意味、言語構造、言語の優位性を考慮した多言語文脈内学習

金子 正弘　Aji Alham Fikri　Timothy Baldwin
MBZUAI
{masahiro.kaneko,alham.fikri,timothy.baldwin}@mbzuai.ac.ae

概要

多言語大規模言語モデル（MLLM）の文脈内学習（ICL）における事例選択では、意味の一致、言語構造の類似、言語の優位性の 3 項目が重要となる。既存研究ではこれら 3 項目を総合的にどのように考慮するべきか明らかにされていない。我々は 3 項目の指標を再定義し、最適なバランスで事例選択する手法を提案する。実験の結果、既存手法と比較して提案手法が最高性能を達成した。

1　はじめに

多言語大規模言語モデル（MLLM）は、さまざまな言語のテキストを学習することで言語間の知識転移を行う [1, 2]。また、学習データから選択された少数の事例から学習する文脈内学習（ICL）により、パラメータ更新なしで様々なタスクにおいて高い性能を実現する [3]。MLLM における ICL では言語間の知識転移を有効に活用するために、多言語な事例候補から事例を選択する手法がある。ICL の性能は使用する事例に大きく依存するため事例の選択方法が重要となる [4]。

MLLM での ICL の事例選択に関する既存研究は、主に (1) 意味の一致、 (2) 言語構造の類似、 (3) 言語の優位性、の 3 項目に着目している。意味の一致を考慮した研究では、推論対象となる入力と意味的に類似する事例を選択する手法が提案され性能改善が報告されている [5, 6]。これは多言語に限らず単言語における ICL でも同様の結果が得られている [3]。言語構造の類似に関しては、形態や文法構造が共通している言語の事例を用いるほど、知識転移において高い効果を発揮することが示されている [7]。これは ICL に限らず、多言語データを用いたモデルの学習で同様の傾向が報告されている [8, 9, 10]。言語の優位性では、言語ごとに推論性能に違いがあり、英語のような高リソース言語のデータを事例として用いることで、低リソース言語の推論の性能が向上することがわかっている [11, 12]。

これら 3 つの項目の個別的な影響については先行研究で検証されているが、総合的に考慮する際の最適な重み付けについては明らかにされていない。そのため、実際のユースケースで各観点をどのように活用すればよいか明確ではなく、MLLM の潜在能力を十分に引き出せていない。3 つの項目を最適に考慮して事例選択するには、選択基準が定量化されていない、明示的に観点が区別されていない、という 2 つの課題が既存研究にはある。1 つ目に関しては、言語構造の類似では言語体系や使用地域 [13, 7]、言語の優位性では MLLM の学習データにおける言語ごとのデータサイズ [11, 5] を参考にヒューリスティックに事例の言語が選定されている。2 つ目に関しては、既存研究は意味の一致と言語構造の類似を明示的に区別しない多言語文埋め込み [2] を使用しており、それぞれの観点のバランスを最適化することができない。

本研究では、MLLM の ICL における意味の一致、言語構造の類似、および言語の優位性の指標を定義し、これらの項目を最適にバランスして事例を選択する手法 (Balanced Multi-Factor ICL; **BMF-ICL**) を提案する。まず、項目ごとに定量化された基準を設けるために、我々は以下のようにそれぞれの項目を定義する：意味の一致では、言語をまたいで意味を比較することに特化した LaBSE [14] の文埋め込みを用いて入力と事例候補との類似度をスコアとする。言語構造の類似では、言語ごとの形態や文法構造の特徴を考慮した言語埋め込み lang2vec [15] を用いて、入力と事例候補それぞれの言語間の類似度をスコアとする。言語の優位性では、事例候補の入力をモデルに与えたときの正解の尤度をスコアとする。これら 3 つのスコアを最適なバランスで考慮して事例を選択するために、項目スコアの重み付け和を最終スコアとし、重み付けを開発データで最適化する。

我々は mCSQA [16] と TYDI [17] の 2 つのデータセットで既存研究と提案手法を比較する。4 つの MLLMs を用いた実験の結果、既存手法と比較して、提案手法が最高精度を達成した。さらに、分析により 3 つの項目全てを考慮することの重要性と提案手法が多言語な事例を選択することを明らかにした。

