M-IFEval: Multilingual Instruction-Following Evaluation

Anonymous ACL submission

Abstract

Instruction following is a core capability of modern Large language models (LLMs), making evaluating this capability essential to understanding these models.

The Instruction Following Evaluation (IFEval) benchmark from the literature does this using objective criteria, offering a measure of LLM performance without subjective AI or human judgement. However, it only includes English instructions, limiting its ability to assess LLMs in other languages.

We propose the Multilingual Instruction Following Evaluation (M-IFEval) benchmark, expanding the evaluation to French, Japanese, and Spanish, with both general and language-specific instructions.

Applying this benchmark to 8 state-of-the-art LLMs, we find that benchmark performance across languages and instruction types can vary widely, underscoring the importance of a multilingual benchmark for evaluating LLMs in a diverse cultural context.

1 Introduction

Large language models (LLMs) have demonstrated amazing accuracy in many fields including medicine (Tian et al., 2024; Frisoni et al., 2024), law (Jiang et al., 2024), and education (Luo et al., 2024). However, their accuracy has also shown to be surprisingly low for some tasks such as reasoning (Tong et al., 2024) and cultural understanding (Wang et al., 2024).

One task of particular importance for LLMs is that of instruction following, where an LLM must carry out the instructions of the user in a "zero-shot" setting (i.e. without necessarily being trained specifically to perform that instruction) (Zhong et al., 2021; Mishra et al., 2022; Wei et al.; Sanh et al., 2022).

Benchmarks such as Instruction-Following Evaluation (IFEval) (Zhou et al., 2023) have proposed ways of evaluating LLMs on instruction following without the need for using an external LLM-as-a-judge (Zheng et al., 2023), which may exhibit self-enhancement bias (Xu et al., 2024).

However, this benchmark is a purely English-based benchmark, raising questions as to the applicability of its results to other languages. While some efforts have been made to make a multilingual version of this benchmark, at present this only extends as far as translating the original prompts into other languages (Qwen, 2024). This approach fails to evaluate aspects of instruction following that are specific to different languages. Specifically developed code is required to understand whether an LLM can, for example, uses the correct punctuation or script for a given language when prompted.

We present Multilingual Instruction Following Evaluation (M-IFEval), a benchmark for evaluating LLM instruction following beyond English. Our benchmark consists of three popular natural languages, French, Japanese, and Spanish, and contains both instructions previously assessed in English as well as novel instructions that are specific to our evaluation languages. We assess 8 state-of-the-art LLMs using M-IFEval and compare their evaluation results to the original English IFEval scores.

Our evaluation results show that while popular models such as GPT4o are better at instruction following in English, we find that models such as o1 and Sonnet perform better at instruction following in the languages of our benchmark. We also highlight that state-of-the-art LLMs achieve surprisingly low scores on some language-specific instructions such as using or not using special characters or scripts.

Our work demonstrates the value of a multilingual benchmark when selecting LLMs for a non-English based task and identifies key areas for improvement in modern LLMs. We make the evaluation code and data for this benchmark available

online[1].

2 Related work

Benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) that contain objective tasks such as natural language inference and semantic similarity have been widely used within the literature to evaluate LLMs (Anil et al., 2023; Le Scao et al., 2023; Dettmers et al., 2023). However, these tasks do not fully represent the realistic usage of LLMs in practical scenarios, for example as conversational agents or decision-making systems.

Other benchmarks such as MT-Bench (Zheng et al., 2023) and InFoBench (Qin et al., 2024) provide a framework for evaluating LLMs in a more practical conversational or instruction-based setting. However, their reliance on subjective AI scoring raises self-enhancement bias concerns (Xu et al., 2024), making them less suitable for model evaluation.

Chatbot Arena (Zheng et al., 2023) involves evaluating LLMs using human users, thus removing the potential for self-enhancement bias. However, Chatbot Arena requires large-scale deployment, making it difficult to replicate this evaluation for local models.

IFEval (Zhou et al., 2023) was designed to evaluate the ability of LLMs to follow instructions in more practical scenarios, making it an effective measure of models intended for real-world usage. Unlike MT-Bench, IFEval uses objective criteria for evaluation, removing the subjectivity inherent in AI-judged systems. However, its primary limitation is that it is currently available only in English, which restricts its applicability for evaluating multilingual models.

