ChatGPT部
Produced by NOB DATA

日本語LLM評価の気になる２点

於 ChatGPT部 #30 ～LT & フリートーク会～

2023年11月18日(土)

太田 博三

背景と問題意識

1つ目：

・10月のJGLUEの評価(大城さん NOB DATA(株))に参加して

・11/15のW&B 東京ミートアップ #8に参加して
（LLMモデルの評価方法 - ）

2つ目：

・英語のLLM評価データを日本語に翻訳したものが、文化的な相違があっても、大きな影響があるのか！？という疑問

※LLMのための日本語インストラクションデータ作成 Riken-aip

気になる2点

1つ目：

・10月のJGLUEの評価(大城さん NOB DATA(株))に参加して NOB DATA

・11/15のW&B 東京ミートアップ #8に参加して
（LLMモデルの評価方法 - ）

→生成タスクの自動評価方法がGPT-4であること

2つ目：

・英語のLLM評価データを日本語に翻訳したものが、文化的な相違があっても、大きな影響があるのか！？という疑問

※LLMのための日本語インストラクションデータ作成 Riken-aip

→日本の文化や文法的な相違があるが、そんな大きな影響があるのか？

→日本語LLM評価を作る必要はあるが、効果は大きいのか？

考察（1点目）

1つ目：
・10月のJGLUEの評価(大城さん NOB DATA(株))に参加して NOB DATA
・11/15のW&B 東京ミートアップ #8に参加して
（LLMモデルの評価方法 - ）

→生成タスクの自動評価方法がGPT-4であること

・

最近の日本語リーダーボード・ベンチマーク

	問題作成	問題数	タスク種別	評価方法
lm-evaluation-harness	×	-	分類・生成	自動
Nejumi	×	-	分類	自動
Rakuda	○	40	生成	自動
Japanese VicunaQA	○	80	生成	自動
Japanese MT-Bench	○	80	生成	自動
ELYZA-tasks-100	○	100	生成	人手・(自動)

・生成タスクで自動評価のタスクは4つのうち、３つでした。
→次のスライドでそのタスクの中身を見てゆきます

最近の日本語リーダーボード・ベンチマーク

	問題作成	問題数	タスク種別	評価方法
lm-evaluation-harness	×	-	分類・生成	自動
Nejumi	×	-	分類	自動
Rakuda	○	40	生成	自動
Japanese VicunaQA	○	80	生成	自動
Japanese MT-Bench	○	80	生成	自動
ELYZA-tasks-100	○	100	生成	人手・(自動)

考察（1点目）

Rakuda (YuzuAI)

日本の地理、政治、歴史、社会に関する40問(人手作成)

- 自動評価(GPT-4)
- ペア比較(2種類の提示順)

Rakudaの問題の例

日本の「三位一体改革」について述べ、その経済に対する影響について解説してください。

戦後の日本政治において最も影響力のあった政治家を一人挙げ、
その貢献について詳しく述べてください。

Rank	Model	Strength	Stronger than the next model at confidence level
1	gpt-4	1472 ± 49	97.5%
2	claude-2	1353 ± 42	89.3%
3	gpt-3.5	1285 ± 37	100.0%

考察（1点目）

最近の日本語リーダーボード・ベンチマーク

	問題作成	問題数	タスク種別	評価方法
lm-evaluation-harness	×	-	分類・生成	自動
Nejumi	×	-	分類	自動
Rakuda	○	40	生成	自動
Japanese VicunaQA	○	80	生成	自動
Japanese MT-Bench	○	80	生成	自動
ELYZA-tasks-100	○	100	生成	人手・(自動)

