# Balanced Multi-Factor In-Context Learning for Multilingual Large Language Models

Anonymous ACL submission

## Abstract

Multilingual large language models (MLLMs) are able to leverage in-context learning (ICL) to achieve high performance by leveraging cross-lingual knowledge transfer without parameter updates. However, their effectiveness is highly sensitive to example selection, particularly in multilingual settings. Based on the findings of existing work, three key factors influence multilingual ICL: (1) semantic similarity, (2) linguistic alignment, and (3) language-specific performance. However, existing approaches address these factors independently, without explicitly disentangling their combined impact, leaving optimal example selection underexplored. To address this gap, we propose balanced multi-factor ICL (**BMF-ICL**), a method that quantifies and optimally balances these factors for improved example selection. Experiments on mCSQA and TYDI across four MLLMs demonstrate that BMF-ICL outperforms existing methods. Further analysis highlights the importance of incorporating all three factors and the importance of selecting examples from multiple languages.

## 1 Introduction

Multilingual large language models (MLLMs) leverage cross-lingual knowledge transfer by learning from text in diverse languages (Conneau and Lample, 2019; Conneau et al., 2020; Xue et al., 2021; Scao et al., 2022). In-context learning (ICL) further enhances performance by selecting a small number of examples from candidate sets, leading to high accuracy on various tasks without parameter updates (Liu et al., 2022). A key approach to enhance the ability for the cross-lingual transfer is to select examples from multilingual candidate pools. Since ICL performance heavily depends on which examples are chosen, the example selection strategy is crucial (Perez et al., 2021; Zhao et al., 2021; Lu et al., 2022; Koike et al., 2024; Hida et al., 2024; Oba et al., 2024).

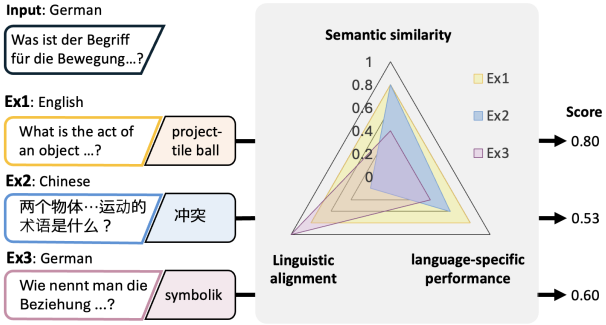



Figure 1: Our proposed method, BMF-ICL, selects multilingual examples for ICL by considering three factors: semantic similarity, linguistic alignment, and language-specific performance.

According to existing work on example selection in multilingual ICL, three main factors influence effectiveness: (1) **semantic similarity**, (2) **linguistic alignment**, and (3) **language-specific performance**. Selecting examples semantically similar to the input often improves performance (Nie et al., 2023; Tanwar et al., 2023; Liu et al., 2022). Using examples from languages that are morphologically and grammatically similar to the target language can lead to stronger knowledge transfer (Johnson et al., 2017; Pires et al., 2019; Yamashita et al., 2020; Winata et al., 2022; Dolicki and Spanakis, 2021). Furthermore, inference performance varies by language, and leveraging data from high-resource languages like English can boost results in low-resource languages (Winata et al., 2021; Etxaniz et al., 2024).

While these three factors are integral to multilingual ICL, existing research typically does not combine them together. Moreover, existing work has two limitations that prevent them from combining these factors: the lack of quantified selection criteria, and the absence of explicit differentiation among the factors. The languages, for example, are often selected heuristically, either based on language groups and geographic regions for lin-

guistic alignment (Nguyen et al., 2024; Winata et al., 2022), or based on per-language data sizes in MLLM training data for language-specific performance (Winata et al., 2021; Nie et al., 2023). Additionally, existing studies typically use multilingual sentence embeddings (Conneau et al., 2020) that do not explicitly distinguish between semantic similarity and linguistic alignment, making it impossible to optimize the balance between them (Nie et al., 2023).

In this study, we propose a method called balanced multi-factor ICL (**BMF-ICL**), a method that defines explicit metrics for semantic similarity, linguistic alignment, and language-specific performance in multilingual in-context learning (MICL), then selects examples by optimally balancing these factors. Figure 1 presents an overview of BMF-ICL, which considers three scores for multilingual example selection. Specifically, we quantify each factor as follows:

**(1) Semantic similarity:** We employ LaBSE (Feng et al., 2022), a language-agnostic sentence embedding model to score the similarity between the input and candidate examples.

**(2) Linguistic alignment:** We use lang2vec (Littell et al., 2017), which captures morphological and grammatical features, to assess how closely the input language aligns with the candidate language.

**(3) Language-specific performance:** We compute the likelihood of producing the correct answer for each language when the MLLM is provided with the candidate example's inputs.

To select examples while balancing these three scores, we take their weighted sum and optimize the weights on development data.

We evaluate both existing approaches and our proposed method on two benchmark datasets, mCSQA (Sakai et al., 2024) and TYDI (Clark et al., 2020). The experimental results across four MLLMs demonstrate that BMF-ICL consistently achieves the highest accuracy compared to existing methods. Further analysis highlights the importance of considering all three factors jointly. Notably, in over 95% of the cases, the proposed method selects examples from two or more languages, demonstrating the performance benefits derived from multilingual data.

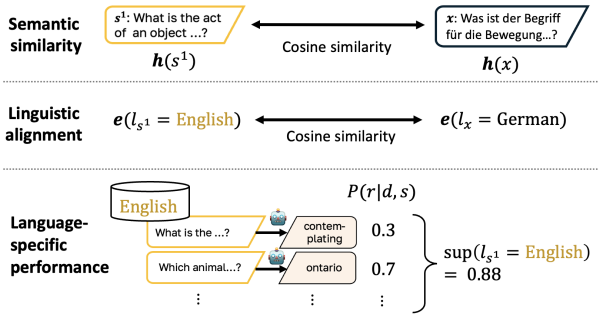



Figure 2: An overview of how BMF-ICL computes semantic similarity, linguistic alignment, and language-specific performance scores to select multilingual examples.

## 2 Balanced Multi-Factor In-Context Learning (BMF-ICL)

We first explain ICL, followed by a discussion of the proposed method for example selection, which takes into account the scores of semantic similarity, linguistic alignment, and language-specific performance, and comprehensively considers these factors. Figure 2 illustrates how BMF-ICL calculates these three scores.

### 2.1 In-Context Learning

Let $x$ be an input text and $y$ be an output text generated by the LLM with parameters $\theta$. In ICL, given a task definition text $d$, the set of examples $\mathcal{E}$, and the input text $x$, the LLM generates $y$ by maximizing the following conditional probability:

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

Our goal is to construct the set of examples $\mathcal{E}$ in a way that maximizes the model's performance or the quality of the generated text. Specifically, $\mathcal{E}$ is composed of $k$ example pairs, drawn from a pool that contains source and reference texts ($\mathcal{S}$, $\mathcal{R}$).

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

Here, $s^{(j)}$ and $r^{(j)}$ ($j = 1, \ldots, k$) are the top $k$ instances ranked by a selection method.

### 2.2 Example Selection via Multi-Factor

We propose a balanced approach to example selection by integrating three key factors for MICL. Specifically, we select the top $k$ instances from the example pool to form $\mathcal{E}$ according to the highest weighted sum of the following three scores:

$$\text{score}^{(j)} = \alpha\, \text{score}^{(j)}_{\text{sem}} + \beta\, \text{score}^{(j)}_{\text{lag}} + \gamma\, \text{score}^{(j)}_{\text{per}}, \quad (3)$$

where $\mathrm{score}^{(j)}_{\mathrm{sem}}$ represents the semantic similarity between the input text and the source text, $\mathrm{score}^{(j)}_{\mathrm{lag}}$ represents the linguistic similarity between the input text and the source text, and $\mathrm{score}^{(j)}_{\mathrm{per}}$ reflects the model's performance in generating the reference text from the source text in the target language. The scalar coefficients $0 \leq \alpha, \beta, \gamma$ satisfy $\alpha + \beta + \gamma = 1$.

