4択クイズを題材にした多肢選択式日本語質問応答データセットの構築

鈴木正敏
東北大学／Studio Ousia

概要

- 日本語言語モデルの評価に利用できる
  多肢選択式の日本語質問応答データセットを作成
  - 国内のクイズ大会で使用された4択クイズの問題が題材
    ⇒ 日本特有の事物に関する知識を多く含む
- 作成したデータセットを用いて
  既存の日本語言語モデルの性能を評価
- 作成したデータセット（一部）と評価スクリプトを公開
  - 🤗Hub: tohoku-nlp/abc-multiple-choice
  - GitHub: cl-tohoku/abc-multiple-choice

目次

背景

- 言語モデルが保持する世界知識を測るために
  高品質な質問応答データセットは不可欠
  - 世界知識＝実世界の事物に関する知識
- ここ1, 2年で日本語の大規模言語モデルが多く登場
  それらモデルがもつ世界知識を評価したい
- しかし, 日本語の高品質な質問応答データセットは稀少
  - 文章読解を伴わずに知識を試すデータが必要
  - 自動評価のためには択一問題の形式が好都合

本研究の目的

日本語の言語モデルがもつ世界知識を評価するための
多肢選択式質問応答データセットを作成する

本研究の貢献

- 国内のクイズ大会で使用された4択クイズの
  問題をもとにデータセットを作成
  - 参加者の知識を試す目的で作られた高品質なデータ
  - 日本特有の事物に関する問題を多く含む
- 作成したデータセットを用いて
  既存の日本語の大規模言語モデルの性能を評価

目次

日本語の質問応答データセット

（日本語の大規模言語モデルの評価に使われている主なもの）

- JCommonsenseQA [Kurihara+ 22, 栗原+ 22]
  - 常識推論能力を評価するための5択問題のデータセット
- JEMHopQA [石井+ 23, 石井+ 24]
  - 多段階の推論を必要とする質問からなるデータセット
- JSQuAD [Kurihara+ 22 , 栗原+ 22], JaQuAD [So+ 22]
  - Wikipedia記事の読解問題のデータセット
- NIILC [関根 03]
  - 百科事典を対象とした質問応答データセット
- JAQKET [鈴木+ 20], AI王データセット[1]
  - クイズ問題を題材に作成した20択 or 一問一答形式のデータセット

⇒ 世界知識を評価するための択一問題のデータセットはほぼない

1. https://sites.google.com/view/project-aio/dataset

多言語の質問応答データセットの難点

- 翻訳の品質に問題
  - 例: X-CSQA [Lin+ 21]

笛が溜まりそうな場所は？
A. パーティー　B. オーケストラ　C. 楽器店　D. マーチングバンド　E. シンフォニー

何かの重さが軽くないとしたら、それは何か？
A. 肝心なこと　B. 闇　C. 重い　D. 煩い　E. 闇

- 日本特有の知識が少なく, 日本語モデルの評価に不向き
  - 例: MKQA [Longpre+ 21]

ツインタワーが建てられるまでどの位の時間がかかりましたか

ラムズはいつスーパーボウルでプレーしましたか

目次

提案データセット

- 競技クイズの大会「abc」[1]で実際に使用された4択クイズの問題を利用
- 第10回（2010年）から第21回（2023年）までの大会で使用された計1,500問からデータセットを作成
  - 第12回大会（2012年）までの問題（450問）は無料で入手可能
    ⇒ データセットを公開: 🤗 tohoku-nlp/abc-multiple-choice
  - 第13回大会（2013年）以降の問題は有料
    ⇒ 1,500問の全データは公開せず, 実験結果のみ報告

