Upper-bound Translation Performance of Llama-2 Under Idealized Setup

Anonymous ACL submission

Abstract

Large Language Models (LLMs) demonstrate state-of-the-art results across multiple tasks, but machine translation remains a challenging task. Our work explores the translation capability of Llama-2-7b-chat and Llama-2-13b-chat under an idealized setup, where all the information needed to generate the correct translation is given to the model. We create an artificial language to help us achieve this goal while also helping us investigate factors affecting these models' performance. Our findings show that Llama-2-13b-chat exhibits strong translation abilities, surpassing 92% of supervised NMT English to XX translations BLEU wise and 85% chrF++ wise. This work underscores the potential of LLMs as translators and gives insight into the necessary resources needed to achieve their full potential.

1 Introduction

Machine Translation (MT) holds a crucial role in bridging socioeconomic gaps (Azzizah, 2015), language documentation, and also language preservation (Abney and Bird, 2010; Bird and Chiang, 2012; Costa-jussà et al., 2022). However, the performance of NMT systems for low-resource languages is still lacking compared to their higher-resource counterpart. Large Language Models (LLMs) and their utilization continue to garner attention in Natural Language Processing (NLP) (Brown et al., 2020; OpenAI, 2023; Touvron et al., 2023a,b) and also Computer Vision (Huang et al., 2023) due to their remarkable performance on various tasks (Bommasani et al., 2022). In machine translation specifically, previous works show that LLMs such as GPT-4 (OpenAI, 2023) exhibit a promising capability for high-resource machine translation. While their performance on low-resource machine translation is left wanting (Robinson et al., 2023; Hendy et al., 2023; Stap and Araabi, 2023; Kadaoui et al., 2023), a question arises: Are they the future of MT?.

Our work attempts to answer the question by tasking both Llama-2-7b-chat and Llama-2-13b-chat[1] (Touvron et al., 2023b) to translate from English to an artificial language we create. Evaluated on FLORES-200[2] (Costa-jussà et al., 2022)'s, using SacreBLEU[3] (Post, 2018), we report that Llama-2-13b-chat is a capable translator, performing higher than 92% of (Costa-jussà et al., 2022)'s NLLB-54b Supervised NMT's English to XX translation BLEU wise and 85% chrF++ wise. We also perform ablation studies which give insights on what affects Llama-2's translation performance on real-life languages.

2 Artificial Language

Reporting the upper-bound performance of Llama-2 as a translator requires an ideal situation. We define this ideal situation by ensuring that the prompts given to the model have enough information to reconstruct the perfect translation. However, this is impossible to achieve using actual languages due to their innate complexity (Bommasani et al., 2022). One word in English may be translated into many different words in other languages depending on factors such as context and the target language's grammatical rules. Because of this, we require an artificial language that we know the exact rules of.

Creating an actual artificial language is hard. Esperanto (of Encyclopaedia Britannica, 2023), the most popular artificial language, took years to be developed. Rather than creating it from scratch, we create our artificial language based on English. Through a combination of character-level bigram noising; artificial compounding; word-order shuffling (Ravfogel et al., 2019); and word-level translation obtained through Google Translate API; we create 20 artificial language variations. We group these 20 variations into five: AB,

[1] Our work is in line with Llama-2's Acceptable Use Policy

[2] Attribution-NonCommercial 4.0 International

[3] SacreBLEU Github Version 2.3.1

C1BA, C2BA, D1BA, and D2BA; with each group consisting of four variations due to word-order shuffling which are: SOV (Subject-Object-Verb), SVO (Subject-Verb-Object), VOS (Verb-Object-Subject), and VSO (Verb-Subject-Object).