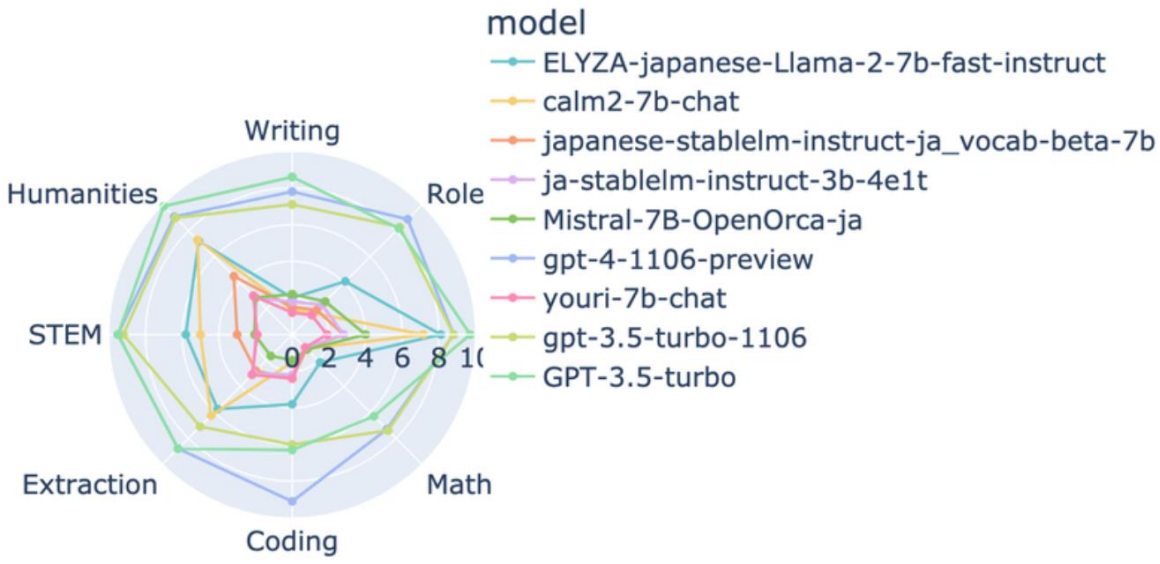
- MT-Benchを翻訳、日本の文化に合うように修正したもの

マルチターン会話能力、指示に従う能力を問う80問
- 8カテゴリ(10問ずつ)

- writing, roleplay, reasoning, math, coding,extraction, など

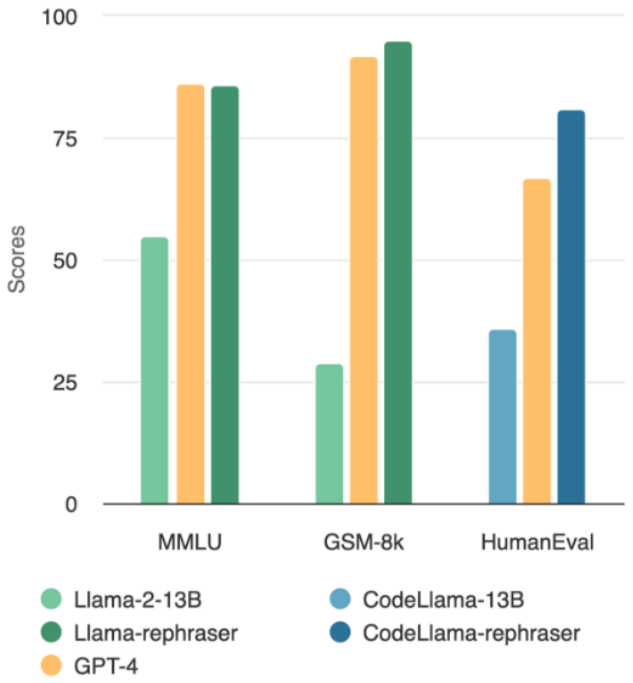
Japanese MT-Benchの問題の例

新入社員へのビジネスメールのエチケットについての指導書を作成してください。敬語の正しい使い方や、日本のビジネス文化での注意点を取り入れてください。



考察（1点目）

・OSSのLlama-rephraser を発表: 13B モデルが
主要ベンチマーク (MMLU/GSK-8K/HumanEval) で
GPT-4 パフォーマンスに到達!しました！



