Dictionary Insertion Prompting for Multilingual Reasoning on Multilingual Large Language Models

Hongyuan Lu[♠*], **Z.L.** [♠*], **Wai Lam**[♡]
[♠]FaceMind Research Asia
[♡]The Chinese University of Hong Kong
hongyuanlu@outlook.com

Abstract

There are two shortages in the current Large Language Models (LLMs) era. The first is short of multilingual models, where most LLMs are English-centric and performance is limited on multilingual reasoning. The second is the place of external knowledge to be used, where most retrieved knowledge is prepended to the user queries (maybe sub-optimal). This paper presents a novel and simple yet effective method called **D**ictionary **I**nsertion **P**rompting (**DIP**). When providing a non-English prompt, DIP looks up a word dictionary and inserts words' English counterparts into the middle of the prompt for LLMs. It then enables better translation into English and better English model thinking steps which leads to obviously better results. We experiment with 10 to 200 languages from FLORES-200.[1] Since there are no adequate datasets, we use the NLLB translator to create synthetic multilingual benchmarks from the existing 4 English reasoning benchmarks such as GSM8K and AQuA. The synthetic benchmarks are translated back into English for quality assurance with manual annotation. Interestingly, the place for injecting the dictionary plays an important factor in the performance gains, and we found that interleaving the dictionary with the original words gives a better performance compared to prepending/appending the dictionary, under the same dictionary constructed.

1 Introduction

In the quick development with large language models (LLMs), there have been quite many popular research areas such as chain-of-thought reasoning (Wang et al., 2023; Wei et al., 2024; Yang et al., 2024), machine translation (Lu et al., 2023; Zhu et al., 2024a,b), code generation (Li et al., 2023; Zhang et al., 2023; Hou et al., 2025), dialogue generation (Li et al., 2022; Lu et al., 2022; Lu and Lam, 2023; Yang et al., 2025; Yang et al., 2026), and even spatial understanding (Hu et al., 2024). Among these, an important research area is multilingual large language models (MLLMs), which consider not only the tasks of machine translation but also reasoning tasks represented in different languages (Huang et al., 2023). This scales the horizon of English-centric LLMs such as popular ChatGPT and enables them to be used by people who mainly speak low-resourced languages. Yet, current methods are usually training-based (Lu et al., 2024; Lim et al., 2024), which usually requires many GPU/TPU computational resources to update model weights from LLMs.

In contrast, this paper investigates how to incorporate a dictionary as an auxiliary knowledge into prompting. In comparison, this is lightweight and flexible, where the dictionary can be customized and replaced in a plug-and-play manner without model training. While the dictionary-based method has been studied on the task of traditional machine translation (Arthur et al., 2016), how to incorporate them into reasoning tasks on MLLMs has been under-studied. This is yet important and needs to be empirically justified whether, and how dictionary-based methods can be better used to improve multilingual reasoning tasks, which obviously enhances LLMs' usefulness in our daily life.

To this end, we propose a novel method called **D**ictionary **I**nsertion **P**rompting (**DIP**). We present the overall algorithm as in Figure 1. By providing a customized dictionary that maps words represented in low-resourced languages into English, DIP inserts the English representation into the original input. This then helps LLMs in pivoting the original input into a complete English representation. This improves the succeeding chain of thought reasoning, which results in a better final output.

By looking deeply into the experimental analysis,

[*] Equal Contribution.

[1]The number of languages varies on the datasets, and we experiment with 200 languages on GSM8K as in Appendix

*Proceedings of the 15th Joint Conference on Lexical and Computational Semantics (*SEM 2026)*, pages 529–544
July 3, 2026 ©2026 Association for Computational Linguistics

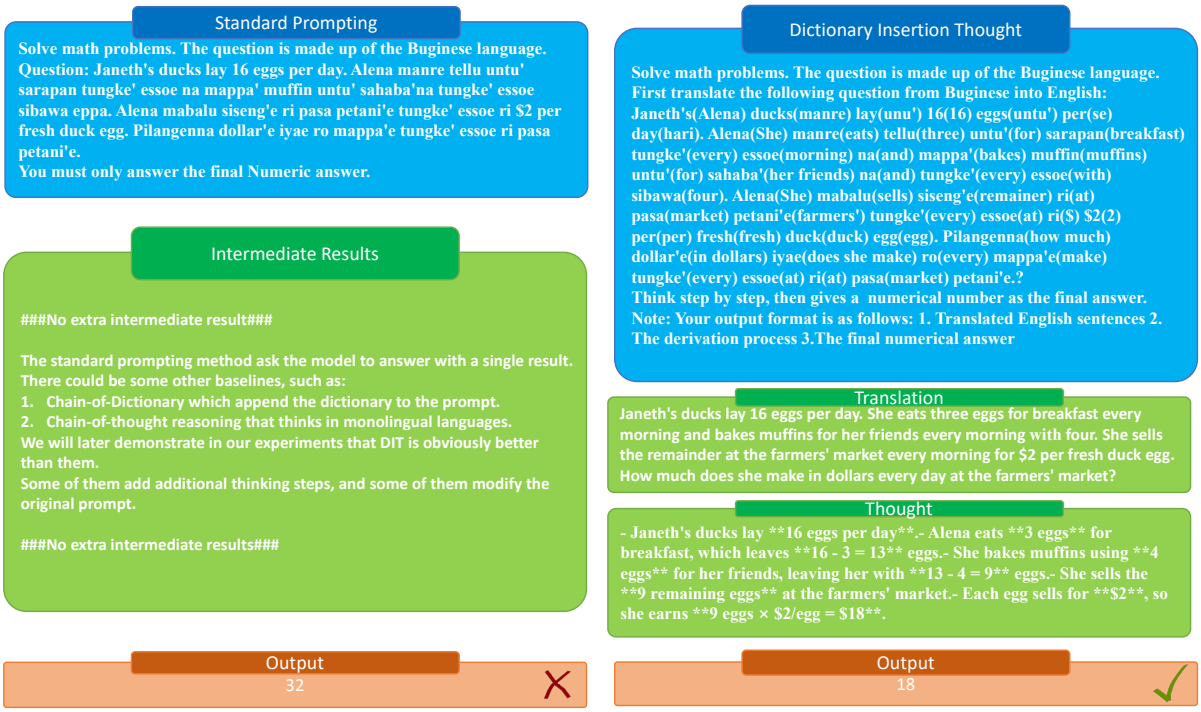

Figure 1: An illustrated comparison of the GSM8K dataset made up in Buginese. Compared to the standard prompting baseline, DIP inserts additional dictionary knowledge written in knowledge in an interleaving manner. This leads to a better intermediate English translation and succeeding English thought. Finally, DIP produces promising results, surpassing several strong baselines.

