M-IFEval: Multilingual Instruction-Following Evaluation

Antoine Dussolle[1,2*], Andrea Cardeña Díaz[1], Shota Sato[1], Peter Devine[1],

[1]Lightblue KK., [2]Arkema,
antoine.dussolle@gmail.com
andrea.cdiaz@hotmail.es
{shota.sato,peter}@lightblue-tech.com

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 low for some tasks such as reasoning (Tong et al., 2024) and cultural understanding (Wang et al., 2024).

One type of 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, among the models tested on the English instruction-following benchmark, widely-used LLMs like GPT4o achieve the highest relative performance. However, for benchmarks in other languages, models such as o1 and Sonnet perform better in instruction following. 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, such as character- and script-level instructions, in modern LLMs. We make the evalu-

[*]Work done at Lightblue KK.

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6176–6191
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

ation code and data for this benchmark available online[1].

2 Related work

Benchmarks like GLUE (Wang et al., 2018), ARC (Clark et al., 2018), SuperGLUE (Wang et al., 2019), Winogrande (Sakaguchi et al., 2019), HellaSwag (Zellers et al., 2019), MMLU (Hendrycks et al., 2021), and others, 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), AlpacaEval (Dubois et al., 2024), Arena-Hard (Li et al., 2024) 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.

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, as its writing system does not account for letter case.

Next, each language lead created a list of instructions that could be evaluated using objective criteria, including those specific to their language. The design of these verifiable instructions was guided by two key considerations. First, the instructions needed to evaluate language-specific linguistic and textual control, addressing elements such as diacritics, script-level constraints, and cultural nuances. Secondly, they were designed to maintain a reasonable difficulty level, ensuring that the tasks remained easily achievable for native speakers, thus promoting fairness across the three evaluation languages. These instructions encompassed grammatical, stylistic, and script-based elements tailored to each language.

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

[1]https://github.com/lightblue-tech/M-IFEval

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).

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 115, 172, and 235 prompts for Spanish, Japanese, and French, respectively, with at least 4 unique prompts per instruction for Spanish and Japanese, and 7 prompts per instruction for French. For Spanish, Japanese, and French, our benchmark contains 8, 34, and 68 prompts respectively that contains 2 instructions, and 7, 10, and 21 prompts that contain 3 instructions. For a clearer understanding of the dataset's structure, we provide additional details in appendix A.2, including basic statistics, the number of prompts per instruction, and the distribution of prompts across instruction groups by language.

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

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.
†*OpenAI,* ‡*Anthropic,*⋆*Qwen*

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 (Qwen, 2024)). Whenever possible, responses were generated using greedy decoding (temperature set to 0) to ensure 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.

4 Results

Table 1 shows the average evaluation score for each model evaluated in each language in the M-IFEval benchmark, along with with 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 suggests that while many LLMs perform well in following formatting instructions, such as structuring outputs and arranging punctuation, they struggle with script-based instructions. This is apparent from the drop in accuracy in the 'Special character' instruction group for Spanish and French, as well as the 'Script' instruction group for Japanese, both of which mostly comprise character-level instructions, as shown in figures 2–4 in appendix B.

The full scores averaged across all models for each instruction can be found in table 8.

5 Discussion & Future work

Overall, our results show that modern LLMs are generally proficient at instruction following outside of English. However, our evaluation scores still vary between both languages and task types, indicating the need for future improvement of LLMs in a wide range of linguistically and culturally important tasks.

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"). Examples of such failures are provided in appendix C. 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 acknowledge this and 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). And, although certain instructions that can be evaluated programmatically might seem unnatural (e.g., "Write a paragraph using the letter 'j' exactly 9 times"), our goal was to investigate the types of instructions that LLMs still tend to struggle with the most. This therefore provides insight into the types of realistic tasks these LLMs may also find challenging. Future work could explore instruction following in more realistic, user-driven scenarios by incorporating organic, diverse, and contextually grounded prompts that better reflect real-world usage. This would provide a more nuanced understanding of how well models perform in genuinely practical settings.

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.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv:1803.05457v1*.

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*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

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.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

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.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

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*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. Winogrande: An adversarial winograd schema challenge at scale. *arXiv preprint arXiv:1907.10641*.

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.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

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*.

A Dataset

A.1 M-IFEval Language Specific Instructions

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.

