Enhancing Multilingual Causal Commonsense Reasoning in LLMs: A Novel Assessment Approach and Strategy

Anonymous ACL submission

Abstract

Commonsense reasoning is crucial for connecting premises to hypotheses by leveraging implicit world knowledge. The XCopa dataset, spanning 11 languages, serves as a benchmark for evaluating cross-lingual transfer capabilities in commonsense reasoning and emphasizes the importance of tapping into implicit knowledge for effective communication in diverse linguistic contexts. Recent advancements in Large Language Models (LLMs), such as Llama2, have made remarkable progress in Causal Commonsense Reasoning, setting new benchmarks. However, multilingual LLMs like XGLM and PolyLM face challenges due to smaller training datasets compared to English-centric LLMs. This work introduces a novel evaluation strategy, G-Evaluation, in the XCopa dataset. While this strategy resulted in decreased accuracy metrics across models, Llama2 showed improved performance, highlighting its adaptability. Despite efforts, multilingual XCopa models still fall behind their English counterparts in accuracy. Models like Llama2 exhibit performance variations across languages, underscoring the need for bridging this gap with Machine Translation (MT). To address this, we propose XTools, a strategy that combines Machine Translation and Automatic Post-Editing tools. By implementing XTools, multilingual accuracy can be elevated to 89.60%, aligning with English performance. Our contributions include redefining the evaluation method with G-Evaluation, introducing XTools for enhancing multilingual capabilities, validating Automatic Post-Editing Tool integration, and showcasing the potential of lightweight models in improving overall performance.

1 Introduction

In the realm of natural language understanding, commonsense reasoning (Davis and Marcus, 2015) plays a crucial role in connecting premises with hypotheses by leveraging implicit world knowledge such as causality, social norms, and emotions. The XCopa (Roemmele et al., 2011; Edoardo M. Ponti and Korhonen, 2020) dataset, spanning 11 languages, serves as a fundamental benchmark for assessing the cross-lingual transfer capabilities of machine learning models in commonsense reasoning, particularly emphasizing the significance of tapping into implicit knowledge for effective communication across diverse linguistic contexts.

The advancement in Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023; Touvron et al., 2023) has marked remarkable progress recently. Some research defines LLM as an Agent (Weng, 2023). While open-source LLMs like Llama2 (Touvron et al., 2023) are primarily trained in English, there is a growing focus on multilingual LLMs such as XGLM (Lin et al., 2022) and PolyLM (Wei et al., 2023). Nevertheless, these multilingual models face challenges due to their relatively smaller training datasets compared to English-centric LLMs. Notably, LLMs have demonstrated exceptional performance in Causal Commonsense Reasoning, continually setting new state-of-the-art benchmarks.

Evaluation of XCopa conventionally involves calculating the perplexity between two given choices based on a premise and a question, with lower perplexity implying the correct answer (Gao et al., 2023). This approach, termed P-Evaluation, is used to compare the predicted answer to the ground truth for accuracy determination. In this work, we introduce a novel evaluation strategy, G-Evaluation, which incorporates random sampling, offering greater flexibility and scalability.

Our experiments with the new evaluation strategy reveal a notable decrease in accuracy metrics across various models on XCopa, highlighting the heightened challenge introduced by our proposed evaluation method. Interestingly, we observed an improvement in the performance of Llama2 under the new evaluation strategy, shedding light on its robustness and adaptability.

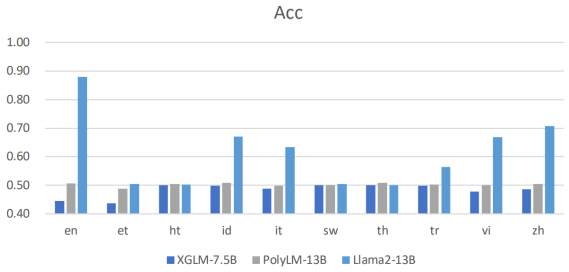

Figure 1: Performance of Multiple LLMs on Multilingual XCopA Dataset

Despite considerable research efforts, the accuracy of multilingual XCopa models still trails behind their English counterparts. As shown in Figure 1, models like Llama2 show significant variation in performance across different languages, with English achieving an 88% accuracy compared to around 50% in non-English test sets. On the other hand, models like XGLM and PolyLM exhibit more balanced performance but hover around 40% to 50% accuracy. Building on prior studies(Edoardo M. Ponti and Korhonen, 2020; Schick et al., 2023), we advocate for leveraging Machine Translation (MT) as a bridge to extend the capabilities of LLMs from English to multiple languages. This paper introduces an enhanced strategy, XTools, which combines Machine Translation (MT) Tool and Automatic Post-Editing (APE) (Raunak et al., 2023; Liang et al., 2023; Koneru et al., 2023) Tool, demonstrating the potential to elevate multilingual accuracy to 89.6%, on par with English performance.

