# Zero-Shot vs. Translation-Based Cross-Lingual Transfer: The Case of Lexical Gaps 

Abteen Ebrahimi ${ }^{1}$ and Katharina von der Wense ${ }^{1,2}$<br>${ }^{1}$ University of Colorado Boulder, ${ }^{2}$ Johannes Gutenberg University Mainz<br>abteen.ebrahimi@colorado.edu


#### Abstract

Cross-lingual transfer can be achieved through two main approaches: zero-shot transfer or machine translation (MT). While the former has been the dominant approach, both have been shown to be competitive. In this work, we compare the current performance and long-term viability of these methods. We leverage lexical gaps to create a multilingual question answering dataset, which provides a difficult domain for evaluation. Both approaches struggle in this setting, though zero-shot transfer performs better, as current MT outputs are not specific enough for the task. Using oracle translation offers the best performance, showing that this approach can perform well long-term, however current MT quality is a bottleneck. We also conduct an exploratory study to see if humans produce translations sufficient for the task with only general instructions. We find this to be true for the majority of translators, but not all. This indicates that while translation has the potential to outperform zero-shot approaches, creating MT models that generate accurate task-specific translations may not be straightforward.


## 1 Introduction

Cross-lingual transfer has helped to develop natural language processing (NLP) systems for a broader set of languages through two main approaches: zero-shot transfer, where a single multilingual model is finetuned on a source language and directly applied to a target language, and translationbased transfer, where data is translated via a machine translation (MT) system before being passed to a downstream model. While zero-shot transfer is more widely used, this has been called into question both for high-resource (Artetxe et al., 2023; Isbister et al., 2021) and low-resource (Ebing and Glavaš, 2023; Ebrahimi et al., 2022) languages.

In this work, we ask the overall question: Which approach has the greater potential in creating NLP
systems which perform at a high level for all languages? To do so, we create a focused question answering (QA) dataset leveraging a phenomena we expect to be challenging for both approaches: lexical gaps, or concepts which are explicitly denoted in one language that can only be expressed as a combination of words in another (Bentivogli and Pianta, 2000). Lexical gaps exist in a variety of domains, such as colors or foods (Khishigsuren et al., 2022), however due to its prevalence across many languages, in this work we focus specifically on kinship terminology: Farsi, e.g., marks the difference between amoo (English: paternal uncle), and daei (English: maternal uncle), while, in English, both relatives are generally just referred to as uncle. Examples in the dataset, created using templates, are simply structured and easy to solve for a human, though accurate translation or recognition of the relevant lexical gaps are required to identify the correct answer. While focusing solely on lexical gaps through this approach reduces example diversity, there is a trade-off as we gain more control in our experiments and the ability for finer-grained analysis.

Evaluating on a diverse set of 5 languages - Catalan, German, Farsi, Hindi, and Vietnamese - we first investigate existing models and find that both translation-based and zero-shot approaches struggle in this setting. However, the zero-shot approach is stronger, and we find that current MT systems do not preserve the required amount of detail for the task, instead falling back to the general translation (e.g., amoo being translated to uncle). Replacing MT outputs with an oracle translation, however, offers the best performance across all methods, showing that with sufficiently accurate translation, this approach can overcome the challenge of lexical gaps.

To see if collecting additional data is a feasible approach for closing the gap between current MT outputs and oracle translations, we conduct

| LGI | Answerable | T | Context Pair (Translated) | Question | Answer Choices | Predicted Answer | Correct Answer |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| 0 | Y | $\checkmark$ | My paternal uncle's name is Sena. <br> My maternal aunt's favorite food is ghormeh sabzi. | What is my paternal uncle's name? | [A] Sena <br> [B] ghormeh sabzi <br> [C] Not Answerable | A | A |
|  | Y | $x$ | My uncle's name is Sena. <br> My aunt's favorite food is ghormeh sabzi. | What is my uncle's name? |  | A | A |
| 0 | N | $\checkmark$ | My paternal uncle's name is Sena. <br> My maternal aunt's favorite food is ghormeh sabzi. | What is my maternal uncle's name? | [A] Sena <br> [B] ghormeh sabzi <br> [C] Not Answerable | C | C |
|  | N | $x$ | My uncle's name is Sena. <br> My aunt's favorite food is ghormeh sabzi. | What is my aunt's name? |  | C | C |
| 1 | Y | $\checkmark$ | My paternal uncle's name is Sena. My maternal uncle's name is Ali. | What is my paternal uncle's name? | [A] Sena <br> [B] Ali <br> [C] Not Answerable | A | A |
|  | Y | $x$ | My uncle's name is Sena. My uncle's name is Ali. | What is my uncle's name? |  | C | A |
| 2 | N | $\checkmark$ | My paternal uncle's name is Sena. <br> My maternal uncle's favorite food is ghormeh sabzi. | What is my maternal uncle's name? | [A] Sena <br> [B] ghormeh sabzi <br> [C] Not Answerable | C | C |
|  | N | $x$ | My uncle's name is Sena. <br> My uncle's favorite food is ghormeh sabzi. | What is my uncle's name? |  | A | C |

Table 1: A full example from the dataset, showing how two context templates, a corresponding question template, and different values can be used to create different examples. Italics represent the slots which were filled by kinship terms or value surface forms (e.g., "Sena", "Ali", "ghormeh sabzi"). The answerable column denotes if the original example is answerable or unanswerable. The T column marks if the translation is specific $(\boldsymbol{\checkmark})$ or general $(\boldsymbol{X})$, and how using the general translation affects the perceived correct answer. LGI denotes the Lexical Gap ID, which depends on the relation between the two kinship entities found in the contexts, and whether they can conflict or not when translated to English.
a case study with Farsi to see if humans produce translations appropriate for the task without any task-specific instructions. Our results are mixed: most, but not all, human translators produce translations close to the oracle which preserve the lexical gap information. Thus, while translation may be a viable long-term approach for cross-lingual transfer, building an appropriate MT system represents a bottleneck, as relying purely on any generally collected parallel data may not be sufficient for a specific downstream task.