we found that the place of insertion is an important factor in improving the performance of DIP. While the prior dictionary-based method usually presents the dictionary in front of the prompt, we found it sub-optimal, and placing the dictionary in an interleaving manner is better. We postulate that such a design puts words with their English counterparts closer, and makes it easy to be understood by LLMs. Experiments also show that better translation quality and thought quality are the key factors to the usefulness of DIP.

We benchmark DIP on about 200 languages from FLORES-200 (NLLB-Team, 2022). Since there are no adequate datasets for covering those languages, we use a high-quality SOTA translator NLLB 3.3B[2] to translate existing arithmetic reasoning benchmarks such as GSM8K (Cobbe et al., 2021) and commonsense reasoning benchmarks such as Date (Srivastava et al., 2022).

In general, we make three key contributions.

- We propose a simple, novel, yet computationally lightweight approach called DIP for better reasoning on multilingual tasks on LLMs.
- Extensive strong results across several benchmarks on ChatGPT and Llama LLMs verify the effectiveness of DIP.
- We investigated further why DIP is useful and found that better intermediate translation and thinking steps are the key.

2 Dictionary Insertion Prompting

Large language models show their promising translation performance when sufficiently pre-trained (Wang et al., 2023; Lu et al., 2023; Tang et al., 2024; Lu et al., 2024a,b, 2025; Lu et al., 2026). Yet, such translation ability usually diminishes in low-resourced languages and it is still an understudied topic for reasoning in those low-resourced languages such as the ones from FLORES-200 (NLLB-Team, 2022).

This paper proposes a novel, yet simple and effective framework called **DIP** (**D**ictionary **I**nsertion **P**rompting) to address these difficulties by integrating the dictionary knowledge into the reasoning process. DIP first looks up a customized dictionary from the original non-English prompt to place its English counterpart accordingly. This is followed by pivoting into English, which succeedingly in-

[2]https://huggingface.co/spaces/Narrativaai/NLLB-Translator

vokes a better English thinking process. The final result is then obviously improved by DIP.

Therefore, DIP is as illustrated in Figure 1:

> (1) Solve the question. The question is made up of the <language> language.
> (2) First translate the following question from <language> into English:
> (3) <question>
> (4) Note: Your output format is as follows:
> 1. Translated English sentences
> 2. The derivation process
> 3.The final numerical answer

, where <language> denotes the language of the non-English question prompts[3] that are written in, and <question> denotes the actual question that is written in those non-English languages.

The reason for choosing English as the pivoting language for DIP on LLMs is due to the fact that English has been primarily used as the pivoting language in traditional machine translation (Utiyama and Isahara, 2007; Wu and Wang, 2007). While other languages can be possibly useful, English is the most high-resourced language on English-centric LLMs, it is intuitively the most helpful, so other attempts are left to future works.

Dictionary Construction To construct the bilingual dictionary mapping between English and the original prompt, we prompt ChatGPT:

> (1) Please provide the translation of the given English sentence into <language>, along with a word-for-word dictionary for each word.
> (2) The output format must be strictly followed:
> 1. Start with 'English:' followed by the English sentence.
> 2. On the next line, start with '<lang>:' followed by the <language> translation.
> 3. On the next line, start with 'dictionary:' followed by each word in the <language> sentence, annotated with its English meaning in parentheses, separated by spaces.
> (3) Now generate translations for the following sentence:
> English: <target>
> <language>: <source>
> dictionary:

[3]We conduct our experiments on non-English languages with English-centric LLMs, and we leave other settings to future work.

, where <language> presents the language that the original synthetic questions are written in, <target> represents the original English sentence which is used to obtain the <source> sentences that are written in <language>.

By prompting LLMs, we obtain a dictionary mapping that could be customized by replacing the arguments with any bilingual corpora.

Note that the pivoting into English and English-based thoughts is indeed optional in DIP and can be pruned for a better trade-off between performance and computational cost, as longer generation usually requires more computational costs.

3 Experimental Setup

3.1 Baselines

We conduct experiments with ChatGPT (GPT-4o-mini), Llama-3.2-1b-instruct (Dubey et al., 2024), and Mixtral-8x7b (Jiang et al., 2024). At the time of writing, all of them are popular and widely used English-centric LLMs which are strong in their multilingual and reasoning capacities. Based on these popular LLMs, we compare DIP to strong baseline methods:

- **Standard Prompting** that directly asks the English-centric LLMs to answer the questions written in those non-English languages.
- **Non-insertion Prompting** prepends/appends the dictionary to the prompt.
- **English Pivoting** that asks the model to translate the question into English before answering (Kim et al., 2019).
- **English Pivot Thought** that asks the model to translate the question into English before answering with chain-of-thought reasoning.

3.2 Datasets

We construct synthetic benchmarks in 200 languages from FLORES-200 (NLLB-Team, 2022) using the NLLB Translators from the following existing benchmarks:

- **GSM8K** is a benchmark of math word problems. We randomly sample 200 instances for each of the 200 languages from FLORES-200 for GSM8K. We also randomly sample 10 low-resourced languages from GSM8K and conduct experiments on the full 1,319 test instances on them (Cobbe et al., 2021).