Our main contributions are:

- Redefinition of a natural evaluation method, G-Evaluation, for assessing LLM's Causal Commonsense Reasoning Ability.
- Introduction of XTools as an effective strategy to enhance LLM's multilingual capabilities.
- Validation of incorporating Automatic Post-Editing Tool within LLM.
- Demonstration of the potential of lightweight solutions like the 7B model in enhancing overall performance.

2 Proposed Assessment Approach: G-Evaluation

Figure 2 displays an example extracted from the XCopa dataset, each comprising a premise, a question, two choices, and their corresponding label. The question field delineates the causal relationship between the premise and the choices, classifying them as "cause" and "effect." The label assigned to the correct choice accentuates the dataset's pivotal role in refining commonsense reasoning abilities across diverse linguistic contexts.

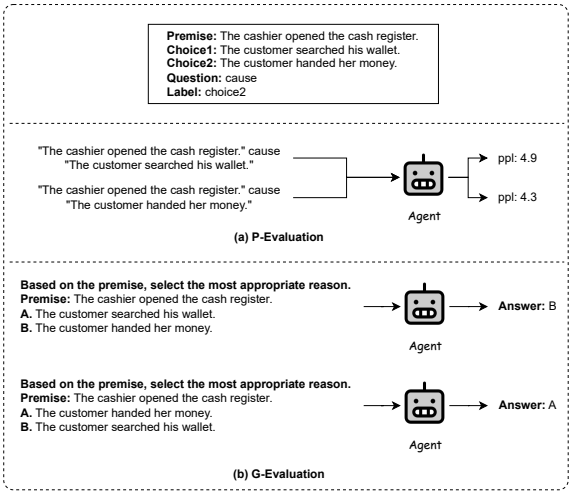

Figure 2: Assessment Approach. (a) P-Evaluation. (b) Proposed G-Evaluation

The traditional evaluation method, P-Evaluation, involves calculating the perplexity between choice1 and choice2 based on the premise and question, where a lower perplexity value indicates the correct answer. Subsequently, this answer is cross-checked with the ground truth to ascertain accuracy, as illustrated in Figure 2(a).

In contrast to the conventional P-Evaluation approach, our novel G-Evaluation method revolutionizes the process by transmuting the question into a natural multiple-choice framework. By designating two options, A and B, to choice1 and choice2 and tasking the Language Model (LLM) to generate the answer, we ensure that the LLM forcefully decodes and outputs the option with the higher probability from A and B. Noteworthy is our practice of randomly assigning either choice to A or B during assessments, introducing a slight variance in accuracy while consistently converging results around a specific value. We believe that such an evaluation method is not only more natural but also provides a more accurate assessment of the LLM's capabilities, as illustrated in Figure 2(b).

3 Proposed Enhanced Strategy: XTools

XTools offers two tools, namely Machine Translation (MT) Tool and Automatic Post-Editing (APE) Tool, to assist Agents in translating multilingual

