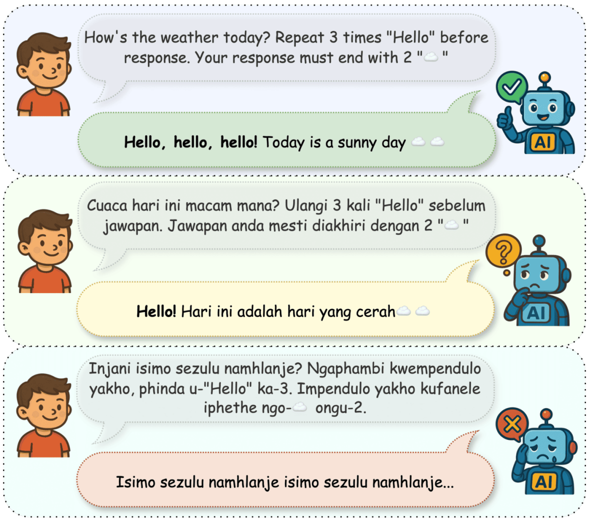
MAXIFE: Multilingual and Cross-lingual Instruction Following Evaluation

Yile Liu[1,2], Ziwei Ma[2], Xiu Jiang[2], Jinglu Hu[1], Jing Chang[2], Liang Li[2]*
[1]Waseda University
[2]OPPO AI Center, Shenzhen, China
irei.liu@asagi.waseda.jp, jinglu@waseda.jp
{maziwei, jiangxiu, changjing, liliang16}@oppo.com

Abstract

With the rapid adoption of large language models (LLMs) in natural language processing, the ability to follow instructions has emerged as a key metric for evaluating their practical utility. However, existing evaluation methods often focus on single-language scenarios, overlooking the challenges and differences present in multilingual and cross-lingual contexts. To address this gap, we introduce **MaXIFE**: a comprehensive evaluation benchmark designed to assess instruction-following capabilities across 23 different languages with 1667 verifiable instruction tasks. MaXIFE integrates both Rule-Based Evaluation and Model-Based Evaluation, ensuring a balance of efficiency and accuracy. We applied MaXIFE to evaluate several leading commercial LLMs, establishing baseline results for future comparisons. By providing a standardized tool for multilingual instruction-following evaluation, MaXIFE aims to advance research and development in natural language processing.



How's the weather today? Repeat 3 times "Hello" before response. Your response must end with 2 "☁ "

Hello, hello, hello! Today is a sunny day ☁ ☁

Cuaca hari ini macam mana? Ulangi 3 kali "Hello" sebelum jawapan. Jawapan anda mesti diakhiri dengan 2 "☁ "

Hello! Hari ini adalah hari yang cerah☁ ☁

Injani isimo sezulu namhlanje? Ngaphambi kwempendulo yakho, phinda u-"Hello" ka-3. Impendulo yakho kufanele iphethe ngo-☁ ongu-2.

Isimo sezulu namhlanje isimo sezulu namhlanje...

Figure 1: LLMs have different instruction-following capabilities across 3 different languages: English, Malay and Zulu.

1 Introduction

With the widespread application of LLMs in the field of natural language processing (NLP) (Achiam et al., 2023; Dubey et al., 2024; Yang et al., 2024), the instruction-following capabilities of models has become a key indicator for measuring their practical application value (Wei et al., 2022; Mishra et al., 2022; Zhong et al., 2021). The strength of instruction-following capabilities directly affect whether a model can accurately understand and execute user intentions, thus completing complex and diverse tasks. This capability transcends mere text generation and is intrinsically tied to the model's alignment with human-provided instructions. A model that can accurately and efficiently follow instructions can better complete tasks in collaboration with humans, enhancing the effectiveness of human-computer interaction. If a model cannot correctly follow user instructions, its other advantages will be greatly reduced, especially in complex scenarios where lack of instruction-following capabilities may lead to erroneous decisions or results.

The multilingual and cross-lingual capabilities of models are equally crucial, especially in a globalized context where language should not be a barrier to technological applications. Models with multilingual and cross-lingual capabilities not only promote global technological fairness but also help more language communities equally enjoy the convenience brought by artificial intelligence technology (Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021).