2 3 項目を最適化した BMF-ICL

入力テキスト x に対してパラメータ θ の LLM が生成した出力テキストを y とする。ICL では、タスク定義文 d、事例集合 $\mathscr{E}$、および入力テキスト x を LLM に与える。

$$y = \underset{\hat{y}}{\operatorname{argmax}} P(\hat{y} \mid d, \mathscr{E}, x; \theta) \tag{1}$$

$\mathscr{E}$ は、ソーステキストと参照テキスト $(s^{(j)} \in \mathcal{S}, r^{(j)} \in \mathcal{R})$ を含む事例候補から、k 個の事例を選択することで構築される。

$$\mathscr{E} = (s^{(i)}, r^{(i)}) \in (\mathcal{S}, \mathcal{R})_{i=1}^{k} \tag{2}$$

$s^{(i)}$ および $r^{(i)}$ は、事例選択手法によって順位付けされた事例候補内の上位 k 個のインスタンスである。

提案手法 BMF-ICL では、事例集合 $\mathscr{E}$ を構築するために、事例候補 $(\mathcal{S}, \mathcal{R})$ から以下 3 つの項目スコアの重み付き和が最も高い上位 k 個のインスタンスを選択する。

$$\text{score}^{(j)} = \alpha\, \text{score}^{(j)}_{\text{sem}} + \beta\, \text{score}^{(j)}_{\text{str}} + \gamma\, \text{score}^{(j)}_{\text{sup}} \tag{3}$$

ここで、$\text{score}^{(j)}_{\text{sem}}$ は入力テキスト x とソーステキスト $s^{(j)}$ 間の意味の一致スコア、$\text{score}^{(j)}_{\text{str}}$ は入力テキスト x とソーステキスト $s^{(j)}$ それぞれの言語同士の言語構造の類似スコア、$\text{score}^{(j)}_{\text{sup}}$ はソーステキスト $s^{(j)}$ から参照テキスト $r^{(j)}$ をモデルが生成する際の言語の優位性を表す。スカラー係数 α, β, γ は $0 \le \alpha, \beta, \gamma \le 1$ と $\alpha + \beta + \gamma = 1$ の条件を満たす。

入力テキストと事例候補内のソーステキスト間の意味の一致度合い $\text{score}^{(j)}_{\text{sem}}$ を計算するために、多言語文埋め込みモデルによってエンコードされた入力テキスト x と j 番目のソーステキスト $s^{(j)}$ の文埋め込みをそれぞれを $\boldsymbol{h}(x)$ と $\boldsymbol{h}(s^{(j)})$ とする。多言語文埋め込みモデルとして我々は LaBSE [14][1] を使用する。LaBSE は大規模対訳データを用いた対照学習によって多言語文埋め込みを学習しており、言語に関係なく一貫してテキスト間の意味的類似度を計算することができる。j 番目のソーステキストに対する意味の一致スコア $\text{score}^{(j)}_{\text{sem}}$ はコサイン類似度を用いて以下のように定義する。

$$\text{score}^{(j)}_{\text{sem}} = \cos(\boldsymbol{h}(x), \boldsymbol{h}(s^{(j)})). \tag{4}$$

入力テキストと事例候補のソーステキスト間の言語構造の類似度を計算するために、入力テキスト x の言語 l_x と j 番目のソーステキスト $s^{(j)}$ の言語 $l_{s^{(j)}}$ に対する言語埋め込みをそれぞれ $\boldsymbol{e}(l_x)$ と $\boldsymbol{e}(l_{s^{(j)}})$ とする。言語埋め込み $\boldsymbol{e}(l_x)$ と $\boldsymbol{e}(l_{s^{(j)}})$ は、類型論的特徴、地理的特徴、系統的特徴などの言語特性を符号化する lang2vec [15][2] から取得する。j 番目のソーステキストに対する言語の類似スコア $\text{score}^{(j)}_{\text{str}}$ はコサイン類似度を用いて以下のように定義される。

$$\text{score}^{(j)}_{\text{str}} = \cos(\boldsymbol{e}(l_x), \boldsymbol{e}(l_{s^{(j)}})) \tag{5}$$