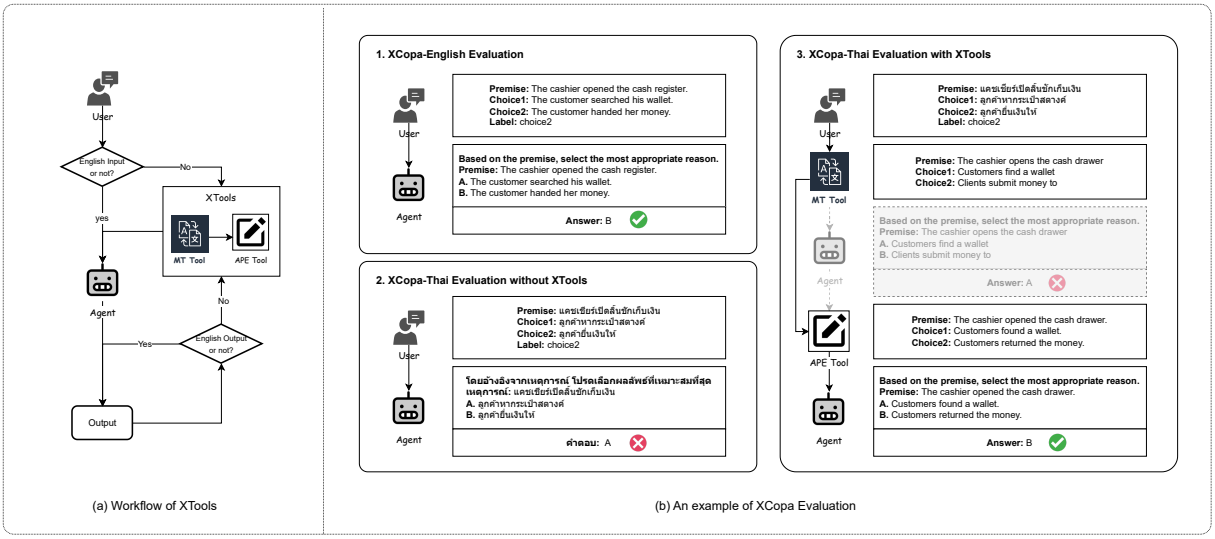

Figure 3: XTools. (a) The Workflow of XTools. (b) An example of dataflow and results on XCopa dataset.

texts into English. The workflow is depicted in Figure 3(a). Furthermore, Figure 3(b) provides an example using the XCopa Thai dataset, illustrating the data flow and results.

3.1 MT Tool: Google Translate

Google Translate is a free online translation service offered by Google that facilitates instant translation across multiple languages and is renowned as one of the best machine translation engines available. In 2017, Google introduced the Transformer model, which was integrated into Google Translate. For machine translation tasks, the Transformer (Vaswani et al., 2017) model comprises an Encoder and a Decoder. The encoder processes the source language sentence into a fixed-length representation, while the decoder generates the target language sentence token by token. These techniques heavily depend on large bilingual parallel corpora to align source sentences with their corresponding translations, with the general belief that translation quality improves as datasets and model sizes increase. **In this study, we specifically focus on Google Translate as our primary machine translation tool.**

3.2 APE Tool: Llama2-ICL-APE

In-Context Learning (ICL) (Min et al., 2022; Dong et al., 2023) is a machine learning approach that emphasizes learning in specific contexts by utilizing relationships and contextual clues across various data formats like text, speech, images, and videos. It focuses on learning through analogies by connecting query questions with relevant instances in a demonstration context presented using natural language templates. Unlike traditional supervised learning, ICL does not require parameter updates through backpropagation, relying instead on a pre-trained language model for predictions without downstream fine-tuning. The goal is for the model to identify patterns in the demonstrations and make accurate predictions based on acquired knowledge.

Llama2-ICL-APE combines Automatic Post-Editing (APE) functionality with the Llama2 model, incorporating ICL technology. Research by (Koneru et al., 2023) indicates that direct fine-tuning of Large Language Models (LLMs) for machine translation purposes could result in performance degradation. However, repurposing LLMs for automatic post-editing can yield more favorable outcomes. Llama2-ICL-APE leverages the robust language model of Llama2, known for its proficiency in handling long sequences, in conjunction with ICL technology to dynamically capture contextual information. This integration enhances the quality of machine translation outputs.

4 Experiment

4.1 Dataset: XCopa

The Cross-lingual Choice of Plausible Alternatives (XCopa) dataset serves as a benchmark for assessing machine learning models' capability to transfer commonsense reasoning skills across different languages. This dataset is a translation and reannotation of the English COPA(Roemmele et al., 2011) and spans 11 languages representing diverse lan-