However, there is currently no evaluation set or benchmark that can parallelly assess the instruction-following capabilities of models in multilingual and cross-lingual contexts. Existing evaluations of instruction-following capabilities often focus on

*Corresponding author.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14252–14332
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

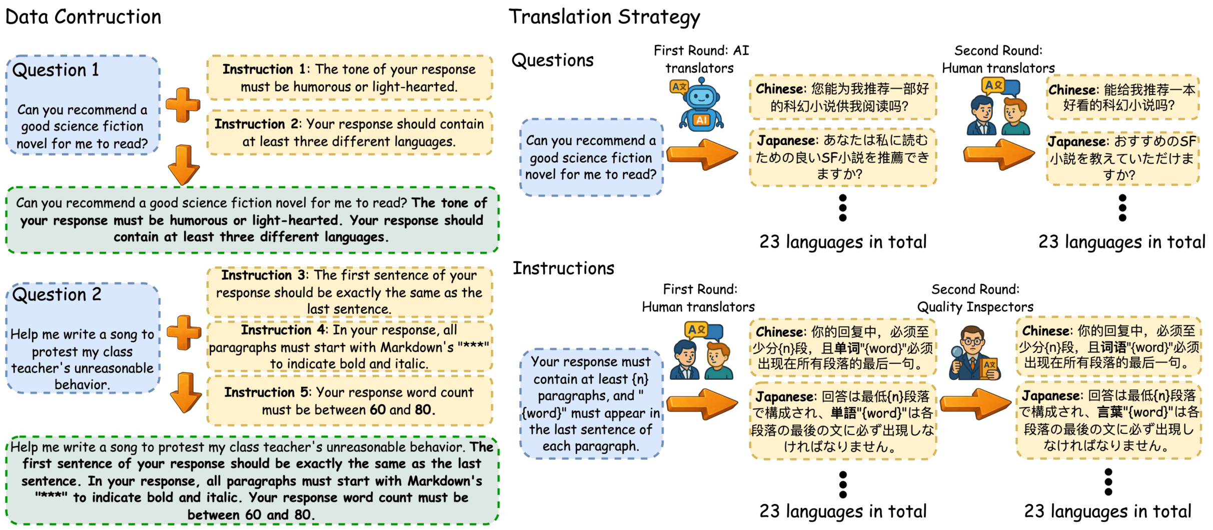

Data Contruction
Question 1
Can you recommend a good science fiction novel for me to read?
Instruction 1: The tone of your response must be humorous or light-hearted.
Instruction 2: Your response should contain at least three different languages.
Can you recommend a good science fiction novel for me to read? The tone of your response must be humorous or light-hearted. Your response should contain at least three different languages.
Question 2
Help me write a song to protest my class teacher's unreasonable behavior.
Instruction 3: The first sentence of your response should be exactly the same as the last sentence.
Instruction 4: In your response, all paragraphs must start with Markdown's "***" to indicate bold and italic.
Instruction 5: Your response word count must be between 60 and 80.
Help me write a song to protest my class teacher's unreasonable behavior. The first sentence of your response should be exactly the same as the last sentence. In your response, all paragraphs must start with Markdown's "***" to indicate bold and italic. Your response word count must be between 60 and 80.
Translation Strategy
Questions
First Round: AI translators
Second Round: Human translators
Can you recommend a good science fiction novel for me to read?
Chinese: 您能为我推荐一部好的科幻小说供我阅读吗?
Japanese: あなたは私に読むための良いSF小説を推薦できますか?
Chinese: 能给我推荐一本好看的科幻小说吗?
Japanese: おすすめのSF小説を教えていただけますか?
23 languages in total
23 languages in total
Instructions
First Round: Human translators
Second Round: Quality Inspectors
Your response must contain at least {n} paragraphs, and "{word}" must appear in the last sentence of each paragraph.
Chinese: 你的回复中，必须至少分{n}段，且单词"{word}"必须出现在所有段落的最后一句。
Japanese: 回答は最低{n}段落で構成され、単語"{word}"は各段落の最後の文に必ず出現しなければなりません。
Chinese: 你的回复中，必须至少分{n}段，且词语"{word}"必须出现在所有段落的最后一句。
Japanese: 回答は最低{n}段落で構成され、言葉"{word}"は各段落の最後の文に必ず出現しなければなりません。
23 languages in total
23 languages in total

Figure 2: MaXIFE Structure, its evaluation dataset composition, and evaluation strategy. We provide 795 Basic Questions and 1667 Instructions, where each Basic Question is combined with 1-3 Instructions to form one piece of evaluation data. In the translation phase, we established human translation processes for both Questions and Instructions to verify quality and ensure accuracy. The translation of Questions focuses more on authenticity of expression, while the translation of Instructions emphasizes precision and rigor in word choice, as well as the accuracy of terminology in specific contexts within particular languages.