言語の判定には fasttext-langdetect [18, 19][3] を用いる。

最後に、言語の優位性のスコアとして、事例候補においてソーステキスト $s^{(j)}$ から参照テキスト $r^{(j)}$ を生成する際の各言語におけるモデルの性能を計算する。対象言語 l_{tgt} のサブセット $(\mathcal{S}_{l_{\text{tgt}}}, \mathcal{R}_{l_{\text{tgt}}})$ を以下のように定義する：

$$(\mathcal{S}_{l_{\text{tgt}}}, \mathcal{R}_{l_{\text{tgt}}}) = \{(s', r' \in (\mathcal{S}, \mathcal{R}) \mid l_{s'} = l_{\text{tgt}}\}. \tag{6}$$

ここで、l_t は事例候補内に含まれる任意の言語をとる。各言語に対するモデルの推論能力は、定義文とソーステキストが与えられた際の参照テキストの対数尤度の平均とする。

$$\text{sup}(l_{\text{tgt}}) = \frac{1}{|\mathcal{S}_{l_{\text{tgt}}}|} \sum_{(s', r') \in \mathcal{S}_{l_{\text{tgt}}}, \mathcal{R}_{l_{\text{tgt}}}} \frac{1}{|r'_i|} \sum_{i=1}^{|r'_i|} \log P(r'_i \mid d, s'; \theta) \tag{7}$$

事例候補内の第 j 番目のインスタンスに対する $\text{score}^{(j)}_{\text{sup}}$ は以下のように計算する。

$$\text{score}^{(j)}_{\text{sup}} = \text{sup}(l_{s^{(j)}}) \tag{8}$$

3 実験

3.1 設定

データセット 多言語データセットの多くは、特定言語のデータセットを他言語に翻訳することで作成されている。このような多言語データセットでは各言語のインスタンスの内容が同一であり、これは言語ごとにデータ分布に違いがある現実的な設定とは乖離がある。さらに、多言語 ICL に期待される言語間での情報の相補性の効果を検証することができ

1) https://huggingface.co/sentence-transformers/LaBSE

2) https://www.cs.cmu.edu/~dmortens/projects/7_project/

3) https://pypi.org/project/fasttext-langdetect/

ない。そのため、本研究では言語ごとに独自に作成された以下の 2 つの多言語データセットを使用する。**mCSQA** [16] は 8 言語の多肢選択式の常識質問応答のためのデータセット、**TYDI** [17] は 11 言語の質問応答データセットである。本研究では、文脈と質問の両方が与えられ、モデルが文脈に基づき回答を生成する gold passage タスクを採用する。

mCSQA と TYDI それぞれ学習データを事例候補とし、開発データで式 3の α、β と γ を探索し、評価データの結果を報告する。どちらも回答を生成させ、完全一致による正解率で評価する。なお、付録 Aに mCSQA と TYDI の言語と統計情報を示した。

プロンプト ICL における事例数として 2、4、8、16 で比較した結果[4]、8 事例を使用した場合に最も高い性能が得られたため、我々は事例数として 8 を採用する。既存研究 [20] およびプロンプトガイドライン[5]を参考にし、mCSQA および TYDI においてそれぞれ付録 Bに示すプロンプトを用いる。

モデル **BLOOMZ** [21]、**XGLM** [22]、**GPT-3.5** [23][6]、および **GPT-4** [24][7]を用いる。

比較手法 ICL には、事例候補として対象言語のインスタンスを用いる設定と、用いない設定の 2 つがある。我々は対象言語のインスタンスを事例候補に用いる設定における比較手法として以下を使う。

- **Random-ICL** は、mCSQA および TYDI における対象言語の事例候補からランダムに選択した 8 つのインスタンスを事例として使う。
- **Translation-ICL** [12] は、対象言語から高リソース言語である英語に翻訳された入力および事例を MLLM に与え ICL を行う。対象言語の事例候補からランダムに事例選択される。
- **Synthetic-ICL** [13] は、対象言語ではソーステキストのみ利用可能な設定を想定している。高リソース言語のインスタンスで擬似参照テキストを生成し、それを対象言語のソーステキストと組み合わせて ICL の事例とする。既存研究に従い、高リソース言語として我々は英語を選択した。事例は英語と対象言語の事例候補からそれぞれランダムにサンプリングされる。