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	17.89	9.10	15.62	11.52	11.30	5.91	14.25	5.84	9.55	9.40	11.04
Non-insertion Prompting	14.94	6.29	13.87	10.08	10.24	6.37	12.89	7.05	10.24	8.19	10.02
English Pivoting	23.65	14.03	19.94	19.79	19.26	10.31	21.91	9.55	17.82	17.44	17.37
English Pivot Thought	61.11	36.77	60.35	57.16	49.13	22.67	60.96	20.85	44.28	35.48	44.88
DIP w/o EP w/o CT	20.39	12.05	22.06	15.47	13.57	12.43	19.03	13.95	16.38	18.20	16.35
DIP w/ EP w/o CT	23.43	16.30	24.94	21.23	18.50	14.86	23.58	19.18	22.14	22.37	20.65
DIP	**67.93**	**46.17**	**80.36**	**67.10**	**53.30**	**43.29**	**68.61**	**60.50**	**63.68**	**68.31**	**61.92**

Table 1: Results for GPT-4o on GSM8K on 10 randomly selected low-resourced languages from FLORES-200. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	53.33	42.00	46.33	48.33	36.67	28.33	53.00	25.33	37.67	40.67	41.17
Non-insertion Prompting	52.67	34.33	45.00	46.00	38.33	29.67	53.00	28.67	45.67	41.33	41.47
English Pivoting	54.67	51.33	51.67	63.00	53.67	40.33	63.67	35.00	57.33	51.67	52.23
English Pivot Thought	61.33	**61.33**	58.67	71.00	60.67	41.67	71.00	36.67	69.33	55.33	58.70
DIP w/o EP w/o CT	66.00	43.00	73.00	65.67	49.00	45.33	62.67	48.67	62.33	61.67	57.73
DIP w/ EP w/o CT	66.67	55.67	75.67	73.67	57.00	55.67	70.00	62.00	67.33	66.00	64.97
DIP	**78.33**	57.00	**89.67**	**76.67**	**65.33**	**65.67**	**77.67**	**71.00**	**78.67**	**75.00**	**73.50**

Table 2: Results for GPT-4o on SVAMP on 10 randomly selected low-resourced languages from FLORES-200. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	4.00	1.00	3.67	3.00	3.00	3.67	3.00	2.67	3.33	3.00	3.03
Non-insertion Prompting	2.67	2.00	3.00	3.00	2.67	1.33	1.67	3.00	1.00	1.00	1.87
English Pivoting	4.00	2.67	9.00	3.33	4.00	3.33	3.33	4.67	3.67	1.00	4.30
English Pivot Thought	4.00	3.67	13.00	2.67	2.33	3.67	6.67	4.67	3.33	3.67	4.80
DIP w/o EP w/o CT	3.33	5.00	4.67	3.67	2.00	4.00	1.67	3.67	1.33	4.00	2.73
DIP w/ EP w/o CT	7.33	4.00	8.00	4.33	3.67	2.67	3.33	4.33	2.00	4.67	4.20
DIP	7.00	**3.67**	12.00	**3.67**	**5.33**	**4.33**	**7.00**	**5.00**	**4.33**	**7.00**	**5.83**

Table 3: Results for Llama-3.2 on SVAMP on 10 randomly selected low-resourced languages from FLORES-200. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	28.00	14.00	36.00	16.33	8.33	12.33	28.33	17.00	27.67	20.67	18.97
Non-insertion Prompting	25.00	23.00	30.67	27.00	23.33	22.67	21.00	29.00	25.00	24.00	22.67
English Pivoting	35.00	10.33	51.33	12.00	7.67	10.33	38.00	18.00	30.67	24.00	22.43
English Pivot Thought	22.67	5.67	27.33	10.33	6.33	9.00	14.00	18.00	20.33	15.33	12.97
DIP w/o EP w/o CT	22.67	17.00	32.67	29.00	21.33	25.33	18.00	39.67	18.00	24.33	22.80
DIP w/ EP w/o CT	27.33	17.67	31.00	36.00	18.33	29.67	17.00	38.33	31.33	40.67	28.00
DIP	**44.00**	**29.00**	**50.33**	**48.00**	**32.00**	**39.67**	**39.00**	**55.67**	**48.33**	**51.67**	**43.27**

Table 4: Results for Mixtral on SVAMP on 10 randomly selected low-resourced languages from FLORES-200. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

- **SVAMP** is a benchmark of math word problems with varying structures. We randomly sample 10 low-resourced languages from FLORES-200 and conduct experiments on the

Methods	# improved	> 5 points	> 10 points	> 20 points	# degraded	> 5 points	> 20 points
Non-insertion Prompting	87/200	9/87	1/87	0/87	113/200	9/113	0/113
English Pivoting	181/200	80/181	8/181	0/181	19/200	0/19	0/19
English Pivot Thought	193/200	8/193	14/193	151/193	7/200	0/7	0/7
DIP w/o EP w/o CT	193/200	84/193	32/193	0/193	7/200	0/7	0/7
DIP w/ EP w/o CT	198/200	86/198	82/198	7/198	2/200	0/2	0/2
DIP	**200/200**	**1/200**	**2/200**	**196/200**	**0/200**	**0/0**	**0/0**

Table 5: Statistics of the changes in accuracy with DIT and other baselines compared to Standard Prompting on GPT-4o with 200 languages on GSM8K. 100% of the directions (200 out of 200) are improved with DIT.

full 1,000 test instances (Patel et al., 2021).

- **AQuA** is a dataset of algebraic word problems. We randomly sample 10 low-resourced languages from FLORES-200 and conduct experiments on their full.
- **Date and Sport** We selected two specialized evaluation sets from the BIG-bench initiative (Srivastava et al., 2022): Date Understanding, which requires inferring a date based on a given context, and Sports Understanding, which involves assessing whether a sports-related sentence is plausible or implausible. We randomly sample 10 low-resourced languages from FLORES-200 and conduct experiments on their full test set.

3.3 Evaluation Metrics

Accuracy is used to evaluate the reasoning tasks. In addition, we use BLEU (Papineni et al., 2002) and chrF++ (Popović, 2015) to evaluate the translation quality as well as the intermediate thinking process. We use the evaluations provided by the sacreBLEU repository with the default signatures.[4]

3.4 Synthetic Generation Quality

Under our setting, almost all the words from the question prompt are assigned a dictionary. In order to ensure the quality of the generated dataset with the 10 languages we study, we employ three experienced human annotators. They are all postgraduate student who are studying for English-relevant degrees. We use GPT to translate the synthetic benchmark back to English and ask them to annotate whether the translated-back English is the same as the original English. With this method, we only preserve the instances that are perfectly agreed upon by all the annotators, and the meanings are preserved. More than 90% of the instances are preserved for each language, and this process ensures a perfect agreement for the final datasets among our annotators.