## 2 Background

### 2.1 Lexical Gaps

Linguistic diversity has been studied extensively by typologists (Comrie, 1989), and for a detailed survey on it's relation to NLP, we refer the reader to Ponti et al. (2019). Lexical gaps, which we can consider a feature of this diversity across languages, and the concept of untranslatability have also been studied (Bella et al., 2022; Bentivogli and Pianta, 2000; Wierzbicka, 2008; Bentivogli et al., 2000; Santos, 1990). Specific to kinship terminology are the works of Khishigsuren et al. (2022) who create a multilingual database of terms and use it to evaluate MT outputs, and Khalilia et al. (2023) who introduce dialect-specific additions.

### 2.2 Cross-Lingual Transfer

Zero-shot cross-lingual transfer can be achieved using embedding models (Ruder et al., 2019), pretrained encoder models (Devlin et al., 2019; Lample and Conneau, 2019; Conneau et al., 2020), or, most recently, through the use of multilingual large language models (LLM; BigScience et al., 2023), ${ }^{1}$ which we use in this work. MT-based approaches have also been shown to be competitive with zeroshot transfer (Ansell et al., 2023; Artetxe et al., 2023; Ebing and Glavaš, 2023; Isbister et al., 2021). In this work, we focus on translate-test, where the target-language evaluation data is first translated to English, then used as input to a task-specific model. For clean comparison, we use the LLMs above as our task-specific model.

## 3 Dataset Construction

We construct a QA dataset around lexical gaps, which consists of simple, factual sentences made via templates. This approach has several benefits: it allows for a fine-grained analysis of lexical gaps, is easily extensible to multiple languages, has a small chance of overlap with the pretraining data, and creates a challenging evaluation for LLMs while remaining trivial for humans. A complete example

[^0]

Table 2: Results for the main experiment. Cells represent accuracies averaged across 3 samples. Columns represent the QA accuracy for each LLM, broken down by LGI, with -A and $-U$ representing answerability. For languages, EN represents the oracle translation, *-N represents NLLB translation, and *-G is for Google Translate. Standard deviations can be found in Table 7.
from the dataset can be found in Table 1 and Appendix A contains a worked example. We describe the overall pipeline below.

For a given language, we start with a set of kinship terms, and consider both their general translation to English, i.e., the translation which is likely to be most common but could create a conflict with respect to the lexical gap, and their specific translation, which preserves full meaning in English. For the templates, we begin with a set of descriptive single-sentence English context templates which each contain two slots, an entity slot, filled with kinship terms (c.f. Table 3), and a value slot, filled from a pool of surface forms linked to the template. Each context template has a corresponding question template, which asks a question which can be answered using only the context.

The question and context templates are then translated to the target language. To prevent an English bias, annotators are asked to create slot values which are natural to the language, and are allowed to modify the English templates to create the most natural target language translation. This creates a parallel set of templates, one in the target language and one in English. While we start with English in the dataset creation, for our experiments
we refer to the English as the oracle translation from the target language, as the entity slot is filled with the specific translation and the meaning is completely preserved. By starting from the same set of English templates for each languages, the dataset remains roughly semantically parallel across the different target languages. Additional details on the translation process can be found in Appendix A.1.

QA examples are created by pairing two context templates together and selecting one of the corresponding question templates. Depending on how the slots are filled, examples can be either answerable, i.e., the answer is found in one of the contexts, or unanswerable, created by switching the entity in the question template. Questions can be unanswerable due to either missing or conflicting information. We categorize examples into three categories, denoted by the lexical gap ID (LGI), which depends on the relationship between the two entities found in the context template pair. LGI 0 denotes examples where the entities do not conflict with respect to a lexical gap, i.e., the general form of the two slots is different. Examples with LGI 1 and 2 are those where the general value is the same, which causes conflict after poor translation. LGI 1 denotes examples which are answerable in the
target language which can become unanswerable after translation, while LGI 2 denotes examples which are unanswerable in the target language and become erroneously answerable if the translation does not use the specific forms. Therefore, while LGI 0 example can be either answerable or unanswerable, they will remain so regardless of the quality of translation; for LGI 1 and 2, poor translation will corrupt the context, and the answerability of the example can change.

The task is framed as a 3-way classification problem, with each value representing an answer choice, along with one option for 'not answerable' (Robinson and Wingate, 2022). Context and answer choice order are shuffled for each example. We collect translations for Farsi, Hindi, Vietnamese, Catalan, and German, with the first three representing languages with different kinship terminologies than English, and the latter two having a terminology which is practically one-to-one with English.

## 4 Zero-Shot Transfer vs. MT

### 4.1 Experimental Setup

In this experiment, we use GPT-3.5 and BLOOMZ 7b1 for the zero-shot approach. We also consider the 7 and 13-billion parameter versions of Llama 2 (Touvron et al., 2023), which is officially English-only. BLOOM-Z was trained using X-P3 (Muennighoff et al., 2023), which contains text from all languages in our experiments except for Farsi. For more information on the performance of these models on other languages and tasks, we refer the reader to Ahuja et al. (2023). For translation, we use Google Translate (Bapna et al., 2022) and No Language Left Behind (NLLB) (NLLB et al., 2022). Additional details for each model and the inference procedure can be found in Appendix B. As the total number of generated examples is large, for each language we take three independent samples of 3000 examples for evaluation, balanced across LGI, to reduce the number of examples to a feasible size.

### 4.2 Results

We present results in Table 2, and discuss our findings below.

Lexical gaps are difficult for zero-shot transfer. For multilingual LLMs, we see a decrease in performance for target language examples which involve lexical gap entities. For BLOOM, comparing LGI 0-A with LGI 1 shows a decrease in performance
of $6.55 \%$ on average across all languages with gaps. For GPT-3.5, the overall stronger model, this effect is less pronounced, with an average difference of $3.12 \%$, and an improvement for Vietnamese. This highlights the inherent difficulty models have with lexical gaps, as even though entities are explicitly referred to differently, models cannot correctly answer the questions.