Model	en	et	ht	id	it	sw	th	tr	vi	zh	Avg
P-Evaluation											
XGLM-7.5B	0.654	0.576	0.510	0.598	0.550	0.502	0.508	0.510	0.576	0.566	0.555
PolyLM-13B	0.692	0.512	0.482	0.618	0.600	0.504	0.498	0.516	0.596	0.622	0.564
Llama2-13B	0.718	0.514	0.492	0.606	0.598	0.490	0.516	0.506	0.560	0.596	0.560
G-Evaluation											
XGLM-7.5B	0.446	0.438	0.500	0.498	0.488	0.500	0.500	0.498	0.478	0.486	0.483
PolyLM-13B	0.506	0.488	0.504	0.508	0.498	0.500	0.508	0.502	0.500	0.504	0.502
Llama2-13B	0.880	0.504	0.502	0.670	0.634	0.504	0.500	0.564	0.668	0.708	0.613

Table 1: Accuracy of Multiple LLMs on Multilingual XCopA Dataset under P-Evaluation and G-Evaluation Strategies.

Model	en	et	ht	id	it	sw	th	tr	vi	zh	Avg(/o en)	Per%
Llama2-13B	0.880	0.504	0.502	0.670	0.634	0.504	0.500	0.564	0.668	0.708	0.584	66.34
Xtools		0.808	0.732	0.816	0.832	0.752	0.730	0.778	0.810	0.838	0.788	89.60
- APE Tool (7B)		0.786	0.718	0.790	0.832	0.722	0.690	0.760	0.780	0.816	0.766	87.05
XTools*		0.802	0.732	0.814	0.832	0.726	0.716	0.788	0.804	0.834	0.783	88.99

Table 2: Accuracy on Multilingual XCopA Dataset using XTools under G-Evaluation Strategy. Note: XTools* represents APE Tool Using Llama2 13B Model as Foundation.

guage families and regions worldwide. The challenge lies in the dataset's requirement for proficiency in both worldly knowledge and the aptitude to generalize across new languages. Additionally, the dataset provides "translate test" data generated via Google Translate for direct use.

4.2 Evaluation Metrics

For the evaluation of Xcopa, we utilize accuracy (ACC) as the evaluation metric. For MT & APE evaluation, we utilize SacreBLEU, which implements BLEU(Papineni et al., 2002), and COMET(Rei et al., 2020) from Unbabel/wmt22-comet-da. SacreBLEU calculates similarity based on n-gram matching, while COMET leverages cross-lingual pretrained models for evaluation.

4.3 Results and Analysis

Evaluation Results under different Approach As shown in Table 1, under the P-Evaluation assessment approach, XGLM and PolyLM demonstrated comparable abilities to Llama2, with PolyLM even achieving slightly higher average accuracy than Llama2. However, upon transitioning to the G-Evaluation evaluation strategy, the accuracy of XGLM and PolyLM sharply declined. In contrast, Llama2 maintained a stable level of accuracy, with the average accuracy increasing from 0.560 to 0.613. Particularly notable was the improvement in performance in English, rising from 0.718 to 0.880. These results indicate the high robustness of Llama2.

Seed	en	et	th
1	0.868	0.786	0.690
2	0.864	0.78	0.692
3	0.866	0.786	0.684
4	0.872	0.778	0.684
5	0.866	0.786	0.692
6	0.868	0.776	0.694
7	0.858	0.776	0.68
8	0.862	0.778	0.694
9	0.87	0.79	0.678
10	0.87	0.778	0.682

Table 3: Accuracy under G-Evaluation Strategy using XTools with Different Random Seeds.

Results of XTools under G-Evaluation As shown in Table 3, for the G-Evaluation, we employed various random seeds to generate the data. It can be observed that while the accuracy may vary slightly, it tends to stabilize around a certain value.

As depicted in Table 2, employing the XTools strategy resulted in a significant improvement in Llama2's average accuracy on non-English tests, rising from 0.584 to 0.788, reaching an impressive 89.60% accuracy level on the English test set. However, when the APE Tool was omitted, the average accuracy decreased by two percentage points to 0.766, emphasizing the necessity and effectiveness of the APE Tool.