4 Math Reasoning Tasks

4.1 GSM8K

GSM8K Table 1 presents the results on GSM8K written in 10 randomly selected low-resourced languages from FLORES-200 (NLLB-Team, 2022) on GPT-4o. The results indicate the effectiveness of DIP in comparison to the baselines. English pivoting does help to improve the Standard Prompting baseline from an average of 11.04 to 17.37. Chain-of-thought reasoning is especially useful, improving the average score from 17.37 on English Pivoting to 44.88 on English Pivot Thought. Compared to the most naive Standard Prompting baseline, DIP obviously improves the average performance from 11.04 to 61.92. On bug_Latn, DIP still obviously improves the strongest baseline English Pivot Thought from 20.85 to 60.50. We also observe the effectiveness of interleaving the dictionary by insertion, which improves traditional appending/prepending previously employed by machine translation, improving the average score from 10.02 on Non-insertion Prompting to 16.35 on DIP w/o EP w/o CT. We postulate such an interleaving manner makes it easy to catch the context between the mapped dictionary pairs, which enhances the model understanding.

Due to space reasons, we leave Figure 2 in the Appendix, which visually presents the performance gap between DIP and the baselines on the full 200 languages in FLORES-200 on GPT-4o. We observe that DIP has quite impressive improvements on lower-resource languages such as the ones in the

[4]https://github.com/mjpost/sacrebleu

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	47.20	47.20	53.60	54.40	61.60	47.20	55.60	45.20	46.40	46.00	50.44
Non-insertion Prompting	48.00	39.20	44.40	43.60	43.60	42.40	43.20	46.00	41.60	43.60	43.56
English Pivoting	49.60	42.80	45.20	44.80	49.60	48.00	47.20	43.20	42.80	43.20	45.64
English Pivot Thought	**74.80**	58.00	63.60	64.00	73.20	52.40	69.60	40.40	56.00	63.60	61.56
DIP w/o EP w/o CT	48.00	38.00	44.40	45.60	47.60	40.40	41.60	38.40	39.20	44.80	42.80
DIP w/ EP w/o CT	49.60	44.80	46.80	48.00	50.40	48.80	45.60	45.20	44.00	46.80	47.00
DIP	72.40	**66.80**	**73.20**	**71.60**	**76.00**	**66.40**	**77.60**	**70.00**	**75.60**	**75.60**	**72.52**

Table 6: Results for GPT-4o on Date Understanding on 10 randomly selected low-resourced languages. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	49.60	56.00	42.80	56.80	55.60	53.20	43.60	54.00	27.20	45.20	48.40
Non-insertion Prompting	48.40	47.60	46.00	51.60	50.00	52.40	41.60	48.80	52.80	54.40	49.36
English Pivoting	48.40	48.80	50.80	48.00	42.80	47.60	52.40	47.60	51.60	45.60	48.36
English Pivot Thought	47.20	50.40	46.80	46.80	46.00	47.20	47.60	46.80	46.80	46.00	47.16
DIP w/o EP w/o CT	46.80	50.00	46.80	55.20	56.80	56.00	42.80	56.80	22.00	52.40	48.56
DIP w/ EP w/o CT	56.40	54.80	64.00	60.40	60.40	**60.40**	56.40	54.80	58.80	54.40	58.08
DIP	**58.80**	**64.40**	**67.20**	**60.80**	**62.40**	59.60	**58.80**	**64.00**	**63.20**	**59.20**	**61.84**

Table 7: Results for GPT-4o on Sports Understanding on 10 randomly selected low-resourced languages. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

fourth row. The improvements with DIP on higher-resourced languages, yet, there are still improvements. Table 5 presents the detailed improvement statistics, and we observe that while traditional English Pivoting and English Pivot Thought make good improvements, they are usually not drastic, with about a 5-10 points increase in accuracy. In comparison, DIP gives a large improvement with 196/200 languages enjoying an improvement with over 20 points in accuracy. This concludes the usefulness of DIP.

4.2 SVAMP

Table 2 presents the results on SVAMP written in 10 randomly selected low-resourced languages from FLORES-200 (NLLB-Team, 2022) on GPT-4o. Similar to the results on GSM8K, there is only one language in which DIP scored less than the English Pivot Thought baseline. On SVAMP, the chain-of-thought reasoning improves less (from 52.23 on English Pivoting to 58.70 on English Pivot Thought) than English pivoting (from 41.17 on Standard Prompting to 52.23 on English Pivoting). This means that the usefulness of these two techniques varies on different tasks. The overall improvement from DIP is obvious, where DIP improves the average score from 58.70 on English Pivot Thought to 73.50. We also observe the effectiveness of interleaving the dictionary by insertion, which improves Non-insertion Prompting, improving the average score from 41.47 on Non-insertion Prompting to 57.73 on DIP w/o EP w/o CT.

Due to limited computational resources, we conduct experiments on SVAMP on Llama-3.2 in Table 3 and Mixtral in Table 4, which shows that the results are consistent as shown on GPT-4o. Results on Llama-3.2 are usually lower than expected since those low-resourced languages are not perfectly supported on the Llama-3.2 we used.