**Semantic similarity** To calculate the semantic similarity between input text $x$ and source texts $s^{(i)}$ in the example pool, let $\boldsymbol{h}(x)$ and $\boldsymbol{h}(s^{(j)})$ be their sentence embeddings. We use LaBSE (Feng et al., 2022)[1] as the multilingual sentence embedding model. LaBSE learns multilingual embeddings through contrastive learning over large-scale parallel data, enabling consistent semantic similarity computation across languages.[2] The semantic similarity score $\mathrm{score}^{(j)}_{\mathrm{sem}}$ for the $j$-th source text is then defined by the cosine similarity:

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

**Linguistic alignment** To calculate the linguistic similarity, let $\boldsymbol{e}(l_x)$ and $\boldsymbol{e}(l_{s^{(j)}})$ be linguistic embeddings corresponding to the languages $l_x$ (the language of $x$) and $l_{s^{(j)}}$ (the language of $s^{(j)}$). We use fasttext-langdetect (Joulin et al., 2016b,a)[3] to detect the languages of the input texts and the source texts. The linguistic embeddings $\boldsymbol{e}(l_x)$ and $\boldsymbol{e}(l_{s^{(j)}})$ are obtained from lang2vec (Littell et al., 2017),[4] which encodes typological, geographical, and phylogenetic properties of languages.

The linguistic similarity score $\mathrm{score}^{(j)}_{\mathrm{lag}}$ for the $j$-th source text is defined as:

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

**Language-specific performance** Finally, we measure the model's performance in each language by evaluating how well it generates the reference text $r^{(j)}$ from the source text $s^{(j)}$. The likelihood of $r^{(j)}$ given $s^{(j)}$ serves as a standard proxy for generation capability in LLM evaluation (Zellers et al., 2019; Alzahrani et al., 2024; Hida et al., 2024) and thus provides a practical signal of language-specific performance. For a target language $l_{\mathrm{tgt}}$, we define the sub-dataset $(\mathcal{S}_{l_{\mathrm{tgt}}}, \mathcal{R}_{l_{\mathrm{tgt}}})$ as follows:

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

Here, $l_{\mathrm{tgt}}$ can be any language present in the candidate examples. We define the model's inference ability for each language $\mathrm{per}(l_t)$ for language $l_t$ as the average log-likelihood of each reference text $r'$ given its corresponding source text $s'$:

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

For the $j$-th source text in the example pool, the performance score $\mathrm{score}^{(j)}_{\mathrm{per}}$ is given by:

$$\mathrm{score}^{(j)}_{\mathrm{per}} = \mathrm{per}(l_{s^{(j)}}) \tag{8}$$

By combining these three scores in Eq. (3), our method aims to select examples that simultaneously capture semantic similarity, linguistic alignment, and model performance, thereby improving the overall effectiveness of in-context learning.

## 3 Experiments

### 3.1 Settings

**Dataset** Many multilingual datasets are constructed by translating a single-language dataset into multiple other languages, resulting in parallel content across languages. This setup diverges from realistic scenarios in which data distributions vary by language and also prevents the assessment of potential synergies gained from multilingual ICL. Therefore, we use two multilingual datasets, each originally developed in its own language rather than through translation.

**mCSQA** (Sakai et al., 2024)[5] contains multilingual commonsense question-answering data in a multiple choice format for 8 languages. **TYDI** (Clark et al., 2020) is a question-answering

[1] https://huggingface.co/sentence-transformers/LaBSE

[2] Although LaBSE is itself a multilingual sentence encoder and may still encode language-specific features, it is trained with a contrastive objective that prioritizes semantic equivalence and does not explicitly model grammatical or typological properties. Hence, we regard it as a practical proxy for semantic similarity. Our goal is not perfect disentanglement between semantics and language-specific structure, but rather to separate the factors as much as feasible; the experiments show that even this partial separation is sufficient to yield consistent gains.

[3] https://pypi.org/project/fasttext-langdetect/

[4] https://www.cs.cmu.edu/~dmortens/projects/7_project/

[5] https://huggingface.co/datasets/yusuke1997/mCSQA

dataset covering 11 typologically diverse languages. We frame the task as gold passage generation, where both the context and question are provided, and the model is required to generate the answer.

Table 4 in Appendix A shows the data size and language group for each language in mCSQA and TYDI. In both datasets, we generate answers and evaluate them based on exact-match accuracy.

**Model** We explore both open-weight and closed-weight models. Specifically, we use **Llama 4** (Meta AI, 2025),[6] **Aya** (Üstün et al., 2024),[7] `gpt-3.5-turbo-0125` (**GPT-3.5**) (Brown et al., 2020), and `gpt-4-turbo-2024-04-09` (**GPT-4**) as multilingual LLMs. We use eight NVIDIA A100 GPUs for our experiments.

**ICL Setup** We use the training sets from mCSQA and TYDI as example pools. To determine the optimal prompt configuration, we vary the number of examples (2, 4, 8, and 16) and test four different prompts. The prompts are based on existing research (Robinson et al., 2022) and prompt guidelines.[8] Across these experiments, mCSQA and TYDI achieved the best performance with 8 examples[9] using the following prompts:

**Prompt for 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:
```

**Prompt for 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:
```

**Weight Selection for BMF-ICL** We explore all combinations of $\alpha$, $\beta$, and $\gamma$ in Eq. Equation 3 from 0 to 1 in increments of 0.1, ensuring $\alpha + \beta + \gamma = 1$. We divide the training sets into four folds and, for each LLM, select the weight combination[10] that attains the best average performance in four-fold cross-validation on the mCSQA and TYDI datasets.[11] These selected weights, together with the 8-example prompt configuration, define our final method for BMF-ICL.

**Baseline** Following previous works (Winata et al., 2021; Etxaniz et al., 2024), we evaluate our approach against a range of baselines under two ICL settings: one where the example candidates include the target language, and one where they do not.[12] The motivation for using cross-lingual prompting is that, when there are no suitable example candidates in the target language, it becomes possible to select helpful examples from other languages. In both settings, we use 8 examples for ICL, consistent with our proposed method.

- **With target-language examples:**
  - *Random-ICL*: We randomly select 8 examples (source text and reference text) from the target language candidate set for each evaluation instance. We report the average scores of three experimental runs.
  - Etxaniz et al. (2024): We translate both the input and examples into English, which is

[6] https://huggingface.co/meta-llama/Llama-4-Maverick-17B-128E-Instruct
[7] https://huggingface.co/CohereForAI/aya-23-8B
[8] https://huggingface.co/docs/transformers/v4.37.0/en/tasks/prompting
[9] The results for 2, 4, and 16 are presented in Appendix E

[10] We show specific weights in Appendix C.
[11] Although the baseline does not tune any weights, both methods draw examples from the same gold-labeled pool. Our method simply performs cross-validation within that pool to tune its weights, and therefore relies on no additional data.
[12] See Appendix F for a performance comparison of heuristic vs. lang2vec language choices. This language-selection–only analysis is omitted from the main results, yet lang2vec still outperforms the heuristic.

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 74.1 | 56.7 | 73.1 | 77.8 | 68.9 | 76.3 | 75.2 | 49.1 |
| | Etxaniz et al. (2024) | – | 57.2 | 74.5 | 79.5 | 67.9 | **79.3** | 78.1 | 48.0 |
| | Nguyen et al. (2024) | – | 57.3 | 73.7 | 78.0 | 69.4 | 77.3 | 79.4 | 49.8 |
| | BMF-ICL | **74.8** | **60.8** † | **76.5** † | **80.6** † | **70.2** † | 79.1 | **80.2** | **51.1** † |
| w/o TL | Non-ICL | 67.5 | 52.3 | 68.1 | 66.7 | 60.9 | 68.8 | 65.7 | 43.9 |
| | Winata et al. (2021) | – | 53.7 | 69.5 | 74.0 | 63.8 | 72.8 | 66.6 | 45.0 |
| | Winata et al. (2022) | 68.2 | 53.7 | 69.7 | 68.8 | 64.2 | 71.1 | 66.5 | 45.0 |
| | Nie et al. (2023) | 69.9 | 54.3 | 69.7 | 72.5 | 65.2 | 71.4 | 67.0 | 45.6 |
| | BMF-ICL | **71.1** † | **55.0** | **70.1** | **74.5** | **66.1** | **74.4** † | **67.8** | **47.6** † |

(a) Llama 4.