我々は対象言語のインスタンスを事例候補に用いない設定における比較手法として以下を使う。

4) 詳細な実験結果は付録 Cを参照。
5) https://huggingface.co/docs/transformers/v4.37.0/en/tasks/prompting
6) gpt-3.5-turbo-0125
7) gpt-4-turbo-2024-04-09

	en	zh	fr	de	ja	nl	pt	ru
Random-ICL	70.8	52.3	69.0	73.8	62.2	71.7	72.5	43.7
Translation-ICL	-	50.1	70.1	75.7	59.5	73.8	74.2	43.0
Synthetic-ICL	-	52.4	69.8	74.4	60.8	72.8	75.0	44.0
BMF-ICL	**71.4**	**55.9**	**72.0**	**76.5**	**63.5**	**74.3**	**76.9**	**45.3**
Non-ICL	61.5	44.9	60.4	56.4	50.7	60.3	57.6	36.9
English-ICL	-	46.5	62.9	64.7	55.0	64.4	60.4	38.4
XLM-ICL	64.8	48.0	63.9	68.4	56.7	64.9	61.3	39.3
BMF-ICL	**66.4**	**49.2**	**65.3**	**70.0**	**58.1**	**67.7**	**62.8**	**41.0**

表 1: mCSQA データセットにおける正解率。

- **Non-ICL** は、事例を用いない zero-shot 手法である。
- **English-ICL** [11] は、MLLM に英語の事例を提示し、対象言語の入力に対して推論を行う。事例は英語の事例候補からランダムにインスタンスを選択される。
- **Multilingual-ICL** [7] は、対象言語を除く多様な言語の事例候補から ICL のためのインスタンスを乱択する。mCSQA と TYDI それぞれで対象言語以外の全ての言語を事例候補とした。
- **XLM-ICL** [5] は、多言語文埋め込み XLM [2] を使い、対象言語の入力に類似した高リソース言語の事例を選択する。既存研究に従い、mCSQA では英語、ドイツ語と中国語、TYDI では英語を高リソース言語とした。

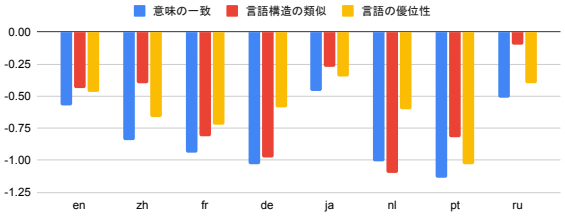
3.2 実験結果

mCSQA と TYDI それぞれのデータセットにおける BLOOMZ、XGLM、GPT-3.5 と GPT-4 の平均スコアを表 1と表 2に示した。上段は対象言語のインスタンスを事例候補に含める設定、下段は含めない設定である。赤い背景は上段では Random-ICL、下段では Non-ICL と比較して性能が低下した既存研究の結果を表す。提案手法は mCSQA と TYDI の全ての言語設定で既存手法と比較して最高性能を達成している。このことから、意味の一致、言語構造の類似と言語の優位性の 3 項目を最適なバランスで考慮することは重要であることがわかる。

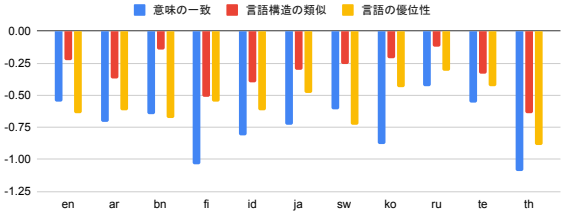
上段では、Translation-ICL と Synthetic-ICL は mCSQA と TYDI の両方で Random-ICL と比較して必ずしも性能が改善するとは限らないことがわかる。下段では、既存手法も提案手法もどちらも全ての言語で Non-ICL と比較して性能が改善した。さらに、上段と下段の結果を比較すると対象言語が事例候補に含まれるか否かで性能の上がり幅が大きく異なるため、対象言語が事例候補にある場合、他言語の恩恵

	en	ar	bn	fi	id	ja	sw	ko	ru	te	th
Random-ICL	68.9	63.4	61.0	70.2	69.1	69.8	61.9	67.3	62.8	62.8	60.3
Translation-ICL	-	61.5	60.0	71.7	69.3	69.1	60.6	66.7	62.0	60.2	60.3
Synthetic-ICL	-	63.2	61.8	72.3	70.0	69.9	61.0	66.9	63.0	60.7	60.9
BMF-ICL	**70.5**	**64.9**	**63.3**	**73.7**	**71.1**	**71.2**	**62.3**	**68.5**	**64.0**	**63.5**	**63.0**
Non-ICL	57.3	54.0	48.2	56.5	55.7	55.6	49.3	57.0	49.0	51.0	50.4
English-ICL	-	57.3	50.7	58.8	59.6	57.6	52.2	58.5	51.5	53.0	51.7
XLM-ICL	61.9	57.4	51.9	59.4	60.2	58.8	53.1	59.8	54.0	53.5	52.1
BMF-ICL	**63.5**	**60.5**	**54.9**	**61.3**	**62.3**	**60.9**	**53.8**	**61.0**	**57.5**	**56.1**	**54.3**