※本発表資料で使用しているクイズ問題例はいずれも公開したデータセットに含まれるものです

1. https://abc-dive.com/portal/

クイズ問題の例

- 日本や世界の事物に関する知識を試す問題が多数

『斜陽』『人間失格』などの作品を残した作家は?			
1. 三島由紀夫	2. 田中英光	3. 有島武郎	**4. 太宰治**

アメリカの大都市・ボストンがある州は?			
1. マサチューセッツ州	2. ニューハンプシャー州	3. メリーランド州	4. コネチカット州

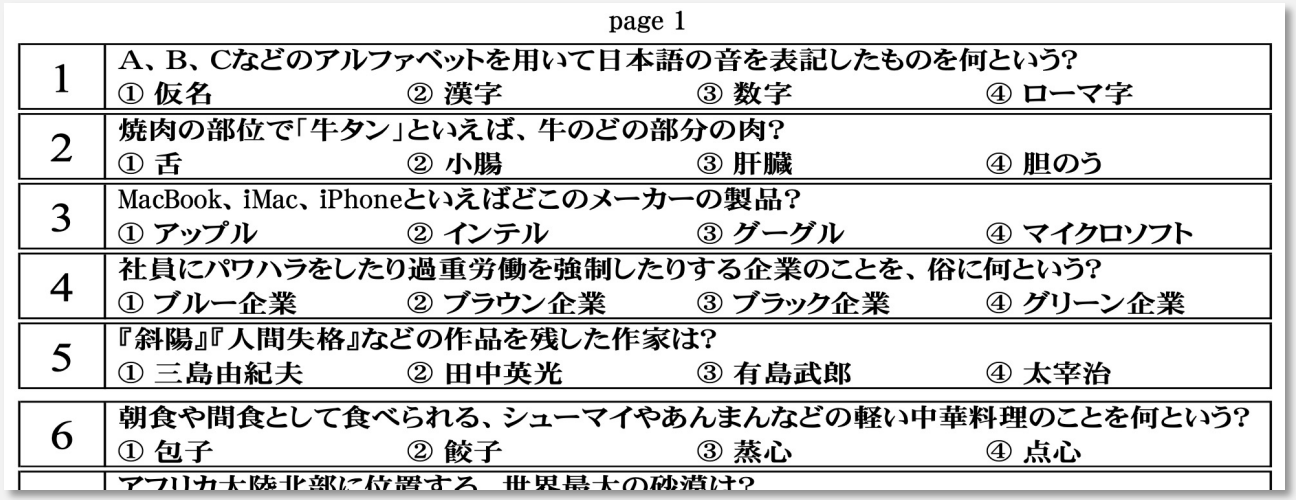
アニメ『サザエさん』で、磯野家の隣に住んでいるのは?			
1. 中島家	2. 花沢家	3. 穴子家	**4. 伊佐坂家**

脚の縞模様が美しいことから「森の貴婦人」とも呼ばれる動物は?			
1. ジャイアントパンダ	**2. オカピ**	3. コビトカバ	4. ボンゴ

神戸市にはその個人美術館もある、劇団「天井桟敷」のポスターで知られる画家・デザイナーは?			
1. 福田繁雄	**2. 横尾忠則**	3. 亀倉雄策	4. 粟津潔

余談: データセット構築の苦労話（？）

- 問題データはPDFで提供されている（一部Excelもあり）
  - Copy & Paste によるテキスト抽出が上手くいかないことも😕



page 1

1	A、B、Cなどのアルファベットを用いて日本語の音を表記したものを何という? ① 仮名　② 漢字　③ 数字　④ ローマ字
2	焼肉の部位で「牛タン」といえば、牛のどの部分の肉? ① 舌　② 小腸　③ 肝臓　④ 胆のう
3	MacBook、iMac、iPhoneといえばどこのメーカーの製品? ① アップル　② インテル　③ グーグル　④ マイクロソフト
4	社員にパワハラをしたり過重労働を強制したりする企業のことを、俗に何という? ① ブルー企業　② ブラウン企業　③ ブラック企業　④ グリーン企業
5	『斜陽』『人間失格』などの作品を残した作家は? ① 三島由紀夫　② 田中英光　③ 有島武郎　④ 太宰治
6	朝食や間食として食べられる、シューマイやあんまんなどの軽い中華料理のことを何という? ① 包子　② 餃子　③ 蒸心　④ 点心
	アフリカ大陸北部に位置する、世界最大の砂漠は?

『「abc the12th / EQIDEN2014」 公式問題集』より引用

- 解決策: スクリーンショットをOCR→目視で確認・修正
  - OCR機能は最新のmacOSに標準で備わっている（便利で高精度）

古くなった問題のフィルタリング

- 時事問題など, 現在では意味が成立しない問題が存在
  - 「今年」などの文言を含むもの, 事実関係が変わったもの, etc.

