Few-shot Prompting or Supervised Tuning? A Comparative Study of LLMs for Linguistically distant Language pairs in BDI

Deepen Naorem, Sanasam Ranbir Singh, Telem Joyson Singh, Priyankoo Sarmah
Indian Institute of Technology Guwahati, Assam, India
{deepennaorem, ranbir, tjoyson, priyankoo}@iitg.ac.in

Abstract

Bilingual Dictionary Induction (BDI) presents significant challenges in distant language pairs, particularly in light of the non-isomorphic nature and complexity of linguistic structures. This paper systematically evaluates the performance of unsupervised, supervised fine-tuning, and few-shot prompting approaches on BDI using Large Language Models (LLMs) on a diverse set of distant language pairs. The unsupervised approach explores the inherent multilingual capabilities of LLMs without fine-tuning, while the supervised fine-tuning method utilizes extensive labeled datasets to train models explicitly for BDI tasks. On the other hand, few-shot prompting leverages minimal examples to elicit accurate responses from the LLMs in a zero-shot or few-shot learning paradigm. Our experimental results reveal that the 5-shot prompting approach outperforms unsupervised and zero-shot settings in all cases and surpasses supervised settings in 82.86% of the cases. Few-shot prompting demonstrates robustness against overfitting, leveraging LLMs' in-context learning and multilingual capabilities, making it particularly effective in target-to-source translation, even for morphologically complex language pairs. At the same time, few-shot prompting in LLM models, such as Llama, remains ineffective for morphologically rich language pairs like En-Mn and En-Ta in source-to-target BDI tasks. These findings suggest that few-shot prompting is a cost-effective and powerful alternative for BDI tasks, with future work enhancing BDI tasks in morphologically rich pairs.

Keywords: Bilingual Dictionary Induction, Large Language Model, Prompting

1. Introduction

Bilingual Dictionary Induction (BDI) serves as a fundamental task in natural language processing (NLP), enabling the automatic creation of bilingual dictionaries by aligning word representations between languages (Ruder et al., 2019). This capability is critical for the development of cross-linguistic resources, particularly for applications such as machine translation (Artetxe et al., 2018b) and cross-lingual information retrieval (Vulić and Moens, 2015). Among widely adopted Cross-lingual word embedding (CLWE) methods, matrix factorization techniques, such as VecMap (Artetxe et al., 2016, 2017), have demonstrated strong performance for pairs of linguistically similar languages. However, their performance deteriorates for linguistically distant language pairs (Patra et al., 2019), such as English-Manipuri (En-Mn) (Naorem et al., 2024a), due to challenges including the non-isomorphic nature, orthographic differences, and significant linguistic divergence. To further enhance BDI performance, the contrastive learning cross-lingual embedding approach (Li et al., 2022) has emerged as a powerful approach that combines static word embeddings (VecMap) with contextual embeddings generated by language models (LM) and large language models (LLMs). However, the contrastive learning approach yields lower performance for linguistically distant and morphologically complex language pairs, such as English-Manipuri (En-Mn) (Naorem et al., 2023, 2024a), compared to similar language pairs, like English-Italian (En-It).

Recent research has explored the potential of large language models (LLMs) to induce BDI using a few-shot prompting approach (Li et al., 2023). This technique involves leveraging the inherent contextual understanding of LLMs to align embeddings with minimal supervision (Brown et al., 2020), often achieving better results than traditional VecMap and contrastive learning methods. However, this study has focused primarily on a limited set of resource-rich and structurally similar language pairs, often neglecting the evaluation of linguistically challenging and distant language pairs, such as English-Manipuri (En-Mn), English-Finnish (En-Fi), English-Turkish (En-Tr), English-Japanese (En-Ja) and English-Tamil (En-Ta). Moreover, evaluation of state-of-the-art LLM models, such as Llama, remains largely unexplored, particularly in the challenging settings mentioned above (Li et al., 2023). Most LLM models possess inherent multilingual properties (Tang et al., 2024) that can also be leveraged in an unsupervised manner for BDI. Apart from the unsupervised approach, two predominant paradigms have emerged in using LLMs for bilingual dictionary induction: few-shot prompting (Li et al., 2023) and supervised fine-tuning using contrastive

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 12042–12053
11–16 May 2026.

learning (Li et al., 2022). Few-shot prompting requires minimal task-specific examples to guide the model. Although few-shot prompting is less resource-intensive (Liu et al., 2023; Cheng et al., 2023), its viability for morphologically complex and distant language pairs still needs to be explored. On the other hand, the supervised fine-tuning approach (Li et al., 2022) involves extensive bilingual dictionary pairs and a much higher computational cost to fine-tune the model for the BDI task. Although both methods hold promise, their comparative effectiveness and practical implications remain underexplored, particularly in linguistically distant pairs and morphologically challenging environments. Motivated by the concerns mentioned above, this paper provides a comprehensive evaluation of LLMs, focusing on key questions such as: *Is few-shot prompting a feasible alternative to supervised fine-tuning methods for distant and morphologically challenging language pairs* ? In summary, the paper has the following contributions.

(i) The paper presents a first-ever comparative analysis of BDI performance using large language models across unsupervised, supervised fine-tuning, and few-shot settings, examining their applicability and limitations in linguistically challenging and distant language pairs.

(ii) This paper shows that the 5-shot prompting approach outperforms unsupervised and zero-shot settings and surpasses supervised settings in 82.86% of the evaluation cases. Experimental results also show that few-shot prompting in LLM models, such as Llama, remains ineffective for morphologically rich language pairs like En-Mn, En-Fi, En-Tr, and En-Ta in source-to-target BDI tasks.

2. Related Work

Previous works on bilingual dictionary induction (BDI) include various methods. The initial mapping-based cross-lingual word embedding model (CLWE) (Mikolov et al., 2013) introduced a regression-based framework to learn a linear mapping function using a bilingual seed dictionary in which word embeddings from different languages are transformed into a shared vector space, facilitating alignment of similar lexical words. Following this, a matrix factorization approach (Artetxe et al., 2016) and its variant (Artetxe et al., 2018a) proposed a closed-form solution commonly known as VecMap.[1] However, both VecMap and a centrality-aligned ridge regression-based orthogonal mapping (Naorem et al., 2024b) struggled to handle morphologically complex words in BDI tasks. An empirical investigation in Vulić et al. (2019) demonstrated that both supervised and unsupervised methods perform poorly in morphologically rich languages, such as Finnish and Turkish, in BDI tasks. Morpheme-based approaches that segment words into their root and suffix components were proposed, leading to slight improvements in BDI performance (Üstün et al., 2019; Chimalamarri et al., 2020). In addition, the work in Czarnowska et al. (2019) introduced 40 morphologically complete dictionaries and highlighted the severe degradation in BDI performance for less frequently inflected words. Later, the method proposed in Czarnowska et al. (2020) used a probabilistic model with morphological awareness that jointly models the translation of the lexeme and the inflectional morphology. With the rise of contrastive learning techniques, cross-lingual words based on contrastive learning (Li et al., 2022) was introduced. However, the contrastive learning approach (Li et al., 2022) showed limitations in handling morphologically rich languages such as Finnish and Turkish. Another important contrastive learning approach that brings the target word with the same root closer enhances the BDI performance for the language pairs where the target language is morphologically rich (Naorem et al., 2025). More recently, a method that leverages LLM proposed a few-shot prompting approach (Li et al., 2023), achieving state-of-the-art BDI scores across many language pairs. However, the few-shot prompting method evaluates only resource-rich and linguistically closer language pairs, neglecting morphologically complex and distant language pairs.