## Oracle translation outperforms zero-shot trans-

 fer. For all languages and models, the average performance of the oracle English translation is greater than zero-shot performance. This difference is more drastic as the similarity of languages decreases from English; oracle translation only shows . $35 \%$ improvement for German when using GPT3.5 , but $23.36 \%, 14.27 \%$, and $26.26 \%$ for Farsi, Hindi, and Vietnamese respectively. Intuitively, using translation is effectively a requirement for the Llama models, particularly for the non-Latin script languages. This further supports prior findings regarding the viability of translation for cross-lingual transfer (Artetxe et al., 2023).
## Performance with MT lags behind oracle trans-

 lation, but not always. MT models are not able to reliably translate lexical gaps, as for LGI 1 and 2, performance on the MT data is generally far less than oracle translation. For example, GPT-3.5 performance on oracle Farsi data (LGI 1) is $98.70 \%$ while the same performance is $66.60 \%$ when NLLB is used for translation. Performance drops to random with Google Translate. This pattern is relatively consistent across languages and categories. However, there are some noticeable cases which differ, e.g., when using NLLB and BLOOM for LGI 2 Farsi. In these cases, poorer or incorrect translations may actually help the downstream QA model perform better for unanswerable questions: the noise added through translation makes detecting unanswerability easier by pushing apart the contexts being compared. A similar effect is observed when comparing NLLB and Google Translate - even among the same language, depending on the type of example and downstream model, neither MT model consistently outperforms the other. This pattern is explicitly prevalent when comparing Llama-7 to 13. Therefore, for successful crosslingual transfer through translation, the choice of translation model should be conditioned on the language and task-specific downstream model.LLMs can still struggle with unanswerable examples. Focusing specifically on performance with the oracle English translations, we see a large spread across LLMs. On the higher end, GPT3.5 gets an average performance of $96.5 \%$ accuracy, while Llama- 7 and 13 get $48.4 \%$ and $58.9 \%$ correct, respectively. BLOOM lies in the middle, with $67.7 \%$ correct. Low performance is due to extremely poor performance on the subset of unanswerable questions. For example, Llama-13 gets $92.8 \%$ of LGI 0-Answerable questions correct, but only $30.2 \%$ of unanswerable ones correct.

## 5 Human Translations of Lexical Gaps

In the previous section, we find that current MT models do not translate lexical gaps with the accuracy necessary to solve the QA task. A common approach to improve model performance is continued pretraining (Gururangan et al., 2020), and we can consider a situation where additional translations are collected for this adaptation. With this study, we are interested in learning how humans translate contexts containing lexical gaps, particularly if there are two entities which, when found independently, are typically translated to the same word. Specifically, we aim to find if (1) for our examples, will humans produce translations close to our oracle translation if they have no knowledge of the downstream task? and (2) If not, how do humans translate sentences which contain conflicting lexical gaps? Namely, do humans follow the 'One Sense Per Discourse' hypothesis (Gale et al., 1992) in our setting, and disambiguate conflicting lexical gaps which arise in the same context pair?

### 5.1 Experimental Setup

We focus on Farsi in this experiment, and use a sample of 15 instances from the dataset. This sample includes single context examples, where the general translation is appropriate, and double context examples with conflicts that require the specific translations to preserve meaning. We control for the order in which participants see the examples. Participants are randomly assigned to two groups; in Group 1, participants see single context examples before double context examples, and vice-versa for Group 2. We provide additional details in Appendix C.

Participants Participants are fluent bilingual speakers of both English and Farsi, many of whom have previous experience with professional transla-
tion. Translators are not given specific instructions; the only guideline is to "preserve the meaning of the passage." In total, 11 people participate, and 5 are assigned to Group 1 and 6 to Group 2.

### 5.2 Results

Of Group 1 participants, 2 of the 5 use the specific English translation, while the others continued using the general translation, even for conflicting sentences. For the two who used specific translations, one used them only for the latter half of examples. In Group 2, 5 of the 6 participants used a specific translation. This indicates that, while the majority of translators naturally lean towards the specific translation for conflicting sentences, i.e., confirming the One Sense per Discourse hypothesis, this is more likely if they are not shown sentences which only need the general translation first.

This variability in human translation represents a hurdle for translation-based cross-lingual transfer, in that the best translation for a given task may not always correlate with how a human naturally translates a given sentence, particularly if that human does not have knowledge of the task. This further confirms the finding that the best translations for cross-lingual transfer are task specific. When researchers are collecting data to train an MT model specifically for cross-lingual transfer, care must be taken that the translations are sufficient for the task.

## 6 Conclusion

In this work, we compare zero-shot and translationbased approaches to cross-lingual transfer with a dataset created around lexical gaps. Using current models, zero-shot transfer offers better performance, and MT models are not capable of sufficiently accurate translations. Using oracle translation reveals the long-term viability of the latter approach, however we find that generally collected translations are not always suitable for the task. As such, collecting task-specific data, or using approaches such as neuro-symbolic models, which can incorporate rules, may be necessary for strong translation-based transfer.

## 7 Ethics Statement

We do not believe that our main research has any ethical concerns. The languages covered by our dataset have a large number of speakers and are not endangered. By using templates which are based around simple, factual statements guarantees that
there are no harmful sentences present in the evaluation. Furthermore, when generating the dataset, we consider all possible slot-surface form combinations, and as such do not bias any particular kinship entity in any way.

For our human study, we received IRB approval before beginning the experiment. All participants were given, and signed, an approved Informed Consent form before they were allowed to begin translating. This form highlighted that we were not able to give the translators the full background information on the study until after the experiment, when they were given an approved debriefing form. The participants were informed that their translations would not be revealed in any way that could be linked back to them, and we ask all participants if they would like their translations to be considered for public release at any point. Participants were not paid due to the short duration of the experiment, and were given the choice of opting-in to being included in the acknowledgments section.