今年8月25日に佐藤健、武井咲主演による実写版映画が公開される、和月伸宏の人気マンガは?			
1.『武装錬金』	2.『エンバーミング』	**3.『るろうに剣心』**	4.『GUN BLAZE WEST』

（出題日: 2012/03/18）

現在15日ある日本の国民の祝日のうち、ちょうど真ん中にあたる8番目の祝日は何?			
1. 昭和の日	2. 憲法記念日	3. みどりの日	**4. こどもの日**

（出題日: 2013/03/24; 2016年に「山の日」が加わり, 2020年に「天皇誕生日」が移動したため, 不成立）

- 解決策: 1,500問全てに目を通し, 成立しない問題を除外

※気合いが必要

- フィルタリング後の問題数は1,340問（89.3%）

目次

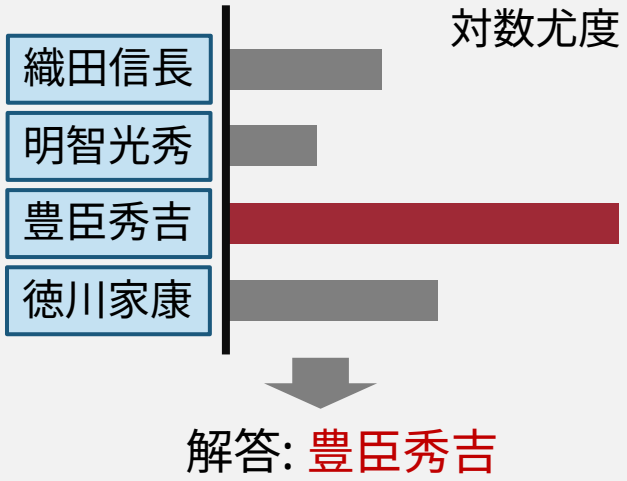
評価方法 (1) 対数尤度

- プロンプトに対して各選択肢の語句を出力させ，
  最も対数尤度が高い選択肢を解答として採用する

与えられた選択肢の中から、最適な答えを選んでください。

質問：**太閤検地や刀狩などの政策を行った、安土桃山時代の武将は?**
選択肢：
- **織田信長**
- **明智光秀**
- **豊臣秀吉**
- **徳川家康**

回答：



- **lm-evaluation-harness**[1] を使用
  - プロンプトは JCommonsenseQA 向けのものと同様に作成
    - プロンプトバージョンは 0.2.1 で実験（参考: Stability AI ブログ[2]）

1. https://github.com/EleutherAI/lm-evaluation-harness/tree/v0.4.1
2. https://ja.stability.ai/blog/japanese-stable-lm-beta

評価方法 (2-1) 選択肢の番号を出力

- プロンプトに対して解答の選択肢の番号を出力させる

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。

指示:
質問と回答の選択肢を入力として受け取り、選択肢から回答を選択してください。なお、回答は選択肢の番号（例：1）でするものとします。 回答となる数値をint型で返し、他には何も含めないことを厳守してください。

入力:
質問：**太閤検地や刀狩などの政策を行った、安土桃山時代の武将は?**
選択肢：1.**織田信長**,2.**明智光秀**,3.**豊臣秀吉**,4.**徳川家康**

応答:

出力
3

- **llm-jp-eval** [Han+ 24][1] を使用
  - 選択肢の番号を答えさせるのは LLM-jp で通常の問題設定
  - プロンプトは JCommonsenseQA 向けのものと同様に作成

1. https://github.com/llm-jp/llm-jp-eval/tree/v1.2.0

評価方法 (2-2) 選択肢の語句を出力

- プロンプトに対して解答の選択肢の語句を出力させる
  - 🤔 選択肢の番号で答えるのは一種の創発的能力なのかも？

以下は、タスクを説明する指示と、文脈のある入力の組み合わせです。要求を適切に満たす応答を書きなさい。

指示:
質問と回答の選択肢を入力として受け取り、選択肢から回答を選択してください。回答の他には何も含めないことを厳守してください。

入力:
質問：**太閤検地や刀狩などの政策を行った、安土桃山時代の武将は?**
選択肢：**織田信長,明智光秀,豊臣秀吉,徳川家康**

応答:

出力
豊臣秀吉

- **llm-jp-eval** を使用
  - 選択肢の番号ではなく語句で答えるようプロンプトを変更

評価方法の詳細

- いずれの評価も 4-shot で実行
  - 例題は過去の大会の筆記問題をもとに作成（詳細は付録を参照）
- 1,500問の全データに対して評価
  - 450問の公開データのみを用いた場合の結果は付録を参照
- 評価指標: 正解率（完全一致）
  - 古くなった問題のフィルタリングの有無の2通りで算出
- 評価に用いたスクリプトは GitHub にて公開
  - https://github.com/cl-tohoku/abc-multiple-choice

評価対象のモデル

- 日本語のモデル
  - cyberagent/calm2-7b
  - elyza/ELYZA-japanese-Llama-2-7b, 13b
  - llm-jp/llm-jp-13b-v1.0
  - stabilityai/japanese-stablelm-base-beta-7b, gamma-7b
  - stockmark/stockmark-13b
  - tokyotech-llm/Swallow-7b-hf, 13b
  - matsuo-lab/weblab-10b
- 英語のモデル
  - meta-llama/Llama-2-7b-hf, 13b
  - mistralai/Mistral-7B-v0.1
- OpenAI API のモデル
  - gpt-3.5-turbo-0125
  - gpt-4-0125-preview

結果

カッコ内は古い問題をフィルタリング後の正解率
太字は1位, 下線は2位を示す

モデル	(1) 対数尤度	(2-1) 番号を出力	(2-2) 語句を出力
calm2-7b	0.622 (0.625)	0.240 (0.243)	0.573 (0.575)
ELYZA-japanese-Llama-2-7b	0.449 (0.453)	0.376 (0.380)	0.405 (0.408)
ELYZA-japanese-Llama-2-13b	0.586 (0.592)	0.523 (0.524)	0.565 (0.569)
llm-jp-13b-v1.0	0.526 (0.521)	0.246 (0.250)	0.448 (0.449)
japanese-stablelm-base-beta-7b	0.498 (0.501)	0.365 (0.364)	0.469 (0.479)
japanese-stablelm-base-gamma-7b	0.719 (0.726)	0.643 **(0.649)**	0.653 (0.664)
stockmark-13b	0.771 (0.762)	0.289 (0.285)	0.000 (0.000)
Swallow-7b-hf	0.754 (0.749)	0.459 (0.455)	0.640 (0.641)
Swallow-13b-hf	**0.812 (0.813)**	**0.649** (0.648)	**0.751 (0.749)**
weblab-10b	0.336 (0.330)	0.270 (0.278)	0.327 (0.325)
Llama-2-7b-hf	0.401 (0.403)	0.267 (0.263)	0.305 (0.311)
Llama-2-13b-hf	0.495 (0.507)	0.378 (0.380)	0.412 (0.427)
Mistral-7B-v0.1	0.383 (0.392)	0.354 (0.365)	0.379 (0.388)
gpt-3.5-turbo-0125	-	0.645 (0.657)	0.699 (0.710)
gpt-4-0125-preview	-	**0.853 (0.861)**	**0.881 (0.890)**

結果

カッコ内は古い問題をフィルタリング後の正解率
太字は1位, 下線は2位を示す

モデル	(1) 対数尤度	(2-1) 番号を出力	(2-2) 語句を出力	
calm2-7b	0.622 (0.625)	0.240 (0.243)	0.573 (0.575)	
ELYZA-japanese-Llama-2-7b	0.449 (0.453)	0.376 (0.380)	0.405 (0.408)	モデルの大型化
ELYZA-japanese-Llama-2-13b	0.586 (0.592)	0.523 (0.524)	0.565 (0.569)	
llm-jp-13b-v1.0	0.526 (0.521)	0.246 (0.250)	0.448 (0.449)	
japanese-stablelm-base-beta-7b	0.498 (0.501)	0.365 (0.364)	0.469 (0.479)	
japanese-stablelm-base-gamma-7b	0.719 (0.726)	0.643 **(0.649)**	0.653 (0.664)	
stockmark-13b	0.771 (0.762)	0.289 (0.285)	0.000 (0.000)	
Swallow-7b-hf	0.754 (0.749)	0.459 (0.455)	0.640 (0.641)	モデルの大型化
Swallow-13b-hf	**0.812 (0.813)**	**0.649** (0.648)	**0.751 (0.749)**	
weblab-10b	0.336 (0.330)	0.270 (0.278)	0.327 (0.325)	
Llama-2-7b-hf	0.401 (0.403)	0.267 (0.263)	0.305 (0.311)	モデルの大型化
Llama-2-13b-hf	0.495 (0.507)	0.378 (0.380)	0.412 (0.427)	
Mistral-7B-v0.1	0.383 (0.392)	0.354 (0.365)	0.379 (0.388)	
gpt-3.5-turbo-0125	-	0.645 (0.657)	0.699 (0.710)	
gpt-4-0125-preview	-	**0.853 (0.861)**	**0.881 (0.890)**	