3. Methodology

Three main approaches have emerged that leverage LLMs in BDI: Unsupervised, Supervised fine-tuning through contrastive learning (Li et al., 2022), and few-shot prompting (Li et al., 2023). Supervised fine-tuning is based on large bilingual dictionary pairs and incurs significantly higher computational costs to tailor the model for contrastive learning in BDI tasks. The unsupervised approach does not require dictionary pairs for leveraging LM/LLM instead, it utilizes inherent multilingual properties in LM/LLM. While few-shot prompting utilizes the in-context learning abilities of LLMs (Liu et al., 2024), requiring only minimal bilingual dictionary pairs. Although all three methods show potential, their comparative analysis and advantages remain under-explored, particularly across distant language pairs and complex morphological environments. The LM/LLM models mentioned in Table 2 are evaluated in supervised, unsupervised, and few-shot settings, as discussed below.

[1] https://github.com/artetxem/vecmap

3.1. Supervised Fine-tuning

If $z = f(x)$ defines a target language dictionary word of a source language dictionary word x, a bilingual dictionary pair is defined as $D = \{(x, z)|x \in X, z \in Z, z = f(x)\}$ where X and Z are the words in the source and target languages, respectively. Given a pair $(x_i, z_i^+) \in D$, x_i and z_i^+ are tokenized using the mBERT tokenizer, giving the following subword sequences: $s_1 x_i .... s_n x_i$, $s_1 z_i^+ .... s_n z_i^+$, $n \geq 1$. The LM/LLM encoding function f_θ takes the sequence as input and gives the average representation of the token in the last transformer layer as the representation of x_i and z_i^+ respectively (Vulić et al., 2020).

For supervised fine-tuning using contrastive learning, positive samples present in D and negative samples not present in D are required. Like in (Li et al., 2022), for a given positive pair $(x_i, z_i^+) \in D$, a hard negative set $S_z^- = \{z_j \mid (x_i, z_i^+) \in D, z_j \neq z_i^+, z_j \notin NN(x_i)\}$ is generated where $NN(x_i)$ is the nearest neighbors of x_i from VecMap embedding of target language excluding z_i^+. Similarly, for the translation of the target-to the source, we generate the set of negative pairs $S_x^- = \{x_j \mid (x_i, z_i^+) \in D, x_j \neq x_i, x_j \notin NN(z_i^+)\}$ where $NN(z_i^+)$ is the nearest neighbor of z_i^+ from VecMap embedding of the source language excluding x_i.

For supervised fine-tuning, we used the state-of-the-art contrastive fine-tuning approach (Li et al., 2022) using a negative pair set S_z^- and S_x^- as described above.

$$loss = \frac{sim(x_i, z_i^+)}{\sum\limits_{z_j \in \{z_i^+\} \cup S_z^-} sim(x_i, z_j) + \sum\limits_{x_j \in S_x^-} sim(x_j, z_i^+)} \quad (1)$$

$$sim(x_i, z_j) = \exp^{\cos(f_\theta(x_i),\, f_\theta(z_j))/\tau} \quad (2)$$

The final contrastive learning objective function that fine-tunes the LM/LLM parameters θ is given in equations 3.

$$\min_\theta - \left[\mathbb{E}_{(x_i, z_i^+) \in D} \log(\text{loss})\right] \quad (3)$$

3.2. Unsupervised Setting

In an unsupervised setting, the off-the-shelf average representation of the token (words) in the last transformer layer, as mentioned above, without fine-tuning, is taken for BDI evaluation. Given a pair $(x_i, z_i^+) \in D$, x_i and z_i^+ are tokenized using the mBERT tokenizer, giving the following subword sequences: $s_1 x_i .... s_n x_i$, $s_1 z_i^+ .... s_n z_i^+$, $n \geq 1$. The LM/LLM encoding function f_θ takes the sequence as input and gives the average representation of the token in the last transformer layer as the representation of x_i and z_i^+ respectively (Vulić et al., 2020).

Table 1: Best Template for Few-shot prompting (Li et al., 2023). For 5-shot $i = 1$ to 5.

n-shot	LLM	Template
0-shot	$mT5_{small}$	The word 'x' in L_z is: <mask>.
	$mT5_{base}$	Translate the word 'x' into L_z: <mask>.
	$XGLM_{564m}$	The L_x word x in L_Z is:
	$mGPT_{1.3B}$	Translate the L_x word x into L_y:
	$Llama-3.2_{1B}$	The L_x word x in L_z is:
5-shot	$mT5_{small}$	[The word x_i in L_z is z_i.]$*5$ The word x in L_z is <mask>.
	$mT5_{base}$	[The word x_i in L_z is z_i.]$*5$ The word x in L_z is <mask>.
	$XGLM_{564M}$	[The word x_i in L_z is z_i.]$*5$ The word x in L_z is.
	$mGPT_{1.3B}$	[The L_x word x_i in L_z is z_i.]$*5$ The L_x word x in L_z is
	$Llama-3.2_{1B}$	[The L_x word 'x_i' in L_z is z_i.]$*5$ The L_x word x in L_z is

3.3. Few-shot Prompting

BDI in the prompting paradigm refers to the task of extracting word-level translation pairs from a large language model (LLM) by designing an appropriate prompt (Li et al., 2023). Let $\ell_s \in \mathcal{V}_s$ denote a source-language word from the source vocabulary $\mathcal{V}_s$, and let $\mathcal{V}_z$ represent the target-language vocabulary. Given a prompt function $\varphi(\cdot)$ and an LLM $\mathcal{L}$, the objective is to obtain a candidate translation

$$\tilde{\ell}_z = \mathcal{L}(\varphi(\ell_s)),$$

such that $\tilde{\ell}_z$ matches the correct target word $\ell_z \in \mathcal{V}_z$. The prompt $\varphi(\cdot)$ steers the LLM $\mathcal{L}$ to retrieve the correct translation across languages. The effectiveness of BDI based on prompting is essentially dependent on the quality of the prompt $\varphi(\cdot)$.