## 8 Limitations

There are several limitations in our work. First, we focus on only 5 languages, 3 of which contain considerable lexical gaps with English. Due to the inherent diversity found across languages, there may be differences, even in our limited domain, of model performance across languages. To account for this, we aim to collect data from a diverse set of languages with different kinship terminologies. We also focus solely on kinship terminology as it is commonly found across languages and is common knowledge among speakers. Performance on other types of lexical gaps may be different, and further experiments using different gaps may be a helpful evaluation. Second, while the use of templates has a number of benefits as highlighted in the paper, it means that the examples used for evaluation are quite simple. While some models do struggle with these examples, the repetition in the dataset and simplicity of the templates means that with finetuning, model performance would likely reach near perfect accuracy. As such, we focus solely on zero-shot evaluation, and believe that this is the best evaluation setup for our data. This approach provides valuable information in how a model is able to generalize to concepts it may not have encountered very often, but are still understood by most humans with just basic knowledge of a given language. In other words, due to the simple nature
of the questions, we would expect a model with strong understanding of a language to perform well on our dataset, and while strong performance may not be generally informative towards the overall strength of the model, we believe that weak performance, which we observe, becomes an important signal. Overall, we believe the trade-off between diversity in examples and stronger control in experiments is valuable due to the finer-grained insights we can gain.

Another limitation arises when comparing the long-term viability of translation vs zero-shot transfer based approaches. For translation, since translation quality typically represents the bottleneck, expected long-term or future performance can be measured by simply replacing the MT outputs with oracle translations. In contrast, it is more difficult to measure the expected performance of zero-shot transfer. Therefore, while we can claim that improvements in translation quality can help overcome the challenge of lexical gaps (although we also find that collecting data for these improvements is not trivial), we cannot claim the same for zero-shot transfer given the experiments in this work. However, this does not mean that zero-shot transfer will never be able to solve lexical gaps.

## Acknowledgments

This work would not have been possible without the translation of templates provided by S. Rehan Ahmed, Minh Duc Bui, George Baker, and Ali Marashian. We would like to thank Atash Nowroozian, Ehsan Meysami, Farzad Soleimannejad, Mehdi Nikzad, Mohammad Kazerouni, Pari, Roozbeh Emami, Shahram Aghamir and all the other bilingual speakers for participating in the Farsi to English translation study. We would also like to thank the reviewers, meta-reviewer, and members of the CU NLP group, particularly Adam Wiemerslage and Maggie Perkoff for their feedback and discussion.

## References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 4232-4267, Singapore. Association for Computational Linguistics.

Alan Ansell, Marinela Parović, Ivan Vulić, Anna Korhonen, and Edoardo Ponti. 2023. Unifying crosslingual transfer across scenarios of resource scarcity. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pages 3980-3995, Singapore. Association for Computational Linguistics.

Mikel Artetxe, Vedanuj Goswami, Shruti Bhosale, Angela Fan, and Luke Zettlemoyer. 2023. Revisiting Machine Translation for Cross-lingual Classification.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Axelrod, Jason Riesa, Yuan Cao, Mia Xu Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apurva Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Hughes. 2022. Building Machine Translation Systems for the Next Thousand Languages.

Gábor Bella, Erdenebileg Byambadorj, Yamini Chandrashekar, Khuyagbaatar Batsuren, Danish Ashgar Cheema, and Fausto Giunchiglia. 2022. Language Diversity: Visible to Humans, Exploitable by Machines.
L. Bentivogli and Emanuelle Pianta. 2000. Looking for lexical gaps.

Luisa Bentivogli, Emanuele Pianta, and Fabio Pianesi. 2000. Coping with Lexical Gaps when Building Aligned Multilingual Wordnets. In Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00), Athens, Greece. European Language Resources Association (ELRA).

BigScience, 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, ZhengXin 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, JanChristoph Kalo, Jekaterina Novikova, Jessica Zosa Forde, Jordan Clive, Jungo Kasai, Ken Kawamura, Liam Hazan, Marine Carpuat, Miruna Clinciu, Najoung Kim, Newton Cheng, Oleg Serikov, Omer Antverg, Oskar van der Wal, Rui Zhang, Ruochen Zhang, Sebastian Gehrmann, Shachar Mirkin, 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, Daniel McDuff, 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. 2023. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Bernard Comrie. 1989. Language Universals and Linguistic Typology: Syntax and Morphology. University of Chicago Press, Chicago, IL.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised Cross-lingual Representation Learning at Scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 84408451, Online. 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.

Benedikt Ebing and Goran Glavaš. 2023. To Translate or Not to Translate: A Systematic Investigation of Translation-Based Cross-Lingual Transfer to LowResource Languages.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating Zero-shot Natural Language Understanding of Pretrained Multilingual Models in Truly Low-resource Languages. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 6279-6299, Dublin, Ireland. Association for Computational Linguistics.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992. One Sense Per Discourse. In Speech and Natural Language: Proceedings of a Workshop Held at Harriman, New York, February 23-26, 1992.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't Stop Pretraining: Adapt Language Models to Domains and Tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8342-8360, Online. Association for Computational Linguistics.
T. Isbister, F. Carlsson, and Magnus Sahlgren. 2021. Should we Stop Training More Monolingual Models, and Simply Use Machine Translation Instead? In Nordic Conference of Computational Linguistics.

Hadi Khalilia, Gábor Bella, Abed Alhakim Freihat, Shandy Darma, and Fausto Giunchiglia. 2023. Lexical diversity in kinship across languages and dialects. Frontiers in Psychology, 14:1229697.

Temuulen Khishigsuren, Gábor Bella, Khuyagbaatar Batsuren, Abed Alhakim Freihat, Nandu Chandran Nair, Amarsanaa Ganbold, Hadi Khalilia, Yamini Chandrashekar, and Fausto Giunchiglia. 2022. Using Linguistic Typology to Enrich Multilingual Lexicons: The Case of Lexical Gaps in Kinship. In Proceedings of the Thirteenth Language Resources and Evaluation Conference, pages 2798-2807, Marseille, France. European Language Resources Association.

Guillaume Lample and Alexis Conneau. 2019. Crosslingual Language Model Pretraining.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev,

Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 15991-16111, Toronto, Canada. Association for Computational Linguistics.

NLLB, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling HumanCentered Machine Translation.