4.3 Ablation Study

The last three rows across all tables above indicate the effectiveness of different components in DIP, where the performance decreases when we remove English pivoting and chain-of-thought. We also observe that the performance of DIP variants is constantly better than or on par with Non-insertion Prompting. This concludes that the position of the dictionary is important, and with an interleaving dictionary only, DIP still surpasses some strong baselines, such as the English pivoting baseline as in Table 2 on SVAMP.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	17.20	16.00	22.80	19.20	22.00	20.40	11.20	21.60	13.20	12.80	17.22
Non-insertion Prompting	16.00	14.40	19.20	16.80	13.20	12.00	15.60	16.40	13.60	15.20	15.16
English Pivoting	20.80	23.20	18.80	24.40	25.60	18.40	**24.40**	20.00	20.00	22.80	21.86
English Pivot Thought	**21.20**	20.40	17.20	24.40	**25.60**	18.80	18.80	18.80	22.40	**24.40**	22.16
DIP w/o EP w/o CT	14.40	20.40	16.80	17.20	20.40	16.80	10.40	17.20	18.00	18.40	19.16
DIP w/ EP w/o CT	20.40	23.60	**22.80**	19.20	20.80	16.00	18.40	15.20	21.20	19.60	21.70
DIP	20.40	**26.80**	20.00	**29.20**	24.00	**22.00**	19.20	**22.80**	**23.60**	20.40	**23.94**

Table 8: Results for Llama-3.2 on Date Understanding on 10 randomly selected low-resourced languages. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
Standard Prompting	22.40	31.60	24.80	29.60	34.80	32.40	18.80	28.80	25.60	21.60	30.92
Non-insertion Prompting	14.40	17.20	18.00	19.20	20.80	26.00	13.20	20.80	11.20	14.40	18.52
English Pivoting	14.40	18.80	15.60	19.60	16.80	18.00	9.20	23.60	12.00	13.60	14.68
English Pivot Thought	21.60	24.00	29.60	25.20	25.60	26.00	18.80	29.60	24.00	21.60	25.01
DIP w/o EP w/o CT	48.00	41.60	43.60	47.60	48.40	43.20	44.00	**51.60**	39.20	41.60	40.64
DIP w/ EP w/o CT	39.60	34.40	36.80	37.60	42.40	41.60	34.00	40.00	32.40	36.00	35.78
DIP	**48.80**	**41.60**	**51.20**	**52.00**	**51.20**	**48.40**	**45.60**	49.60	**45.20**	**57.20**	**49.08**

Table 9: Results for Mixtral on Date Understanding on 10 randomly selected low-resourced languages. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
English Pivoting	47.70 / 69.28	40.94 / 60.11	56.04 / 76.59	49.25 / 67.31	46.74 / 65.09	22.35 / 43.43	41.93 / 64.48	26.68 / 52.47	34.14 / 53.31	33.31 / 58.05	39.91 / 61.01
DIP	**57.43 / 75.81**	**54.11 / 71.80**	**72.93 / 84.47**	**69.63 / 81.91**	**61.61 / 75.82**	**46.86 / 67.54**	**58.59 / 76.32**	**71.59 / 84.97**	**62.24 / 78.75**	**55.09 / 72.34**	**61.00 / 76.97**

Table 10: Evaluations on translation quality on 10 randomly selected low-resourced languages on GSM8K on GPT-4o, evaluated in BLEU / chrF++ scores.

Model	kaz_Cyrl	nso_Latn	srp_Cyrl	xho_Latn	ibo_Latn	tum_Latn	asm_Beng	bug_Latn	ckb_Arab	azb_Arab	Average
English Pivot Thought	4.99 / 21.32	4.58 / 20.24	6.13 / 23.10	**3.73 / 19.34**	5.19 / 21.80	3.65 / 17.61	**6.13 / 23.48**	4.45 / 19.46	4.68 / 20.40	**5.52** / 21.52	4.91 / 20.83
DIP	**5.91 / 23.16**	**4.84 / 20.85**	**7.96 / 25.38**	2.79 / 16.47	**5.89 / 22.80**	**5.16 / 20.94**	5.97 / 23.47	**6.94 / 24.26**	**5.54 / 22.40**	5.30 / **22.87**	**5.63 / 22.26**

Table 11: Evaluations on thinking quality on 10 randomly selected low-resourced languages on GSM8K on GPT-4o, evaluated in BLEU / chrF++ scores.

4.4 Commonsense Reasoning Tasks

Date Understanding Table 6 presents the results on Date Understanding written in 10 randomly selected low-resourced languages from FLORES-200 (NLLB-Team, 2022) on GPT-4o. Similar to the results on math tasks, DIP is obviously better than all the baselines, except for on kaz_Cyrl. We also note that the experiments on Llama-3.2 in Table 8 and Mixtral in Table 9 all give the same conclusion that DIP is obviously effective.

On Llama-3.2, it is obviously better than the baselines, improving it from 17.22 on average on Standard Prompting to 23.94 on average with DIP. In contrast, on Mixtral, DIP gives a better improvement than on Llama, by to 49.08 on DIP, which probably indicate that DIP can gives a better improvement when the applied LLMs are stronger in performance. Nevertheless, the performance improvement from DIP is consistent. We also note that for some languages on Llama-3.2, DIP can be a bit lower than some baselines, such as in Table 8 on ibo_Latn. However, this does not affect the overall final conclusion that DIP is obviously better than

the baselines, according to the obvious averaged performance improvement with DIP as in Table 8.

Sports Understanding Table 7 presents the results on Sports Understanding written in 10 randomly selected low-resourced languages from FLORES-200 (NLLB-Team, 2022). Similar to the results on math tasks, DIP is obviously better than all the baselines, on all the languages. This suggests the consistent improvement and the usefulness of DIP in the task of Sports Understanding.

Ablation Study The last three rows indicate the effectiveness of different components in DIP, where the performance decreases when we remove English pivoting and chain-of-thought. All those components are important to DIP.

4.5 Translating Performance

In order to study the effectiveness of DIP, we conduct deeper analysis as in Table 10 and Table 11. We found that there are two main reasons why DIP is useful. We found that DIP usually gives a better English pivoting performance as in Table 10. This succeedingly gives a better thinking process under the evaluations as demonstrated in Table 11.

As in Table 10, we see that the translation performance has improved from 39.91 in BLEU to 61.00 in BLEU, which is a large improvement. This is also consistent with chrF++ scores, improving from 61.01 to 76.97. This represents the usefulness of DIP in terms of the intermediate translation.

In Table 11, we see that while the thinking process can be diverse, there is a clear improvement between DIP and the baselines in terms of their thinking process to the original ground truth in GSM8K. This also indicates that the synthetic benchmark has reasonable quality, as the translated-back English aligns well with the original English dataset.