We adopted the standard prompt template commonly used by methods leveraging autoregressive LLMs for few-shot prompting (Li et al., 2023). This study utilized the mask-filling-style-inspired template with a span-corruption objective for the encoder-decoder model mT5. Here, <mask> tokens are used as placeholders for target words, as proposed in prompting based BDI method (Li et al., 2023). The study also used a GPT-style template (Li et al., 2023) that leverages a causal language modeling objective for the XGLM, mGPT, and Llama decoder-only models. The GPT-style prompt induces LLMs to produce the target word immediately after the repeated input sequence. We perform both zero-shot and few-shot prompting using the best templates as proposed in Li et al. (2023). The template details are given in Table 1.

Table 2: Model Details

LM/LLM	No of Parameters
mBERT	110 Millions
IndicBERTv2-MLM-only	278 Millions
mT5-small	300 Millions
mT5-base	580 Millions
ByT5-base	580 Millions
XGLM	564 Millions
mBART-large	610 Millions
mGPT	1.3 Billion
Llama-3.2	1 Billion

Table 3: Statistics of data, **LP**: Language Pairs

LP	Platform	sentences		words		unique words	
En-Mn	Sangai Express +Poknafam+PMI	En 129546	Mn 181553	En 3.5M	Mn 3.3M	En 15247	Mn 24449
En-It	European Parliament	En 1.90 M	It 1.90M	En 49.6M	It 47.4M	En 151,017	It 219,976
En-Fi	European Parliament	En 1.92 M	Fi 1.92M	En 47.4M	Fi 32.2M	En 151017	Fi 219976
En-Hi	CILT,IIT Bombay	En 1.6M	Hi 1.6M	En 23.8M	Hi 24.6M	En 238,765	Hi 392,634
En-Ja	opensubtitles.org kitsunekko.net	En 2.8 M	Ja 2.8M	En 23.6M	Ja 21.5M	En 154,276	Ja 138,487
En-Ta	AI for Bharat	En 442776	Ta 442776	En 10.3M	Ta 8.4M	En 79518	Ta 314452
En-Tr	MaCoCu-tr-en 2.0	En 1.6 M	Tr 1.6M	En 55.0M	Tr 51.5M	En 411397	Tr 884161

3.4. Model used in Evaluation

For our study, we consider two pre-trained Language Models (LM): mBERT (Devlin et al., 2019) and IndicBERTv2-MLM-only (Doddapaneni et al., 2023). We also take seven multilingual Large Language models (LLM): $mT5_{small}$ (Xue et al., 2021), $mT5_{base}$ (Xue et al., 2021), $ByT5_{base}$ (Xue et al., 2022), $XGLM_{564M}$ (Lin et al., 2022), $mBART_{large}$ (Tang et al., 2020), $mGPT_{1.3B}$ (Shliazhko et al., 2024), and $Llama-3.2_{1B}$ (et al., 2024). In addition to LLMs, we also take the contrastive learning approach (Li et al., 2022) for the evaluation of BDI as it gives better performance than VecMap. For contrastive fine-tuning of LM/LLM, we used pre-trained $mBERT_{base}$ (En-It, En-Fi, En-Hi, En-Tr, En-Ja, and En-Ta) (Devlin et al., 2019) and IndicBERTv2-MLM-only (En-Mn) (Doddapaneni et al., 2023). $mT5_{small}$, $mT5_{base}$, $ByT5_{base}$, $XGLM_{564M}$, $mBART_{large}$, $mGPT_{1.3B}$, and $Llama-3.2_{1B}$, are trained in the continued training approach (Gupta et al., 2023) to incorporate Manipuri data. $Llama-3.2_{1B}$ is also trained in the continued training process (Gupta et al., 2023) to incorporate Finnish, Turkish, Japanese, and Tamil languages. The details of the LLM models are shown in Table 2.

4. Dataset

For the continued training process (Gupta et al., 2023), this study considers five language pairs: English-Manipuri (En-Mn), English-Finnish (En-Fi), English-Turkish (En-Tr), English-Tamil (En-Ta), and English-Japanese (En-Ja). For En-Fi, the Europarl[2] parallel corpus (Koehn, 2005) extracted from the proceedings of the European Parliament is used. The parallel corpus provided in MaCoCu-tr-en 2.0 (Bañón et al., 2023) is used for En-Tr. The En-Ta parallel corpus from the Bharat Parallel Corpus Collection (BPCC), AI4BHARAT[3] is used. For En-Ja, a parallel sub-title corpus (Pryzant et al., 2018) extracted from the conversational dialogue is used for English-Japanese.

Manipuri, also known as Meiteilon, is a low-resource Tibeto-Burman language widely spoken in Manipur, a state in Northeast India. It stands out as a distant language from English due to its unique language family and intricate morphological characteristics (morphologically rich) and sentence structures (subject-object-verb) (Singh and Bandyopadhyay, 2010a,b; Choudhury et al., 2004). Due to the limited size of the parallel corpus for En–Mn, we use the comparable corpus employed in Naorem et al. (2024a). The comparable corpus is extracted from two prominent online news article platforms in Manipuri: Sangai Express[4] and Poknafam[5]. En-Mn comparable corpus is also extracted from news updates posted on PMIndia[6] (Haddow and Kirefu, 2020). We also consider structurally similar language pairs like En-It using the Europarl parallel corpus. For English-Hindi (En-Hi), data generated by the Center for Indian Languages Technology, IIT Bombay (Kunchukuttan et al., 2018) are used. This study considers the bilingual dictionary available at the Directorate of Language Planning and Implementation, Government of Manipur [7] for the En-Mn language pair and the MUSE[8] library for the En-It, En-Tr, En-Fi, En-Hi, En-Ja, and En-Ta language pairs. The details of the data set are given in Table 3.

5. Experimental Setup

5.1. Contrastive learning parameter

The hyper-parameter values are N_{iter}=5, N_{neg}=50, N_{iter} is the number of iterations in VecMap used

[2] https://www.statmt.org/europarl/
[3] https://ai4bharat.iitm.ac.in/bpcc/
[4] https://www.thesangaiexpress.com/index.html
[5] http://www.poknapham.in
[6] https://www.pmindia.gov.in/en/
[7] https://www.dlpi.mn.gov.in/en/
[8] https://github.com/facebookresearch/MUSE