Edoardo Maria Ponti, Helen O'Horan, Yevgeni Berzak, Ivan Vulić, Roi Reichart, Thierry Poibeau, Ekaterina Shutova, and Anna Korhonen. 2019. Modeling Language Variation and Universals: A Survey on Typological Linguistics for Natural Language Processing. Computational Linguistics, 45(3):559-601.

Joshua Robinson and David Wingate. 2022. Leveraging Large Language Models for Multiple Choice Question Answering. In The Eleventh International Conference on Learning Representations.

Sebastian Ruder, Ivan Vulić, and Anders Søgaard. 2019. A Survey of Cross-lingual Word Embedding Models. Journal of Artificial Intelligence Research, 65:569631.

Diana Santos. 1990. Lexical gaps and idioms in machine translation. In COLING 1990 Volume 2: Papers Presented to the 13th International Conference on Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu,

Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and FineTuned Chat Models.

Anna Wierzbicka. 2008. Why There Are No 'Colour Universals' in Language and Thought. The Journal of the Royal Anthropological Institute, 14(2):407425.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38-45, Online. Association for Computational Linguistics.

## Appendix

## A Dataset Construction

We illustrate the dataset construction through a specific example for Farsi. Let the specific kinship term for Context 1 (refer to this as Entity 1) be "my father's brother", with a specific translation of paternal uncle and a general translation of uncle. The kinship term for Context 2 (Entity 2) will be "my mother's brother", with specific translation maternal uncle and general translation uncle. Since the two general translations are equal, pairing these two entities together will yield an example in either LGI 1 or 2.

Let the Context 1 template be: "My [person]'s favorite food is [food]." and let the Context 2 template be: "My [person]'s name is [name].". Here, [person] represents the entity slot while [food] and [name] represent value slots. Context 1 has a corresponding question template (refer to this as Question 1): "What is my [person]'s favorite food?" and Context 2 has the question (Question 2): "What is my [person]'s name?". We can create the following LGI 2 examples, i.e., one which is unanswerable and becomes answerable, by filling Question 1 with Entity 2, e.g.,:

My paternal uncle's favorite food is [food].
My maternal uncle's name is [name].
What is my maternal uncle's favorite food?
with answer choices being the objects used to fill the value slots, i.e. the [food], [name] slots, as well as a 'Not answerable. ' option. Using the specific translation, we easily see that the question is unanswerable, as "my maternal uncle's favorite food" is never discussed. However, if we poorly translate both "paternal uncle" and "maternal uncle" to "uncle", we become able to erroneously answer the question.

To create an LGI 1 example, i.e. one which is answerable but becomes unanswerable, we need to use the same Contexts for both people, as if the contexts were different, the question would remain solvable (by only replying with the relevant value).

My paternal uncle's name is [name1].
My maternal uncle's name is [name2].
What is my maternal uncle's name?
Here, the answer is clearly namel, however
this becomes ambiguous with poor translation and therefore unanswerable.

## A. 1 Annotation Process

While using templates for creating examples is relatively simple in some languages such as Farsi and English, it becomes more complicated when moving to languages which require morphological changes within the templates, depending on the values used to fill empty slots (e.g., German or Vietnamese). We address these issues in two ways: first, by simply creating a different template for each combination of required morphological changes, or second, by defining a post-processing map which modifies the sentences after slots have been filled to correct any erroneous terms within the templates. For the first approach, we add metadata to the slot values (e.g., gender information for names and kinship terms) as well as the templates and use this information to constrain which templates can be used with which values. For the second approach, the slot meta-data is used to trigger specific transforms which correct the grammar.

Since each language is different, we required multiple rounds of translation and verification with each annotator for each language, which restricted the total number of languages we were able to include. As such, we aimed to select languages which maximized diversity across kinship terminology, language family, and script.

## B Experimental Setup

## B. 1 LLMs

We focus on three LLMs: GPT-3.5, ${ }^{2}$ Llama 2 (Touvron et al., 2023), and BLOOM-Z (BigScience et al., 2023). All main results are zero-shot evaluations of the models, and we do no additional finetuning.

GPT-3.5 For GPT-3.5, we rely on the OpenAI API. We use the gpt-3.5-turbo-0301 model for all of our results.

Llama We use the 7-billion and 13-billion parameter versions of Llama, finetuned for chatcompletion, in our work. In our preliminary results, we also experiment with using the text-completion Llama models however chose not to continue with them due to extremely poor performance. We use the official Llama implementation ${ }^{3}$ for inference.

[^1]BLOOM For BLOOM, we use the 7-billion parameter version through the Huggingface (Wolf et al., 2020) implementation.

## B. 2 Machine Translation

Google Translate We use the Google Translate (Bapna et al., 2022) API to collect translations. We specify both the source and target language.

NLLB For NLLB (NLLB et al., 2022), we use the Huggingface implementation. For the translations which we use in the main results, we use the nllb-200-distilled-1.3B version with default settings.

## B. 3 Extracting Question Answering Responses

Prompt and Hyperparameter Tuning We tune the input prompts, answer formats, system prompt (when applicable), and temperature setting for each LLM separately. We include the final prompts we use in Table 4. All models performed best with a temperature of 0.3 , except for GPT- 3.5 where 0.6 was better. For tuning, we create an independent English-only version of our dataset which uses English-only entities and a new set of templates to ensure that the test set evaluations are truly held out.

Model Inputs As input to the model, we include the prompt, context, question, and the three possible answer choices. We then pick the most likely answer choice decoded by the model as the predicted answer. We rely on a single 40GB Nvidia V100 GPU for all of our experiments.

## C Human Evaluation

## C. 1 Experimental Setup

We specifically consider 5 types of samples, each with varying levels of conflict. The specific examples (translated to what we are calling the oracle translation) used in the study are shown in Table 5. Type 1 contains single context sentences with an entity that never conflicts, Type 2 contains single context sentences with entities that could conflict, Type 3 contains double context sentences whose entities do not conflict, Type 4 contains double context sentences whose entities do conflict, and Type 5 is the same as Type 4, but includes a question about one of the entities. The specific orderings of example types which participants were shown was
$1,2,3,4,5$ (i.e., single context sentences before double context sentences) for Group 1, and 1,5,4,3,2 (i.e. conflicting double context sentences with a specific question before non-conflicting sentences) for Group 2.