The model evaluation of the instruction-tuned Qwen 2.5 (Qwen, 2024) extended IFEval to support multilingual settings by translating 100 examples per language and removing instructions that were not applicable to a given language. While this approach allows for deterministic evaluation of LLMs on multilingual data, it may neglect language-specific aspects of instruction-following.

To overcome the limitations of existing benchmarks, we propose a new, multilingual version of IFEval that is not simply a translation of previous datasets but contains language-specific instructions for novel evaluation.

[1] ANONYMIZED

3 Method

This section details how we constructed the M-IFEval benchmark and how we used it to evaluate various state-of-the-art LLMs.

We first chose the languages that would be included in our benchmark. We chose French, Japanese, and Spanish as our research team included one native speaker in each language, with each acting as language lead for their respective language.

Our team consulted the list of instructions included in the original English language IFEval benchmark and considered any that were not applicable to their language. Following this, we removed the word number length constraint instruction and the case change instructions (all uppercase, all lowercase, and frequency of all capital words) for Japanese.

Next, each language lead created a list of instructions that could be evaluated using objective criteria, including instructions that are specific to their language. These instructions include grammatical, stylistic, and script-based instructions across the three evaluation languages.

The **Spanish-specific instructions** consisted of three special character-based instructions (ñ frequency, accent frequency, and ü frequency) and two punctuation-based instructions (using grammatically correct question marks and exclamation marks).

The **French-specific instructions** consisted of three special character-based instructions (forbidding use of œ/ç, forbidding accents, and adding the correct accents to a given text), and two content based instructions (not using Arabic numerals in the response and using the informal direct way of addressing someone).

The **Japanese-specific instructions** consisted of seven script based instructions (only/do not use katakana, only/do not use hiragana, use at least/most N kanji, include furigana, write all numbers as kanji), two format based instructions (responses must be a numbered list of N items, responses must include at least N taigen-dome[2]), and one instruction each of a length based instruction (use at least/most N characters), a start/end based instruction (end all sentences with です/ます), and a punctuation based instruction (do not use 。 - a Japanese period).

[2] A Japanese grammatical structure where a sentence ends with a noun or noun phrase (Hayashi and Matsubara, 2007)

The list of all language-specific instructions can be found in table 3 in the appendix.

This resulted in a list of instructions for each language which were a mix of instructions from the original IFEval benchmark and instructions that were specific to their language.

Each language lead then developed a function for each instruction that evaluates whether a response did or did not correctly follow the given instruction. These functions were then added to the evaluation codebase of the original IFEval benchmark.

Prompts that instruct the LLM to follow at least one instruction were then developed by the language leads in a similar way to the original IFEval work. Prompts were developed by generating multiple example prompts using a state-of-the-art LLM before selecting and editing prompts manually to obtain a list of multiple prompts per instruction. As with the original IFEval, we constructed prompts with one, two, and three instructions contained within the same prompt. The correct evaluation arguments (e.g. specifying 2 if the prompt specifies 2 sentences in the sentence counting instruction) were then manually added to each instruction.

This process resulted in 172, 115, and 235 prompts for Japanese, Spanish, and French, respectively, with at least 4 unique prompts per instruction for Spanish and Japanese, and 7 prompts per instruction for French. For Japanese, Spanish, and French, our benchmark contains 34, 8, and 68 prompts respectively that contains 2 instructions, and 10, 7, and 21 prompts that contain 3 instructions.

Responses to these prompts were then generated using all the state-of-the-art LLMs that we had access to, consisting of 4 versions of OpenAI's GPT (GPT4o, GPT4o Mini, o1, o1 Mini), 3 versions of Anthropic's Claude 3.5 (Opus, Sonnet, and Haiku), and the largest multilingual open source LLM that we could run in 4 bits on a single 40GB A100 GPU (Qwen 2.5 32B Instruct GPTQ Int4). To generate responses, the temperature was set to 0 where possible in the interests of reproducibility.

These responses were then evaluated using our evaluation code and we report the average score in each language for each model. We separately report the scores only of the average language-specific instructions for each language and model. As with the original IFEval work, we calculate both the strict and loose scores for each instruction. We report the strict scores in the main document and report the loose scores in the appendix.