表 2: TYDI データセットにおける正解率。



(a) mCSQA データセットにおける正解率の変化。



(b) TYDI データセットにおける正解率の変化。

図 1: アブレーションの結果。

により性能改善することは難しいことがわかる。

4 分析

4.1 3 項目のアブレーション

提案手法の 3 項目のそれぞれの重要性を調査するために、アブレーション実験を行う。図 1は、mCSQA と TYDI それぞれにおける対象言語を事例候補に含む設定での言語ごとのアブレーション結果である。例えば、「意味の一致」の結果では式 3の α を 0 とし β と γ はそのままの値を用いる。

結果から、全ての言語においていずれの項目を抜いても性能が低下しており、3 項目を考慮して ICL の事例を選択することが重要であることがわかる。mCSQA では 7/8 言語、TYDI では 8/11 言語が最も「意味の一致」の項目を抜いた時に性能が低下していることから、単言語設定 [3] と同様に多言語設定においても近い問題を提示することが性能向上の寄与する。同じ言語グループが存在する en-de-nl や fr-pt では「言語構造の一致」が重要な役割を果たしており、他言語と比較してより性能が低下する。

	mCSQA				TYDI			
	総合	意味	構造	優位性	総合	意味	構造	優位性
1	0.05	0.00	0.10	0.15	0.01	0.01	0.01	0.05
2	0.19	0.01	**0.34**	**0.31**	0.14	0.13	0.11	0.17
3	**0.31**	**0.40**	0.23	0.27	0.28	0.18	**0.30**	**0.29**
4	0.24	0.32	0.16	0.19	**0.30**	0.20	0.28	0.23
5	0.11	0.17	0.07	0.05	0.13	**0.29**	0.20	0.10
6	0.06	0.07	0.06	0.02	0.08	0.10	0.08	0.09
7	0.03	0.03	0.03	0.01	0.05	0.07	0.01	0.06
8	0.01	0.00	0.01	0.00	0.01	0.02	0.01	0.01

表 3: BMF-ICL の 8 事例内においてそれぞれの言語タイプ数を持つインスタンスの割合。

4.2 事例内における言語タイプ数

提案手法が言語間の知識転移によって性能改善を達成していることを示す。まず、mCSQA と TYDI のインスタンスごとに、ICL の 8 事例に含まれる言語タイプの数を調査する。その結果を基に、それぞれの評価データ内で異なる言語タイプ数を持つインスタンスの割合を算出する。この時、3 項目の影響も調査するために式 3における対象項目の重みを 1 とし、他の重みを 0 にしたときの結果を報告する。

表 3 は、mCSQA と TYDI それぞれにおいて対象言語を事例候補に含めた場合に、ICL で選択された事例の言語タイプ数ごとの割合を表す。総合は全ての重みを考慮した時の結果であり、意味、構造と優位性はそれぞれの項目だけを考慮した時の結果である。総合の結果では、mCSQA では 3 言語、TYDI では 4 言語の割合が最も多く、提案手法により多言語な事例が選択されている。さらに、意味の一致性が他 2 項目と比較して最も割合が多い言語タイプ数が大きいことから、意味の一致が特に事例の言語的多様性に貢献している。

5 おわりに

我々は MLLM の ICL のために意味の一致、言語構造の類似、言語の優位性を最適なバランスで考慮する事例選択手法を提案した。提案手法は既存の事例選択手法と比較して最高性能を達成した。さらに、3 項目全てを考慮することが重要であり、選択事例が言語的に多様であることを示した。

参考文献

[1] Alexis Conneau and Guillaume Lample. Cross-lingual language model pretraining. **Advances in neural information processing systems**, Vol. 32, , 2019.