## C. 2 Participants

Participants were bilingual speakers of Farsi and English close to the authors of the paper. They were not paid for their translations, as it only required a short period of participation (10-15 minutes).

## D Additional Experiments with Translation

## D. 1 Finetuning MT Models

Here we are interested in learning if MT models can learn to correctly translate lexical gaps through finetuning. We partition the data by selecting a subset of kinship terms and context templates for Farsi, Hindi, and Vietnamese, and hold out examples which contain them as an evaluation set. Focusing only on the context pairs, we use the remaining examples in the target language and English version to create parallel data for finetuning. We also consider single context sentences, which instead of a context pair (which is a double context), contains only one sentence. In these cases, we use the general translation since there are no entities we need to disambiguate. To measure performance, we count the number of examples which contain correctly translated entities.

For finetuning NLLB, we use a batch size of 32 , warmup ratio of 0.1 , and a learning rate of $2 \mathrm{e}-$ 6 , which we tune by hand as rates used by prior works were too large for our data. We train independent models for $25,50,100,250,500,1000$, 2500,5000 , and 10000 steps, and measure performance by counting the number of examples which contain correctly translated entities. We present results in Figure 1 and Table 6.

We see that NLLB is able to translate the seen lexical gaps easily, however, performance for unseen entities only shows improvements in the single context case. Double context performance either never improves, or degrades quickly. This indicates that the model does not learn, in general, to use the specific translation in the double context setting.

## D. 2 Using LLMs for Translation

As LLMs also have the ability to produce translations, and due to their strong performance in other
tasks, they may offer a stronger alternative to standalone MT models for our task. To experiment wit this, we consider a 1 -shot setting where model is asked to translate a double context sentence, and provide a different double context translation as the example.

For example, if we want to get translations for the Farsi input "My paternal uncle's name is Ali. My maternal uncle's name is Sena", we may use "My maternal aunt's name is Mojghan. My paternal aunt's name is Sheyda." as the example in the prompt, in order to show the model that it should use the specific form of each lexical gap in the translation.

We find that GPT-3.5 is unable to handle the lexical gaps correctly, only translating both entities to their specific form correctly $12 \%$ of the time. However, the results are promising; it correctly translates at least one gap correctly $40 \%$ of the time, and gets the general form correct in $79 \%$ of examples. Notably, in a manual review, the model does attempt to make translations specific, however these are generally incorrect and directly copied from the reference example in the prompt. The model also uses relative terms to differentiate between conflicting entities.

## E Tables and Figures

| English | Catalan | German | Farsi (transl.) | Hindi (transl.) | Vietnamese |
| :---: | :---: | :---: | :---: | :---: | :---: |
| Uncle | oncle | Onkel | amoo (paternal uncle) <br> daei (maternal uncle) | taoo (older paternal uncle) chacha (younger paternal uncle) mama | Chú <br> Câu |
| Aunt | tia | Tante | khaleh (maternal aunt) ameh (paternal aunt) | mowsee phuphu | $\begin{aligned} & \text { Dì } \\ & \text { Cô } \end{aligned}$ |
| Grandmother | àvia | Großmutter | madar bozorg | nani (maternal grandmother) <br> dadi (paternal grandmother) | Bà ngoại <br> Bà nội |
| Grandfather | avi | Großvater | pedar bozorg | nana (maternal grandfather) dada (paternal grandfather) | Ông ngoại Ông nội |
| Mother <br> Father | mare <br> pare | Mutter Vater | madar pedar | maa <br> pita | $\begin{aligned} & \text { Mẹ } \\ & \text { Bố } \end{aligned}$ |
| Sister <br> Brother | germana <br> germà | Schwester <br> Bruder | khahar <br> baradar | choti bahan (younger sister) badi bahan (older sister) chota bhai (younger brother) bada bhai (older brother) | Em gái <br> Chị gái <br> Em trai <br> Anh trai |
| Son <br> Daughter | fill <br> filla | Sohn <br> Tochter | pesar dokthar | beta <br> beti | Con trai Con gái |
| Cousin | cosí (female cousin) <br> cosina (male cousin) | Cousin Cousine | dokhtar khaleh (maternal aunt's daughter) dokhtar ameh (paternal aunt's daughter) dokhtar daei (maternal uncle's daughter) dokhtar amoo (paternal uncle's daughter) pesar khaleh (maternal aunt's son) pesar ameh (paternal aunt's son) pesar daei (maternal uncle's son) pesar amoo (paternal uncle's son) | mowseri bahan phupheri bahan mameri bahan chacheri bahan mowsera bhai phuphera bhai mamera bhai chachera bhai |  |

Table 3: The kinship terms we use, aligned horizontally for each language. Here, the gaps between languages can be seen as empty lines in a given row. Farsi and Hindi terms are transliterated to Latin script.

| Model | Prompt Type | Prompt |
| :---: | :---: | :---: |
| GPT-3.5 | System Prompt Input Prompt <br> Answer Prompt | Answer the question using the given context. Respond with the letter corresponding to the correct answer choice. If the answer is unclear, reply wih the option for 'not answerable'. Here is the example:\n Answer: |
| Llama 7 | System Prompt Input Prompt <br> Answer Prompt | Answer the question using the given context. Respond with the letter corresponding to the correct answer choice. I'm trying to answer a question in \{language\} using only the information found in the context below. If the answer is unclear, reply with the option for 'not answerable'. Here is the information and question: $\backslash n$ Answer: |
| Llama-13 | System Prompt <br> Input Prompt <br> Answer Prompt | Respond with the letter corresponding to the correct answer choice. <br> Answer the question written in \{language\} below using the information in the context. If the answer is unclear, reply with the option for 'not answerable'. Here is the information and question: $\backslash n$ <br> Answer: |
| BLOOM-Z | System Prompt Input Prompt <br> Answer Prompt | N/A <br> If the answer is unclear, reply wih the option for 'not answerable'. Here is the example:\n <br> Answer: |

Table 4: Prompts selected for each LLM using the independent English development set.