In the aforementioned experiments, we utilized the Llama2 7B model as the APE Tool. Additionally, we also assessed the results of using the

Model	en	et	ht	id	it	sw	th	tr	vi	zh
Llama2-13B	0.880	0.504	0.502	0.670	0.634	0.504	0.500	0.564	0.668	0.708
Google Translate Tool		0.786	0.718	0.790	0.832	0.722	0.690	0.760	0.780	0.816
LLM-based MT		0.616	0.578	0.808	0.836	0.546	0.612	0.702	0.800	0.808
+ APE Tool		0.616	0.582	0.808	0.836	0.538	0.610	0.704	0.800	0.808

Table 4: Accuracy Comparison between Llama2-ICL-MT and Google Translate

Lang	Metic	MT	+APE(7B)	+APE(13B)
et	COMET	0.883	0.892	0.892
	BLEU	52.3851	55.3852	55.0876
ht	COMET	0.799	0.809	0.819
	BLEU	41.4114	47.6386	47.3801
it	COMET	0.899	0.909	0.907
	BLEU	52.9644	55.7427	55.7090
th	COMET	0.829	0.860	0.866
	BLEU	22.0342	37.1617	38.9785
zh	COMET	0.913	0.923	0.922
	BLEU	51.3654	57.4918	57.8066

Table 5: MT Metrics for Translations at Different Stages

Llama2 13B model as the APE Tool, denoted as XTools* in Table 2. Surprisingly, despite employing a larger APE model, the accuracy remained consistent. This suggests that the APE Tool itself is lightweight, and the 7B model size is sufficient.

Furthermore, we evaluated the translation quality after applying the APE Tool following machine translation methodology. As shown in Table 5, we observed enhancements in the COMET scores post APE, validating the effectiveness of the APE Tool.

5 Ablation Study: How about using LLM-based MT as Tool rather than traditional MT like Google Translate?

In recent years, Language Model (LM)-based approaches have gained attention in the field of machine translation (Jiao et al., 2023; Zeng et al., 2023; Chen et al., 2023; Xu et al., 2023; Yang et al., 2023; Zhang et al., 2023). One line of LLM-based methods focuses on zero-shot or few-shot translation by incorporating in-context learning(Hendy et al., 2023). By conditioning the LLM on a source sentence, the model can generate translations in the target language without explicitly using parallel data. This approach has shown promising results in enabling translation for language pairs with limited or no parallel resources. Another approach involves using a small amount of high-quality bilingual parallel data to construct translation-guiding instructions. These instructions explicitly define the translation behavior by providing source-language consistent cues during the supervised fine-tuning (SFT) process. By utilizing these specially crafted instructions, the LM can be fine-tuned to perform translation more accurately and robustly.

When considering the use of LLM-based Machine Translation (MT) tools such as Llama2-ICL-MT compared to traditional ones like Google Translate, we conducted experiments employing the Llama2 and ICL technologies to construct translation requests in a five-shot manner, defining the approach as Llama2-ICL-MT. The results, as shown in Table 4, reveal that on high-resource languages like Italian and Chinese, the accuracy of Llama2-ICL-MT is comparable to Google Translate, with some instances even showing higher accuracy. However, a significant disparity is noted on low-resource languages.

Furthermore, we observed that following the usage of Llama2-ICL-MT, additional optimization using APE did not further enhance the translation quality. This finding underscores the effectiveness of LLM-based MT tools in certain scenarios.

6 Related Work: Large Language Models

Foundation Model Foundation Model, a product of pre-training, is a prominent type of Large Language Model. It has gained substantial recognition in recent years for its impressive capabilities in natural language processing tasks. The most prevalent architectural framework for such models is the Transformer, which employs a series of self-attention mechanisms to process input text efficiently.

Among the state-of-the-art Large Language Models, notable examples include GPT-3(Brown et al., 2020) and Llama2(Touvron et al., 2023). These models have been widely lauded for their exceptional proficiency in understanding and generating natural language text. They showcase the remarkable potential of Foundation Models, pushing the boundaries of language processing and setting new benchmarks in various applications.

Instruct/Chat Model Instruct/Chat Model, a variant of Large Language Models, is specifically developed through the process of Supervised Fine-Tuning (SFT). Unlike Foundation Models, which are pre-trained, Instruct/Chat Models undergo additional supervised training to enhance their performance in specific tasks such as instruction following or conversational dialogue.

Supervised Fine-Tuning involves training the model on labeled datasets, where human annotators provide examples of desired input-output behavior. This approach enables Instruct/Chat Models to learn task-specific skills and exhibit improved performance in situations that require language understanding, generation, and interaction.

7 Conclusion

In conclusion, the significance of commonsense reasoning in multilingual contexts cannot be understated. The XCopa dataset has shed light on the challenges and opportunities presented by cross-lingual transfer learning in the realm of implicit knowledge utilization. While recent advancements in Large Language Models have propelled the field forward, there remains a notable discrepancy in performance between English-centric models and their multilingual counterparts.