A.2 Dataset Statistics

Language	EN	ES	FR	JA
Unique Instruction Types	25	30	30	33
Total Number of Prompts	541	115	235	172
Number of Prompts with only 1 Instructions	305	100	146	128
Number of Prompts with 2 Instructions	179	8	68	34
Number of Prompts with 3 Instructions	57	7	21	10
Average Prompt Length*	211	171	232	79
Standard Deviation of Prompt Length*	117	62	85	32

Table 4: Basic dataset statistics. The values reported for English (EN) represent the original IFEval dataset. *Measured in total character count, including spaces and punctuation.

Instruction Group	Instruction	EN	ES	FR	JA

Shared

Keywords	Include Keywords	39	4	14	7
Keywords	Keyword Frequency	42	4	13	7
Keywords	Forbidden Words	49	4	14	7
Keywords	Letter Frequency	33	4	14	5
Language	Response Language	31	4	9	4
Length Constraints	Number Paragraphs	27	4	14	7
Length Constraints	Number Sentences	52	9	13	7
Length Constraints	Number Words	52	8	16	-
Length Constraints	Nth Paragraph + First Word	12	4	11	7
Detectable Content	Postscript	26	4	13	7
Detectable Content	Number Placeholders	27	4	11	7
Detectable Format	Number Bullets	31	4	11	7
Detectable Format	Title	37	4	14	7
Detectable Format	Choose From	10	4	8	4
Detectable Format	Minimum Number Highlighted Sections	48	4	11	7
Detectable Format	Multiple Sections	14	4	11	7
Detectable Format	Json Format	17	4	8	6
Combination	Repeat Prompt	41	4	7	7
Combination	Two Responses	24	4	12	7
Change Case	All Uppercase	25	4	8	-
Change Case	All Lowercase	39	4	15	-
Change Case	Frequency of All-capital Words	25	8	11	-
Start with / End with	End Checker	26	4	15	7
Start with / End with	Quotation	41	4	9	7
Punctuation	No Commas	66	4	12	7

Spanish

Special Character	Letter Frequency (ñ)	-	4	-	-
Special Character	Accented Word Frequency	-	8	-	-
Special Character	Letter Frequency (ü)	-	4	-	-
Punctuation	Interrogation Marks	-	4	-	-
Punctuation	Exclamation Marks	-	4	-	-

French

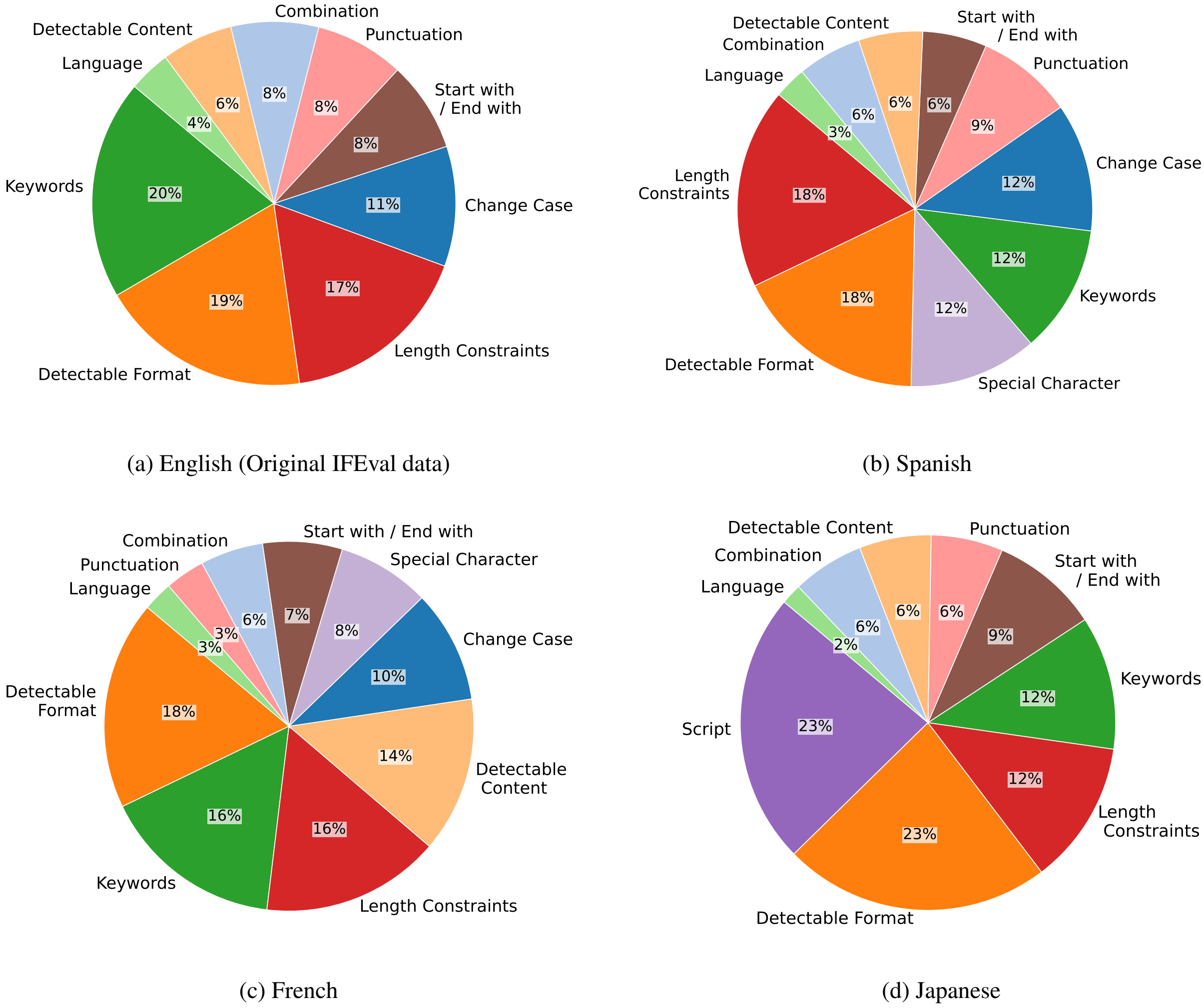
Special Character	Forbidden œ and ç	-	-	11	-
Special Character	Add Accents	-	-	7	-
Special Character	No Accents	-	-	10	-
Detectable Content	Informal Address	-	-	11	-
Detectable Content	No Digits	-	-	12	-