結果

カッコ内は古い問題をフィルタリング後の正解率
太字は1位, 下線は2位を示す

モデル	(1) 対数尤度	(2-1) 番号を出力	(2-2) 語句を出力
calm2-7b	0.622 (0.625)	0.240 (0.243)	0.573 (0.575)
ELYZA-japanese-Llama-2-7b	0.449 (0.453)	0.376 (0.380)	0.405 (0.408)
ELYZA-japanese-Llama-2-13b	0.586 (0.592)	0.523 (0.524)	0.565 (0.569)
llm-jp-13b-v1.0	0.526 (0.521)	0.246 (0.250)	0.448 (0.449)
japanese-stablelm-base-beta-7b	0.498 (0.501)	0.365 (0.364)	0.469 (0.479)
japanese-stablelm-base-gamma-7b	0.719 (0.726)	0.643 **(0.649)**	0.653 (0.664)
stockmark-13b	0.771 (0.762)	0.289 (0.285)	0.000 (0.000)
Swallow-7b-hf	0.754 (0.749)	0.459 (0.455)	0.640 (0.641)
Swallow-13b-hf	**0.812 (0.813)**	**0.649** (0.648)	**0.751 (0.749)**
weblab-10b	0.336 (0.330)	0.270 (0.278)	0.327 (0.325)
Llama-2-7b-hf	0.401 (0.403)	0.267 (0.263)	0.305 (0.311)
Llama-2-13b-hf	0.495 (0.507)	0.378 (0.380)	0.412 (0.427)
Mistral-7B-v0.1	0.383 (0.392)	0.354 (0.365)	0.379 (0.388)
gpt-3.5-turbo-0125	-	0.645 (0.657)	0.699 (0.710)
gpt-4-0125-preview	-	**0.853 (0.861)**	**0.881 (0.890)**

独自に収集したコーパスも利用

コーパスを独自に精錬

結果

カッコ内は古い問題をフィルタリング後の正解率
太字は1位, 下線は2位を示す

モデル	(1) 対数尤度	(2-1) 番号を出力	(2-2) 語句を出力
calm2-7b	0.622 (0.625)	0.240 (0.243)	0.573 (0.575)
ELYZA-japanese-Llama-2-7b	0.449 (0.453)	0.376 (0.380)	0.405 (0.408)
ELYZA-japanese-Llama-2-13b	0.586 (0.592)	0.523 (0.524)	0.565 (0.569)
llm-jp-13b-v1.0	0.526 (0.521)	0.246 (0.250)	0.448 (0.449)
japanese-stablelm-base-beta-7b	0.498 (0.501)	0.365 (0.364)	0.469 (0.479)
japanese-stablelm-base-gamma-7b	0.719 (0.726)	0.643 **(0.649)**	0.653 (0.664)
stockmark-13b	0.771 (0.762)	0.289 (0.285)	0.000 (0.000)
Swallow-7b-hf	0.754 (0.749)	0.459 (0.455)	0.640 (0.641)
Swallow-13b-hf	**0.812 (0.813)**	**0.649** (0.648)	**0.751 (0.749)**
weblab-10b	0.336 (0.330)	0.270 (0.278)	0.327 (0.325)
Llama-2-7b-hf	0.401 (0.403)	0.267 (0.263)	0.305 (0.311)
Llama-2-13b-hf	0.495 (0.507)	0.378 (0.380)	0.412 (0.427)
Mistral-7B-v0.1	0.383 (0.392)	0.354 (0.365)	0.379 (0.388)
gpt-3.5-turbo-0125	-	0.645 (0.657)	0.699 (0.710)
gpt-4-0125-preview	-	**0.853 (0.861)**	**0.881 (0.890)**