Model name	EN	ES	FR	JA	Mean
o1	86.7	**92.7**	**91.3**	75.7	**86.6**
Opus	87.3	90.5	87.0	75.7	84.4
Sonnet	88.1	87.6	88.1	**77.0**	84.2
o1 Mini	83.9	92.0	88.4	69.5	83.3
GPT4o	**88.6**	89.8	87.8	70.4	82.7
GPT4o Mini	86.0	85.4	85.5	65.9	78.9
Qwen 2.5 32B I.	86.0	82.5	81.7	65.9	76.7
Haiku	77.3	78.8	78.3	61.9	73.0

Table 1: Average strict scores of M-IFEval for each language for each model evaluated, sorted by the mean combined Spanish, French, and Japanese scores.

Model name	ES	FR	JA	Mean
o1	**75.0**	**96.1**	61.4	**77.5**
Sonnet	66.7	90.2	**70.5**	75.8
Opus	62.5	90.2	64.8	72.5
GPT4o	58.3	80.4	55.7	64.8
o1 Mini	66.7	72.5	50.0	63.1
Qwen 2.5 32B I.	54.2	78.4	54.5	62.4
Haiku	54.2	80.4	52.3	62.3
GPT4o Mini	58.3	68.6	47.7	58.2

Table 2: Average strict scores of M-IFEval for each language only on the instructions that are specific to that language, sorted by the mean combined Spanish, French, and Japanese scores.

4 Results

Table 1 shows the average evaluation score for each model evaluated in each language in the M-IFEval benchmark compared to the English scores in the original IFEval benchmark.

We observe that while GPT4o and Sonnet are the top two models for English M-IFEval, o1 and Opus have the highest score on average for the three languages in our benchmark. We also observe a greater spread in scores between the best and worse performing models in our evaluation compared to the English IFEval, with the best and worst scores on the original English benchmark having a difference of 11.3 percentage points, while we observe differences of 13.9, 13.0, and 15.1 for Spanish, French, and Japanese respectively.

Table 2 shows the average scores only on instructions that are unique to each language. We observe that while the o1 model attains the greatest scores on Spanish and French benchmarks, Sonnet achieves markedly higher scores on the Japanese benchmark. When we analysed the scores only of instructions that had been included in the original

IFEval benchmark (i.e. instructions not unique to the language), we found that o1 achieves a score on the Japanese benchmark of 84.8 while Sonnet achieves a score of 81.2.

When we analysed specific instructions with the lowest average evaluation scores across all models that we tested, we found that the 10 instructions with the lowest scores all were language-specific instructions such as forbidding "œ/ç", forbidding katakana, or specifying the frequency of the "ñ" character. The average scores for these three instruction types across all models was 60.2%, 14.3%, and 0.0%, respectively. Conversely, we observe that LLMs attain high scores in following instructions such as adding accents to French text, adding Spanish question marks/exclamantion marks, and making both French and Spanish text uppercase/lowercase. This indicates that while many LLMs display high performance in following instructions such as formatting outputs a certain way and arranging punctuation, they are not as proficient at following script-based instructions.

The full scores averaged across all models for each instruction can be found in table 6.

5 Discussion & Future work

Overall, our results show that modern LLMs are generally proficient at instruction following outside of English. We find that the overall average evaluation scores of LLMs in English are similar for Spanish and French. However, we find that LLM instruction following performance falls in languages less similar to English, such as Japanese.

Our results show that Sonnet is more proficient at Japanese-specific instructions compared to o1, whereas o1 is more proficient at Spanish and French specific instructions. This could indicate that o1 has been trained on more Spanish and French data, or linguistically similar languages that confer cross lingual generalisation (Snæbjarnarson et al., 2023; Muennighoff et al., 2023), while Sonnet may have been trained on more Japanese data.

Moreover, we find that the highest performing model in our English evaluation was neither o1 nor Sonnet, but GPT4o. This highlights the need for multilingual LLM evaluations to select the best model for a target language, as no single LLM excels in all languages.

Our results also show that LLMs generally achieve poor performance on seemingly simple language-specific tasks such as restricting usage of a given script (e.g. "write your answer without using any katakana") or controlling for the amount of times a certain special character is used (e.g. "use the 'ñ' character exactly 5 times in your response"). This contrasts with high English scores for similar tasks (e.g. "use the letter c at least 60 times in your response"). This may reflect a gap between LLM performance in English to that of other languages which has been observed in other tasks (Ahuja et al., 2024; Jin et al., 2024).