[2] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 8440–8451, Online, July 2020. Association for Computational Linguistics.

[3] Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In **Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures**, pp. 100–114, Dublin, Ireland and Online, May 2022. Association for Computational Linguistics.

[4] Ethan Perez, Douwe Kiela, and Kyunghyun Cho. True few-shot learning with language models. In **Advances in Neural Information Processing Systems**, Vol. 34, pp. 11054–11070. Curran Associates, Inc., 2021.

[5] Ercong Nie, Sheng Liang, Helmut Schmid, and Hinrich Schütze. Cross-lingual retrieval augmented prompt for low-resource languages. In **Findings of the Association for Computational Linguistics: ACL 2023**, pp. 8320–8340, Toronto, Canada, July 2023. Association for Computational Linguistics.

[6] Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. Multilingual LLMs are better cross-lingual in-context learners with alignment. In **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 6292–6307, Toronto, Canada, July 2023. Association for Computational Linguistics.

[7] Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. Cross-lingual few-shot learning on unseen languages. In **Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 777–791, Online only, November 2022. Association for Computational Linguistics.

[8] Melvin Johnson, Mike Schuster, Quoc V. Le, Maxim Krikun, Yonghui Wu, Zhifeng Chen, Nikhil Thorat, Fernanda Viégas, Martin Wattenberg, Greg Corrado, Macduff Hughes, and Jeffrey Dean. Google's multilingual neural machine translation system: Enabling zero-shot translation. **Transactions of the Association for Computational Linguistics**, Vol. 5, pp. 339–351, 2017.

[9] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual BERT? In **Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics**, pp. 4996–5001, Florence, Italy, July 2019. Association for Computational Linguistics.

[10] B lażej Dolicki and Gerasimos Spanakis. Analysing the impact of linguistic features on cross-lingual transfer. **arXiv preprint**, 2021.

[11] Genta Indra Winata, Andrea Madotto, Zhaojiang Lin, Rosanne Liu, Jason Yosinski, and Pascale Fung. Language models are few-shot multilingual learners. In **Proceedings of the 1st Workshop on Multilingual Representation Learning**, pp. 1–15, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[12] Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. Do multilingual language models think better in English? In **Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)**, pp. 550–564, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[13] Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. Democratizing LLMs for low-resource languages by leveraging their English dominant abilities with linguistically-diverse prompts. In **Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 3501–3516, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[14] Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. Language-agnostic BERT sentence embedding. In **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 878–891, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[15] Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers**, pp. 8–14, Valencia, Spain, April 2017. Association for Computational Linguistics.

[16] Yusuke Sakai, Hidetaka Kamigaito, and Taro Watanabe. mCSQA: Multilingual commonsense reasoning dataset with unified creation strategy by language models and humans. In **Findings of the Association for Computational Linguistics ACL 2024**, pp. 14182–14214, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

[17] J. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 454–470, 2020.

[18] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. **arXiv preprint**, 2016.

[19] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. Fasttext.zip: Compressing text classification models. **arXiv preprint arXiv:1612.03651**, 2016.

[20] Joshua Robinson, Christopher Rytting, and David Wingate. Leveraging large language models for multiple choice question answering. **ArXiv**, Vol. abs/2210.12353, , 2022.

[21] BigScience. Bloom: A 176b-parameter open-access multilingual language model. **ArXiv**, Vol. abs/2211.05100, , 2022.

[22] Meta AI. Few-shot learning with multilingual generative language models. In **Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing**, pp. 9019–9052, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[23] OpenAI. Language models are few-shot learners. In **Advances in Neural Information Processing Systems**, Vol. 33, pp. 1877–1901. Curran Associates, Inc., 2020.

[24] OpenAI. Gpt-4 technical report. **arXiv preprint**, 2023.