→OpenAI の汚染除去方法に従いましたが、データ汚染の証拠は
見つかりませんでした。
→何かトリックを使ったのではないか？という疑惑がありました。
→ 評価データが学習に使われている可能性がある と考えられます！
eg.テスト セットを書き直すだけで、大幅に高いベンチマーク
パフォーマンスが上がるようです。
テストサンプルを言い換えたり、別の言語に翻訳したりするだけでできる。

cf. “Catch me if you can! How to beat GPT-4 with a 13B model” [Blog]
Shuo Yangらの寄稿(2023 年 11 月 14 日)

まとめ：考察（1点目）

・自動評価がGPT-4になっているため、GPT-4のためのGPT-4の書き写しになっているようでした。

・翻訳や一部言い換えで、高いパフォーマンスの数値に到達できるが、
本当に、これで日本語LLMの評価になっているのでしょうかという疑問も残りました。

※個人的には、GPT-4を超えるOSSも
出てきて欲しいです。
→では、どうしたらよいのでしょうか？

→→日本語に置き換えると、
ネイティブよりも大きな影響があるのかでしょうか？
（次のスライドへ：考察（２点目））

最近の日本語リーダーボード・ベンチマーク

	問題作成	問題数	タスク種別	評価方法
lm-evaluation-harness	×	-	分類・生成	自動
Nejumi	×	-	分類	自動
Rakuda	○	40	生成	自動
Japanese VicunaQA	○	80	生成	自動
Japanese MT-Bench	○	80	生成	自動
ELYZA-tasks-100	○	100	生成	人手・(自動)

考察（２点目）

2つ目：

・英語のLLM評価データを日本語に翻訳したものが、文化的な相違があっても、大きな影響があるのか！？という疑問

※LLMのための日本語インストラクションデータ作成 Riken-aip

→日本の文化や文法的な相違があるが、そんな大きな影響があるのか？

→日本語LLM評価を作る必要はあるが、効果は大きいのか？

英語から日本語へ翻訳する際の注意点

１）慣用句の認識："Kick the bucket"（亡くなる）を「バケツを蹴る」

２）文化の有無：アメリカの「Thanksgiving」（感謝祭）は日本にはない

３）同音異義語や多義語の誤訳：“bat”は「コウモリ」も「バット」も意味しますが、文脈によって日本語訳が異なる。

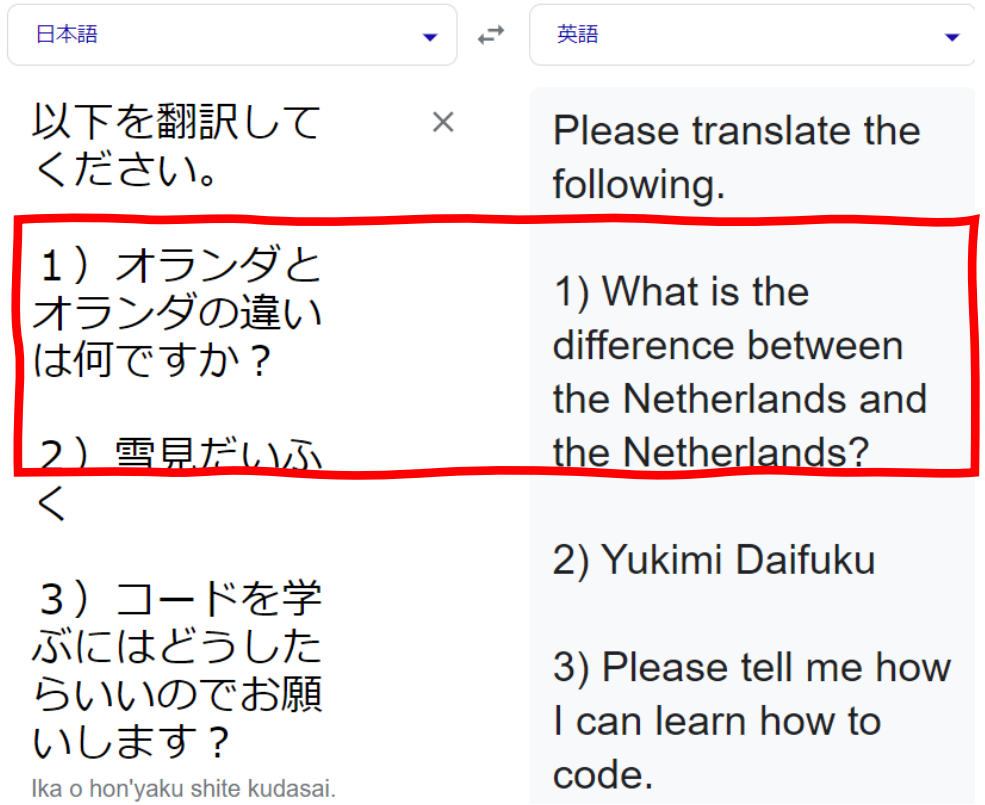
https://chat.openai.com/share/0993a999-1b23-4a1e-8b61-2de470cf3fe7