Llama2-7B から継続学習

Mistral-7B から継続学習

目次

すべてのモデルが正答した問題例
（対数尤度による評価）

- キーワードや固有名詞から答えを容易に連想できる問題が多い

「ネビュラ賞」「ヒューゴー賞」「星雲賞」といえば、どんなジャンルの文学作品に与えられる賞?			
1. 歴史小説	2. 推理小説	3. 詩	**4. SF**
JR四国の本社がある都市は?			
1. 高松市	2. 松山市	3. 徳島市	4. 高知市
日本の歴史区分で「戦後」といったら、普通何という戦争の終結後を指す?			
1. 日清戦争	2. 日露戦争	3. 第一次世界大戦	**4. 第二次世界大戦**
正式には「国際連合教育科学文化機関」という国連の専門機関は何?			
1. UNHCR	**2. UNESCO**	3. UNICEF	4. UNCTAD
欧米のレストランなどで、食事などのサービスに感謝して支払うお金を何と言う?			
1. アップ	**2. チップ**	3. ノップ	4. リップ
MacBook、iMac、iPhoneといえばどこのメーカーの製品?			
1. アップル	2. インテル	3. グーグル	4. マイクロソフト
社員にパワハラをしたり過重労働を強制したりする企業のことを、俗に何という?			
1.ブルー企業	2. ブラウン企業	**3. ブラック企業**	4. グリーン企業
奈良県や和歌山県などを含む、日本の半島は?			
1. 三浦半島	**2. 紀伊半島**	3. 渥美半島	4. 大隅半島

すべてのモデルが誤答した問題例
（対数尤度による評価）

- 視覚情報,フレーズに関する知識, 多段階推論などが必要な問題が多い

文字遊びの「へのへのもへじ」で、鼻を表しているひらがなは?			
1. へ	2. の	**3. も**	4. じ
未熟であることを指す慣用句で「黄色い」といわれるのは何?			
1. くちばし	2. 尻	3. 声	4. おなか
毛利衛は「地球に」、バイロンは「愛に」、パスツールは「科学に」ないといったものは何?			
1. 未来	**2. 国境**	3. 平和	4. 絶望
名前に「島」という漢字が付く日本の県で、県庁所在地名に「島」という漢字が付かないのは?			
1. 鹿児島県	**2. 島根県**	3. 徳島県	4. 福島県
日本の文化勲章にデザインされている花は何?			
1. 菊	2. 桐	**3. 橘**	4. 蘭
南米の国・チリの、立法上の首都はどこ?			
1. サンティアゴ	2. ビニャデルマル	3. サンフェルナンド	**4. バルパライソ**
「ラスパイレス指数」に名を残す経済学者エティエンヌ・ラスパイレスはどこの国の人?			
1. アメリカ	2. イギリス	**3. ドイツ**	4. フランス
上田敏が訳した一節「神、そらに知ろしめす すべて世は事も無し」が有名な詩人は誰?			
1. カール・ブッセ	2. ポール・ヴェルレーヌ	**3. ロバート・ブラウニング**	4. ルコント・ド・リール

まとめ

- 日本語言語モデルの評価に利用できる
  多肢選択式の日本語質問応答データセットを作成
  - 国内のクイズ大会で使用された4択クイズの問題が題材
    ⇒ 日本特有の事物に関する知識を多く含む
- 作成したデータセットを用いて
  既存の日本語言語モデルの性能を評価
- 作成したデータセット（一部）と評価スクリプトを公開
  - 🤗Hub: tohoku-nlp/abc-multiple-choice
  - GitHub: cl-tohoku/abc-multiple-choice

謝辞

- 本研究で公開したデータセットのクイズ問題は abc/EQIDEN 実行委員会[1]より
  研究目的での利用および再配布の許諾を得たものです. 記して感謝いたします.

- データセットに含まれるクイズ問題の著作権は abc/EQIDEN 実行委員会に帰属します.