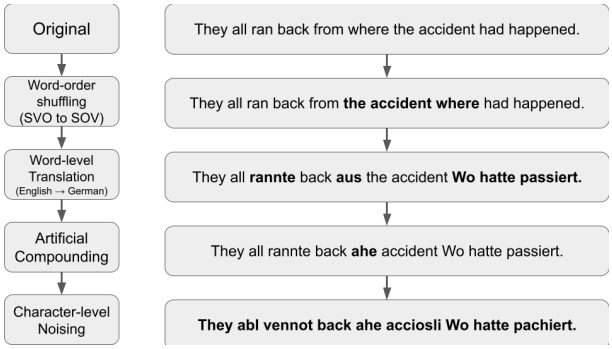

Figure 1: Example case of transforming an English sentence to the artificial language for C/DxBA variations. AB variation skips the **Word-level Translation** step.

AB variation is obtained through performing word-order shuffling followed by artificial compounding and character-level bigram noising. Word-order shuffling is conducted following the algorithm of (Ravfogel et al., 2019). To perform artificial compounding, We choose the top 5% most often occurring word-level bigrams from eng-simple_wikipedia_2021 (100K) leipzig corpora (Goldhahn et al., 2012) and perform word-level bigram blending, transforming two words into one (e.g. Motor + Hotel = Motel). To perform character-level bigram noising, we analyze and take the top 5% most common character-level bigrams found in nltk's list of English words corpus (Bird et al., 2009) and create a random mapping between them. The results of combining these approaches give us four variations. An illustration is provided in Figure 1

C/DxBA variations, referring to four variation groups other than AB, is obtained through adding word-level translation step in the middle of AB variation. We reuse the same mapping we mention in the AB section above. We randomly choose 20% of all words that appear in the corpus and translate them to 4 different languages: German (C1BA), Portuguese (C2BA), Afrikaans (D1BA), and Galician (D2BA). These four languages are chosen on two factors which are language family (Eberhard et al., 2023) and language representation in Llama-2 (Touvron et al., 2023b). German and Portuguese are represented in Llama-2, while German and Afrikaans share the same language family as English. Meanwhile, Portuguese and Galician share the same language family but are not related to English. The result of combining these approaches gives us the last 16 variations. An illustration of this process is provided in Figure 1.

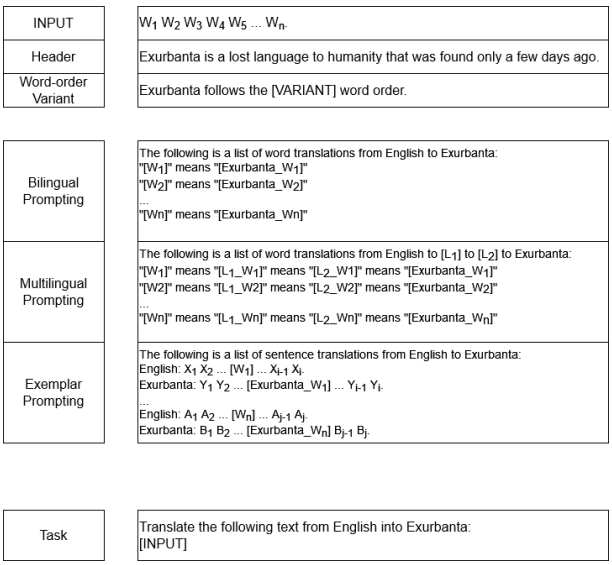

Figure 2: Prompting template used for Bilingual, Multilingual, and Exemplar prompting. **Header** is required to stop the model from refusing to perform the translation task. **[VARIANT]** depends on the word-order variation. A complete prompt consists of **Header**, **Word-order Variant**, One of the prompting methods, and the **Task** section.

We utilize in-context learning to task Llama-2 in providing translation from English to all the variations of our artificial language, which we name Exurbanta. We conduct the experiments using three main approaches which are bilingual, multilingual, and exemplar prompting. The template for each prompting approach is visible in Figure 2.

3 Related Works

While previous works show the remarkable capabilities of LLMs in multiple tasks (Brown et al., 2020; OpenAI, 2023; Bommasani et al., 2022), translation is one where LLMs still falls behind on. This is especially true for low-resource language translation (Brown et al., 2020; Zhu et al., 2023), where their performance is far below that of Supervised NMT (Costa-jussà et al., 2022). Many have shown that LLMs translation can improve through clever usage of in-context learning (Brown et al., 2020; Agrawal et al., 2022; Moslem et al., 2023; Lu et al., 2023), but questions of the potential of LLMs as translator are not yet explored.