Future work could consider exactly why the performance of LLMs varies for different languages. Previous work has investigated the effect of different language mixtures on downstream tasks (Üstün et al., 2024; Wei et al., 2023), so experiments involving different mixes of multilingual pre-training data and fine-tuning data could possibly show the effect of training data on instruction following performance.

Experiments using a byte-level tokenizer (Xue et al., 2022) could possibly answer the question of why script or character based instructions are so hard to follow for modern token-level LLMs.

6 Conclusion

In this paper, we have presented M-IFEval, a multilingual benchmark which evaluates the instruction following abilities of LLMs in three non-English languages: French, Japanese, and Spanish.

Our results show that while GPT4o achieves the greatest instruction following performance on the English IFEval benchmark, we find that other models, o1 and Sonnet, achieve higher scores on M-IFEval.

This finding highlights the importance of multilingual evaluation in assessing a model's instruction following abilities.

We also identify several types of instructions for which the average LLM performance was surprisingly low. This includes specifying the usage/non usage of a certain script and specifying the frequency of a certain amount of non-English characters.

This work contributes a new benchmark to the field of multilingual evaluation of LLMs and provides observations for what these models can and cannot do in the context of multilingual instruction following.

7 Limitations

One of the limitations of this work is that our benchmark only considers instructions that can be objectively evaluated using simple string checking code. This means that our evaluation does not include any of the large group of possible instructions which would require more intricate analysis to evaluate upon (e.g. translation quality, fact checking, question answering). We leave more detailed evaluation of LLMs on specific tasks to other benchmarks such as XNLI (Conneau et al., 2018), XQuad (Artetxe et al., 2020), and Flores (Costa-jussà et al., 2022).

Another limitation of this work is that we only consider three non-English languages in our evaluation. Moreover, these three languages were all relatively high-resource languages, and since we observe a gap between English and our evaluation languages, we may observe an even greater gap for low resource languages. Future work could include adding more languages to our benchmark, particularly low resource languages. This could entail adding more language-specific instructions (e.g. converting "Boko", or Latin, script in Hausa to "Ajami", or Arabic, script (Abdulmumin, 2014)) to further identify if there are any other tasks in which LLMs perform particularly poorly.

A final limitation of this work is that we only evaluate over 8 state-of-the-art LLMs in our evaluation when other LLMs such as Gemini (Reid et al., 2024) are also available. This was done due to a combination of technical, financial, and document-space limitations, and so the main contributions of our paper are that we demonstrate that relative instruction following performance is not uniform across all languages for a given LLM, and that some of the top performing LLMs available still cannot perform basic tasks such as controlling special character usage. We leave it for future work to use this benchmark to compare their models against others.

References

SA Abdulmumin. 2014. A survey of historical prevalence of hausa language in contemporary literacy. *ZAHIRA–Journal of Historical Research, Dept. of History, Ahmadu Bello University, ZARIA Nigeria*, 5(4).

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Mohamed Ahmed, Kalika Bali, et al. 2024. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2598–2637.

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the cross-lingual transferability of monolingual representations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314*.

Giacomo Frisoni, Alessio Cocchieri, Alex Presepi, Gianluca Moro, and Zaiqiao Meng. 2024. To generate or to retrieve? on the effectiveness of artificial contexts for medical open-domain question answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9878–9919, Bangkok, Thailand. Association for Computational Linguistics.

Yukiko Hayashi and Shigeki Matsubara. 2007. Sentence-style conversion of japanese news article for text-to-speech application. In *Proceedings of 7th International Symposium on Natural Language Processing*, pages 257–262.

Hang Jiang, Xiajie Zhang, Robert Mahari, Daniel Kessler, Eric Ma, Tal August, Irene Li, Alex Pentland, Yoon Kim, Deb Roy, and Jad Kabbara. 2024. Leveraging large language models for learning complex legal concepts through storytelling. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7194–7219, Bangkok, Thailand. Association for Computational Linguistics.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of

large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 2627–2638, New York, NY, USA. Association for Computing Machinery.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

Haohao Luo, Yang Deng, Ying Shen, See-Kiong Ng, and Tat-Seng Chua. 2024. Chain-of-exemplar: Enhancing distractor generation for multimodal educational question generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7978–7993, Bangkok, Thailand. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, et al. 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.