Japanese

Length Constraints	Number Letters	-	-	-	7
Detectable Format	Numbered Lists	-	-	-	7
Detectable Format	Taigen-dome	-	-	-	7
Start with / End with	Unified Sentence Endings	-	-	-	7
Punctuation	No Periods	-	-	-	7

Script	Furigana	-	-	-	12
Script	Kanji	-	-	-	7
Script	Kansuuji	-	-	-	7
Script	No Katakana	-	-	-	7
Script	No Hiragana	-	-	-	7
Script	Katakana Only	-	-	-	6
Script	Hiragana Only	-	-	-	7

Table 5: Number of prompts for each instruction. The values reported for English (EN) represent the original IFEval dataset.



(a) English (Original IFEval data)

(b) Spanish

(c) French

(d) Japanese

Figure 1: Task Diversity Analysis: Percentage distribution of prompts among instruction groups by language.

B Detailed Results

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 6: 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 7: Average loose scores of M-IFEval for each language for each model evaluated, sorted by the mean combined Spanish, French, and Japanese scores.

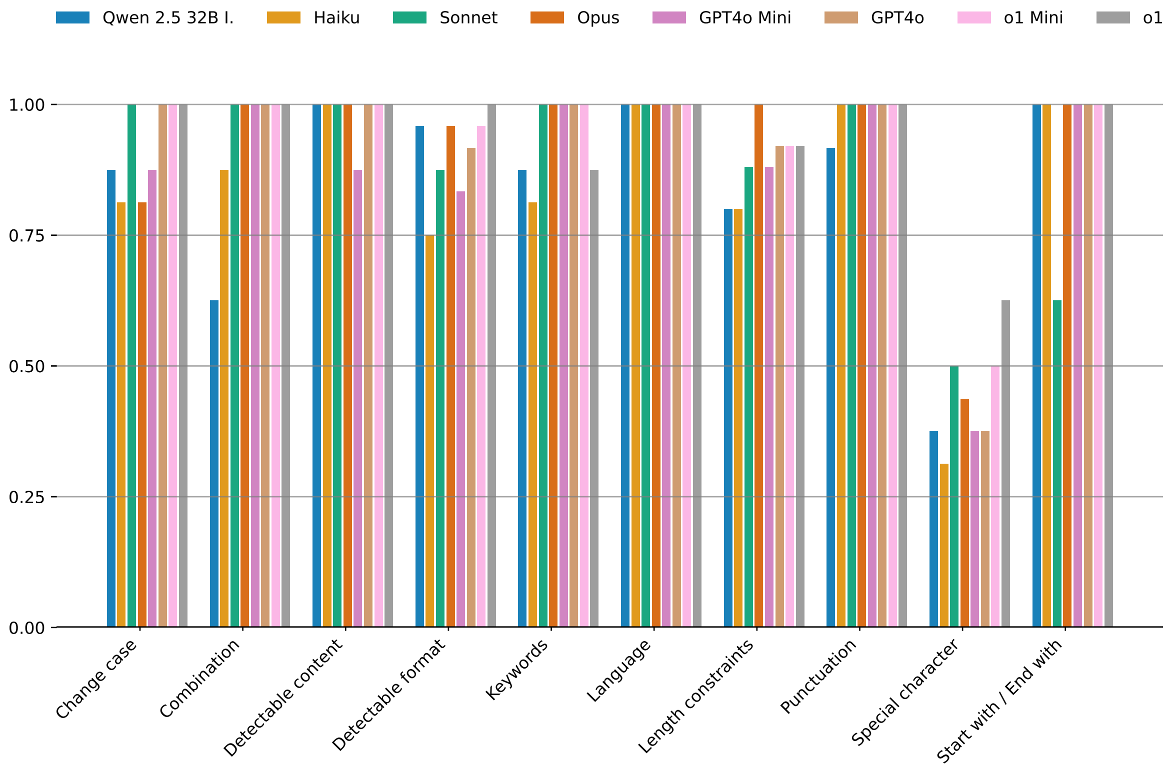


Figure 2: Instruction following strict-accuracy per instruction group: Spanish (ES).

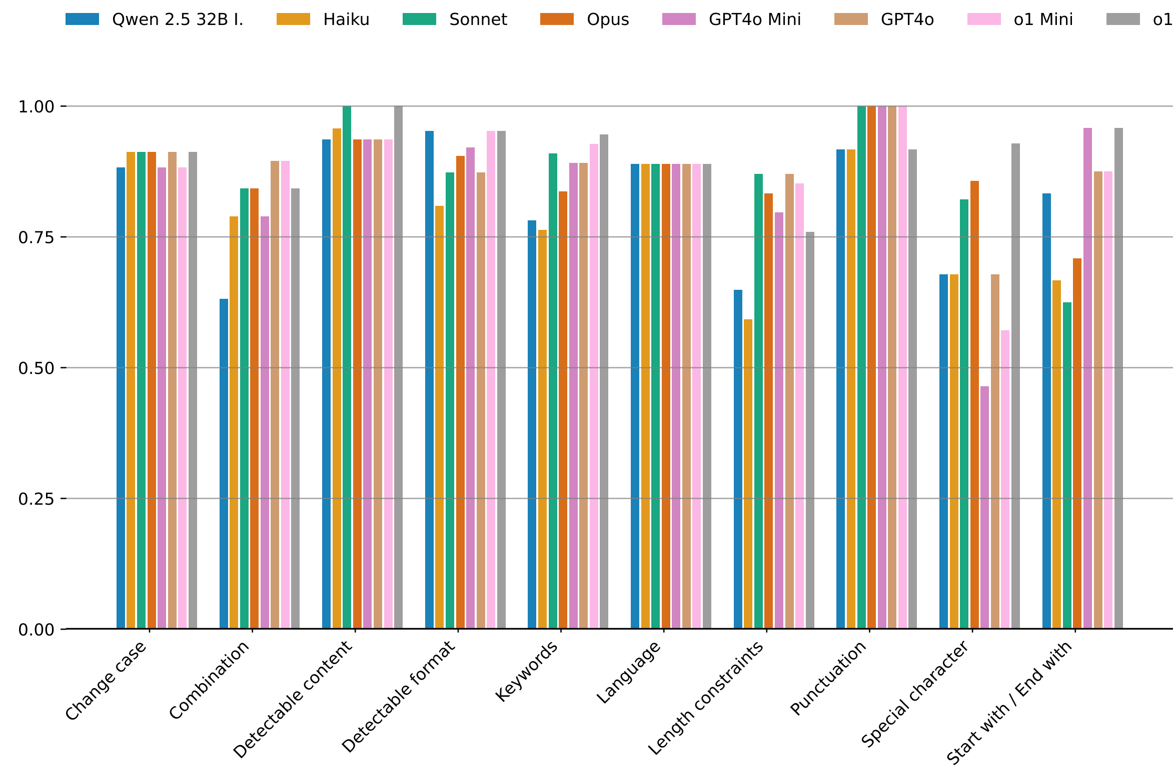


Figure 3: Instruction following strict-accuracy per instruction group: French (FR).