Order	Code	BLEU_bp	BLEU_in	BLEU_out	BLEU_near	BLEU_far	BLEU_exemplar
SOV	AB	**41.39**	5.35	5.63	5.00	3.53	23.58
	C1BA	**42.48**	8.61	8.09	6.06	7.97	19.47
	C2BA	**42.38**	8.43	8.05	5.75	8.76	19.10
	D1BA	**42.82**	7.71	7.94	6.22	7.47	20.54
	D2BA	**41.82**	7.80	7.50	5.37	8.35	19.17
SVO	AB	**46.53**	5.91	6.17	5.68	3.95	25.60
	C1BA	**47.55**	9.78	8.40	6.33	9.04	22.64
	C2BA	**47.77**	9.57	8.59	6.01	9.43	21.90
	D1BA	**47.64**	8.72	8.52	6.12	8.28	22.01
	D2BA	**46.75**	8.54	8.20	5.29	8.50	21.57
VOS	AB	**38.36**	5.19	5.27	4.71	3.23	23.62
	C1BA	**41.74**	8.71	7.43	5.56	8.28	18.24
	C2BA	**41.58**	8.22	7.61	5.67	8.26	16.89
	D1BA	**41.52**	7.57	7.40	5.47	7.42	17.70
	D2BA	**41.00**	7.74	7.09	5.06	7.65	18.13
VSO	AB	**40.89**	5.48	5.32	4.78	3.28	23.73
	C1BA	**42.35**	8.82	7.60	5.53	8.08	18.45
	C2BA	**42.22**	8.33	7.78	5.55	8.36	17.38
	D1BA	**42.21**	7.41	7.34	5.64	7.28	17.91
	D2BA	**41.76**	7.58	7.39	5.05	7.76	16.87

Table 1: BLEU Scores for English to Exurbanta Translation on Llama-2-13b-chat.

4 Results

We evaluate Llama-2's translation performance using six prompting experiments: **bp** for bilingual, **in; out; near; far** for multilingual prompting where each refers to two languages **in** or **out** of Llama-2's training data while **near** and **far** refers to two languages in the same language family as English or not, and **exemplar** for exemplar prompting. The BLEU score of each prompting experiment tested on Llama-2-13b-chat is visible in Table 1. The chrF++ score alongside Llama-2-7b-chat's performance are attached in appendix A.

SVO languages are the easiest to translate to. Our experiments revealed that the English to Exurbanta translation performance of Llama-2 depends on what word order rules govern Exurbanta. Llama-2 consistently has the easiest time translating into the SVO (Subject-Verb-Object) word order variant while having the hardest time translating into the VOS (Verb-Object-Subject) word order variant. We contribute this to Llama-2's training data consisting of mostly English tokens.

Bilingual Prompting Performs Best. We observe that bilingual prompting emerged as the most effective prompting strategy, followed second by exemplar prompting which has much lower performance. multilingual prompting emerged as the worst-performing method, resulting in BLEU scores below 10. We report on possible causes in Section 5.

Limited Impact of Language Vocabulary. We also show that Llama-2-13b-chat performance did not drop when translating into the C/DxBA variations. Note that C/DxBA variations have 20% of its English vocabulary translated into another language while AB variation is not given this additional vocabulary complexity. Surprisingly, Llama-2-13b-chat performs better when translating English into the artificial language with additional vocabulary complexity.

Approach	Statistic	BLEU	chrF++
Ours	Avg Bilingual	43.03	57.80
Supervised	Mean	27.10	45.31
	Min	2.70	9.80
	Max	58.40	70.80

Table 2: The average performance of our best-performing prompting approach on Llama-2-13b-chat compared to NLLB-54b (Costa-jussà et al., 2022) Supervised NMT performance on English to XX direction.