	言語グループ	mCSQA 学習	mCSQA 開発	mCSQA 評価	TYDI 学習	TYDI 開発	TYDI 評価
Arabic (ar)	Semitic	-	-	-	23.0k	1.3k	1.4k
Bengali (bn)	Indo-Aryan	-	-	-	10.7k	0.3k	0.3k
Chinese (zh)	Sinitic	12.2k	1.5k	1.5k	-	-	-
English (en)	Germanic	10.9k	1.3k	1.3k	9.2k	1.0k	1.0k
Finnish (fi)	Finnic	-	-	-	15.2k	2.0k	2.0k
French (fr)	Romance	8.0k	1.0k	1.0k	-	-	-
German (de)	Germanic	12.5k	1.5k	1.5k	-	-	-
Indonesian (id)	Malayo-Polynesian	-	-	-	14.9k	1.8k	1.8k
Japanese (ja)	Japonic	11.7k	1.4k	1.4k	16.2k	1.7k	1.7k
Kiswahili (sw)	Bantu	-	-	-	17.6k	2.2k	2.2k
Korean (ko)	Koreanic	-	-	-	10.9k	1.6k	1.7k
Dutch (nl)	Germanic	12.2k	1.5k	1.5k	-	-	-
Portuguese (pt)	Romance	12.7k	1.5k	1.5k	-	-	-
Russian (ru)	Slavic	6.6k	0.8k	0.8k	12.8k	1.6k	1.6k
Telugu (te)	Dravidian	-	-	-	24.5k	2.4k	2.5k
Thai (th)	Tai	-	-	-	11.3k	2.2k	2.2k

表 4: mCSQA と TYDI それぞれのデータサイズと言語グループ。

A　データセットの統計情報

表 4は mCSQA と TYDI の言語ごとの言語グループとデータサイズを表している。

B　プロンプト

mCSQA と TYDI それぞれのプロンプトとして我々は以下を用いる。

mCSQA

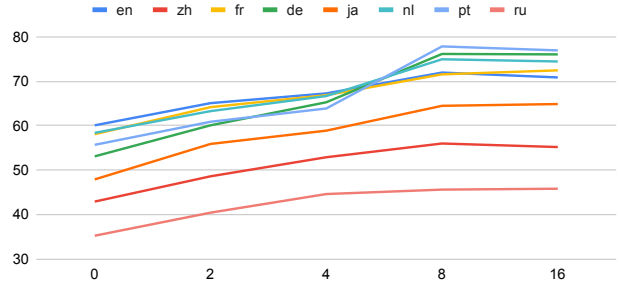
```
Answer the question.
Question: [Question of Example 1]
a. [Choice A of Example 1]
b. [Choice B of Example 1]
c. [Choice C of Example 1]
d. [Choice D of Example 1]
e. [Choice E of Example 1]
Answer: [Answer of Example 1]

:
:
Question: [Question of Example 8]
a. [Choice A of Example 8]
b. [Choice B of Example 8]
c. [Choice C of Example 8]
d. [Choice D of Example 8]
e. [Choice E of Example 8]
Answer: [Answer of Example 8]
Question: [Question of Input]
a. [Choice A of Input]
b. [Choice B of Input]
c. [Choice C of Input]
d. [Choice D of Input]
e. [Choice E of Input]
Answer:
```

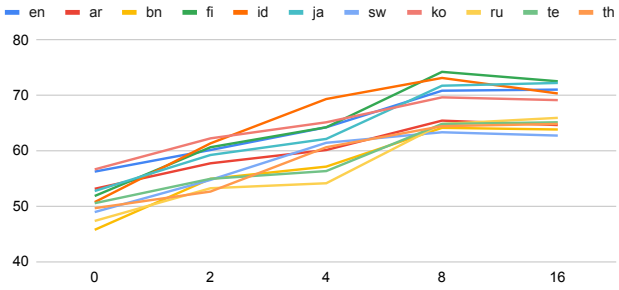
TYDI

```
Answer the question using the context.
Context: [Context of Example 1]
Question: [Question of Example 1]
Answer: [Answer of Example 1]
Context: [Context of Example 2]
Question: [Question of Example 2]
Answer: [Answer of Example 2]

:
:
Context: [Context of Example 7]
Question: [Question of Example 7]
Answer: [Answer of Example 7]
Context: [Context of Example 8]
Question: [Question of Example 8]
Answer: [Answer of Example 8]
Context: [Context of Input]
Question: [Question of Input]
Answer:
```



(a) mCSQA。



(b) TYDI。

図 2: 事例数ごとの性能。

C　事例数ごとの性能

図 2は、対象言語を事例に含む設定における mCSQA と TYDI それぞれの開発データでの事例数を 0, 2, 4, 8, 16 で変更した時の性能を表している。mCSQA では 5/8 言語で、TYDI では 7/11 言語で事例数が 8 で最も性能が高くなった。