| Type | Examples |
| :--- | :--- |
| Type 1 | My brother's favorite food is ash-e reshteh. <br> My grandmother's shirt is blue. <br> My sister likes watching soccer. |
| Type 2 | My paternal uncle's shirt is blue. <br> My maternal aunt has a pet cat. <br> My paternal aunt's daughter was born in 2004. |
| Type 3 | My maternal aunt's shirt is white. My sister was born in 1998. <br> My paternal uncle's son is out for a run. My maternal aunt is cooking khoresht tonight. <br> My grandmother is sleeping in the living room. My mom is out for a run. |
| Type 4 | My paternal uncle's shirt is blue. My maternal uncle's shirt is red. <br> My paternal aunt's son is playing soccer. My paternal uncle's daughter is watching TV. <br> My maternal aunt's favorite color is purple. My paternal aunt was born in 1983. |
| Type 5 | My paternal uncle's shirt is blue. My maternal uncle's shirt is red. What color is my paternal uncle's shirt? <br> My paternal aunt was born in 1983. My maternal aunt's favorite color is purple. What is my paternal aunt's favorite color? <br> My maternal aunt's daughter is out for a run. My paternal uncle's son is cooking zereshk polo and morgh. Where is my maternal aunt's daughter? |

Table 5: English translations of the Farsi examples shown to human annotators for the case study.


Figure 1: Main result figure for finetuning NLLB. The x -axis marks the total number of steps the model is finetuned for, on a log scale. The $y$-axis marks the accuracy of translations, calculated by totaling the number of pairs which contain correctly translated entities.

| Lang | Seen | Type | 0 | 25 | 50 | 100 | 250 | 500 | 1000 | 2500 | 5000 | 10000 |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
| FA | Unseen | U-All-1 | 51.59\% | 51.59\% | 51.59\% | 58.73\% | 56.35\% | 58.73\% | 96.83\% | 81.75\% | 73.02\% | 71.43\% |
|  |  | U-All-2 | 8.18\% | 10.60\% | 12.48\% | 15.16\% | 1.70\% | 0.00\% | 0.00\% | 0.00\% | 0.00\% | 0.00\% |
|  |  | U-Entity-1 | 48.00\% | 51.00\% | 52.00\% | 53.00\% | 50.00\% | 53.50\% | 91.00\% | 74.00\% | 70.00\% | 70.00\% |
|  |  | U-Entity-2 | 7.37\% | 11.26\% | 14.09\% | 18.79\% | 7.98\% | 0.81\% | 0.00\% | 0.00\% | 0.00\% | 0.00\% |
|  |  | U-Template-1 | 79.88\% | 79.88\% | 81.11\% | 81.11\% | 88.54\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% |
|  |  | U-Template-2 | 55.09\% | 57.37\% | 58.52\% | 59.50\% | 74.08\% | 80.78\% | 77.68\% | 80.47\% | 84.22\% | 93.11\% |
|  | Seen | Single Context | 76.54\% | 80.77\% | 80.77\% | 80.77\% | 93.27\% | 94.23\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% |
|  |  | Double Context | 56.41\% | 59.14\% | 59.73\% | 59.77\% | 90.15\% | 97.63\% | 99.87\% | 100.00\% | 100.00\% | 100.00\% |
| HI | Unseen | U-All-1 | 61.29\% | 61.29\% | 62.90\% | 61.29\% | 77.42\% | 75.00\% | 51.61\% | 48.39\% | 48.39\% | 48.39\% |
|  |  | U-All-2 | 14.65\% | 14.37\% | 14.22\% | 14.27\% | 17.72\% | 5.91\% | 1.47\% | 7.23\% | 17.25\% | 18.62\% |
|  |  | U-Entity-1 | 60.00\% | 60.00\% | 60.00\% | 60.00\% | 62.00\% | 44.00\% | 50.00\% | 50.00\% | 50.00\% | 50.00\% |
|  |  | U-Entity-2 | 13.33\% | 13.33\% | 13.33\% | 13.33\% | 11.33\% | 0.52\% | 3.70\% | 17.33\% | 20.00\% | 20.00\% |
|  |  | U-Template-1 | 63.41\% | 64.30\% | 68.29\% | 78.49\% | 84.26\% | 95.12\% | 98.67\% | 100.00\% | 100.00\% | 100.00\% |
|  |  | U-Template-2 | 37.46\% | 40.29\% | 43.49\% | 52.56\% | 56.67\% | 79.95\% | 86.33\% | 93.54\% | 96.39\% | 98.19\% |
|  | Seen | Single Context | 62.96\% | 62.96\% | 65.93\% | 74.32\% | 87.16\% | 98.27\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% |
|  |  | Double Context | 37.68\% | 40.45\% | 44.37\% | 53.87\% | 63.24\% | 96.21\% | 99.56\% | 100.00\% | 100.00\% | 100.00\% |
| VI | Unseen | U-All-1 | 91.88\% | 96.25\% | 97.50\% | 100.00\% | 100.00\% | 98.75\% | 88.12\% | 54.37\% | 53.75\% | 53.75\% |
|  |  | U-All-2 | 47.87\% | 64.40\% | 75.35\% | 79.08\% | 70.20\% | 45.55\% | 25.70\% | 8.84\% | 9.20\% | 9.20\% |
|  |  | U-Entity-1 | 82.27\% | 85.00\% | 88.64\% | 95.45\% | 100.00\% | 99.09\% | 100.00\% | 47.27\% | 45.45\% | 45.45\% |
|  |  | U-Entity-2 | 51.56\% | 63.78\% | 68.59\% | 73.04\% | 74.15\% | 56.32\% | 33.90\% | 5.77\% | 5.65\% | 5.65\% |
|  |  | U-Template-1 | 81.94\% | 87.50\% | 90.97\% | 97.22\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% |
|  |  | U-Template-2 | 49.84\% | 65.88\% | 75.20\% | 79.94\% | 92.97\% | 95.18\% | 95.75\% | 99.06\% | 98.96\% | 98.75\% |
|  | Seen | Single Context | 83.29\% | 91.90\% | 95.19\% | 98.73\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% | 100.00\% |
|  |  | Double Context | 61.17\% | 75.74\% | 79.52\% | 81.27\% | 93.36\% | 96.39\% | 99.54\% | 100.00\% | 100.00\% | 100.00\% |