Our study introduced the G-Evaluation strategy to assess the performance of multilingual models on the XCopa dataset, revealing both strengths and areas for improvement. The versatility demonstrated by models like Llama2 underscores the potential for adaptation and enhancement across diverse linguistic landscapes. However, the need for bridging the performance gap through strategies like Machine Translation and Automatic Post-Editing tools is evident.

The proposed XTools strategy stands out as a promising approach to elevate multilingual model accuracy, showcasing the feasibility of reaching parity with English models. By leveraging lightweight models in conjunction with efficient tools, our research paves the way for improved cross-lingual commonsense reasoning capabilities.

Looking ahead, continued efforts to refine evaluation methods, optimize model training datasets, and integrate innovative approaches like XTools will be instrumental in advancing the field of multilingual commonsense reasoning. As we strive towards more effective communication and understanding across languages, the journey towards enhancing multilingual model performance remains an exciting and evolving frontier in natural language processing research.

8 Limitations

Our study is limited by a focus on the XCopa dataset for evaluating multilingual commonsense reasoning models, potentially overlooking performance variations in other benchmarks. The effectiveness of our XTools strategy may be hindered by the quality of Machine Translation and Automatic Post-Editing tools. Evaluation metrics may not fully capture multilingual model capabilities, and the rapid pace of NLP advancements could risk our findings becoming outdated. Additionally, computational resource requirements may restrict the scalability of our proposed strategies for researchers with limited resources.

References

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

Yijie Chen, Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2023. Improving translation faithfulness of large language models via augmenting instructions.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy

Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113.

Ernest Davis and Gary Marcus. 2015. Commonsense reasoning and commonsense knowledge in artificial intelligence. *Commun. ACM*, 58(9):92–103.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

Olga Majewska Qianchu Liu Ivan Vulić Edoardo M. Ponti, Goran Glavaš and Anna Korhonen. 2020. XCOPA: A multilingual dataset for causal commonsense reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2023. A framework for few-shot language model evaluation.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are GPT models at machine translation? A comprehensive evaluation. *CoRR*, abs/2302.09210.

Wenxiang Jiao, Jen tse Huang, Wenxuan Wang, Zhiwei He, Tian Liang, Xing Wang, Shuming Shi, and Zhaopeng Tu. 2023. Parrot: Translating during chat using large language models tuned with human translation and feedback.

Sai Koneru, Miriam Exel, Matthias Huck, and Jan Niehues. 2023. Contextual refinement of translations: Large language models for sentence and document-level post-editing.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate.

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 T. 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, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9019–9052. Association for Computational Linguistics.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 11048–11064. Association for Computational Linguistics.

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, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL.

Vikas Raunak, Amr Sharaf, Yiren Wang, Hany Hassan Awadallah, and Arul Menezes. 2023. Leveraging gpt-4 for automatic translation post-editing.

Ricardo Rei, Craig Stewart, Ana C. Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2685–2702. Association for Computational Linguistics.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools.

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 fine-tuned chat models.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz

Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, Tianxiang Hu, Shangjie Li, Binyuan Hui, Bowen Yu, Dayiheng Liu, Baosong Yang, Fei Huang, and Jun Xie. 2023. Polylm: An open source polyglot large language model. *CoRR*, abs/2307.06018.

Lilian Weng. 2023. Llm-powered autonomous agents. *lilianweng.github.io*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2023. A paradigm shift in machine translation: Boosting translation performance of large language models.

Wen Yang, Chong Li, Jiajun Zhang, and Chengqing Zong. 2023. Bigtranslate: Augmenting large language models with multilingual translation capability over 100 languages.

Jiali Zeng, Fandong Meng, Yongjing Yin, and Jie Zhou. 2023. Tim: Teaching large language models to translate with comparison.

Shaolei Zhang, Qingkai Fang, Zhuocheng Zhang, Zhengrui Ma, Yan Zhou, Langlin Huang, Mengyu Bu, Shangtong Gui, Yunji Chen, Xilin Chen, and Yang Feng. 2023. Bayling: Bridging cross-lingual alignment and instruction following through interactive translation for large language models.

A Example Appendix

This is an appendix.