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 61.1 | 37.1 | 59.0 | 63.9 | 46.6 | 60.1 | 65.0 | 33.6 |
| | Etxaniz et al. (2024) | - | 37.0 | 61.5 | 66.5 | 46.0 | **64.0** | 66.7 | 32.9 |
| | Nguyen et al. (2024) | - | 38.8 | 60.1 | 62.9 | 47.0 | 63.9 | **68.4** | 34.0 |
| | BMF-ICL | **61.7** | **42.9** † | **63.0** † | **67.8** † | **47.2** | 63.7 | 67.9 | **35.0** † |
| w/o TL | Non-ICL | 44.9 | 25.0 | 44.3 | 32.7 | 25.1 | 39.1 | 37.7 | 22.9 |
| | Winata et al. (2021) | - | 29.5 | 48.9 | **55.1** | 35.5 | 47.0 | 40.9 | 26.8 |
| | Winata et al. (2022) | 47.5 | 30.0 | 49.0 | 40.2 | 35.5 | 45.7 | 39.7 | 25.6 |
| | Nie et al. (2023) | 49.0 | **31.3** | 48.6 | 50.3 | 36.9 | 49.1 | 40.3 | 27.7 |
| | BMF-ICL | **52.1** † | **31.3** | **50.6** † | 54.5 | **37.5** | **53.3** † | **42.0** † | **30.3** † |

(b) Aya.

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 79.7 | 65.1 | 79.1 | 83.8 | 78.5 | 83.3 | 79.5 | 55.7 |
| | Etxaniz et al. (2024) | - | 65.8 | 80.1 | 85.1 | 77.3 | **85.8** | 83.0 | 54.5 |
| | Nguyen et al. (2024) | - | 65.2 | 79.5 | 84.5 | 79.0 | 83.1 | 84.1 | 56.6 |
| | BMF-ICL | **80.4** | **68.5** † | **82.3** † | **86.1** † | **80.1** † | 85.7 | **85.5** † | **58.0** † |
| w/o TL | Non-ICL | 77.2 | 64.0 | 78.3 | 81.2 | 76.3 | 81.6 | 77.7 | 52.9 |
| | Winata et al. (2021) | - | 64.1 | 78.4 | 82.1 | 76.0 | **83.8** | 77.6 | 52.8 |
| | Winata et al. (2022) | 77.0 | 63.9 | 78.5 | 81.0 | 76.5 | 82.0 | 78.0 | 53.3 |
| | Nie et al. (2023) | 78.8 | 64.1 | **78.7** | 82.0 | 77.3 | 81.0 | 78.5 | 53.3 |
| | BMF-ICL | **79.2** | **65.1** † | 78.5 | **83.1** † | **78.4** † | 83.5 | **78.8** | **55.0** † |

(c) GPT-3.5.

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 82.2 | 68.8 | 79.3 | 83.5 | 78.8 | 83.5 | 80.0 | 53.3 |
| | Etxaniz et al. (2024) | - | 67.5 | 79.9 | 84.7 | 76.2 | 84.8 | 81.9 | 52.9 |
| | Nguyen et al. (2024) | - | 70.1 | 81.1 | 84.3 | 78.1 | 84.0 | 83.3 | 53.7 |
| | BMF-ICL | **83.3** † | **71.7** † | **82.5** † | **85.0** | **80.8** † | **86.5** † | **86.6** † | **55.8** † |
| w/o TL | Non-ICL | 79.7 | 66.0 | 75.4 | 81.9 | 76.0 | 82.3 | 78.1 | 50.3 |
| | Winata et al. (2021) | - | 65.2 | 76.0 | 83.0 | 75.1 | 82.8 | 79.0 | 49.7 |
| | Winata et al. (2022) | 78.2 | 65.6 | 75.8 | 82.3 | 76.5 | 82.6 | 79.0 | 50.5 |
| | Nie et al. (2023) | **81.9** | 67.7 | 78.1 | **83.2** | 77.0 | 82.5 | 79.2 | 51.9 |
| | BMF-ICL | 81.3 | **69.0** † | **79.6** † | **83.2** | **78.3** † | **83.0** | **80.3** † | **53.0** † |

(d) GPT-4.

Table 1: Results for baseline ICL methods and our method on mCSQA across the four LLMs. Red and blue indicate scores lower and higher than Random-ICL or Non-ICL. The top half of each table shows results with the target language in the example pool ("w/ TL"), and the bottom half without ("w/o TL"). † indicates statistically significant differences between the highest and second highest score in each LLM according to McNemar's test ($p < 0.01$).

dominant in the model's training data, and feeds them into the MLLM for ICL.[13]

  - Nguyen et al. (2024): This baseline generates pseudo-reference texts in the target language by leveraging examples from a high-resource language, then pairs them with the original target-language source text to create ICL examples.[14]

- **Without target-language examples:**
  - *Non-ICL*: A zero-shot baseline that provides only the input text (no examples).
  - Winata et al. (2021): This baseline provides English examples to the MLLM while performing inference on the input in the target language. We randomly sample examples

[13] Following the original translation setting, we use four examples from the FLORES-200 dataset (Costa-jussà et al., 2022), prepending each sentence with its language name (e.g., *English: Mary did not slap the green witch.*).

[14] Following their approach, we use English as the high-resource language and randomly sample examples from the candidate set.

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 70.9 | 64.8 | 63.6 | 70.7 | 70.3 | 71.9 | 63.4 | 68.8 | 63.5 | 64.1 | 60.6 |
| | Etxaniz et al. (2024) | – | 63.3 | 62.6 | 72.0 | 70.8 | 71.2 | 62.4 | 68.1 | 63.0 | 61.3 | 60.3 |
| | Nguyen et al. (2024) | – | 64.7 | 64.4 | 73.0 | 70.9 | 71.8 | 63.2 | 67.8 | 63.9 | 61.7 | 61.0 |
| | BMF-ICL | **72.0** † | **65.6** | **66.2** † | **74.3** † | **72.2** † | **73.7** † | **64.2** | **69.2** | **65.1** † | **64.9** | **63.7** † |
| w/o TL | Non-ICL | 62.8 | 57.9 | 51.5 | 62.0 | 61.2 | 61.2 | 54.2 | 63.3 | 53.3 | 54.4 | 55.6 |
| | Winata et al. (2021) | – | 59.4 | 52.3 | 63.6 | 63.0 | 62.4 | 54.7 | 63.7 | 54.3 | 55.1 | 56.4 |
| | Winata et al. (2022) | 62.6 | 60.8 | 52.8 | 62.6 | 65.0 | 63.2 | 55.3 | 64.1 | 55.7 | 55.7 | 56.6 |
| | Nie et al. (2023) | 65.5 | 60.5 | 53.5 | 63.8 | 65.4 | 63.8 | 55.7 | 64.9 | 59.1 | 56.0 | 56.8 |
| | BMF-ICL | **66.3** | **64.9** † | **55.8** † | **65.9** † | **67.9** † | **65.6** † | **55.9** | **65.5** | **61.4** † | **58.0** † | **58.7** † |

(a) Llama 4.

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 59.3 | 57.0 | 56.3 | 67.0 | 62.0 | 63.4 | 58.7 | **61.7** | 59.3 | 60.1 | 54.3 |
| | Etxaniz et al. (2024) | - | 56.3 | 55.5 | 69.1 | 61.1 | 62.2 | 56.0 | 60.3 | 60.1 | 53.1 | 52.7 |
| | Nguyen et al. (2024) | - | 57.8 | 56.0 | 70.1 | 62.4 | 63.1 | 57.3 | 61.1 | 59.6 | 53.9 | 54.0 |
| | BMF-ICL | **62.5** † | **59.9** † | **59.5** † | **71.5** † | **63.5** † | **65.8** † | **60.0** † | 61.2 | **60.9** | **61.3** † | **58.9** † |
| w/o TL | Non-ICL | 40.5 | 47.5 | 31.5 | 45.6 | 37.8 | 33.9 | 36.6 | 45.7 | 38.0 | 37.5 | 40.1 |
| | Winata et al. (2021) | - | 50.2 | 32.5 | 48.4 | 40.3 | 35.4 | 37.8 | 46.0 | 39.7 | 39.0 | 41.4 |
| | Winata et al. (2022) | 45.1 | 53.3 | 33.1 | 47.0 | 43.9 | 36.1 | 39.1 | 46.7 | 40.5 | 40.4 | 42.5 |
| | Nie et al. (2023) | 49.0 | 51.7 | 34.2 | 48.2 | 44.7 | 36.6 | 39.6 | 48.8 | 50.1 | 40.0 | 42.2 |
| | BMF-ICL | **50.6** † | **55.4** † | **40.2** † | **52.6** † | **46.8** † | **40.1** † | **40.3** | **50.4** † | **53.8** † | **46.4** † | **48.5** † |

(b) Aya.