Table 4: The results of evaluation (P@5) on BDI task over LM/LLM

	LM/LLM	En → Mn	Mn → En	En → It	It → En	En → Fi	Fi → En	En → Hi	Hi → En	En → Tr	Tr → En	En → Ja	Ja → En	En → Ta	Ta → En
Unsupervised	mBERT/IndicBert	00.44	00.44	19.71	18.43	01.31	01.57	01.71	01.32	13.74	13.74	04.00	03.22	06.54	06.25
	$mT5_{small}$	**00.89**	**00.46**	21.00	22.71	**03.97**	**05.68**	01.43	01.18	**14.40**	14.97	02.14	02.78	**06.97**	**08.40**
	$mT5_{base}$	00.14	00.28	10.28	10.28	00.57	00.57	00.86	00.88	12.43	12.57	03.00	03.37	06.14	06.28
	$ByT5_{base}$	00.14	00.00	18.00	27.00	01.43	03.00	00.43	00.44	12.14	**15.28**	01.71	01.76	05.57	05.57
	$XGLM_{564M}$	00.14	00.00	14.28	14.43	00.86	01.00	00.43	00.44	12.43	12.43	01.71	01.76	06.14	06.14
	$mBART_{large}$	00.57	00.43	17.71	18.00	01.57	01.71	**02.14**	**02.21**	13.14	12.86	05.43	03.08	06.43	06.86
	$mGPT_{1.3B}$	00.28	00.14	**25.71**	**28.71**	01.28	01.86	00.43	00.44	13.86	13.14	**06.28**	**07.62**	06.14	06.14
	$Llama-3.2_{1B}$	00.00	00.00	15.86	17.14	01.00	01.14	00.43	00.44	13.28	13.43	01.71	03.22	06.14	06.16
Supervised	mBERT/IndicBert	09.03	09.54	61.28	65.86	**18.46**	**23.11**	12.86	17.26	26.74	**32.43**	25.28	26.39	12.60	17.03
	$mT5_{small}$	07.86	09.43	41.14	47.28	12.57	16.86	05.14	07.82	21.28	24.00	30.14	29.91	15.43	16.71
	$mT5_{base}$	**36.14**	**35.43**	49.71	56.71	01.14	00.71	00.86	00.59	12.00	12.00	44.57	42.81	21.43	23.00
	$ByT5_{base}$	00.57	01.00	21.00	33.71	02.57	04.57	00.43	00.44	12.71	17.00	02.14	02.49	05.57	05.57
	$XGLM_{564M}$	00.71	00.71	40.14	45.43	14.71	18.57	13.14	12.83	24.86	24.43	04.57	06.01	13.28	22.71
	$mBART_{large}$	14.57	16.00	**63.57**	**66.86**	11.43	09.43	**41.28**	**42.77**	16.86	16.14	41.57	35.78	**28.86**	**30.70**
	$mGPT_{1.3B}$	00.14	00.00	40.00	43.28	02.71	03.14	00.28	00.29	18.43	18.28	32.86	29.62	06.14	06.14
	$Llama-3.2_{1B}$	23.28	28.57	44.86	48.14	10.86	11.71	26.86	30.09	**30.28**	30.86	**48.14**	**43.11**	06.28	07.14
0-shot	$mT5_{small}$	00.00	00.28	09.14	20.43	00.71	02.43	00.43	01.91	11.57	13.86	01.71	13.49	**06.14**	06.86
	$mT5_{base}$	00.00	**03.57**	25.28	28.43	03.28	07.43	00.43	14.90	12.41	**17.86**	01.57	**41.35**	05.28	14.28
	$XGLM_{564M}$	00.28	00.57	22.28	24.14	07.43	09.28	00.28	09.59	12.71	14.57	04.57	11.43	03.57	07.57
	$mGPT_{1.3B}$	00.00	02.43	36.57	39.28	**08.28**	**14.43**	07.71	19.47	**20.14**	13.43	**47.43**	28.15	04.43	**20.00**
	$Llama-3.2_{1B}$	**05.71**	01.57	**48.00**	**53.00**	00.14	03.71	**26.00**	**56.19**	03.86	15.43	00.28	09.38	00.14	01.57
5-shot	$mT5_{small}$	05.71	08.28	38.86	56.43	14.57	33.57	16.57	25.37	26.28	44.43	37.71	46.63	17.57	33.00
	$mT5_{base}$	**14.43**	32.43	59.00	73.28	**32.00**	55.14	36.14	56.78	42.86	61.28	52.86	65.39	**36.28**	36.17
	$XGLM_{564M}$	07.43	11.57	40.14	57.71	29.57	52.28	32.71	44.10	35.71	50.86	27.28	40.47	25.57	33.86
	$mGPT_{1.3B}$	08.00	20.71	64.00	**83.71**	31.14	**60.86**	21.43	56.34	**54.57**	**76.43**	**62.71**	**77.71**	13.28	**45.57**
	$Llama-3.2_{1B}$	09.71	**58.71**	**67.14**	80.28	02.43	19.28	**46.71**	**65.63**	21.43	36.28	01.43	33.87	00.57	18.14

in contrastive learning. N_{neg} is the number of negative samples for a positive pair. The AdamW (Loshchilov and Hutter, 2019) optimizer with a learning rate of 2e-5 is used to fine-tun mBERT/IndicBert, $XGLM_{564M}$, $mBART_{large}$, $mGPT_{1.3B}$ and $Llama-3.2_{1B}$. For fine-tuning $mT5_{small}$, $mT5_{base}$, and $ByT5_{base}$, a learning rate of 6e-5 is used . LM/LLM WEs are fine-tuned for five epochs with $\tau = 0.1$.

5.2. Continued Training

mT5: Continued training is performed using the same span-corruption objective as in the original mT5 pretraining, allowing the decoder to reconstruct the masked spans through token prediction. Optimization is carried out using AdamW with a learning rate of $5{\times}10^{-5}$ and a weight decay of 0.01. A linear learning-rate scheduler with 1% warmup is applied. The model is trained for 1 epoch. The original sentencepiece tokenizer is used with the vocabulary extension to incorporate the proper token for the BDI task. To reduce catastrophic forgetting and maintain multilingual transfer, the corresponding English corpus (parallel/comparable) is also included in the continue training process.

ByT5: We used the same span-corruption denoising objective. Optimization is performed using AdamW with a learning rate of $1{\times}10^{-4}$ and weight decay at 0.01. A linear learning-rate scheduler with 1% warmup is applied. Training is done with an epoch of 1. To reduce catastrophic forgetting and maintain multilingual transfer, the corresponding English corpus (parallel/comparable) is also included in the continue training process. Since ByT5 works on BPE, an extension of the vocabulary is not required.