考察（２点目）

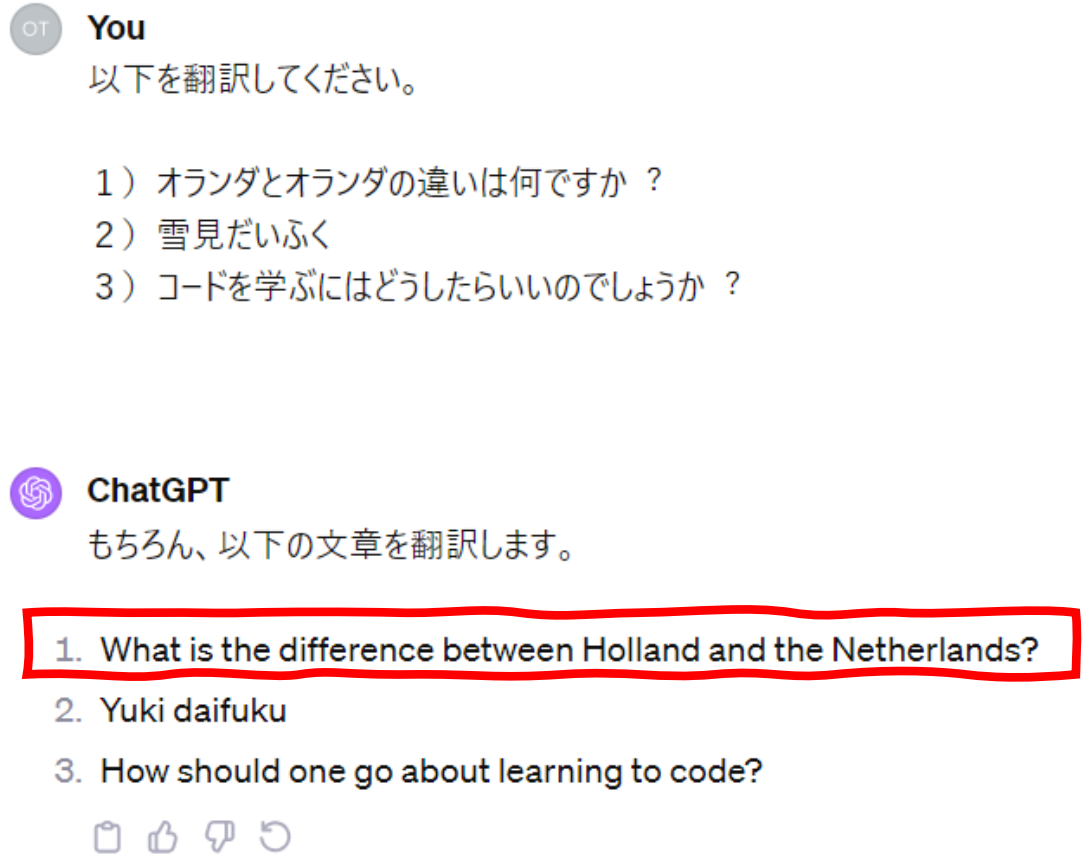
2つ目：
・英語のLLM評価データを日本語に翻訳したものが、文化的な相違があっても、大きな影響があるのか！？という疑問
※LLMのための日本語インストラクションデータ作成 Riken-aip