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 75.9 | **68.2** | 66.8 | 72.3 | 73.8 | 75.6 | 65.4 | 71.9 | 65.3 | 65.8 | 63.3 |
| | Etxaniz et al. (2024) | - | 66.3 | 65.7 | 73.2 | 75.0 | 75.0 | 65.2 | 71.5 | 64.2 | 64.8 | 63.5 |
| | Nguyen et al. (2024) | - | 67.7 | 68.0 | 74.2 | 74.6 | 75.5 | 65.7 | 70.7 | 65.8 | 65.1 | 64.0 |
| | BMF-ICL | 76.1 | 68.0 | **69.1** † | **75.5** † | **76.0** † | **77.1** † | **66.0** | **72.6** | **66.9** † | **66.5** | **65.8** † |
| w/o TL | Non-ICL | 72.3 | 62.3 | 60.1 | 69.0 | 71.3 | 72.9 | 61.7 | 70.8 | 59.9 | 61.6 | 62.3 |
| | Winata et al. (2021) | - | 63.3 | 60.8 | 70.1 | 72.8 | 74.0 | 62.0 | 71.3 | 60.6 | 62.0 | 62.9 |
| | Winata et al. (2022) | 70.1 | 64.0 | 61.3 | 69.3 | 74.1 | 74.8 | 62.2 | 71.5 | 62.2 | 62.2 | 62.6 |
| | Nie et al. (2023) | 72.6 | 64.2 | 61.7 | 70.5 | 74.3 | 75.4 | **62.6** | 71.8 | 63.0 | 62.8 | 63.0 |
| | BMF-ICL | **73.0** | **68.9** † | **62.5** | **71.6** † | **76.9** † | **76.6** † | **62.6** | **72.0** | **64.6** † | **63.0** | **63.1** |

(c) GPT-3.5.

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | Random-ICL | 80.1 | 69.2 | 65.1 | 73.0 | 75.1 | 76.0 | 64.0 | 72.0 | 66.4 | 66.0 | 65.0 |
| | Etxaniz et al. (2024) | - | 68.3 | 65.5 | 73.8 | 74.7 | 75.5 | 63.0 | 72.8 | 65.0 | 66.6 | 65.5 |
| | Nguyen et al. (2024) | - | 69.3 | **66.0** | 74.0 | 76.0 | 75.9 | 63.5 | 71.7 | **67.8** | 65.4 | 66.3 |
| | BMF-ICL | **80.7** | **71.5** † | **66.0** | **75.7** † | **76.6** | **76.4** | **64.2** | **74.8** † | 67.6 | **66.3** | **67.1** |
| w/o TL | Non-ICL | 77.3 | 63.0 | 62.2 | 70.7 | 74.3 | 72.9 | 62.0 | 70.1 | 62.7 | 64.1 | 61.7 |
| | Winata et al. (2021) | - | 63.9 | 64.0 | 71.6 | 74.6 | 73.2 | 62.5 | 71.1 | 63.9 | 65.0 | 62.7 |
| | Winata et al. (2022) | 78.6 | 64.1 | 64.8 | 71.2 | 74.9 | 74.2 | 62.8 | 70.7 | 64.5 | 64.7 | 62.2 |
| | Nie et al. (2023) | 79.1 | 64.5 | 65.0 | 71.1 | 74.7 | 74.1 | 63.1 | 71.3 | 64.0 | 65.3 | 63.1 |
| | BMF-ICL | **80.5** † | **66.8** † | **66.1** † | **72.5** | 75.1 | **75.9** † | **63.7** | **71.8** | **65.7** † | **65.8** | **64.5** † |

(d) GPT-4.

Table 2: Results for baseline ICL methods and our method on TYDI across the four LLMs. Red and blue indicate scores lower and higher than Random-ICL or Non-ICL. The top half of each table shows results with the target language in the example pool ("w/ TL"), and the bottom half without ("w/o TL"). † indicates statistically significant differences between the highest and second highest score in each LLM according to McNemar's test ($p < 0.01$).

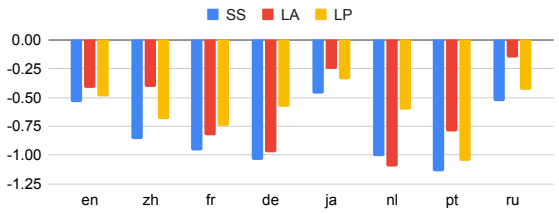
from the English candidate set.

- Winata et al. (2022): We randomly select examples for ICL from a pool across various languages excluding the target language.
- Nie et al. (2023): We use XLM-based (Conneau et al., 2020) multilingual sentence embeddings to select examples in high-resource languages similar to the input text in a low-resource language. Following previous work, we use English, German, and Chinese as high-resource languages for mCSQA, and English and Arabic for TYDI.

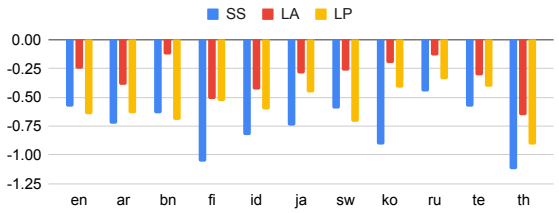
### 3.2 Experimental Results

Table 1 and Table 2 present the test set results on the mCSQA and TYDI datasets, for the four LLMs. The upper half of each table represents the setting where instances of the target language are included in the example candidates, while the lower half represents the setting where they are not included.

Comparisons with baselines show that BMF-ICL consistently achieves the best performance on mCSQA (in **28/32** cases) when target-language examples are included and also (in **27/32** cases) when they are not. On TYDI, it attains the high-




(a) Accuracy variations on the mCSQA dataset.




(b) Accuracy variations on the TYDI dataset.

Figure 3: Results of the ablation study. Semantic similarity, linguistic alignment, and language-specific performance are denoted as SS, LA, and LP, respectively.

| | mCSQA | | | | TYDI | | | |
|---|---|---|---|---|---|---|---|---|
| | All | SS | LA | LP | All | SS | LA | LP |
| 1 | 0.04 | 0.00 | 0.10 | 0.11 | 0.01 | 0.01 | 0.01 | 0.04 |
| 2 | 0.18 | 0.01 | **0.32** | **0.31** | 0.13 | 0.11 | 0.11 | 0.18 |
| 3 | **0.32** | **0.43** | 0.25 | 0.27 | 0.28 | 0.19 | **0.30** | **0.29** |
| 4 | 0.25 | 0.31 | 0.15 | 0.21 | **0.31** | 0.21 | 0.28 | 0.21 |
| 5 | 0.10 | 0.16 | 0.08 | 0.07 | 0.13 | **0.30** | 0.18 | 0.11 |
| 6 | 0.07 | 0.07 | 0.06 | 0.02 | 0.08 | 0.10 | 0.09 | 0.09 |
| 7 | 0.03 | 0.03 | 0.03 | 0.01 | 0.05 | 0.07 | 0.02 | 0.06 |
| 8 | 0.01 | 0.00 | 0.01 | 0.00 | 0.01 | 0.01 | 0.01 | 0.01 |

Table 3: The proportion of instances with each language type count within the 8 examples of BMF-ICL. *All* represents the results of a method that considers three factors with optimized weights. SS, LA, and LP represent results considering only semantic similarity, linguistic alignment, and language-specific performance, respectively.

est performance in **39/44** cases under the target-language-included setting and in **42/44** cases under the target-language-excluded setting. These findings confirm that BMF-ICL achieves state-of-the-art performance overall. Its improvements are observed across all four tested LLMs, indicating that BMF-ICL is not heavily model-dependent. Furthermore, BMF-ICL does not exhibit poor performance on any particular language. Notably, the method is more effective in cases where the target language is not included among the example candidates. In contrast, existing approaches are not always better than the Random-ICL or Non-ICL baselines, particularly in the target-language-included setting, underscoring the challenge of achieving stable improvements without quantitative optimization.

# 4 Analysis

## 4.1 Ablation Study

To assess the importance of each factor in BNF-ICL, we conducted an ablation study by removing one factor at a time. For example, to assess the impact of semantic similarity, we set $\alpha = 0$ in Equation 3 and optimize the remaining weights $\beta$ and $\gamma$.[15] Figure 3 displays the ablation results under the setting where the target language is included among the candidate examples for both mCSQA and TYDI, averaged over the four MLLMs.

[15]The results with all weights set to 1 are shown in Appendix D.

Across all languages, performance declines whenever one of the three factors is excluded, indicating the importance of jointly considering all three. In mCSQA, 7 out of 8 languages, and in TYDI, 8 out of 11 languages, exhibited the largest drop when the semantic similarity factor was removed. This suggests that, as in monolingual settings (Liu et al., 2022), providing semantically similar examples contributes to performance improvement in multilingual settings. Additionally, within closely-related language groups such as en–de–nl or fr–pt, linguistic alignment plays a crucial role leading to pronounced performance declines compared to other languages when removed.

