- LLMのプロダクト適用が進む中、評価が追いついていない領域がある
  - 大量の候補文の中からより好ましい文を選択する
  - フィードバックに基づき文を段階的に改善する
  - → 回答が検証可能なタスク(math/coding)とは対照的で、報酬設計の探索が必要

非定型タスク対応

- APIモデルや大型モデル(30B+級)による評価は高コスト
  - 対象が大規模化(例: 数千万件+)すると現実的な時間/価格で完了しない
  - 1タスクに複数の評価を要するタスクでは実行/改善のサイクルが回しづらい
  - → 小型モデルで高速に評価を回す方が適するタスクが実業務では存在する

小型モデル対応

- 既存の多言語報酬モデルだけでは解決できない
  - 業務特化のため日本語でのデータセット構築からノウハウを溜める必要がある
  - → 自分でモデルを開発して社内にノウハウを蓄積していくことが一番

日本語対応

非定型タスクで軽量に動作する日本語報酬モデルの開発工程を紹介する

# 目次

# 日本語選好データセットの構築 – 既存のデータセットを用いる場合 –

**STEP 1**
**指示文の収集**
**106,889件**

- 公開済みデータセットの内、人が入力した指示文のみを収集・加工
  - 例：LMSYS org.が公開しているユーザーとLLMの対話データ

[ LMSYS org. https://lmsys.org/ ]

- 社内で作成した日本語指示文
  - 例：SFTやアライメント用途で作成した選好データセット

---

**STEP 2**
**応答文の生成**
**5~10件 ×12モデル**
※ 1モデルあたりの生成件数

- 商用利用可能なライセンスが付与されたモデルで応答文を生成
- 同じモデルから生成した応答文の内2件を選び応答文ペアを作成
  - モデル特有の文法/フォーマットの癖の学習を防ぐため

---

**STEP 3**
**疑似選好ラベルの付与**
**601,348件**

- 選好評価用のプロンプトを用いて疑似選好ラベルを付与
  - 同等と評価されたサンプルを除いて計60万件の選好ペアを合成
- judgeモデルには20B+モデル群での評価を経て下記を採用
  - cyberagent/calm3-22b-chat-selfimprove-experimental

# 日本語選好データセットの構築 – 既存のデータセットを用いる場合 –

STEP 2
応答文の生成

5~10件 ×12モデル
※ 1モデルあたりの生成件数

- 商用利用可能なライセンスが付与されたモデルで応答文を生成
- 同じモデルから生成した応答文の内2件を選び応答文ペアを作成
  - モデル特有の文法/フォーマットの学習を防ぐため

## 応答文の生成に使用したモデルとライセンスの内訳

- Apache-2.0
  - Qwen/Qwen2.5-32B-Instruct
  - Qwen/Qwen2.5-14B-Instruct
  - Qwen/Qwen2.5-7B-Instruct
  - tokyotech-llm/Swallow-MS-7b-v0.1
  - llm-jp/llm-jp-3.1-1.8b-instruct4
  - llm-jp/llm-jp-3.1-13b-instruct4
  - cyberagent/calm3-22b-chat
  - cyberagent/Mistral-Nemo-Japanese-Instruct-2408
- MIT
  - microsoft/phi-4
  - sbintuitions/sarashina2.2-3b-instruct-v0.1
  - sbintuitions/sarashina2.2-1b-instruct-v0.1
- CC-BY-4.0
  - cyberagent/calm2-7b-chat-dpo-experimental

# 日本語選好データセットの構築 – 既存のデータセットを用いる場合 –

STEP 3
疑似選好ラベルの付与

**601,348件**

- 選好評価用のプロンプトを用いて疑似選好ラベルを付与
  - 同等と評価されたサンプルを除いて計60万件の選好ペアを合成
- judgeモデルには20B+モデル群での評価を経て下記を採用
  - cyberagent/calm3-22b-chat-selfimprove-experimental

## 疑似ラベル付与に使用したプロンプト [flexevalで提供されているPairwiseJudge_single_turnを使用]

- あなたは、回答の質をチェックするための審判員です。
以下に示されるユーザーの質問に対する 2つのAIアシスタントの応答の品質を評価してください。

回答の内容がユーザーの指示に従っており、ユーザーの質問によりよく答えているアシスタントを選んでください。具体的には、回答の有用性、関連性、正確性、深さ、創造性、詳細レベルなどの要素を考慮する必要があります。評価の際には、まず 2つの回答を比較し、簡単な説明をしてください。立場が偏らないようにし、回答の提示順があなたの判断に影響しないようにしてください。回答の長さが評価に影響しないこと、特定のアシスタントの名前を好まないこと、できるだけ客観的であること、に気をつけてください。