Table 6: Main results for the finetuning results of NLLB.

| Model <br> Lang. |  | GPT-3.5 |  |  |  | Llama-7 |  |  |  | Llama-13 |  |  |  | BLOOM-Z |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | 0-A | 0-U | 1 | 2 | 0-A | 0-U | 1 | 2 | 0-A | 0-U | 1 | 2 | 0 | 0-U | 1 | 2 |
| CA | CA | 1.16 | 0.54 | 1.25 | 0.95 | 1.79 | 1.15 | 2.51 | 0.61 | 2.56 | 1.84 | 1.97 | 1.58 | 0.17 | 1.78 | 0.17 | 1.63 |
|  | EN | 0.22 | 0.08 | 0.17 | 0.35 | 0.76 | 0.72 | 1.95 | 0.91 | 0.49 | 0.63 | 0.46 | 1.68 | 0.09 | 1.15 | 0.35 | 1.38 |
|  | CA-N | 1.20 | 0.14 | 2.10 | 1.19 | 1.21 | 1.69 | 2.42 | 1.46 | 0.97 | 0.61 | 2.27 | 1.52 | 0.47 | 1.60 | 0.30 | 1.46 |
|  | CA-G | 0.71 | 0.41 | 0.79 | 0.44 | 1.61 | 0.71 | 3.14 | 0.61 | 1.29 | 0.17 | 0.30 | 0.36 | 1.15 | 0.73 | 1.54 | 0.72 |
| DE | DE | 0.89 | 0.29 | 0.17 | 0.51 | 0.66 | 1.29 | 2.18 | 0.06 | 1.12 | 0.81 | 0.52 | 0.36 | 0.75 | 2.06 | 2.05 | 0.56 |
|  | EN | 0.35 | 0.50 | 0.17 | 0.30 | 0.72 | 0.32 | 1.88 | 0.76 | 1.11 | 0.81 | 1.14 | 0.06 | 0.10 | 0.62 | 0.35 | 0.59 |
|  | DE-N | 0.83 | 0.27 | 3.09 | 1.21 | 0.97 | 0.28 | 4.01 | 0.61 | 1.00 | 1.57 | 0.46 | 0.15 | 0.78 | 1.94 | 0.52 | 0.12 |
|  | DE-G | 0.79 | 0.22 | 1.09 | 0.66 | 1.39 | 0.74 | 1.55 | 1.18 | 0.96 | 1.06 | 2.01 | 0.64 | 0.51 | 1.25 | 2.97 | 0.21 |
| FA | FA | 19 | 4.22 | 0.55 | 2.79 | 2.52 | 1.20 | 0.25 | 0.20 | 1.29 | 0.25 | 0.61 | 0.06 | 2.51 | 1.96 | 1.42 | 1.72 |
|  | EN | 0.39 | 0.31 | 0.52 | 1.06 | 2.69 | 0.42 | 0.53 | 1.70 | 1.73 | 2.21 | 1.33 | 1.15 | 0.27 | 1.74 | 0.32 | 0.36 |
|  | FA-N | 1.36 | 1.67 | 1.15 | 1.62 | 0.97 | 1.77 | 0.60 | 0.26 | 1.55 | 2.61 | 1.23 | 0.71 | 0.67 | 0.91 | 1.59 | 0.35 |
|  | FA-G | 7.07 | 2.77 | 1.61 | 1.59 | 9.28 | 0.89 | 1.38 | 2.23 | 3.78 | 4.52 | 0.81 | 0.81 | 5.75 | 4.91 | 1.79 | 0.95 |
| HI | HI | 1.21 | 1.62 | 0.44 | 0.64 | 3.09 | 1.50 | 0.93 | 1.45 | 2.09 | 0.38 | 1.85 | 0.12 | 0.38 | 1.41 | 1.15 | 0.95 |
|  | EN | 0.42 | 0.84 | 0.23 | 0.57 | 3.38 | 2.13 | 0.64 | 1.30 | 1.11 | 2.49 | 0.32 | 0.68 | 0.17 | 1.53 | 0.50 | 1.33 |
|  | HI-N | 2.53 | 0.64 | 0.53 | 1.58 | 2.85 | 2.59 | 0.67 | 0.21 | 1.58 | 3.26 | 0.72 | 2.14 | 0.58 | 4.83 | 0.93 | 2.14 |
|  | HI-G | 2.61 | 0.57 | 1.33 | 1.37 | 4.29 | 2.55 | 1.51 | 0.62 | 1.82 | 2.25 | 2.04 | 1.55 | 1.11 | 3.23 | 1.10 | 1.10 |
| VI | vi | 20.81 | 4.15 | 0.15 | 0.70 | 1.05 | 2.18 | 0.91 | 2.17 | 0.90 | 1.80 | 2.96 | 0.23 | 0.48 | 0.93 | 0.75 | 2.18 |
|  | EN | 0.54 | 1.00 | 0.17 | 0.92 | 2.09 | 1.17 | 0.47 | 0.85 | 1.19 | 2.19 | 0.30 | 0.68 | 0.47 | 2.65 | 0.21 | 2.20 |
|  | VI-N | 2.33 | 1.57 | 1.25 | 3.30 | 1.74 | 1.41 | 2.00 | 1.08 | 1.18 | 2.01 | 0.80 | 2.10 | 0.27 | 2.70 | 1.63 | 0.90 |
|  | VI-G | 0.45 | 0.38 | 0.71 | 1.80 | 0.84 | 1.86 | 2.00 | 0.60 | 1.17 | 2.76 | 1.59 | 0.75 | 0.75 | 0.40 | 0.76 | 0.89 |

Table 7: Standard deviations for the main results.


[^0]:    ${ }^{1}$ https://chat.openai.com/

[^1]:    ${ }^{2}$ https://chat.openai.com/
    ${ }^{3}$ https://github.com/facebookresearch/llama