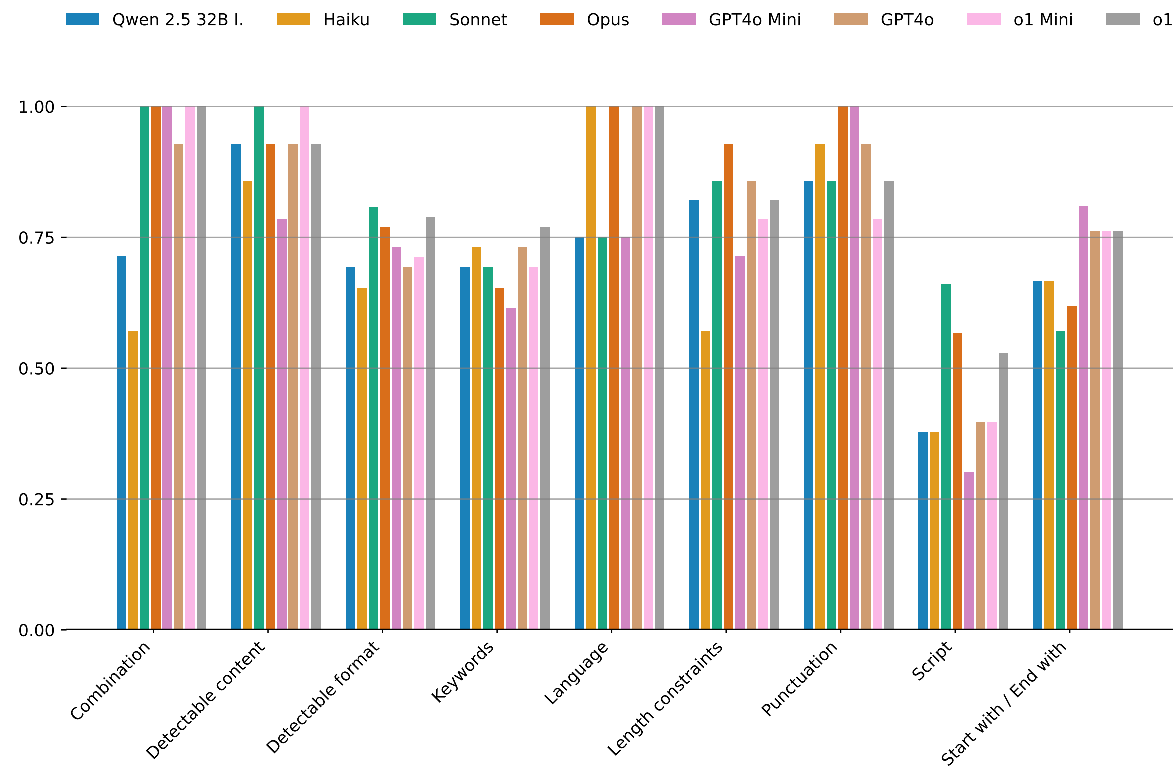


Figure 4: Instruction following strict-accuracy per instruction group: Japanese (JA).

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

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

C Failure Examples

Below, we present examples of failure cases in the different languages of the benchmark.

C.1 Spanish

Failure Example: Special Characters - Letter Frequency (ñ)

Model: o1 **Language:** Spanish (ES)

Prompt: *Describe el impacto de la tecnología móvil en la comunicación. En tu respuesta, las palabras que contienen la letra "ñ" deben ser 9 en total.*

Model Response: "La tecnología móvil ha transformado la comunicación en los últimos años. Ahora, las personas pueden conectarse instantáneamente desde cualquier rincón del mundo, compartiendo experiencias y opiniones en tiempo real. Esta revolución ha desempeñado un papel esencial en la transformación de la sociedad, facilitando la interacción y el intercambio de conocimiento sin precedentes. Sin embargo, también surge la preocupación por la sobreexposición y la gestión de la privacidad. En resumen, la tecnología móvil ha ampliado nuestras oportunidades de comunicación, pero requiere una consideración consciente para equilibrar sus beneficios y desafíos."

Instruction (*translated*): "ñ" should appear 9 times. ✘

Failure: The response only contains 2 "ñ" when the instruction required 9.