XGLM: For XGLM, causal language modeling objective is used, where the model predicts the next token autoregressively from left to right over the input sequence. Optimization is carried out using AdamW with a learning rate of $5{\times}10^{-5}$ and a weight decay of 0.01. A linear learning-rate scheduler with 1% warmup is applied and the model is trained for 1 epoch. The original XGLM tokenizer, based on subword segmentation, is retained with vocabulary extension. To reduce catastrophic forgetting, the corresponding English corpus (parallel/comparable) is also included during continued training.

mBART: Continued training is performed using the same denoising autoencoding objective as in its original pretraining, where portions of the input sequence are corrupted through text infilling, and the decoder reconstructs the original sequence from the corrupted input. Optimization is carried out using AdamW with a learning rate of $5{\times}10^{5}$ and a weight decay of 0.01. A linear learning-rate scheduler with 1% warmup is applied and the model is trained for 1 epoch. The original SentencePiece tokenizer is retained with a vocabulary extension. To reduce catastrophic forgetting, the corresponding English corpus (parallel/comparable) is also included during continued training process.

mGPT, Llama: For mGPT, continued training is performed using a causal language modeling objective. Optimization is carried out using AdamW with a learning rate of $5{\times}10^{5}$ and a weight decay of 0.01. A linear learning-rate scheduler with 1% warmup is applied and the model is trained for 1 epoch. The BPE tokenizer is used without vocabulary extension. To reduce catastrophic forgetting and maintain multilingual transfer, the corresponding English corpus (parallel/comparable) is also included during continued training process.

5.3. Few-shot prompting

We consider a zero-shot to 10-shot prompting. Like in (Li et al., 2023), we set the beam size to

5 for all LLMs. For encoder-decoder models, the maximum sequence length is fixed at 5. In contrast, decoder-only models are set to 5 plus the input sequence length, since they first replicate the input before generating new content. For encoder-decoder LLMs, the evaluation batch size is set to 100 for smaller and 8 for larger models.

5.4. BDI Evaluation

The LM/LLM models are evaluated in Bilingual Dictionary Induction (BDI) at P@5 (Precision at 5). A training dictionary of 3500 and 700 testing pairs is used for all language pairs. Since English-Manipuri have limited (4200) dictionary pairs, we take 3500 as training pairs and 700 as testing pairs for all the language pairs in our experiment.

6. Results and Discussion

Unsupervised: For En-Mn, Mn-En, En-Fi, Fi-En, En-Ta, and Ta-En $mT5_{small}$ gives the highest score of 00.89, 00.46, 03.97, 05.68, 06.97, and 08.40, respectively. $mGPT_{1.3B}$ gives a better score of 25.71, 28.71, 06.28, and 07.62 for En-It, It-En, En-Ja and Ja-En, respectively. In En-Hi and Hi-En, the highest performance is shown by $mBART_{large}$ with 02.14 and 02.21, respectively. In the case of En-Tr, $mT5_{small}$ produces a higher performance score of 14.40. However, $ByT5_{base}$ performs better in Tr-En with a score of 15.28. Although no particular LLM model outperforms in all language pairs, all the LLM models perform less in linguistic distant and morphologically complex pairs than linguistically similar pairs like En-It in unsupervised settings. Details of BDI results are shown in Table 4.

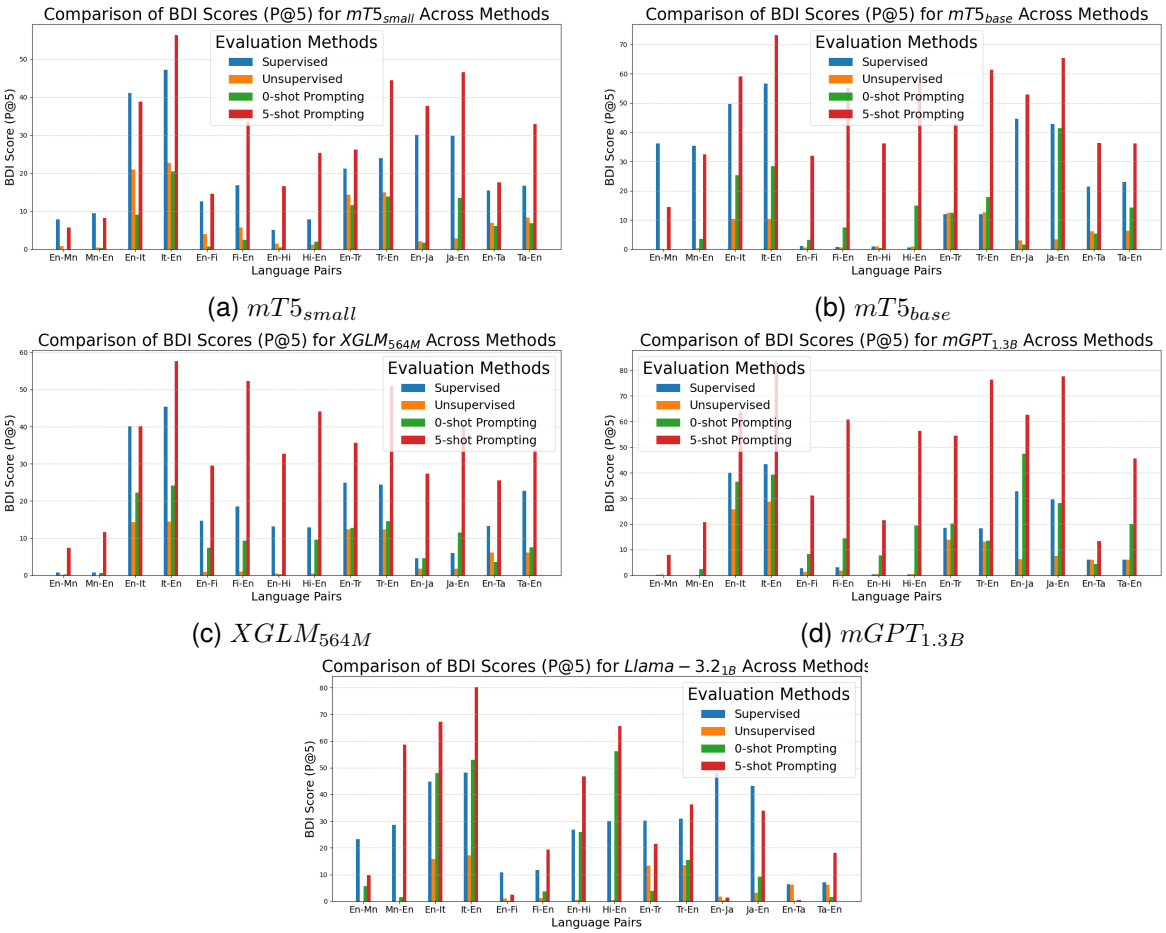
Supervised: For En-Mn and Mn-En, $mT5_{base}$ gives a higher score of 36.14 and 35.43, respectively. $mBART_{large}$ gives the highest score of 63.57, 66.86, 41.28, 42.77, 28.86 and 30.70 in En-It, It-En, En-Hi, Hi-En, En-Ta and Ta-En, respectively. For En-Fi and Fi-En LM models, mBERT performs better than all the LLM models. In the case of En-Tr, the highest score is given by $Llama-3.2_{1B}$ with a score of 30.28, but mBERT gives the highest score of 32.43 in Tr-En. $Llama-3.2_{1B}$ gives the highest performance in En-Ja and Ja-En with a score of 48.14 and 43.11, respectively. Like in an unsupervised setting, there is no particular LLM model that outperforms in all language pairs, but all the LLM models give much lesser performance in linguistically distant and morphologically complex pairs as compared to linguistically similar pairs like En-It in an unsupervised setting. Details of BDI results are shown in Table 4.