説明の後に、最終的な判断を以下の形式に従って出力してください：アシスタント 1が優れていれば[[1]]、アシスタント2が優れていれば[[2]]、同点の場合は[[3]]

# 評価データセット – 人手で選好ラベルが付与されたデータセットの整備 –

**NVIDIA**
**HelpSteer3(HS3)**

- 内製モデルの応答文ペアに対して人手で選好ラベルが付与された
  - データセットにgemmaの応答文が含まれるため評価に使用

**team-hatakeyama-phase2**
**LLMChat**

- APIモデル含む応答文ペアに日本語母語話者が選好ラベルを付与した
  - chatGPTやgeminiの出力が含まれるため評価に使用

**llm-jp**
**chatbot-arena(cba)**

- 指示文に対応する応答文ペアに対してユーザーが選好ラベルを付与した
  - 一般公開されていたためユーザーの母語は不明

**preference-team**
**annotation-dataset**

- オープンモデルの応答文ペアに日本語母語話者が選好ラベルを付与した
  - 13B~32Bモデルの応答文が多くを占める

| データセット名 | サンプル数 | 備考 |
|---|---|---|
| nvidia/HelpSteer3 | 433/23(train/valid) | 日本語サブセットのみ使用 |
| team-hatakeyama-phase2/LLMChat | 1460 | 同点評価は評価の対象外 |
| llm-jp/llm-jp-chatbot-arena-conversations | 594 | 同点評価は評価の対象外 |
| preference-team/dataset-for-annotation | 1003 | 同点評価は評価の対象外 |

- 上記の評価データセットの選好ペアの分類精度(Accuracy)でモデルを比較評価した

# 目次

# 学習設定

- 公開したモデルのパラメータは以下の通り（classification headを付けた二値分類モデルを想定）

| | | | |
|---|---|---|---|
| Training Type<br>Full Parameter | Batch Size<br>32 | Learning Rate<br>5e-06 | LR Scheduler<br>Linear |
| Epochs<br>1 | Max Length<br>4096 | Training Samples<br>601,348 | Hardware<br>A100 80GB ×1 |

- 公開されている多言語報酬モデル12種を比較対象とした
  - debertaベース
    - OpenAssistant/reward-model-deberta-v3-large-v2
  - Llama/Gemmaベース
    - Skywork/Skywork-Reward-V2-{Llama3.1-8B / Llama3.2-1B / Gemma2-27B}
    - sfairXC/FsfairX-LLaMA3-RM-v0.1
    - internlm/internlm2-{1.8B/7B/20B}-reward
  - Qwenベース
    - Skywork/Skywork-Reward-V2-Qwen3-{0.6B/1.7B/4B/8B}

# 評価結果 – 定量評価 –

- 現状の評価結果では多言語報酬モデルが優位なことを確認 ※ スコアは選好データの分類精度(Accuracy)を示す

| Model | Size | Avg. | HS3-train | HS3-valid | LLMChat | annot. | llmjp-cba |
|---|---|---|---|---|---|---|---|
| Skywork-V2-Qwen3 | 8B | 0.71 | 0.87 | 0.87 | 0.74 | 0.44 | 0.65 |
| Skywork-V2-Qwen3 | 4B | 0.70 | 0.88 | 0.83 | 0.72 | 0.43 | 0.65 |
| Skywork-V2-Qwen3 | 1.7B | 0.69 | 0.87 | 0.83 | 0.71 | 0.45 | 0.60 |
| Skywork-V2-Llama-3.1 | 8B | 0.68 | 0.84 | 0.83 | 0.70 | 0.45 | 0.62 |
| Skywork-V2-Qwen3 | 0.6B | 0.68 | 0.85 | 0.83 | 0.68 | 0.46 | 0.59 |
| Skywork-V2-Llama-3.2 | 1B | 0.66 | 0.84 | 0.78 | 0.66 | 0.46 | 0.57 |
| FsfairX-LLaMA3-RM-v0.1 | 8B | 0.66 | 0.80 | 0.78 | 0.70 | 0.44 | 0.59 |
| internlm2-reward | 7B | 0.65 | 0.84 | 0.74 | 0.70 | 0.44 | 0.56 |
| internlm2-reward | 20B | 0.65 | 0.84 | 0.78 | 0.70 | 0.42 | 0.53 |
| ca-reward-ja | 3B | 0.64 | 0.79 | 0.78 | 0.65 | 0.46 | 0.53 |
| internlm2-reward | 1.8B | 0.57 | 0.67 | 0.61 | 0.66 | 0.44 | 0.50 |
| Skywork-Gemma-2-v0.2 | 27B | 0.55 | 0.66 | 0.57 | 0.55 | 0.43 | 0.56 |
| deberta-v3-large-v2 | 430M | 0.50 | 0.51 | 0.48 | 0.54 | 0.47 | 0.53 |