Yiwei Qin, Kaiqiang Song, Yebowen Hu, Wenlin Yao, Sangwoo Cho, Xiaoyang Wang, Xuansheng Wu, Fei Liu, Pengfei Liu, and Dong Yu. 2024. InFoBench: Evaluating instruction following ability in large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 13025–13048, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Qwen. 2024. Qwen 2.5: A party of foundation models.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2022. Multitask prompted training enables zero-shot task generalization. In *ICLR 2022-Tenth International Conference on Learning Representations*.

Vésteinn Snæbjarnarson, Annika Simonsen, Goran Glavaš, and Ivan Vulić. 2023. Transfer to a low-resource language via close relatives: The case study on Faroese. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 728–737, Tórshavn, Faroe Islands. University of Tartu Library.

Yuanhe Tian, Ruyi Gan, Yan Song, Jiaxing Zhang, and Yongdong Zhang. 2024. ChiMed-GPT: A Chinese medical large language model with full training regime and better alignment to human preferences. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7156–7173, Bangkok, Thailand. Association for Computational Linguistics.

Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. 2024. Can LLMs learn from previous mistakes? investigating LLMs' errors to boost for reasoning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3065–3080, Bangkok, Thailand. Association for Computational Linguistics.

Ahmet Üstün, Viraat Aryabumi, Zheng Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15894–15939, Bangkok, Thailand. Association for Computational Linguistics.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in neural information processing systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Wenxuan Wang, Wenxiang Jiao, Jingyuan Huang, Ruyi Dai, Jen-tse Huang, Zhaopeng Tu, and Michael Lyu. 2024. Not all countries celebrate thanksgiving: On the cultural dominance in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6349–6384, Bangkok, Thailand. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Xiangpeng Wei, Haoran Wei, Huan Lin, Tianhao Li, Pei Zhang, Xingzhang Ren, Mei Li, Yu Wan, Zhiwei Cao, Binbin Xie, et al. 2023. Polylm: An open source polyglot large language model. *arXiv preprint arXiv:2307.06018*.

Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Wang. 2024. Pride and prejudice: LLM amplifies self-bias in self-refinement. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15474–15492, Bangkok, Thailand. Association for Computational Linguistics.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Ruiqi Zhong, Kristy Lee, Zheng Zhang, and Dan Klein. 2021. Adapting language models for zero-shot learning by meta-tuning on dataset and prompt collections. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2856–2878.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Instruction Group	Instruction	Description
Spanish		
Special Characters	Letter Frequency (ñ)	"ñ" should appear {N} times
Special Characters	Accented Word Frequency	include at least/most {N} words with accents
Special Characters	Letter Frequency (ü)	"ü" should appear {N} times
Punctuation	Interrogation Marks	Include at least one question
Punctuation	Exclamation Marks	Include at least one exclamation
French		
Special Characters	Forbidden œ and ç	Do not use {char} characters
Special Characters	No Accents	Do not use accents
Special Characters	Add Accents	Add the correct accents to the given text
Detectable Content	Informal Address	Speak directly and informally to the user
Detectable Content	No Digits	Do not use Arabic numerals
Japanese		
Length Constraints	Number Letters	Use at least/most {N} characters
Detectable Format	Numbered Lists	Include a numbered list of exactly {N} items
Detectable Format	Taigen-dome	Include exactly {N} taigen-dome
Start with / End with	Unified Sentence Endings	All sentences must end in {ending}
Punctuation	No Periods	Do not use Japanese periods
Script	Furigana	Furigana must follow all kanji
Script	Kanji	Include at least/most {N} kanji characters
Script	Kansuuji	All numbers must be written with Kanji
Script	No Katakana	Do not include any katakana characters
Script	No Hiragana	Do not include any hiragana characters
Script	Katakana Only	Only use katakana characters
Script	Hiragana Only	Only use hiragana characters

Table 3: Full list of added instructions in Spanish, French, and Japanese.

Model name	ES	FR	JA	Mean
Sonnet	75.0	96.1	**76.1**	**82.4**
o1	**79.2**	**100.0**	63.6	80.9
Opus	62.5	96.1	65.9	74.8
Haiku	58.3	88.2	59.1	68.6
GPT4o	58.3	82.4	63.6	68.1
o1 Mini	66.7	72.5	55.7	65.0
Qwen 2.5 32B I.	54.2	78.4	58.0	63.5
GPT4o Mini	58.3	70.6	58.0	62.3

Table 4: Average loose scores of M-IFEval for each language only on the instructions that are specific to that language, sorted by the mean combined Spanish, French, and Japanese scores.