Zero-shot: For En-Mn, $Llama-3.2_{1B}$ gives the highest performance score of 05.71, but $mT5_{base}$ gives the best score of 03.57 in Mn-En. $Llama-3.2_{1B}$ gives the highest En-It, It-En, En-Hi, Hi-En score. For En-Fi and Fi-En, $mGPT_{1.3B}$ outperformed all the remaining LLM models. In the case of En-Tr and En-Ja, $mGPT_{1.3B}$ gives the highest score. $mT5_{base}$ outperform the remaining LLM models in Tr-En and Ja-En. For En-Ta, $mT5_{small}$ gives the highest score. On the other hand, $mGPT_{1.3B}$ gives the best score in Ta-En. Similar to unsupervised and supervised, all the LLM models perform much less in linguistic distant and morphologically complex pairs than in linguistically similar pairs like En-It in unsupervised settings. Details of BDI results are shown in Table 4.

5-shot: For En-Mn $mT5_{small}$ gives the highest performance score of 14.43, but $Llama-3.2_{1B}$ gives the best score of 58.71 in Mn-En. $Llama-3.2_{1B}$ gives the highest En-It, En-Hi, and Hi-En scores. $mGPT_{1.3B}$ outperformed all the remaining LLM models in It-En, Fi-En, En-Tr, Tr-En, En-Ja, Ja-En and Ta-En. In the case of En-Fi and En-Ta, $mT5_{base}$ gives the highest score. $Llama-3.2_{1B}$ and $mGPT_{1.3B}$ outperform all LLM models in 78.57% (11 cases out of 14 cases) of the time. Similarly, the 5-shot approach performs less in linguistic distant and morphologically complex language pairs like En-Mn, Mn-En, En-Ta, and Ta-En. Details of BDI results are shown in Table 4.

Supervise Fine-tuning vs Unsupervised vs 0-shot vs 5-shot: Fig 1, shows that 5-shot prompting outperformed both unsupervised and zero-shot prompting 100% (in all cases) of the time. 5-shot prompting outperforms supervised setting in 82.86% (58 cases out of 70 cases) of the time. The supervised setting outperforms the 5-shot prompt in $mT5_{small}$, $mT5_{base}$, and $Llama-3.2_{1B}$ for En-Mn. For Mn-En and Ta-En, the 5-shot prompting approach outperforms supervised settings in $Llama-3.2_{1B}$. The performance differences in Mn-En/Ta-En and En-Mn/En-Ta tasks using Llama models can be attributed to the model's bias toward high-resource languages like English. While effective for En-Mn and En-Ta, supervised fine-tuning approaches may over-fit the training data, reducing their ability to generalize to unseen examples. In contrast, few-shot prompting leverages the model's inherent multilingual and contextual understanding, avoiding over-fitting and preserving the pretrained multilingual alignment. This makes the few-shot prompting approach results higher than supervised in the case of target-to-source translation: Mn-En and Ta-En. A similar trend is also observed in $XGLM_{564M}$

Morphologically Complex Settings: On the other hand, the supervised fine-tuning approach outperforms 5-shot prompting for En-Mn, En-Fi, En-Tr, and En-Ta in $Llama-3.2_{1B}$. The na-



(a) $mT5_{small}$ (b) $mT5_{base}$

(c) $XGLM_{564M}$ (d) $mGPT_{1.3B}$

(e) $Llama - 3.2_{1B}$

Fig. 1: Supervised Fine-tuning vs Unsupervised vs 0-shot vs 5-shot

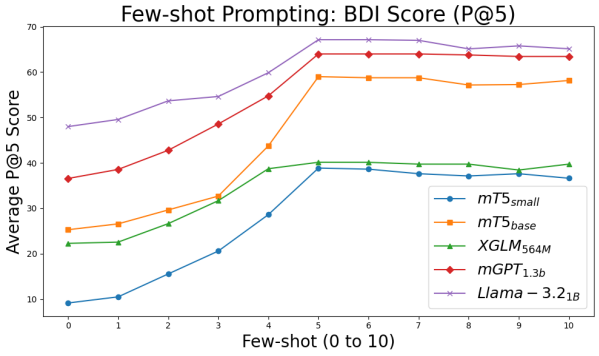


Fig. 2: BDI scores averaged over 14 BDI directions with respect to n-shot (0 to 10)

ture of the task in the translation of En-Mn, En-Fi, En-Tr, and En-Ta is inherently more complex, as Manipuri, Finnish, Turkish, and Tamil are morphologically rich languages, making fine-tuned models more advantageous. In Mn-En, Fi-En, Tr-En, and Ta-En, the BDI task is often simpler, finding a direct English equivalent for a Manipuri/Finnish/Turkish/Tamil word), making it more amenable to a few-shot prompting. This observation suggests that few-shot prompting in the Llama model is still ineffective for morphologically rich language pairs in the BDI task. Morphologically rich languages generate multiple inflected words. For example, a Manipuri root word can produce multiple inflected forms with different meanings. LLMs trained in subword-level tokenization (e.g., BPE) generally split a word inconsistently, leading to incorrect representation (not capturing morphological information). This makes it difficult for LLM models to learn a good alignment between morphologically complex words and their English translations. In addition, LLMs are overwhelmingly trained in English and other resources-rich data. Low-resource and morphologically rich languages are underrepresented, and LLM models struggle to learn morphological information effectively.

Best n-shot: For all the language pairs in our study, BDI is evaluated for 0 to 10 shots. The average BDI score on 14 BDI directions for $mT5_{small}$, $mT5_{base}$, $XGLM_{564M}$, $mGPT_{1.3B}$, and $Llama - 3.2_{1B}$ are reported in Fig 2. Fig 2 shows that the BDI performance averaged over 14 BDI directions

Table 5: Error analysis of 5-shot prompting on language pairs where the target language is morphologically rich.

	English word	$Llama-3.2_{1B}$	$XGLM_{564M}$	Ground truth
En-Mn	**name**	মিং name	মিং name	**ммиң**
	population	মী man	মীগী for people	**мишиң**
	nobody	কনা who	কনাগী whose	**кнамтта**
	june	জুন june	জুন June	**жун**
	nine	মাপন nine	৯ nine	**мапн**
En-Fi	**finally**	lopulta eventually	lopettaa end	**loppu**
	increase	lisää more	lisäämään to increase	**lisäys**
	account	tilinpäätös financial statements	tili account	**tiliä**
	growth	kasvu growth	kasvu growth	**kasvu**
	control	ohjaus control	hallita to control	**ohjaus**
En-Tr	**history**	tarihin of history	tarihte in history	**tarih**
	future	gelecektir will come	gelecek future	**gelecek**
	media	medyada in the media	medyaya to the media	**medya**
	academic	akademik academic	akademik academic	**akademik**
	away	uzakta away	uzak away	**uzakta**
En-Ta	**gas**	வா come on	வா come on	**வாயு**
	hours	மணி bell	மணி bell	**மணிநேரம்**
	existing	இரு be	உள்ளது is	**இருக்கும்**
	team	அணி team	அணி team	**அணி**
	eye+	கண் eye	கண்கள் eyes	**கண்**

starts to saturate at 5-shot prompt.