5 Related Work

Multilingual Tasks on Large Language Models There has been limited research conducted on effective methods for prompting English-centric large language models on non-English tasks, such as the standard cross-lingual tasks such as machine translation. Most of the existing research focused on evaluating the translation performance of English-centric LLMs, using prompts such as 'Translate to language_name: text' (Brown et al., 2020; Lin et al., 2022; Le Scao et al., 2022; Zhang et al., 2022). Different prompt formats are explored (Reynolds and McDonell, 2021; Wang et al., 2023), Furthermore, Garcia and Firat (2022) have investigated the potential need for prompts for regulating the formality or specific dialect of the generation. Finally, Agrawal et al. (2022) and Vilar et al. (2022) focused on identifying appropriate in-context examples to improve machine translation quality with LLMs. In addition to machine translation which has already scaled to over 200 languages from FLORES-200 (NLLB-Team, 2022), there is also a trend in solving non-English reasoning tasks on English-centric LLMs (Huang et al., 2023), yet, the number of languages studied are usually insufficient, with about tens of languages.

Dictionary-based Method for Multilingual Language Models This research is relevant to the idea of lexical restrictions in the task of machine translation. This can be divided into either hard constraints (Hokamp and Liu, 2017; Post and Vilar, 2018) or soft constraints (Song et al., 2019; Dinu et al., 2019; Chen et al., 2021).

There have been several works that explored using dictionaries in supervised machine translation. Zhang and Zong (2016) enhance neural machine translation (NMT) by integrating a bilingual dictionary that incorporates less common or unseen words found in the bilingual training data. Arthur et al. (2016) improve the translation of rare words by augmenting the system with discrete translation lexicons and leveraging the attention vector to identify the relevant lexical probabilities. Hämäläinen and Alnajjar (2020) employs a dictionary to generate synthetic parallel data, thereby enhancing the training of NMT models. While most of previous work has focused on using dictionaries for the task of machine translation, doing multilingual reasoning tasks is under-studied.

In contrast, DIP is the first work that exploits the use of a dictionary in terms of reasoning tasks in non-English languages on English-centric LLMs.

6 Conclusions

In conclusion, our proposed method, DIP, provides an effective solution to multilingual reasoning by inserting English counterparts for non-English prompts, experimental analysis indicates that DIP enhances translation accuracy and reasoning capabilities within English-centric LLMs. Through extensive experiments on approximately 200 languages using synthetic multilingual benchmarks created from existing benchmarks such as GSM8K

and SVAMP, DIP has demonstrated substantial improvement in multilingual math and commonsense reasoning tasks across various LLMs. We also found that interleaving the dictionaries plays an important factor in the final performance.

Limitations

This paper presents an analysis of 200 languages only. However, there are more than thousands of languages around the world. The paper can be further extended by including more languages as well as more analysis.

Ethical Statement

We honour and support the ACL ARR Code of Ethics. There is no ethical issue known to us. Well-known and widely used LLMs are used in our work, which is subjected to generating offensive context. However, the above-mentioned issues are widely known to commonly exist for LLMs. Any content generated does not reflect the view of the authors.

References

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context Examples Selection for Machine Translation. *arXiv e-prints*, arXiv:2212.02437.

Philip Arthur, Graham Neubig, and Satoshi Nakamura. 2016. Incorporating discrete translation lexicons into neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1557–1567, Austin, Texas. Association for Computational Linguistics.

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.

Guanhua Chen, Yun Chen, Yong Wang, and Victor O. K. Li. 2021. Lexical-constraint-aware neural machine translation via data augmentation. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training Verifiers to Solve Math Word Problems. *arXiv e-prints*, arXiv:2110.14168.

Georgiana Dinu, Prashant Mathur, Marcello Federico, and Yaser Al-Onaizan. 2019. Training neural machine translation to apply terminology constraints. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3063–3068, Florence, Italy. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ron-

nie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The Llama 3 Herd of Models. *arXiv e-prints*, arXiv:2407.21783.

Xavier Garcia and Orhan Firat. 2022. Using natural language prompts for machine translation. *arXiv e-prints*, arXiv:2202.11822.

Mika Hämäläinen and Khalid Alnajjar. 2020. A template based approach for training nmt for low-resource uralic languages - a pilot with finnish. In *Proceedings of the 2019 2nd International Conference on Algorithms, Computing and Artificial Intel-*

ligence, ACAI '19, page 520–525, New York, NY, USA. Association for Computing Machinery.

Chris Hokamp and Qun Liu. 2017. Lexically constrained decoding for sequence generation using grid beam search. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1535–1546, Vancouver, Canada. Association for Computational Linguistics.

Ruijie Hou, Yueyang Jiao, Hanxu Hu, Yingming Li, Wai Lam, Huajian Zhang, and Hongyuan Lu. 2025. LNE-blocking: An efficient framework for contamination mitigation evaluation on large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3512–3528, Suzhou, China. Association for Computational Linguistics.

Hanxu Hu, Hongyuan Lu, Huajian Zhang, Yun-Ze Song, Wai Lam, and Yue Zhang. 2024. Chain-of-symbol prompting for spatial reasoning in large language models. In *First Conference on Language Modeling*.

Haoyang Huang, Tianyi Tang, Dongdong Zhang, Xin Zhao, Ting Song, Yan Xia, and Furu Wei. 2023. Not all languages are created equal in LLMs: Improving multilingual capability by cross-lingual-thought prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12365–12394, Singapore. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts. *arXiv e-prints*, arXiv:2401.04088.