## 4.2 The Diversity of Languages in the Examples

BMF-ICL demonstrates improved performance through cross-linguistic knowledge transfer, as clarified in our study. For further analysis, we examined how many distinct language types are included among the 8 examples in each ICL prompt. Based on these counts, we then computed the proportion of test instances associated with each distinct language-type count. We report results for two settings: (i) BMF-ICL optimized with all three factors, and (ii) ablation settings where only one factor is set to 1 in Eq. (3), while the others are set to 0, allowing us to assess each factor's impact on language diversity.

Table 3 shows the distribution of distinct language-type counts for examples selected by BMF-ICL when the target language is included among the candidates. The All column represents results considering all weights, while SS, LA, and PL correspond to each factor used in isolation. Un-

der All, the most frequently selected diversity level is three languages in mCSQA and four in TYDI, indicating that BMF-ICL naturally favors multilingual examples. Furthermore, the semantic similarity factor (SS) tends to yield higher language diversity than the other two factors, underscoring its particular importance for encouraging more varied language selection.

## 5 Related Work

Etxaniz et al. (2024) showed that translating low-resource language inputs and examples into English improves LLM performance compared to direct inference in the original language. This approach leverages the English-centric training of most LLMs, but may not fully capture linguistic, cultural, or societal norms. Additionally, using translated examples for ICL risks information loss or distortion, as LLMs struggle with accurately conveying cultural or societal nuances (Yao et al., 2023; Tenzer et al., 2024; Intrator et al., 2024).

Winata et al. (2021) discovered that providing English examples for ICL improves LLM inference for both English and non-English tasks, though English was heuristically chosen. Winata et al. (2022) showed that randomly sampling from a multilingual dataset outperforms selecting examples based on geographical or linguistic proximity. However, the role of semantic alignment and language-specific capacity in example selection remains unclear in the original work.

Nie et al. (2023) introduced a method that uses multilingual sentence embeddings (Conneau et al., 2020) to select examples in high-resource language similar to the input text in low-resource language. The multilingual sentence embeddings do not explicitly distinguish between semantic and linguistic similarity, making it impossible to adjust their optimal balance for ICL examples. Moreover, this study focuses on only masked language models such as mBERT (Devlin et al., 2019) and XLM (Conneau et al., 2020) rather than LLMs.

To leverage unlabeled datasets in low-resource languages, Nguyen et al. (2024) address the data scarcity in low-resource languages using instances from diverse high-resource languages as ICL examples to create synthetic data from unlabeled datasets in low-resource languages, which are then used as ICL examples in a low-resource setting. However, this method does not consider the similarity between the input and example texts.

The following studies have proposed MICL methods specialized for binary classification tasks. Tanwar et al. (2023) proposed a method that uses multilingual sentence embeddings (Reimers and Gurevych, 2020) to retrieve similar texts in another language as examples for ICL in a cross-lingual setting. This method explicitly presents cross-lingual label correspondences (e.g., *In French, "bad" means "mal"*). Cahyawijaya et al. (2024) introduced query alignment for ICL, selecting examples from parallel data with source texts that match the input language and target texts in high-resource languages. This method used multilingual sentence embeddings (Reimers and Gurevych, 2019, 2020) to measure the similarity between the input text and the source texts in the parallel data, selecting semantically similar texts as examples. The labels from the high-resource language are used directly, avoiding translation errors. Unlike these existing studies, which focus on binary classification tasks, our study applies ICL methods to more general generative tasks.

Qin et al. (2023) introduced a method that processes inputs in languages other than English by using the prompt *Let's think in English step by step!* to enable step-by-step reasoning in English. This method consistently improves the performance in languages other than English. Shi et al. (2023) also demonstrates that step-by-step reasoning enhances the multilingual capabilities of MLLMs. Unlike our research, which focuses on multilingual knowledge transfer through examples in ICL, this study emphasizes multilingual knowledge transfer within the reasoning process.

## 6 Conclusion

In this paper, we propose BMF-ICL, an approach for multilingual example selection in ICL for MLLMs. BMF-ICL quantifies and balances three key factors: semantic similarity, linguistic alignment, and language-specific performance. By leveraging LaBSE (Feng et al., 2022) embeddings for semantic similarity, lang2vec (Littell et al., 2017) for linguistic alignment, and MLLM likelihoods for language-specific performance, BMF-ICL optimally selects examples through a weighted scoring mechanism. Experimental results on the mCSQA and TYDI datasets, using four different MLLMs, demonstrated that BMF-ICL consistently achieves higher accuracy than existing methods.

## Limitations

We demonstrated the effectiveness of the proposed method by conducting large-scale experiments in various languages; however, this does not guarantee performance improvements in all languages. As future work, it would be worthwhile to validate the method on a broader range of tasks beyond question answering. On the other hand, since there are not many multilingual datasets created from scratch for each language, this is an aspect that needs to be considered from the dataset creation stage.

## Ethical Considerations

mCSQA (Sakai et al., 2024) is a dataset that reflects common sense across different cultures, and our experimental results indicate that the proposed method enhances the understanding of common sense within each culture by leveraging multilingual information. Therefore, it also has the potential to positively impact safety-related tasks such as social biases, morality, and ethics (Kaneko et al., 2022, 2024; Kaneko and Baldwin, 2024; Anantaprayoon et al., 2023; Hämmerl et al., 2023), where multicultural factors play a significant role.

## References


Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan Alrashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, et al. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13787–13805.

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. 2023. Evaluating gender bias of pre-trained language models in natural language inference by considering all labels. arXiv preprint arXiv:2309.09697.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. ArXiv, abs/2005.14165.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. LLMs are few-shot in-context low-resource language learners. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 405–433, Mexico City, Mexico. Association for Computational Linguistics.

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

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

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. arXiv preprint arXiv:2207.04672.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Błażej Dolicki and Gerasimos Spanakis. 2021. Analysing the impact of linguistic features on cross-lingual transfer. arXiv preprint arXiv:2105.05975.

Julen Etxaniz, Gorka Azkune, Aitor Soroa, Oier Lacalle, and Mikel Artetxe. 2024. 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), pages 550–564, Mexico City, Mexico. Association for Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic BERT sentence embedding. In Proceedings of the 60th Annual Meeting of the Association

for Computational Linguistics (Volume 1: Long Papers), pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Katharina Hämmerl, Bjoern Deiseroth, Patrick Schramowski, Jindřich Libovický, Constantin Rothkopf, Alexander Fraser, and Kristian Kersting. 2023. Speaking multiple languages affects the moral bias of language models. In Findings of the Association for Computational Linguistics: ACL 2023, pages 2137–2156, Toronto, Canada. Association for Computational Linguistics.

Rem Hida, Masahiro Kaneko, and Naoaki Okazaki. 2024. Social bias evaluation for large language models requires prompt variations. arXiv preprint arXiv:2407.03129.

Yotam Intrator, Matan Halfon, Roman Goldenberg, Reut Tsarfaty, Matan Eyal, Ehud Rivlin, Yossi Matias, and Natalia Aizenberg. 2024. Breaking the language barrier: Can direct inference outperform pre-translation in multilingual LLM applications? In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers), pages 829–844, Mexico City, Mexico. Association for Computational Linguistics.

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. 2017. Google's multilingual neural machine translation system: Enabling zero-shot translation. Transactions of the Association for Computational Linguistics, 5:339–351.

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

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016b. Bag of tricks for efficient text classification. arXiv preprint arXiv:1607.01759.

Masahiro Kaneko and Timothy Baldwin. 2024. A little leak will sink a great ship: Survey of transparency for large language models from start to finish. arXiv preprint arXiv:2403.16139.

Masahiro Kaneko, Danushka Bollegala, and Timothy Baldwin. 2024. Eagle: Ethical dataset given from real interactions. arXiv preprint arXiv:2402.14258.

Masahiro Kaneko, Aizhan Imankulova, Danushka Bollegala, and Naoaki Okazaki. 2022. Gender bias in masked language models for multiple languages. arXiv preprint arXiv:2205.00551.

Ryuto Koike, Masahiro Kaneko, and Naoaki Okazaki. 2024. How you prompt matters! Even task-oriented constraints in instructions affect LLM-generated text detection. In Findings of the Association for Computational Linguistics: EMNLP 2024, pages 14384–14395, Miami, Florida, USA. Association for Computational Linguistics.