1. https://abc-dive.com/portal/

参考文献

- [Han+ 24] Namgi Han, 植田暢大, 大嶽匡俊, 勝又智, 鎌田啓輔, 清丸寛一, 児玉貴志, 菅原朔, Bowen Chen, 松田寛, 宮尾祐介, 村脇有吾, 劉弘毅. llm-jp-eval: 日本語大規模言語モデルの自動評価ツール. In NLP, 2024.
- [Kurihara+ 22] Kentaro Kurihara, Daisuke Kawahara, and Tomohide Shibata. JGLUE: Japanese General Language Understanding Evaluation. In LREC, 2022.
- [Lin+ 21] Bill Yuchen Lin, Seyeon Lee, Xiaoyang Qiao, and Xiang Ren. Common Sense Beyond English: Evaluating and Improving Multilingual Language Models for Commonsense Reasoning. In ACL, 2021.
- [Longpre+ 21] Shayne Longpre, Yi Lu, and Joachim Daiber. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. TACL, Vol. 9, pp. 1389–1406, 2021.
- [So+ 22] ByungHoon So, Kyuhong Byun, Kyungwon Kang, and Seongjin Cho. JaQuAD: Japanese Question Answering Dataset for Machine Reading Comprehension. arXiv preprint arXiv:2202.01764, 2022.
- [石井+ 23] 石井愛, 井之上直也, 関根聡. 根拠を説明可能な質問応答システムのための日本語マルチホップQAデータセット構築. In NLP, 2023.
- [石井+ 24] 石井愛, 井之上直也, 鈴木久美, 関根聡. JEMHopQA:日本語マルチホップQAデータセットの改良. In NLP, 2024.
- [栗原+ 22] 栗原健太郎, 河原大輔, 柴田知秀. JGLUE: 日本語言語理解ベンチマーク. In NLP, 2022.
- [鈴木+ 20] 鈴木正敏, 鈴木潤, 松田耕史, 西田京介, 井之上直也. JAQKET: クイズを題材にした日本語QAデータセットの構築. In NLP, 2020.
- [関根 03] 関根聡. 百科事典を対象とした質問応答システムの開発. In NLP, 2003.

付録

プロンプトに使用した例題

- 「abc」第9回大会の筆記問題（最初の4問）を元に作成

1. 「A」「B」「C」のうち、JIS規格で鉛筆の芯の固さに用いられているアルファベットはどれでしょう?
2. 柳刃、菜切、出刃などの種類がある、まな板と共に使われる台所用品は何でしょう?
3. 記号「m」で表される、「リットル」や「メートル」の頭について、「1000 分の1」を意味する言葉は何でしょう?
4. 音楽バンドで歌を担当する人のことを英語で何というでしょう?

『abc ～the ninth～』より引用

次のうち、JIS規格で鉛筆の芯の固さに用いられているアルファベットは?			
1. A	**2. B**	3. C	4. D
柳刃、菜切、出刃などの種類がある、まな板と共に使われる台所用品は?			
1. 包丁	2. 鍋	3. 箸	4. やかん
記号「m」で表される、「リットル」や「メートル」の頭について、「1000分の1」を意味する言葉は?			
1. マイクロ	2. マクロ	**3. ミリ**	4. メガ
音楽バンドで歌を担当する人のことを英語で何という?			
1. キーボード	2. パーカッション	3. ベース	**4. ボーカル**

公開データのみを用いた実験結果

カッコ内は古い問題をフィルタリング後の正解率
太字は1位, 下線は2位を示す

モデル	(1) 対数尤度	(2-1) 番号を出力	(2-2) 語句を出力
calm2-7b	0.587 (0.594)	0.227 (0.234)	0.531 (0.536)
ELYZA-japanese-Llama-2-7b	0.407 (0.421)	0.338 (0.353)	0.382 (0.386)
ELYZA-japanese-Llama-2-13b	0.571 (0.581)	0.493 (0.503)	0.562 (0.569)
llm-jp-13b-v1.0	0.529 (0.520)	0.224 (0.231)	0.436 (0.442)
japanese-stablelm-base-beta-7b	0.464 (0.482)	0.324 (0.330)	0.442 (0.457)
japanese-stablelm-base-gamma-7b	0.671 (0.688)	**0.596** (0.609)	0.613 (0.632)
stockmark-13b	0.740 (0.739)	0.271 (0.272)	0.000 (0.000)
Swallow-7b-hf	0.713 (0.706)	0.433 (0.439)	0.580 (0.591)
Swallow-13b-hf	**0.749 (0.746)**	0.593 (**0.614**)	**0.704 (0.695)**
weblab-10b	0.389 (0.401)	0.244 (0.254)	0.360 (0.363)
Llama-2-7b-hf	0.384 (0.388)	0.296 (0.294)	0.289 (0.302)
Llama-2-13b-hf	0.473 (0.492)	0.373 (0.386)	0.382 (0.396)
Mistral-7B-v0.1	0.362 (0.368)	0.340 (0.358)	0.342 (0.343)