Yunsu Kim, Petre Petrov, Pavel Petrushkov, Shahram Khadivi, and Hermann Ney. 2019. Pivot-based transfer learning for neural machine translation between non-English languages. 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 866–876, Hong Kong, China. Association for Computational Linguistics.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, 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 Laurençon, Yacine Jernite, Julien Launay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph 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, María Grandury, Mario Šaško, 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 Luis López, Rui Ribeiro, Salomey Osei, Sampo Pyysalo, Sebastian Nagel, Shamik Bose, Shamsuddeen Hassan Muhammad, Shanya Sharma, Shayne Longpre, Somaieh Nikpoor, Stanislav 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, Davut Emre Taşar, 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 Nayak, Ryan Teehan, Samuel Albanie, Sheng Shen, Srulik Ben-David, Stephen H. Bach, Taewoon Kim, Tali Bers, Thibault Fevry, Trishala Neeraj, Urmish Thakker, Vikas Raunak, Xiangru Tang, Zheng-Xin Yong, Zhiqing Sun, Shaked Brody, Yallow 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 François Lavallée, 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, Aurélie Névéol, Charles Lovering, Dan Garrette, Deepak Tunuguntla, Ehud Reiter, Ekaterina Taktasheva, Ekaterina Voloshina, Eli Bogdanov, Genta Indra Winata, Hailey Schoelkopf, Jan-Christoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawa-

mura, 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, Shani Pais, Tatiana Shavrina, Thomas Scialom, Tian Yun, Tomasz Limisiewicz, Verena Rieser, Vitaly Protasov, Vladislav Mikhailov, Yada Pruksachatkun, Yonatan Belinkov, Zachary Bamberger, Zdeněk Kasner, Alice Rueda, Amanda Pestana, Amir Feizpour, Ammar Khan, Amy Faranak, Ana Santos, Anthony Hevia, Antigona Unldreaj, Arash Aghagol, Arezoo Abdollahi, Aycha Tammour, Azadeh HajiHosseini, Bahareh Behroozi, Benjamin Ajibade, Bharat Saxena, Carlos Muñoz Ferrandis, Danish Contractor, David Lansky, Davis David, Douwe Kiela, Duong A. Nguyen, Edward Tan, Emi Baylor, Ezinwanne Ozoani, Fatima Mirza, Frankline Ononiwu, Habib Rezanejad, Hessie Jones, Indrani Bhattacharya, Irene Solaiman, Irina Sedenko, Isar Nejadgholi, Jesse Passmore, Josh Seltzer, Julio Bonis Sanz, Livia 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, Nour Fahmy, Olanrewaju Samuel, Ran An, Rasmus Kromann, Ryan Hao, Samira Alizadeh, Sarmad Shubber, Silas Wang, Sourav Roy, Sylvain Viguier, Thanh Le, Tobi Oyebade, Trieu Le, Yoyo Yang, Zach Nguyen, Abhinav Ramesh Kashyap, Alfredo Palasciano, Alison Callahan, Anima Shukla, Antonio Miranda-Escalada, Ayush Singh, Benjamin Beilharz, Bo Wang, Caio Brito, Chenxi Zhou, Chirag Jain, Chuxin Xu, Clémentine Fourrier, Daniel León Periñán, Daniel Molano, Dian Yu, Enrique Manjavacas, Fabio Barth, Florian Fuhrimann, Gabriel Altay, Giyaseddin Bayrak, Gully Burns, Helena U. Vrabec, Imane Bello, Ishani Dash, Jihyun Kang, John Giorgi, Jonas Golde, Jose David Posada, Karthik Rangasai Sivaraman, Lokesh Bulchandani, Lu Liu, Luisa Shinzato, Madeleine Hahn de Bykhovetz, Maiko Takeuchi, Marc Pàmies, Maria A Castillo, Marianna Nezhurina, Mario Sänger, Matthias Samwald, Michael Cullan, Michael Weinberg, Michiel De Wolf, Mina Mihaljcic, Minna Liu, Moritz Freidank, Myungsun Kang, Natasha Seelam, Nathan Dahlberg, Nicholas Michio Broad, Nikolaus Muellner, Pascale Fung, Patrick Haller, Ramya Chandrasekhar, 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 Bharati, Tanmay Laud, Théo Gigant, Tomoya Kainuma, Wojciech Kusa, Yanis Labrak, Yash Shailesh Bajaj, Yash Venkatraman, Yifan Xu, Yingxin Xu, Yu Xu, Zhe Tan, Zhongli Xie, Zifan Ye, Mathilde Bras, Younes Belkada, and Thomas Wolf. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv e-prints*, arXiv:2211.05100.

Kun Li, Tianhua Zhang, Liping Tang, Junan Li, Hongyuan Lu, Xixin Wu, and Helen Meng. 2022. Grounded dialogue generation with cross-encoding re-ranker, grounding span prediction, and passage dropout. In *Proceedings of the Second DialDoc Workshop on Document-grounded Dialogue and Conversational Question Answering*, pages 123–129, Dublin, Ireland. Association for Computational Linguistics.

Peng Li, Tianxiang Sun, Qiong Tang, Hang Yan, Yuanbin Wu, Xuanjing Huang, and Xipeng Qiu. 2023. CodeIE: Large code generation models are better few-shot information extractors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15339–15353, Toronto, Canada. Association for Computational Linguistics.

Zheng Wei Lim, Nitish Gupta, Honglin Yu, and Trevor Cohn. 2024. Mufu: Multilingual Fused Learning for Low-Resource Translation with LLM. *arXiv e-prints*, arXiv:2409.13949.

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.

Hongyuan Lu, Haoyang Huang, Shuming Ma, Dongdong Zhang, Wai Lam, Zhaochuan Gao, Anthony Aue, Arul Menezes, and Furu Wei. 2023. TRIP: Accelerating document-level multilingual pre-training via triangular document-level pre-training on parallel data triplets. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7845–7858, Singapore. Association for Computational Linguistics.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Furu Wei, and Wai Lam. 2024a. Revamping multilingual agreement bidirectionally via switched back-translation for multilingual neural machine translation. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 264–275, St. Julian's, Malta. Association for Computational Linguistics.

Hongyuan Lu and Wai Lam. 2023. PCC: Paraphrasing with bottom-k sampling and cyclic learning for curriculum data augmentation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 68–82, Dubrovnik, Croatia. Association for Computational Linguistics.

Hongyuan Lu, Wai Lam, Hong Cheng, and Helen Meng. 2022. Partner personas generation for dialogue response generation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human*

Language Technologies, pages 5200–5212, Seattle, United States. Association for Computational Linguistics.

Hongyuan Lu, Zixuan Li, Zefan Zhang, and Wai Lam. 2025. SLoW: Select low-frequency words! automatic dictionary selection for translation on large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 898–913, Suzhou, China. Association for Computational Linguistics.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2023. Chain-of-Dictionary Prompting Elicits Translation in Large Language Models. *arXiv e-prints*, arXiv:2305.06575.