Model name	EN	ES	FR	JA	Mean
Sonnet	**93.0**	**94.9**	**94.8**	**85.0**	**91.5**
o1	89.1	**94.9**	93.6	77.4	88.6
Opus	92.6	91.2	92.8	77.9	87.3
GPT4o	91.2	92.0	90.4	76.1	86.2
o1 Mini	86.8	92.7	89.6	72.6	84.9
GPT4o Mini	88.7	89.1	89.3	71.7	83.3
Haiku	85.3	86.1	89.3	70.8	82.1
Qwen 2.5 32B I.	88.0	84.7	84.6	70.4	79.9

Table 5: Average loose scores of M-IFEval for each language for each model evaluated, sorted by the mean combined Spanish, French, and Japanese scores.

Instruction Family	Instruction Name	Languages	Score
Special characters	Letter Frequency (ñ)	ES	0.0
Script	No Katakana	JA	14.3
Script	Furigana	JA	14.6
Special characters	Letter Frequency (ü)	ES	15.6
Start with / End with	Unified Sentence Endings	JA	33.9
Script	No Hiragana	JA	35.7
Script	Hiragana Only	JA	48.2
Special characters	Forbidden œ and ç	FR	60.2
Script	Katakana Only	JA	60.4
Special characters	No Accents	FR	63.8
Detectable format	Taigen-dome	JA	64.3
Detectable format	JSON Format	EN, ES, FR, JA	67.5
Length constraints	Nth Paragraph First Word	EN, ES, FR, JA	67.6
Keywords	Letter Frequency	EN, ES, FR, JA	71.2
Change cases	French Uppercase	FR	73.4
Script	Kanji	JA	75.0
Length constraints	Number Sentences	EN, ES, FR, JA	76.4
Combination	Repeat prompt	EN, ES, FR, JA	78.2
Detectable format	Number Bullets	EN, ES, FR, JA	79.2
Special characters	Accented Word Frequency	ES	79.7
Change cases	Capital Word Frequency	EN, ES, FR	80.1
Length constraints	Number Paragraphs	EN, ES, FR, JA	80.5
Length constraints	Number Words	EN, ES, FR	82.2
Keywords	Forbidden Words	EN, ES, FR, JA	84.1
Keywords	Keyword Frequency	EN, ES, FR, JA	85.2
Start with / End with	Quotation	EN, ES, FR, JA	85.2
Change cases	English Uppercase	EN	86.0
Change cases	English Uppercase	EN	88.1
Combination	Two Responses	EN, ES, FR, JA	88.3
Length constraints	Number Letters	JA	89.3
Keywords	Include Keywords	EN, ES, FR, JA	89.6
Punctuation	No Commas	EN, ES, FR, JA	90.2
Detectable content	Informal Address	FR	90.9
Letters	Kansuuji	JA	91.1
Start with / End with	End Checker	EN, ES, FR, JA	91.3
Detectable format	Title	EN, ES, FR, JA	92.3
Punctuation	No Periods	JA	92.9
Detectable content	Number Placeholder	EN, ES, FR, JA	92.9
Detectable format	Choose From	EN, ES, FR, JA	93.3
Detectable format	Minimum Number Highlighted Sections	EN, ES, FR, JA	94.8
Language	Response Language	EN, ES, FR, JA	95.6
Detectable format	Numbered Lists	JA	96.4
Punctuation	Exclamation Marks	ES	96.9
Detectable format	Multiple Sections	EN, ES, FR, JA	97.6
Detectable content	Postscript	EN, ES, FR, JA	97.8
Special characters	Add Accents	FR	98.2
Punctuation	Question Marks	ES	100.0
Change cases	Spanish Lowercase	ES	100.0
Change cases	Spanish Uppercase	ES	100.0

Instruction Family	Instruction Name	Languages	Score
Change cases	French Lowercase	FR	100.0
Detectable content	No Digits	FR	100.0

Table 6: The percentage strict instruction following score for each instruction, averaged over all 8 models evaluated.