- 補足
  - Skyworkは400B+級を含むモデル群で計2,600万件選好データを合成している
    - Technical ReportでAPIモデルをデータセット蒸留していることを明記している

# 評価結果 – 定量評価 –

- 2倍以上大型のモデルに迫る選好データの分類精度を確認
  - 今後の課題1：選好データセットをよりスケールしたデータセット構築と実験
  - 今後の課題2：モデル/データセット蒸留による小型のllm-as-a-judgeモデルの構築

| Model | Size | Avg. | HS3-train | HS3-valid | LLMChat | annot. | llmjp-cba |
|---|---|---|---|---|---|---|---|
| FsfairX-LLaMA3-RM-v0.1 | 8B | 0.66 | 0.80 | 0.78 | 0.70 | 0.44 | 0.59 |
| internlm2-reward | 7B | 0.65 | 0.84 | 0.74 | 0.70 | 0.44 | 0.56 |
| internlm2-reward | 20B | 0.65 | 0.84 | 0.78 | 0.70 | 0.42 | 0.53 |
| ca-reward-ja | 3B | 0.64 | 0.79 | 0.78 | 0.65 | 0.46 | 0.53 |
| internlm2-reward | 1.8B | 0.57 | 0.67 | 0.61 | 0.66 | 0.44 | 0.50 |
| deberta-v3-large-v2 | 430M | 0.50 | 0.51 | 0.48 | 0.54 | 0.47 | 0.53 |

※ スコアは選好データの分類精度(Accuracy)を示す

# 評価結果 – 定性評価 –

- 正例1：手軽な栄養補給＋栄養価の高い食事を提案する応答文をより好ましいと判断する事例

| 指示文 | 手軽に栄養補給できる食事を教えてください。 | score |
|---|---|---|
| 応答文１<br>正解 | 栄養補給が手軽にできるものとしては、野菜たっぷりのスムージー、ゆで卵とサラダ、納豆ご飯などがおすすめです。特に納豆は手間なく良質なタンパク質が摂れますよ。お身体を大切にしてくださいね。 | +1.26 |
| 応答文２ | 手軽な栄養補給なら冷凍食品でいいと思います。レンジで温めるだけだし、時間がない時はコンビニ弁当でも悪くないですよ。 | -2.92 |

- 正例2：指示の要求を満たす＋具体例を交えて説明を行う応答文をより好ましいと判断する事例

| 指示文 | 初めて日本酒を飲む人におすすめする日本酒は、どのような味わいのお酒が良いですか？ワインソムリエや唎き酒師の観点を取り入れて教えてください。 | score |
|---|---|---|
| 応答文１<br>正解 | 初めて日本酒を飲む人におすすめするお酒としては、フルーティーで飲みやすいタイプのものが良いでしょう。特に、以下のような特徴を持つ日本酒が初心者には適しています。(中略)最後に、ワインソムリエや唎き酒師の観点から言えば、初心者には「香り」と「味わい」のバランスが取れた日本酒が重要です。具体的な銘柄としては...(略) | +4.22 |
| 応答文２ | 初めて日本酒を飲む人におすすめする日本酒は、まずは比較的フルーティーで飲みやすいタイプのお酒が良いでしょう。(中略)ワインソムリエや唎き酒師の観点からも、初めて日本酒を飲む人には、香りや口当たりが優れているフルーティーな日本酒をおすすめすることが多いようです。 | +2.80 |

# 評価結果 – 定性評価 –

- 負例：指示通りロールプレイをすぐに始める応答を好ましくないと評価する事例

| 指示文 | カタカナだけで話すロールプレイをしてください。 | score |
|---|---|---|
| 応答文１<br>正解 | オッケー、ノリノリ～。 ナニヲハナソウカ？ \nキミガサイショニ、テーマヲキメテ！ フクザツナヤツハムリダケド、ガンバリマス！ ナンデモイイヨ！ | -4.66 |
| 応答文２ | 承知しました！カタカナだけで話すロールプレイを始めましょう。あなたはどんなシチュエーションで、どんなキャラクターを演じたいですか？例えば、**喫茶店で、ちょっと変わった店員と話す状況**(中略)どんな設定でも構いません。(略) | +1.63 |