Patrick Littell, David R. Mortensen, Ke Lin, Katherine Kairis, Carlisle Turner, and Lori Levin. 2017. 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, pages 8–14, Valencia, Spain. Association for Computational Linguistics.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. 2022. 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, pages 100–114, Dublin, Ireland and Online. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Meta AI. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. https://ai.meta.com/blog/llama-4-multimodal-intelligence/. Accessed: 2025-05-17.

Xuan-Phi Nguyen, Mahani Aljunied, Shafiq Joty, and Lidong Bing. 2024. 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), pages 3501–3516, Bangkok, Thailand. Association for Computational Linguistics.

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

Daisuke Oba, Masahiro Kaneko, and Danushka Bollegala. 2024. In-contextual gender bias suppression for large language models. In Findings of the Association for Computational Linguistics: EACL 2024, pages 1722–1742.

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

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

Libo Qin, Qiguang Chen, Fuxuan Wei, Shijue Huang, and Wanxiang Che. 2023. Cross-lingual prompting: Improving zero-shot chain-of-thought reasoning across languages. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 2695–2709, Singapore. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 4512–4525, Online. Association for Computational Linguistics.

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

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

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ili'c, Daniel Hesslow, Roman Castagn'e, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurenccon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa Etxabe, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris C. Emezue, Christopher Klamm, Colin Leong, Daniel Alexander van Strien, David Ifeoluwa Adelani, Dragomir R. Radev, Eduardo Gonz'alez Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady ElSahar, Hamza Benyamina, Hieu Trung Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jorg Frohberg, Josephine Tobing, Joydeep Bhattacharjee, Khalid Almubarak, Kimbo Chen, Kyle Lo, Leandro von Werra, Leon Weber, Long Phan, Loubna Ben Allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, Mar'ia Grandury, Mario vSavsko, Max Huang, Maximin Coavoux, Mayank Singh, Mike Tian-Jian Jiang, Minh Chien Vu, Mohammad A. Jauhar, Mustafa Ghaleb, Nishant Subramani, Nora Kassner, Nurulaqilla Khamis, Olivier Nguyen, Omar Espejel, Ona de Gibert, Paulo Villegas, Peter Henderson, Pierre Colombo, Priscilla Amuok, Quentin Lhoest, Rheza Harliman, Rishi Bommasani, Roberto L'opez, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, S. Longpre, Somaieh Nikpoor, S. Silberberg, Suhas Pai, Sydney Zink, Tiago Timponi Torrent, Timo Schick, Tristan Thrush, Valentin Danchev, Vassilina Nikoulina, Veronika Laippala, Violette Lepercq, Vrinda Prabhu, Zaid Alyafeai, Zeerak Talat, Arun Raja, Benjamin Heinzerling, Chenglei Si, Elizabeth Salesky, Sabrina J. Mielke, Wilson Y. Lee, Abheesht Sharma, Andrea Santilli, Antoine Chaffin, Arnaud Stiegler, Debajyoti Datta, Eliza Szczechla, Gunjan Chhablani, Han Wang, Harshit Pandey, Hendrik Strobelt, Jason Alan Fries, Jos Rozen, Leo Gao, Lintang Sutawika, M Saiful Bari, Maged S. Al-Shaibani, Matteo Manica, Nihal V. Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Févry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiang Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Y Uri, Hadar Tojarieh, Adam Roberts, Hyung Won Chung, Jaesung Tae, Jason Phang, Ofir Press, Conglong Li, Deepak Narayanan, Hatim Bourfoune, Jared Casper, Jeff Rasley, Max Ryabinin, Mayank Mishra, Minjia Zhang, Mohammad Shoeybi, Myriam Peyrounette, Nicolas Patry, Nouamane Tazi, Omar Sanseviero, Patrick von Platen, Pierre Cornette, Pierre Franccois Lavall'ee, Rémi Lacroix, Samyam Rajbhandari, Sanchit Gandhi, Shaden Smith, Stéphane Requena, Suraj Patil, Tim Dettmers, Ahmed Baruwa, Amanpreet Singh, Anastasia Cheveleva, Anne-Laure Ligozat, Arjun Subramonian, Aur'elie N'ev'eol, Charles Lovering, Daniel H Garrette, Deepak R. Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Xiangru Tang, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, S. Osher Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdenvek Kasner, Zdeněk Kasner, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ananda Santa Rosa Santos, Anthony Hevia, Antigona Unl-

dreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ayoade Ajibade, Bharat Kumar Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David M. Lansky, Davis David, Douwe Kiela, Duong Anh Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatim Tahirah Mirza, Frankline Ononiwu, Habib Rezanejad, H.A. Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jan Passmore, Joshua Seltzer, Julio Bonis Sanz, Karen Fort, Lívia Dutra, Mairon Samagaio, Maraim Elbadri, Margot Mieskes, Marissa Gerchick, Martha Akinlolu, Michael McKenna, Mike Qiu, Muhammed Ghauri, Mykola Burynok, Nafis Abrar, Nazneen Rajani, Nour Elkott, Nourhan Fahmy, Olanrewaju Samuel, Ran An, R. P. Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas L. Wang, Sourav Roy, Sylvain Viguier, Thanh-Cong Le, Tobi Oyebade, Trieu Nguyen Hai Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Kumar Singh, Benjamin Beilharz, Bo Wang, Caio Matheus Fonseca de Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel Le'on Perin'an, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Iman I.B. Bello, Isha Dash, Ji Soo Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthi Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, María Andrea Castillo, Marianna Nezhurina, Mario Sanger, Matthias Samwald, Michael Cullan, Michael Weinberg, M Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patricia Haller, Patrick Haller, Renata Eisenberg, Robert Martin, Rodrigo Canalli, Rosaline Su, Ruisi Su, Samuel Cahyawijaya, Samuele Garda, Shlok S Deshmukh, Shubhanshu Mishra, Sid Kiblawi, Simon Ott, Sinee Sang-aroonsiri, Srishti Kumar, Stefan Schweter, Sushil Pratap Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yashasvi Bajaj, Y. Venkatraman, Yifan Xu, Ying Xu, Yu Xu, Zhee Xao Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. Bloom: A 176b-parameter open-access multilingual language model. ArXiv, abs/2211.05100.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language models are multilingual chain-of-thought reasoners. In The Eleventh International Conference on Learning Representations.

Eshaan Tanwar, Subhabrata Dutta, Manish Borthakur, and Tanmoy Chakraborty. 2023. 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), pages 6292–6307, Toronto, Canada. Association for Computational Linguistics.

Helene Tenzer, Stefan Feuerriegel, and Rebecca Piekkari. 2024. Ai machine translation tools must be taught cultural differences too. Nature, 630(8018):820–820.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, et al. 2024. Aya model: An instruction finetuned open-access multilingual language model. arXiv preprint arXiv:2402.07827.

Genta Winata, Shijie Wu, Mayank Kulkarni, Thamar Solorio, and Daniel Preotiuc-Pietro. 2022. 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), pages 777–791, Online only. Association for Computational Linguistics.

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

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 483–498, Online. Association for Computational Linguistics.

Ikumi Yamashita, Satoru Katsumata, Masahiro Kaneko, Aizhan Imankulova, and Mamoru Komachi. 2020. Cross-lingual transfer learning for grammatical error correction. In Proceedings of the 28th International Conference on Computational Linguistics, pages 4704–4715.

Binwei Yao, Ming Jiang, Diyi Yang, and Junjie Hu. 2023. Benchmarking llm-based machine translation on cultural awareness. arXiv preprint arXiv:2305.14328.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In International conference on machine learning, pages 12697–12706. PMLR.

| | Language group | mCSQA Train | mCSQA Valid | mCSQA Test | TYDI Train | TYDI Valid | TYDI Test |
|---|---|---|---|---|---|---|---|
| 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 |

Table 4: Dataset statistics and language groups for mCSQA and TYDI.

## A Dataset Statistics

Table 4 shows the data size and language groups for mCSQA (Sakai et al., 2024) and TYDI (Clark et al., 2020).

## B Prompts

The following are four candidate prompts for mCSQA in English; each instruction is translated into the corresponding target language.