We conducted a comparison of the average performance achieved by our best-performing prompting approach across all experiments with (Costa-jussà et al., 2022)'s NLLB-54b Supervised NMT models. Our findings show that although Llama-2-13b-chat did not surpass NLLB-54b's best performance in terms of both BLEU and chrF++ metrics, it remains a proficient translator in an idealized setup. On average, Llama-2-13b-chat outperforms NLLB-54b in 92% of English to XX translation

performance in terms of BLEU and 85% in terms of chrF++.

5 Ablation On Artificial Language Translation

Code	Experiment	BLEU	chrF++
AB	No Mask	46.53	61.33
	Mask Noun	*36.73*	*53.95*
	Mask Verb	44.41	59.83
	Mask Adj	45.09	60.49
	Mask Other	**46.64**	**61.38**
D1BA	No Mask	47.64	61.17
	Mask Noun	*36.05*	*52.21*
	Mask Verb	45.56	59.38
	Mask Adj	46.20	60.21
	Mask Other	**47.91**	**61.32**

Table 3: Impact of masking 50% of the resources by word type on Llama-2-13b-chat, tested for SVO word order. *Italic* and **Bold** implies *lowest* and **highest** performance grouped by **Code** respectively. **No Mask** results are our obtained upper-bound performance.

We examine the impact of masking half of the nouns, verbs, adjectives, or other word types from the bilingual prompt given to Llama-2-13b-chat. We observed that when the prompts only have access to half the nouns, its performance drops by almost 10 BLEU, as reported in Table 3. This huge decline is not repeated for verbs and adjectives, with a performance drop of around 2 BLEU. Surprisingly, when information on half of the other word types are not given, the performance slightly increases. This highlights the importance of quality for noun translations when prompting Llama-2 for translation.

Code	Experiment	avg-BLEU	avg-chrF++
AB	Ideal	41.79	57.83
	Random	33.70	52.20
D1BA	Ideal	43.54	58.23
	Random	32.12	48.88

Table 4: Impact of shuffling the word-level translation given in the prompts, averaged across all word orders. All experiments ran on Llama-2-13b-chat.

We examine the impact of shuffling the word-level translation in bilingual prompting. Our results show a sharp decrease in Llama-2-13b-chat translation performance, visible in Table 4. This indicates that random shuffling might introduce complexity to the prompt, which confuses the model, hindering Llama-2-13b-chat's translation capability. This problem may be mitigated by using a more robust model.

Code	Experiment	avg-BLEU	avg-chrF++
AB	Exemplar 7b-chat	7.94	32.27
	Exemplar 13b-chat	24.13	45.73
	5-shot Random	7.03	33.14
	Hybrid	24.84	47.32
	Kitchen-sink	23.41	49.32
D1BA	Exemplar 7b-chat	7.23	29.98
	Exemplar 13b-chat	19.54	40.43
	5-shot Random	5.11	37.45
	Hybrid	24.19	43.71
	Kitchen-sink	12.03	37.45

Table 5: Ablation study on the performance of Exemplar prompting for English to Exurbanta, averaged on all word orders.

Additional studies done on exemplar prompting indicate that Llama-2-13b-chat has a much higher reasoning capability compared to its 7b counterpart. Shown in Table 5, Llama-2-13b-chat shows the capability of extracting information on an artificial language it has never seen, only through examples. Meanwhile, its 7b counterpart fails with this, even though both models are given identical prompts.

We also observed that Llama-2, despite it being training with 4K context windows, faces challenges with long prompts. In our **kitchen-sink** experiment, combining bilingual prompting with exemplar prompting led to inferior performance compared to using only one of these methods individually. However, when using a **hybrid** approach, utilizing both bilingual and exemplar prompting by dividing the required information equally between them, we observed a slight performance improvement compared to using only exemplar prompting.

6 Conclusion

In this study, we have examined the upper-bound performance of Llama-2-7b-chat and Llama-2-13b-chat within an idealized setup, shedding light on the potential LLMs have as translators. Our findings demonstrate that in an idealized setup, both models performs well, with Llama-2-13b-chat surpassing 92% of Supervised NMT English to XX translations BLEU wise and 85% chrF++ wise. While these results are promising, we acknowledge that they have yet to outperform the best-performing Supervised NMT systems. Our investigation has also revealed the models' inability to handle long prompts. Our ablation experiments suggest that the problem persists even for Llama-2-13b-chat.