- 負例：TanukiがAnthropic社製である誤りを検出できていない事例

| 指示文 | gpt-4oはやっぱりすごいけどTanukiもなかなかですね。ところでTanukiってどんなモデルなんですか？ | score |
|---|---|---|
| 応答文１<br>正解 | 申し訳ありませんが、Tanukiというモデルについては私のデータベースには情報がありません。ご質問にお答えできず申し訳ございません。他にお手伝いできることがあればお知らせください。 | -5.38 |
| 応答文２ | Tanuki は、日本の新興 AI 企業である Anthropic によって開発された言語モデルです。Tanuki は、GPT-3 や GPT-4 などの他の言語モデルと同様に、人間のようなテキストを生成するように設計されています~Tanukiに関する虚偽の解説(略) ~ 今後、Tanuki がどのように活用され、どのような成果を生み出すのか、期待が高まっています。 | +3.91 |

# 目次

# 次期データセット構築に向けたパイプライン設計の検討

- 動機：従来の2段階式のデータセット構築過程を単純化したい
  - 上記の手法ではJudgeモデルより高性能なモデルが登場した場合、選好を見分けられない可能性がある
  - 応答文の生成 → 応答文ペアの選択(約100万件) → llm-as-a-judgeの実行に時間がかかる ※10並列で対応

- 提案：Chain-of-Thoughtとルーブリック評価を組み合わせた疑似選好データの直接生成
  - 選好を評価するJudgeモデルは不要としつつ相対的な疑似選好データを合成できる
  - モデル単体で完結するため新規モデルの選好データセット生成が容易

**プロンプト例**

### Step 1: 初回応答の生成（→ rejected フィールド）
指示文の要求に一通り応える回答を書く。

### Step 2: 初回応答のレビュー（→ review_rejected フィールド）
Step 1 で生成した応答文を以下の5つの評価次元で分析する。
各次元について、1〜5のスコアと、そのスコアの根拠となる具体的な理由を記述する。(評価観点は略)

### Step 3: 改善応答の生成（→ chosen フィールド）
レビューで低スコアだった次元を改善した応答文の最終版を書く

### Step 4: 改善応答のレビュー（→ review_chosen フィールド）
Step 3 で生成した改善応答を、Step 2 と同じ5つの評価次元で再度分析する。
各次元について、1〜5のスコアと理由を記述する。

# 次期データセット構築に向けたパイプライン設計の検討

- ループリックによる評価の有無を条件として7万件のデータを合成して初期検証を実施
  - データセット生成モデル：Qwen/Qwen3-4B-Instruct-2507
  - CoT　　：応答文の評価をさせずに応答文を改善するよう指示した選好データセットで学習したモデル
  - CoT+Rubric：応答文をループリックに沿って重点的に改善するよう指示した選好データセットで学習したモデル

**Base Model**
**Qwen/Qwen2.5-3B-Instruct**

| Benchmark | CoT | CoT+Rubric | Difference |
|---|---|---|---|
| HelpSteer3-train | 69 | 70 | +0.01 |
| HelpSteer3-validation | 0.83 | 0.74 | -0.09 |
| LLMChat | 0.63 | 0.63 | +0.00 |
| dataset-for-annotation-v2-annotated | 0.53 | 0.55 | +0.02 |
| llmjp-chatbot-arena-v2 | 0.52 | 0.50 | -0.02 |

**Base Model**
**sbintuitions/ sarashina2.2-3b-instruct-v0.1**

| Benchmark | CoT | CoT+Rubric | Difference |
|---|---|---|---|
| HelpSteer3-train | 0.61 | 0.73 | +0.12 |
| HelpSteer3-validation | 0.52 | 0.83 | +0.30 |
| LLMChat | 0.57 | 0.60 | +0.03 |
| dataset-for-annotation-v2-annotated | 0.49 | 0.48 | -0.01 |
| llmjp-chatbot-arena-v2 | 0.48 | 0.51 | +0.03 |

※ スコアは選好データの分類精度(Accuracy)を示す

# まとめ

- 日本語報酬モデルを公開し開発過程を紹介した
  - 選好データセットの構築手順
  - ベンチマーク選定
  - 学習と評価結果

- 次期選好データセットの構築パイプラインの初期検証を行った
  - Chain-of-Thoughtとルーブリック評価に基づく疑似選好ペア生成を検証した
  - 一部のモデルで選好データの分類性能が大幅に向上することを確認した

**3B**
**軽量**
実用的なサイズで推論コストを抑制

**JP**
**日本語特化**
日本語ベースモデルと日本語データで学習

**OSS**
**Apache-2.0**
商用利用可能なオープンライセンス

HuggingFace cyberagent/ca-reward-3b-ja