Is few-shot prompting a feasible alternative to supervised fine-tuning methods?: In $mT5_{small}$, 5-shot prompting outperformed supervised fine-tuning except for En-Mn, Mn-En, and En-It. 5-shot prompting outperformed supervised fine-tuning in most BDI directions except En-Mn and Mn-En for the $mT5_{base}$ LLM model. In $XGLM_{564M}$, 5-shot prompting outperformed supervised fine-tuning in all BDI directions. Similarly, 5-shot prompting outperformed supervised fine-tuning in all BDI directions for $mGPT_{1.3B}$. Evaluation of $Llama-3.2_{1B}$ shows a slightly different result with supervised fine-tuning outperforming 5-shot prompting in En-Mn, En-Fi, En-Tr, En-Ja, Ja-En and En-Ta. The under-performance of the 5-shot prompting in $Llama-3.2_{1B}$ in the above-mentioned BDI directions might be due to many factors such as the fewer resources in the Mn, Fi, Tr, Ja and Ta languages in the continued training process of $Llama-3.2_{1B}$. The biased nature of the Llama model to resource-rich Languages like English is also a contributing factor. Another factor may be complex morphological understanding in source-to-target BDI tasks such as En-Mn, En-Fi, En-Tr, and En-Ta. On the other

hand, target-to-source BDI directions like Mn-En, Fi-En, Tr-En, and Ta-Mn are much simpler tasks and thus give better results than supervised fine-tuning by mitigating over-fitting on the training dictionary. While supervised fine-tuning methods are effective for morphologically complex languages, few-shot prompting strikes a balance by leveraging inherent multi-lingual contextual understanding and avoiding over-fitting issues. Few-shot prompting is preferable for practical applications like BDI, where annotated training data is expensive.

7. Error Analysis

The results of the prompting approach (Li et al., 2023) on the BDI task are chosen for error analysis. The English word, the predicted translations, the English meaning of the predicted translation, and the ground truth translation are given in Table 5. The English meanings of the predicted translations are obtained using FinnWordNet, TurkishWordNet, TamilWordNet for En-Fi, En-Tr, and En-Ta, respectively. For En-Mn, the author, being a native Manipuri speaker and fluent in English, translated the English word into Manipuri by himself. We report the results generated by $Llama-3.2_{1B}$ for error analysis. The Few-shot prompting fails to provide the correct translation of the English words *name*, *population*, and *nobody*. The predicted word মিং (informal translation of the name) is different from the ground truth মমিং, which is inflected with the prefix ম. In the case of the predicted word মী and কনা, these predicted words represent root words with the English translation *man* and *who*, respectively. These translations differ significantly from the ground truth মীশিং (population) and কনামত্তা (nobody). মীশিং is inflected with the suffix শিং and কনামত্তা is inflected with মত্তা. In the case of En-Fi translation, the predicted word *lopulta* (eventually), *lisää* (more), and *tilinpäätös* (financial statements) have slightly different meanings from their respective ground truth. *lopulta*, *lisää*, and *tilinpäätös* are inflected with the suffix *lta*, *ä*, and *npäätös* respectively which changes the meaning slightly. For En-Tr translation, the predicted word *tarihin* (of history), *gelecektir* (will come), and *medyada* (in the media) have slightly different meanings from their respective ground truth. *tarih*, *gelecek*, and *medya* are inflected with the suffix *in*, *tir*, and *da* respectively which changes the meaning slightly. Similarly, in En-Ta translation, the predicted word வா (come on), மணி (bell), and இரு (be) have different meanings from their respective ground truth. The ground truths வாயு, மணிநேரம் and இருக்கும் are inflected with the suffix யு, நேரம், and க்கும், which changes the meaning.

8. Conclusion

This paper evaluates the performance of various large language models (LLMs) across diverse distant language pairs and task settings (unsupervised, supervised fine-tuning, zero-shot and 5-shot prompting) in intrinsic BDI tasks. The results show that all LLMs consistently perform better on linguistically similar pairs (e.g., En-It) than distant and morphologically complex pairs (e.g., En-Mn, En-Fi, En-Ta). Morphologically rich languages like Manipuri, Finnish, Turkish, and Tamil pose significant challenges, particularly in unsupervised and zero-shot settings. The 5-shot prompting approach outperforms unsupervised and zero-shot settings in all cases and even surpasses supervised settings in 82.86% of cases. Few-shot prompting demonstrates robustness against over-fitting, leveraging LLMs' In-context learning multilingual capabilities, making it particularly effective in target-to-source translation even for morphologically complex language pairs. At the same time, few-shot prompting in LLM models like Llama is still ineffective for morphologically rich language pairs like En-Mn and En-Ta in source-to-target BDI tasks. While supervised fine-tuning methods are effective, especially for morphologically complex languages, few-shot prompting strikes a balance by leveraging inherent multi-lingual contextual understanding and avoiding over-fitting issues. Few-shot prompting is preferable for practical applications like BDI, where annotated training data is scarce or expensive. From the analysis, the performance of BDI saturates at 5-shot prompting, indicating diminishing returns beyond this point.

9. Limitations Section

A key limitation of this paper is that English is consistently included as one language in every evaluated pair. BDI evaluation between two low-resource and linguistically unrelated languages is generally more challenging than involving English. Consequently, the in-context learning approach adopted here may not generalize as effectively to non-English language pairs, where alternative methods such as supervised contrastive learning could prove more suitable. A detailed comparison of computation time between supervised fine-tuning and in-context learning would provide a clearer picture of the advantages of in-context learning. A comparison between a few-shot prompting and finetuning with the same few-shot example may solidify the strength of LLM in BDI.

10. Acknowledgements

We sincerely thank anonymous reviewers for the time and effort they dedicated to the review process. We also thank the Indian Institute of Technology, Guwahati, for the fellowship supporting my research.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2016. Learning principled bilingual mappings of word embeddings while preserving monolingual invariance. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2289–2294, Austin, Texas. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2017. Learning bilingual word embeddings with (almost) no bilingual data. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Vancouver, Canada. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018a. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, Melbourne, Australia. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho. 2018b. Unsupervised neural machine translation. In *International Conference on Learning Representations*.