Limitations

While our study provides valuable insights into the upper-bound performance of Llama-2 models in an idealized setup, several limitations should be considered when interpreting the results:

Lack of Human Evaluation. We did not conduct human evaluation on the prompts used in our artificial language experiments. Although we believe that we provide ideal prompts to the model, human evaluation would have added an important dimension to our study. It would have allowed us to gauge how proficient human are when provided with the same prompts, providing a benchmark for the model's performance.

Limited Generalizability. Our experiments were conducted by creating an artificial language, Exurbanta, which simplifies the complexity of real languages. Real languages exhibit nuances and variations that were not fully captured by our artificial language approach. One word in English can have multiple translations in different contexts and languages, meanwhile our artificial language do not have this nuance.

Model Choice. We focused our experiments on Llama-2-7b-chat and Llama-2-13b-chat, and our findings are specific to these models. We did not explore the potential differences in performance when using larger and more robust models like Llama-2-70b-chat. It is possible that some issues, such as handling long prompts, are less relevant for larger models. Future research is needed to investigate the translation performance of these larger models in an idealized setup.

References

Steven Abney and Steven Bird. 2010. The human language project: Building a universal corpus of the world's languages. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 88–97.

Sweta Agrawal, Chunting Zhou, Mike Lewis, Luke Zettlemoyer, and Marjan Ghazvininejad. 2022. In-context examples selection for machine translation.

Yuni Azzizah. 2015. Socio-economic factors on indonesia education disparity. *International Education Studies*, 8:218.

Steven Bird and David Chiang. 2012. Machine translation for language preservation. In *Proceedings of COLING 2012: Posters*, pages 125–134.

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural Language Processing with Python*.

Rishi Bommasani, Drew A. Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S. Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, Erik Brynjolfsson, Shyamal Buch, Dallas Card, Rodrigo Castellon, Niladri Chatterji, Annie Chen, Kathleen Creel, Jared Quincy Davis, Dora Demszky, Chris Donahue, Moussa Doumbouya, Esin Durmus, Stefano Ermon, John Etchemendy, Kawin Ethayarajh, Li Fei-Fei, Chelsea Finn, Trevor Gale, Lauren Gillespie, Karan Goel, Noah Goodman, Shelby Grossman, Neel Guha, Tatsunori Hashimoto, Peter Henderson, John Hewitt, Daniel E. Ho, Jenny Hong, Kyle Hsu, Jing Huang, Thomas Icard, Saahil Jain, Dan Jurafsky, Pratyusha Kalluri, Siddharth Karamcheti, Geoff Keeling, Fereshte Khani, Omar Khattab, Pang Wei Koh, Mark Krass, Ranjay Krishna, Rohith Kuditipudi, Ananya Kumar, Faisal Ladhak, Mina Lee, Tony Lee, Jure Leskovec, Isabelle Levent, Xiang Lisa Li, Xuechen Li, Tengyu Ma, Ali Malik, Christopher D. Manning, Suvir Mirchandani, Eric Mitchell, Zanele Munyikwa, Suraj Nair, Avanika Narayan, Deepak Narayanan, Ben Newman, Allen Nie, Juan Carlos Niebles, Hamed Nilforoshan, Julian Nyarko, Giray Ogut, Laurel Orr, Isabel Papadimitriou, Joon Sung Park, Chris Piech, Eva Portelance, Christopher Potts, Aditi Raghunathan, Rob Reich, Hongyu Ren, Frieda Rong, Yusuf Roohani, Camilo Ruiz, Jack Ryan, Christopher Ré, Dorsa Sadigh, Shiori Sagawa, Keshav Santhanam, Andy Shih, Krishnan Srinivasan, Alex Tamkin, Rohan Taori, Armin W. Thomas, Florian Tramèr, Rose E. Wang, William Wang, Bohan Wu, Jiajun Wu, Yuhuai Wu, Sang Michael Xie, Michihiro Yasunaga, Jiaxuan You, Matei Zaharia, Michael Zhang, Tianyi Zhang, Xikun Zhang, Yuhui Zhang, Lucia Zheng, Kaitlyn Zhou, and Percy Liang. 2022. On the opportunities and risks of foundation models.