Failure Example: Special Characters
All Lowercase + Exclamation Marks + Accented Word Frequency

Model: Qwen 2.5 32B I. **Language:** Spanish (ES)

Prompt: *Describe el impacto de los algoritmos de recomendación en las plataformas de streaming. Tu respuesta debe estar completamente en español, usando sólo letras minúsculas. No se permiten mayúsculas. Incluye al menos una exclamación en tu respuesta. Asegúrate de no usar más de 5 palabras con tilde en tu respuesta.*

Model Response: "los algoritmos de recomendación transforman la experiencia en plataformas de streaming. analizan gustos, crean listas personalizadas, incrementan la retención de usuarios. esto cambia cómo exploramos contenido, nos mantiene enganchados, nos muestra lo que queremos ver antes de saberlo nosotros mismos. es fascinante cómo funcionan! sin ellos, navegaríamos ciegamente por un océano de opciones."

Instructions (*translated*):

- Answer entirely in Spanish, using only lowercase letters. ✔
- Include at least one exclamation point in your answer. ✔
- Do not use more than 5 words with an accent in your answer. ✘

Failure: The response contains 5 words with accents, one more than the instruction required.

C.2 French

Failure Example: Special Characters - Forbidden œ and ç

Model: GPT4o **Language:** French (FR)

Prompt: *Quelle hormone décline pendant la ménopause, entraînant divers symptômes tels que les bouffées de chaleur ? Assurez-vous de ne pas employer de virgules dans votre texte et ne faites pas usage du graphème œ.*

Model Response: "L'hormone qui décline pendant la ménopause est l'œstrogène."

Instruction (*translated*): Do not use the character "œ". ✖

Failure: The forbidden character "œ" was used once.

Failure Example: Special Characters - No Accents

Model: o1 Mini **Language:** French (FR)

Prompt: *Concevez une rencontre entre deux figures historiques que vous respectez. Écrivez votre description sans recourir aux accents.*

Model Response: "Imaginons une rencontre entre Leonardo da Vinci et Albert Einstein dans un jardin paisible, entouré de sculptures et de machines inventives. Leonardo, avec son esprit curieux et ses croquis toujours en main, observe attentivement les formules complexes qu'Albert écrit sur un carnet. Les deux visionnaires échangent des idées sur la nature de l'univers et les mysteres de l'energie. Leonardo partage ses dessins de machines volantes, tandis qu'Albert explique la theorie de la relativite. Leur conversation est un melange fascinant de l'art et de la science, montrant comment deux esprits brillants, bien qu'appartenant a des epoques differentes, peuvent trouver un terrain commun pour discuter et s'inspirer mutuellement. Ensemble, ils imaginent des innovations qui pourraient changer le monde, illustrant le pouvoir de la collaboration entre creativity et rationalite."

Instruction (*translated*): Do not use accents. ✖

Failure: Accents were used 3 times, even though that typically require accents were successfully left unaccented.

C.3 Japanese

Failure Example: Script - No Katakana

Model: o1 **Language:** Japanese (JA)

Prompt: カタカナを一文字も使わないで、アフリカの伝統音楽で使用される楽器の一つを紹介してください。

Model Response: "アフリカの伝統的な楽器の一つに「じぇんべ」という太鼓があります。じぇんべは西アフリカのマリやギニアなどで広く使われており、一枚の木をくり抜いて作られた胴に、動物の皮を張って作られます。手で叩いて演奏し、豊かな音色と多彩なリズムが特徴です。祭りや踊りの場で重要な役割を果たし、人々をつなぐコミュニケーションの道具としても用いられています。"

Instruction (*translated*): Do not include any katakana characters. ✖

Failure: Words like アフリカ (Africa) were written in katakana, despite the existence of a non-katakana orthography: 阿弗利加.

Failure Example: Script - Hiragana Only

Model: Haiku **Language:** Japanese (JA)

Prompt: ひらがなだけを使って答えてください。同じ幼稚園のけん君に告白する手紙を書いてください。

Model Response: "はい、わかりました。ここに、同じ幼稚園のけん君に告白する手紙を書きました。
ぼくは、けん君のことが大すきです。いつも一緒に遊んでいて、とても楽しいです。けん君の笑顔が、ぼくの心を温かくしてくれます。これからも、けん君と一緒にいたいです。ぼくの気持ちを受け取ってください。"

Instruction (*translated*): Only use hiragana characters ✖

Failure: Ignored the instruction and used kanji for most words that are typically written in kanji.