Marta Bañón, Malina Chichirau, Miquel Esplà-Gomis, Mikel L. Forcada, Aarón Galiano-Jiménez, Cristian García-Romero, Taja Kuzman, Nikola Ljubešić, Rik van Noord, Leopoldo Pla Sempere, Gema Ramírez-Sánchez, Peter Rupnik, Vít Suchomel, Antonio Toral, and Jaume Zaragoza-Bernabeu. 2023. Turkish-english parallel corpus MaCoCu-tr-en 2.0. Slovenian language resource repository CLARIN.SI.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D 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 Ziegler, Jeffrey Wu, Clemens Winter, Chris 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. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Zhoujun Cheng, Jungo Kasai, and Tao Yu. 2023. Batch prompting: Efficient inference with large language model APIs. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 792–810, Singapore. Association for Computational Linguistics.

Santwana Chimalamarri, Dinkar Sitaram, and Ashritha Jain. 2020. Morphological segmentation to improve crosslingual word embeddings for low resource languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, 19(5).

Sirajul Islam Choudhury, Leihaorambam Sarbajit Singh, Samir Borgohain, and Pradip Kumar Das. 2004. Morphological analyzer for manipuri: design and implementation. In *Asian Applied Computing Conference*, pages 123–129. Springer.

Paula Czarnowska, Sebastian Ruder, Ryan Cotterell, and Ann Copestake. 2020. Morphologically aware word-level translation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2847–2860, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Paula Czarnowska, Sebastian Ruder, Edouard Grave, Ryan Cotterell, and Ann Copestake. 2019. Don't forget the long tail! a comprehensive analysis of morphological generalization in bilingual lexicon induction. 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 974–983, Hong Kong, China. Association for Computational Linguistics.

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.

Sumanth Doddapaneni, Rahul Aralikatte, Gowtham Ramesh, Shreya Goyal, Mitesh M. Khapra, Anoop Kunchukuttan, and Pratyush

Kumar. 2023. Towards leaving no Indic language behind: Building monolingual corpora, benchmark and models for Indic languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12402–12426, Toronto, Canada. Association for Computational Linguistics.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models.

Kshitij Gupta, Benjamin Th'erien, Adam Ibrahim, Mats L. Richter, Quentin G. Anthony, Eugene Belilovsky, Irina Rish, and Timothée Lesort. 2023. Continual pre-training of large language models: How to (re)warm your model? *ArXiv*, abs/2308.04014.

Barry Haddow and Faheem Kirefu. 2020. PMIndia – A Collection of Parallel Corpora of Languages of India. *arXiv e-prints*, page arXiv:2001.09907.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: papers*, pages 79–86.

Anoop Kunchukuttan, Pratik Mehta, and Pushpak Bhattacharyya. 2018. The IIT Bombay English-Hindi parallel corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Yaoyiran Li, Anna Korhonen, and Ivan Vulić. 2023. On bilingual lexicon induction with large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9577–9599, Singapore. Association for Computational Linguistics.

Yaoyiran Li, Fangyu Liu, Nigel Collier, Anna Korhonen, and Ivan Vulić. 2022. Improving word translation via two-stage contrastive learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4353–4374, Dublin, Ireland. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, Veselin Stoyanov, and Xian Li. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2023. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Comput. Surv.*, 55(9).

Yuhan Liu, Xiuying Chen, Gao Xing, Ji Zhang, and Rui Yan. 2024. IAD: In-context learning ability decoupler of large language models in meta-training. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8535–8545, Torino, Italia. ELRA and ICCL.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Tomás Mikolov, Quoc V. Le, and Ilya Sutskever. 2013. Exploiting similarities among languages for machine translation. *CoRR*, abs/1309.4168.

Deepen Naorem, Okram Jimmy Singh, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2023. English-manipuri cross-lingual embedding: A preliminary study. In *2023 International Conference on Asian Language Processing (IALP)*, pages 74–79.

Deepen Naorem, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024a. Embarking on a preliminary exploration: Cross-lingual embedding in english-manipuri. *International Journal of Asian Language Processing*, 34(02):2450007.

Deepen Naorem, Sanasam Ranbir Singh, and Priyankoo Sarmah. 2024b. Improving linear orthogonal mapping based cross-lingual representation using ridge regression and graph centrality. *Computer Speech & Language*, 87:101640.

Deepen Naorem, Sanasam Ranbir Singh, Telem Joyson Singh, and Priyankoo Sarmah. 2025. Mace: Morphology aware cross-lingual embedding using contrastive learning. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3124–3136.

Barun Patra, Joel Ruben Antony Moniz, Sarthak Garg, Matthew R. Gormley, and Graham Neubig. 2019. Bilingual lexicon induction with semi-supervision in non-isometric embedding spaces. In *Proceedings of the 57th Annual Meeting of*

the Association for Computational Linguistics, pages 184–193, Florence, Italy. Association for Computational Linguistics.

Reid Pryzant, Youngjoo Chung, Dan Jurafsky, and Denny Britz. 2018. JESC: Japanese-English subtitle corpus. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A survey of cross-lingual word embedding models. *J. Artif. Int. Res.*, 65(1):569–630.

Oleh Shliazhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. mgpt: Few-shot learners go multilingual. *Transactions of the Association for Computational Linguistics*, 12:58–79.

Thoudam Doren Singh and Savaji Bandyopadhyay. 2010a. Statistical machine translation of English-Manipuri using morpho-syntactic and semantic information. In *Proceedings of the 9th Conference of the Association for Machine Translation in the Americas: Student Research Workshop*, Denver, Colorado, USA. Association for Machine Translation in the Americas.

Thoudam Doren Singh and Sivaji Bandyopadhyay. 2010b. Manipuri-English bidirectional statistical machine translation systems using morphology and dependency relations. In *Proceedings of the 4th Workshop on Syntax and Structure in Statistical Translation*, pages 83–91, Beijing, China. Coling 2010 Organizing Committee.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715, Bangkok, Thailand. Association for Computational Linguistics.

Y. Tang, C. Tran, Xian Li, Peng-Jen Chen, Naman Goyal, Vishrav Chaudhary, Jiatao Gu, and Angela Fan. 2020. Multilingual translation with extensible multilingual pretraining and finetuning. *ArXiv*, abs/2008.00401.

Ahmet Üstün, Gosse Bouma, and Gertjan van Noord. 2019. Cross-lingual word embeddings for morphologically rich languages. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2019)*, pages 1222–1228, Varna, Bulgaria. INCOMA Ltd.

Ivan Vulić, Goran Glavaš, Roi Reichart, and Anna Korhonen. 2019. Do we really need fully unsupervised cross-lingual embeddings? 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 4407–4418, Hong Kong, China. Association for Computational Linguistics.

Ivan Vulić and Marie-Francine Moens. 2015. Monolingual and cross-lingual information retrieval models based on (bilingual) word embeddings. In *Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval*, pages 363–372.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

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.