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.

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, Loïc 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: Scal-

ing human-centered machine translation. *CoRR*, abs/2207.04672.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, 26 edition. SIL International, Dallas, Texas.

Dirk Goldhahn, Thomas Eckart, and Uwe Quasthoff. 2012. Building large monolingual dictionaries at the leipzig corpora collection: From 100 to 200 languages. In *International Conference on Language Resources and 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.

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Johan Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. 2023. Language is not all you need: Aligning perception with language models.

Karima Kadaoui, Samar M. Magdy, Abdul Waheed, Md Tawkat Islam Khondaker, Ahmed Oumar El-Shangiti, El Moatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. Tarjamat: Evaluation of bard and chatgpt on machine translation of ten arabic varieties.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models.

Yasmin Moslem, Rejwanul Haque, John D. Kelleher, and Andy Way. 2023. Adaptive machine translation with large language models.

Editors of Encyclopaedia Britannica. 2023. Esperanto. `https://www.britannica.com/topic/Esperanto`. Accessed on 2023-10-10.

OpenAI. 2023. Gpt-4 technical report.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Shauli Ravfogel, Yoav Goldberg, and Tal Linzen. 2019. Studying the inductive biases of rnns with synthetic variations of natural languages.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high- (but not low-) resource languages.

David Stap and Ali Araabi. 2023. ChatGPT is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 163–167, Toronto, Canada. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. Llama: Open and efficient foundation language models.

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. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Wenhao Zhu, Hongyi Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2023. Multilingual machine translation with large language models: Empirical results and analysis.

A Artificial Language Translation

Order	**Code**	BLEU_**bp**	BLEU_**in**	BLEU_**out**	BLEU_**near**	BLEU_**far**	BLEU_**exemplar**
SOV	AB	**34.43**	6.52	5.75	7.52	5.92	7.80
	C1BA	**26.26**	6.96	5.57	5.28	7.03	7.68
	C2BA	**24.92**	6.07	5.23	5.66	6.52	7.62
	D1BA	**26.74**	6.39	6.07	5.32	7.09	7.56
	D2BA	**25.61**	5.94	5.38	5.34	6.26	7.15
SVO	AB	**38.06**	6.96	5.63	7.90	6.17	8.23
	C1BA	**29.34**	7.25	5.56	5.49	7.39	7.45
	C2BA	**27.63**	6.53	5.49	5.99	6.70	7.87
	D1BA	**29.31**	6.62	6.42	5.75	7.56	7.58
	D2BA	**28.17**	6.22	5.60	5.75	6.71	7.68
VOS	AB	**32.80**	5.53	5.22	6.50	5.08	7.85
	C1BA	**25.34**	6.56	5.45	5.08	6.96	6.89
	C2BA	**24.25**	5.91	5.08	5.61	6.40	6.83
	D1BA	**25.75**	5.98	5.92	5.04	7.02	6.63
	D2BA	**23.92**	5.49	5.43	5.27	6.22	6.50
VSO	AB	**33.42**	5.79	5.41	6.87	4.98	7.89
	C1BA	**26.03**	6.39	5.43	5.16	6.94	7.47
	C2BA	**24.45**	5.75	5.11	5.55	6.28	7.21
	D1BA	**26.58**	5.78	5.81	5.20	6.90	7.15
	D2BA	**24.78**	5.45	5.31	5.38	6.23	6.91

Table 6: BLEU Scores for English to Exurbanta Translation on **Llama-2-7b-chat**. Bilingual prompting achieves the highest performance, denoted as **bp**. Multilingual prompting with L1 (German) and L2 (Portuguese) yields **in**. Multilingual prompting with L1 (Afrikaans) and L2 (Galician) yields **out**. Multilingual prompting with L1 (German) and L2 (Afrikaans) yields **near**. Multilingual prompting with L1 (Portuguese) and L2 (Galician) yields **ar**. Exemplar prompting results are represented as **exemplar**.