→日本の文化や文法的な相違があるが、そんな大きな影響があるのか？
→日本語LLM評価を作る必要はあるが、効果は大きいのか？

・Google翻訳で確認してみると…



間違い

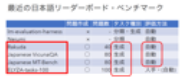


正解

→(日本語の)同音多義語の知識(情報)が必要。

考察（２点目）

考察（1点目）



- MT-Benchを翻訳、日本の文化に合うように修正したもの

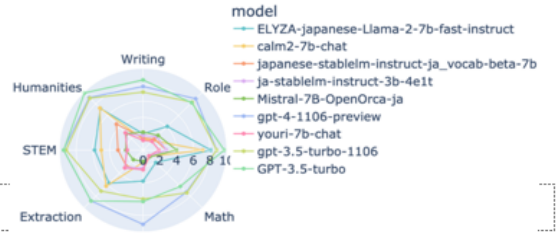
マルチターン会話能力、指示に従う能力を問う80問
- 8カテゴリ(10問ずつ)

- writing, roleplay, reasoning, math, coding,extraction, など

- テキストを入力

Japanese MT-Benchの問題の例

新入社員へのビジネスメールのエチケットについての指導書を作成してください。敬語の正しい使い方や、日本のビジネス文化での注意点を取り入れてください。



- Japanese MT-Benchの「日本の文化に合うように修正したもの(スライド6枚目)」は、以下の3点の修正のようです。

英語から日本語へ翻訳する際の注意点

1）慣用句の認識："Kick the bucket"（亡くなる）を「バケツを蹴る」

2）文化の有無：アメリカの「Thanksgiving」（感謝祭）は日本にはない

3）同音異義語や多義語の誤訳：“bat”は「コウモリ」も「バット」も意味しますが、文脈によって日本語訳が異なる。

⇒主に英語圏のLLMを日本語LLMにするインパクトはそれほど大きくはないようです。

※左の3点以外にあれば、教えてください！

まとめ

気になる2点

1つ目：
・10月のJGLUEの評価(大城さん NOB DATA(株))に参加して NOB DATA
・11/15のW&B 東京ミートアップ #8に参加して
（LLMモデルの評価方法 - ）

→生成タスクの自動評価方法がGPT-4であること

2つ目：
・英語のLLM評価データを日本語に翻訳したものが、文化的な相違があっても、大きな影響があるのか！？という疑問
※LLMのための日本語インストラクションデータ作成 Riken-aip

→日本の文化や文法的な相違があるが、そんな大きな影響があるのか？
→日本語LLM評価を作る必要はあるが、効果は大きいのか？

- 1つ目：OSSのLLMは、生成タスクの自動評価がGPT-4で行われているので、スコアをあげるなら、評価データセットの言い換えや翻訳をするとできそうです。
- 2つ目：異文化の差異や慣用句の差異は、確かに見受けられたが、ボリューム感はそれほど大きくはないように思われました。

→日本語LLM評価は若干、出来レースなフェーズに入った感じがしました。
※OSSがGPT-4を抜く日は来るのでしょうか～

ご清聴ありがとうございました

参考文献・URL一覧

1. LLMモデルの評価方法 - W&B 東京ミートアップ #8 - connpass
https://wandb.connpass.com/event/300670/
2. NOB DATA株式会社 https://nobdata.co.jp/
3. LLMのための日本語インストラクションデータ作成プロジェクト – RIKEN-AIP, LIAT
4. The Rakuda Ranking of Japanese AI https://yuzuai.jp/benchmark
5. Nejumi LLMリーダーボード | LLM_evaluation_Japan – Weights & Biases
https://wandb.ai/wandb/LLM_evaluation_Japan/reports/Nejumi-LLM---Vmlldzo0NTUzMDE2?accessToken=u1ttt89al8oo5p5j12eq3nldxh0378os9qjjh14ha1yg88nvs5irmuao044b6eqa
6. The Rakuda Ranking of Japanese AI https://yuzuai.jp/benchmark