Hongyuan Lu, Haoran Yang, Haoyang Huang, Dongdong Zhang, Wai Lam, and Furu Wei. 2024b. Chain-of-dictionary prompting elicits translation in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 958–976, Miami, Florida, USA. Association for Computational Linguistics.

Hongyuan Adam Lu, Z. L., Victor Wei, Zefan Zhang, Zhao Hong, Qiqi Xiang, Bowen Cao, and Wai Lam. 2026. Adam's Law: Textual Frequency Law on Large Language Models. *arXiv e-prints*, arXiv:2604.02176.

Yinquan Lu, Wenhao Zhu, Lei Li, Yu Qiao, and Fei Yuan. 2024. LLaMAX: Scaling Linguistic Horizons of LLM by Enhancing Translation Capabilities Beyond 100 Languages. *arXiv e-prints*, arXiv:2407.05975.

NLLB-Team. 2022. No language left behind: Scaling human-centered machine translation.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Arkil Patel, Satwik Bhattamishra, and Navin Goyal. 2021. Are NLP models really able to solve simple math word problems? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2080–2094, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Matt Post and David Vilar. 2018. Fast lexically constrained decoding with dynamic beam allocation for neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1314–1324, New Orleans, Louisiana. Association for Computational Linguistics.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI EA '21, New York, NY, USA. Association for Computing Machinery.

Kai Song, Yue Zhang, Heng Yu, Weihua Luo, Kun Wang, and Min Zhang. 2019. Code-switching for enhancing NMT with pre-specified translation. 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 449–459, Minneapolis, Minnesota. ACL.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew Dai, Andrew La, Andrew Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, César Ferri Ramírez, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Chris Waites, Christian Voigt, Christopher D. Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele

Rodola, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A. Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germán Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Wang, Gonzalo Jaimovitch-López, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Shevlin, Hinrich Schütze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B. Simon, James Koppel, James Zheng, James Zou, Jan Kocoń, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh D. Dhole, Kevin Gimpel, Kevin Omondi, Kory Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros Colón, Luke Metz, Lütfi Kerem Şenel, Maarten Bosma, Maarten Sap, Maartje ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramírez Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L. Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael A. Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan A. Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan LeBras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Ruslan Salakhutdinov, Ryan Chi, Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel S. Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima, Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven T. Piantadosi, Stuart M. Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsu Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Ramasesh, Vinay Uday Prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Yiding Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2022. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv e-prints*, arXiv:2206.04615.

Tianyi Tang, Hongyuan Lu, Yuchen Jiang, Haoyang Huang, Dongdong Zhang, Xin Zhao, Tom Kocmi, and Furu Wei. 2024. Not all metrics are guilty: Improving NLG evaluation by diversifying references. 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 6596–6610, Mexico City, Mexico. Association for Computational Linguistics.

Masao Utiyama and Hitoshi Isahara. 2007. A comparison of pivot methods for phrase-based statistical machine translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491, Rochester, New York. Association for Computational Linguistics.

David Vilar, Markus Freitag, Colin Cherry, Jiaming Luo, Viresh Ratnakar, and George Foster. 2022. Prompting PaLM for Translation: Assessing Strategies and Performance. *arXiv e-prints*, arXiv:2211.09102.

Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2717–2739, Toronto, Canada. Association for Computational Linguistics.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a Good NLG Evaluator? A Preliminary Study. *arXiv e-prints*, arXiv:2303.04048.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2024. Chain-of-thought prompting elicits reasoning in large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Hua Wu and Haifeng Wang. 2007. Pivot language approach for phrase-based statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 856–863, Prague, Czech Republic. Association for Computational Linguistics.

Hao Yang, Hongyuan Lu, Dingkang Yang, Wenliang Yang, Peng Sun, Xiaochuan Zhang, Jun Xiao, Kefan He, Wai Lam, Yang Liu, and Xinhua Zeng. 2026. Stephanie2: Thinking, Waiting, and Making Decisions Like Humans in Step-by-Step AI Social Chat. *arXiv e-prints*, arXiv:2601.05657.

Hao Yang, Hongyuan Lu, Xinhua Zeng, Yang Liu, Xiang Zhang, Haoran Yang, Yumeng Zhang, Shan Huang, Yiran Wei, and Wai Lam. 2025. Stephanie: Step-by-step dialogues for mimicking human interactions in social conversations. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 153–166, Albuquerque, New Mexico. Association for Computational Linguistics.

Haoran Yang, Yumeng Zhang, Jiaqi Xu, Hongyuan Lu, Pheng-Ann Heng, and Wai Lam. 2024. Unveiling the generalization power of fine-tuned large language models. 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 884–899, Mexico City, Mexico. Association for Computational Linguistics.

Jiajun Zhang and Chengqing Zong. 2016. Bridging Neural Machine Translation and Bilingual Dictionaries. *arXiv e-prints*, arXiv:1610.07272.

Kechi Zhang, Zhuo Li, Jia Li, Ge Li, and Zhi Jin. 2023. Self-edit: Fault-aware code editor for code generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 769–787, Toronto, Canada. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. OPT: Open Pre-trained Transformer Language Models. *arXiv e-prints*, arXiv:2205.01068.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024a. Multilingual machine translation with large language models: Empirical results and analysis. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2765–2781, Mexico City, Mexico. Association for Computational Linguistics.

Wenhong Zhu, Hongkun Hao, Zhiwei He, Yun-Ze Song, Jiao Yueyang, Yumeng Zhang, Hanxu Hu, Yiran Wei, Rui Wang, and Hongyuan Lu. 2024b. CLEAN–EVAL: Clean evaluation on contaminated large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 835–847, Mexico City, Mexico. Association for Computational Linguistics.

Figure 2: An illustrated comparison of the GSM8K dataset with 200 languages from FLORES-200 on six baselines/variants and DIT. While DIP has good improvements in the higher-resourced languages, the performance improvement of DIT is especially obvious in low-resource languages in the last row of the graph. EP denotes the English pivoting translation process, and CT denotes chain-of-thought reasoning steps.