**Prompt 1 for 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:
```

**Prompt 2 for mCSQA**

```
Provide a response to 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:
```

**Prompt 3 for mCSQA**

```
Please 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:
```

**Prompt 4 for mCSQA**

```
Respond to 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:
```

The following are four candidate prompts for TYDI in English; each instruction is translated into the corresponding target language.

**Prompt 1 for 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:
```

**Prompt 2 for TYDI**

```
Provide an answer to the question based on 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:
```

**Prompt 3 for TYDI**

```
Please give an answer to the question using the provided
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:
```

**Prompt 4 for TYDI**

```
Please answer the question by utilizing 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:
```

## C The Weights of the Three Factors

Table 5 shows weights for semantic similarity with weight $\alpha$, linguistic alignment with weight $\beta$, and language-specific performance with weight $\gamma$ in Equation 3 in BMF-ICL for each language in the mCSQA and TYDI datasets.

## D BMF-ICL with Uniform Weights

Table 6 shows the extent to which performance degrades when the weights $\alpha$, $\beta$, and $\gamma$ of BMF-ICL are all set to 1, difference compared to the performance of MBF-ICL with optimized weights. From the experimental results, it can be observed that optimizing the weights across all settings contributes to performance improvement.

## E Few-shot Results with 2, 4, and 16 Examples

Table 7 and Table 8 show BMF-ICL performance for the 2-, 4-, and 16-shot settings on both mCSQA and TYDI.

| | Aya | | | Llama 4 | | | GPT-3.5 | | | GPT-4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Language** | $\alpha$ | $\beta$ | $\gamma$ | $\alpha$ | $\beta$ | $\gamma$ | $\alpha$ | $\beta$ | $\gamma$ | $\alpha$ | $\beta$ | $\gamma$ |
| Chinese | 0.4 | 0.4 | 0.2 | 0.5 | 0.3 | 0.2 | 0.6 | 0.2 | 0.2 | 0.7 | 0.1 | 0.2 |
| English | 0.5 | 0.3 | 0.2 | 0.6 | 0.2 | 0.2 | 0.7 | 0.1 | 0.2 | 0.8 | 0.0 | 0.2 |
| French | 0.4 | 0.3 | 0.3 | 0.5 | 0.2 | 0.3 | 0.6 | 0.1 | 0.3 | 0.7 | 0.0 | 0.3 |
| German | 0.4 | 0.5 | 0.1 | 0.5 | 0.4 | 0.1 | 0.6 | 0.3 | 0.1 | 0.7 | 0.2 | 0.1 |
| Japanese | 0.4 | 0.3 | 0.3 | 0.5 | 0.2 | 0.3 | 0.6 | 0.1 | 0.3 | 0.7 | 0.0 | 0.3 |
| Dutch | 0.4 | 0.4 | 0.2 | 0.5 | 0.3 | 0.2 | 0.6 | 0.2 | 0.2 | 0.7 | 0.1 | 0.2 |
| Portuguese | 0.6 | 0.1 | 0.3 | 0.7 | 0.0 | 0.3 | 0.8 | 0.0 | 0.2 | 0.9 | 0.0 | 0.1 |
| Russian | 0.5 | 0.3 | 0.2 | 0.6 | 0.2 | 0.2 | 0.7 | 0.1 | 0.2 | 0.8 | 0.0 | 0.2 |

(a) mCSQA.

| | Aya | | | Llama 4 | | | GPT-3.5 | | | GPT-4 | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Language** | $\alpha$ | $\beta$ | $\gamma$ | $\alpha$ | $\beta$ | $\gamma$ | $\alpha$ | $\beta$ | $\gamma$ | $\alpha$ | $\beta$ | $\gamma$ |
| Arabic | 0.5 | 0.4 | 0.1 | 0.6 | 0.3 | 0.1 | 0.7 | 0.2 | 0.1 | 0.8 | 0.1 | 0.1 |
| Bengali | 0.4 | 0.1 | 0.5 | 0.5 | 0.0 | 0.5 | 0.6 | 0.0 | 0.4 | 0.7 | 0.0 | 0.3 |
| English | 0.6 | 0.3 | 0.1 | 0.7 | 0.2 | 0.1 | 0.8 | 0.1 | 0.1 | 0.9 | 0.0 | 0.1 |
| Finnish | 0.4 | 0.3 | 0.3 | 0.5 | 0.2 | 0.3 | 0.6 | 0.1 | 0.3 | 0.7 | 0.0 | 0.3 |
| Indonesian | 0.4 | 0.2 | 0.4 | 0.5 | 0.1 | 0.4 | 0.6 | 0.0 | 0.4 | 0.7 | 0.0 | 0.3 |
| Japanese | 0.6 | 0.2 | 0.2 | 0.7 | 0.1 | 0.2 | 0.8 | 0.0 | 0.2 | 0.9 | 0.0 | 0.1 |
| Kiswahili | 0.4 | 0.4 | 0.2 | 0.6 | 0.2 | 0.2 | 0.6 | 0.2 | 0.2 | 0.7 | 0.1 | 0.2 |
| Korean | 0.5 | 0.4 | 0.1 | 0.7 | 0.2 | 0.1 | 0.7 | 0.2 | 0.1 | 0.8 | 0.1 | 0.1 |
| Russian | 0.5 | 0.4 | 0.1 | 0.7 | 0.2 | 0.1 | 0.7 | 0.2 | 0.1 | 0.8 | 0.1 | 0.1 |
| Telugu | 0.4 | 0.2 | 0.4 | 0.6 | 0.0 | 0.4 | 0.6 | 0.0 | 0.4 | 0.7 | 0.0 | 0.3 |
| Thai | 0.4 | 0.2 | 0.4 | 0.6 | 0.0 | 0.4 | 0.6 | 0.0 | 0.4 | 0.7 | 0.0 | 0.3 |

(b) TYDI.

Table 5: The weights for semantic similarity (SS) with weight $\alpha$, linguistic alignment (LA) with weight $\beta$, and language-specific performance (LP) with weight $\gamma$ in BMF-ICL for every language with each LLM in the mCSQA and TYDI datasets.

## F Lang2vec vs. Heuristic Language Choice

To assess whether our continuous, lang2vec truly offers an advantage over simpler heuristics, we ran a focused ablation on mCSQA for German and Dutch. Because existing studies (Winata et al., 2021; Etxaniz et al., 2024) almost invariably choose English as the example pool's language, we run our mCSQA experiments with English and target languages that are linguistically close to it, specifically German and Dutch. Table 9 shows that using continuous lang2vec similarity consistently outperforms the heuristic setting.

| | Llama 4 | Aya | GPT-3.5 | GPT-4 |
|---|---|---|---|---|
| Chinese | -5.2 | -3.2 | -2.4 | -2.3 |
| English | -5.4 | -4.2 | -3.0 | -2.6 |
| French | -4.9 | -3.3 | -2.7 | -2.5 |
| German | -4.6 | -3.3 | -2.3 | -1.9 |
| Japanese | -5.0 | -2.9 | -2.6 | -2.6 |
| Dutch | -4.4 | -2.6 | -2.4 | -2.1 |
| Portuguese | -5.2 | -3.0 | -2.2 | -2.1 |
| Russian | -4.7 | -3.6 | -2.3 | -1.8 |

(a) mCSQA.

| | Llama 4 | Aya | GPT-3.5 | GPT-4 |
|---|---|---|---|---|
| Arabic | -6.6 | -5.1 | -3.9 | -3.2 |
| Bengali | -6.2 | -4.4 | -3.7 | -3.5 |
| English | -.7.0 | -6.6 | -4.8 | -4.3 |
| Finnish | -5.9 | -4.2 | -3.7 | -3.3 |
| Indonesian | -5.5 | -4.7 | -4.0 | -4.1 |
| Japanese | -6.3 | -5.3 | -3.6 | -4.0 |
| Kiswahili | -5.2 | -4.7 | -2.9 | -2.7 |
| Korean | -6.1 | -5.2 | -4.0 | -4.3 |
| Russian | -5.9 | -4.6 | -4.1 | -3.6 |
| Telugu | -4.9 | -3.9 | -3.2 | -3.0 |
| Thai | -5.3 | -4.4 | -3.6 | -3.3 |

(b) TYDI.

Table 6: Performance drop of BMF-ICL when using uniform weights instead of optimized weights.