Order	Code	chrF++_bp	chrF++_in	chrF++_out	chrF++_near	chrF++_far	chrF++_exemplar
SOV	AB	**51.02**	26.26	23.42	25.96	23.99	31.97
	C1BA	**44.47**	25.03	21.62	21.68	23.92	30.51
	C2BA	**43.89**	23.47	21.20	23.76	22.80	30.51
	D1BA	**44.57**	23.25	23.44	21.92	24.34	30.27
	D2BA	**44.41**	22.97	21.39	23.11	22.62	29.74
SVO	AB	**54.08**	26.56	23.19	26.13	24.03	32.68
	C1BA	**46.64**	25.33	21.73	21.74	24.16	30.50
	C2BA	**46.15**	23.93	21.58	23.91	23.02	30.88
	D1BA	**46.93**	23.60	23.60	22.11	24.84	30.45
	D2BA	**46.08**	23.03	21.44	23.34	23.06	30.60
VOS	AB	**49.54**	25.78	23.32	25.57	23.58	32.28
	C1BA	**43.09**	24.27	21.37	21.60	23.53	29.48
	C2BA	**42.95**	23.05	20.99	23.70	22.53	29.40
	D1BA	**43.59**	22.68	23.26	21.49	24.12	29.36
	D2BA	**42.65**	22.22	21.44	22.92	22.23	29.35
VSO	AB	**49.79**	25.83	23.48	25.35	23.03	32.15
	C1BA	**43.79**	24.10	21.35	21.57	23.62	30.47
	C2BA	**43.23**	22.87	20.97	23.48	22.38	30.05
	D1BA	**44.29**	22.40	23.11	21.63	24.04	29.86
	D2BA	**43.15**	22.08	21.14	22.89	22.37	29.72

Table 7: chrF++ Scores for English to Exurbanta Translation on **Llama-2-7b-chat**. Bilingual prompting resulted in the best translation.

Order	Code	chrF++_bp	chrF++_in	chrF++_out	chrF++_near	chrF++_far	chrF++_exemplar
SOV	AB	**57.12**	27.83	25.56	26.31	23.12	45.21
	C1BA	**57.06**	30.88	27.30	25.32	28.33	40.23
	C2BA	**57.12**	30.08	26.99	25.63	28.56	39.45
	D1BA	**57.79**	29.58	27.87	25.35	28.39	40.54
	D2BA	**56.69**	29.84	26.84	25.20	28.91	39.65
SVO	AB	**61.33**	28.74	26.18	27.16	23.68	47.11
	C1BA	**60.77**	31.92	27.40	25.41	29.41	42.70
	C2BA	**60.80**	31.28	27.19	25.88	29.30	41.80
	D1BA	**61.17**	30.79	28.44	25.17	29.26	42.43
	D2BA	**60.06**	30.47	27.09	25.00	28.96	41.86
VOS	AB	**55.91**	27.73	25.24	26.08	22.78	45.22
	C1BA	**56.35**	30.86	26.72	24.80	28.68	39.43
	C2BA	**56.23**	29.82	26.50	25.68	28.35	38.66
	D1BA	**56.64**	29.94	27.60	24.78	28.56	39.33
	D2BA	**56.33**	29.85	26.46	24.95	28.30	39.31
VSO	AB	**56.99**	28.35	25.37	26.10	22.67	45.41
	C1BA	**56.83**	31.05	26.76	24.63	28.60	39.70
	C2BA	**57.02**	29.88	26.40	25.56	28.33	39.03
	D1BA	**57.32**	29.53	27.45	24.81	28.50	39.45
	D2BA	**56.60**	29.58	26.44	24.70	28.34	38.31

Table 8: chrF++ Scores for English to Exurbanta Translation on **Llama-2-13b-chat**. Bilingual prompting resulted in the best translation.