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 71.8 | 57.8 | 73.5 | 77.6 | 67.2 | 76.1 | 77.2 | 48.1 |
| | 4-shot | 73.8 | 59.8 | 75.5 | 79.6 | 69.2 | 78.1 | 79.2 | 50.1 |
| | 16-shot | 76.3 | 62.3 | 78.0 | 82.1 | 71.7 | 80.6 | 81.7 | 52.6 |
| w/o TL | 2-shot | 69.1 | 53.0 | 68.1 | 72.5 | 64.1 | 72.4 | 65.8 | 45.6 |
| | 4-shot | 70.6 | 54.5 | 69.6 | 74.0 | 65.6 | 73.9 | 67.3 | 47.1 |
| | 16-shot | 72.1 | 56.0 | 71.1 | 75.5 | 67.1 | 75.4 | 68.8 | 48.6 |

(a) Llama 4

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 58.7 | 39.9 | 60.0 | 64.8 | 44.2 | 60.7 | 64.9 | 32.0 |
| | 4-shot | 60.7 | 41.9 | 62.0 | 66.8 | 46.2 | 62.7 | 66.9 | 34.0 |
| | 16-shot | 63.2 | 44.4 | 64.5 | 69.3 | 48.7 | 65.2 | 69.4 | 36.5 |
| w/o TL | 2-shot | 50.1 | 29.3 | 48.6 | 52.5 | 35.5 | 51.3 | 40.0 | 28.3 |
| | 4-shot | 51.6 | 30.8 | 50.1 | 54.0 | 37.0 | 52.8 | 41.5 | 29.8 |
| | 16-shot | 53.1 | 32.3 | 51.6 | 55.5 | 38.5 | 54.3 | 43.0 | 31.3 |

(b) Aya

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 77.4 | 65.5 | 79.3 | 83.1 | 77.1 | 82.7 | 82.5 | 55.0 |
| | 4-shot | 79.4 | 67.5 | 81.3 | 85.1 | 79.1 | 84.7 | 84.5 | 57.0 |
| | 16-shot | 81.9 | 70.0 | 83.8 | 87.6 | 81.6 | 87.2 | 87.0 | 59.5 |
| w/o TL | 2-shot | 77.2 | 63.1 | 76.5 | 81.1 | 76.4 | 81.5 | 76.8 | 53.0 |
| | 4-shot | 78.7 | 64.6 | 78.0 | 82.6 | 77.9 | 83.0 | 78.3 | 54.5 |
| | 16-shot | 80.2 | 66.1 | 79.5 | 84.1 | 79.4 | 84.5 | 79.8 | 56.0 |

(c) GPT-3.5

| | | en | zh | fr | de | ja | nl | pt | ru |
|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 80.3 | 68.7 | 79.5 | 82.0 | 77.8 | 83.5 | 83.6 | 52.8 |
| | 4-shot | 82.3 | 70.7 | 81.5 | 84.0 | 79.8 | 85.5 | 85.6 | 54.8 |
| | 16-shot | 84.8 | 73.2 | 84.0 | 86.5 | 82.3 | 88.0 | 88.1 | 57.3 |
| w/o TL | 2-shot | 79.3 | 67.0 | 77.6 | 81.2 | 76.3 | 81.0 | 78.3 | 51.0 |
| | 4-shot | 80.8 | 68.5 | 79.1 | 82.7 | 77.8 | 82.5 | 79.8 | 52.5 |
| | 16-shot | 82.3 | 70.0 | 80.6 | 84.2 | 79.3 | 84.0 | 81.3 | 54.0 |

(d) GPT-4

Table 7: BMF-ICL performance with 2, 4, and 16 shots on mCSQA.

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 69.0 | 62.6 | 63.2 | 71.3 | 69.2 | 70.7 | 61.2 | 66.2 | 62.1 | 61.9 | 60.7 |
| | 4-shot | 71.0 | 64.6 | 65.2 | 73.3 | 71.2 | 72.7 | 63.2 | 68.2 | 64.1 | 63.9 | 62.7 |
| | 16-shot | 73.5 | 67.1 | 67.7 | 75.8 | 73.7 | 75.2 | 65.7 | 70.7 | 66.6 | 66.4 | 65.2 |
| w/o TL | 2-shot | 64.3 | 62.9 | 53.8 | 63.9 | 65.9 | 63.6 | 53.9 | 63.5 | 59.4 | 56.0 | 56.7 |
| | 4-shot | 65.8 | 64.4 | 55.3 | 65.4 | 67.4 | 65.1 | 55.4 | 65.0 | 60.9 | 57.5 | 58.2 |
| | 16-shot | 67.3 | 65.9 | 56.8 | 66.9 | 68.9 | 66.6 | 56.9 | 66.5 | 62.4 | 58.5 | 59.7 |

(a) Llama 4

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 59.5 | 56.9 | 56.5 | 68.5 | 60.5 | 62.8 | 57.0 | 58.2 | 57.9 | 58.3 | 55.9 |
| | 4-shot | 61.5 | 58.9 | 58.5 | 70.5 | 62.5 | 64.8 | 59.0 | 60.2 | 59.9 | 60.3 | 57.9 |
| | 16-shot | 64.0 | 61.4 | 61.0 | 73.0 | 65.0 | 67.3 | 61.5 | 62.7 | 62.4 | 62.8 | 60.4 |
| w/o TL | 2-shot | 48.6 | 53.4 | 38.2 | 50.6 | 44.8 | 38.1 | 38.3 | 48.4 | 51.8 | 44.4 | 46.5 |
| | 4-shot | 50.1 | 54.9 | 39.7 | 52.1 | 46.3 | 39.6 | 39.8 | 49.9 | 53.3 | 45.9 | 48.0 |
| | 16-shot | 51.6 | 56.4 | 41.2 | 53.6 | 47.8 | 41.1 | 41.3 | 51.4 | 54.8 | 47.4 | 49.5 |

(b) Aya

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 73.1 | 65.0 | 66.1 | 72.5 | 73.0 | 74.1 | 63.0 | 69.6 | 63.9 | 63.5 | 62.8 |
| | 4-shot | 75.1 | 67.0 | 68.1 | 74.5 | 75.0 | 76.1 | 65.0 | 71.6 | 65.9 | 65.5 | 64.8 |
| | 16-shot | 77.6 | 69.5 | 70.6 | 77.0 | 77.5 | 78.6 | 67.5 | 74.1 | 68.4 | 68.0 | 67.3 |
| w/o TL | 2-shot | 71.0 | 66.9 | 60.5 | 69.6 | 74.9 | 74.6 | 60.6 | 70.0 | 62.6 | 61.0 | 61.1 |
| | 4-shot | 72.5 | 68.4 | 62.0 | 71.1 | 76.4 | 76.1 | 62.1 | 71.5 | 64.1 | 62.5 | 62.6 |
| | 16-shot | 74.0 | 69.9 | 63.5 | 72.6 | 77.9 | 77.6 | 63.6 | 73.0 | 65.6 | 64.0 | 64.1 |

(c) GPT-3.5

| | | en | ar | bn | fi | id | ja | sw | ko | ru | te | th |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| w/ TL | 2-shot | 77.7 | 68.5 | 63.0 | 72.7 | 73.6 | 73.4 | 61.2 | 71.8 | 64.6 | 63.3 | 64.1 |
| | 4-shot | 79.7 | 70.5 | 65.0 | 74.7 | 75.6 | 75.4 | 63.2 | 73.8 | 66.6 | 65.3 | 66.1 |
| | 16-shot | 82.2 | 73.0 | 67.5 | 77.2 | 78.1 | 77.9 | 65.7 | 76.3 | 69.1 | 67.8 | 68.6 |
| w/o TL | 2-shot | 78.5 | 64.8 | 64.1 | 70.5 | 73.1 | 73.9 | 61.7 | 69.8 | 63.7 | 63.8 | 62.5 |
| | 4-shot | 80.0 | 66.3 | 65.6 | 72.0 | 74.6 | 75.4 | 63.2 | 71.3 | 65.2 | 65.3 | 64.0 |
| | 16-shot | 81.5 | 67.8 | 67.1 | 73.5 | 76.1 | 76.9 | 64.7 | 72.8 | 66.7 | 66.8 | 65.5 |

(d) GPT-4

Table 8: BMF-ICL performance with 2, 4, and 16 shots on TYDI.

| Model | Language | **lang2vec** | **heuristic** |
|---|---|---|---|
| Llama 4 | de | 75.3$^*$ | 70.1 |
| | nl | 70.1$^*$ | 63.6 |
| Aya | de | 54.5$^*$ | 53.3 |
| | nl | 53.3$^*$ | 52.0 |
| GPT-3.5 | de | 83.1$^*$ | 81.9 |
| | nl | 83.5$^*$ | 82.0 |
| GPT-4 | de | 83.2 | 82.8 |
| | nl | 83.0$^*$ | 82.0 |

Table 9: Comparison of linguistic-alignment strategies (lang2vec vs. heuristic English selection) on mCSQA. $*$ denotes a statistically significant improvement over the heuristic according to McNemar's test ($p < 0.01$).