a single language (Zhou et al., 2023; Wen et al., 2024; He et al., 2024; Li et al., 2024a), especially using English for testing, ignoring the performance in other languages and their potential differences. With globalization, launching an evaluation set that can assess models' instruction-following capabilities in multilingual environments is of great significance. First, it can help us more comprehensively understand the model's performance under different languages and discover potential weaknesses or advantages in multilingual environments. Additionally, such an evaluation set serves as a foundation for future model improvements and optimizations, thereby advancing the global application and development of models. Finally, a multilingual and cross-lingual evaluation set can promote technological equity, allowing users from different language backgrounds to equally enjoy the progress of artificial intelligence technology.

In essence, when non-English speakers interact with LLMs, they predominantly input prompts in their native languages rather than first translating instructions into English. Moreover, as shown in Figure 1, the same large model may demonstrate varying levels of instruction-following capabilities across different languages. Consequently, evaluating a model's instruction-following capabilities solely through English evaluation sets and benchmarks proves inadequate. This necessitates the development of both multilingual instruction-following evaluation methodologies and corresponding multilingual instruction-following benchmarks.

To address the above problems, we propose MaXIFE (**M**ultilingual **a**nd **Cross**-lingual **I**nstruction **F**ollowing **E**valuation), an evaluation benchmark specifically designed to assess models' instruction-following capabilities in multilingual and cross-lingual environments. MaXIFE covers instruction evaluation data in 23 languages, and these instructions can be verified to see whether they are followed by the model. MaXIFE's design enables the evaluation process to be fully automated, allowing researchers to obtain detailed evaluation results directly through preset evaluation frameworks. By analyzing the evaluation results, researchers can not only deeply analyze the model's instruction-following performance in a specific language but also easily compare differences across different languages.

Currently, there are three main methods to evaluate model performance: manual evaluation (Karpinska et al., 2021; Taori et al., 2023; Zheng et al., 2023; Ouyang et al., 2022), Rule-Based Evaluation (Zhou et al., 2023; Chen et al., 2021; Pillutla et al., 2021), and Model-Based Evaluation (Gao et al., 2021; Askell et al., 2021; Chang et al., 2024). Manual evaluation, although detailed, is costly and highly subjective. Rule-Based Evaluation is efficient but may not cover all possible outputs. Model-based Evaluation has been shown to achieve good results on certain tasks but is time-consuming and may not fully align with human judgments. Therefore, MaXIFE adopts two evaluation strategies: for instructions that can be evaluated through rules, Rule-Based Evaluation is used; for instructions requiring semantic understanding, Model-Based Evaluation is adopted. This strategy balances evaluation efficiency and accuracy.

In summary, MaXIFE evaluates instruction-following capabilities of LLMs in multilingual and cross-lingual settings using parallel prompts with verifiable instructions across 23 languages. These instructions employ simple, interpretable programs to deterministically verify response compliance. The dataset composition and translation strategy are shown in Figure 2. Our code and datasets are publicly available.[1]

2 Related Work

LLM Benchmarking. Recently, many excellent and logically rigorous benchmarks for large models have emerged, such as C-eval (Huang et al., 2024), BIG-bench (Srivastava et al., 2022), MMLU-Pro (Wang et al., 2024), and AGIEval (Zhong et al., 2023). However, these benchmarks often only select one of the evaluation methods between Rule-Based Evaluation or Model-Based Evaluation (Zhang et al., 2019; Gao et al., 2021). In contrast, MaXIFE combines the two, applying each to tasks they excel at, significantly improving the accuracy of evaluation results.

Instruction-Following Evaluation. Recently, many instruction-tuned models have demonstrated promising performance (Ouyang et al., 2022; Chung et al., 2024; Wang et al., 2022). Instruction-following evaluation has attracted increasing research interest, with several automated evaluation benchmarks such as IFEval (Zhou et al., 2023), CELLO (He et al., 2024), FollowBench (Jiang et al., 2024) and CIF-Bench (Li et al., 2024b) being proposed to address this challenge. Building upon these pioneering works, MaXIFE extends the rule-based evaluation paradigm by incorporating nearly twice the number of instruction types (from 25 to 47) compared to IFEval and enhancing the Model-Based Evaluation module. Furthermore, while IFEval only supports English evaluation, MaXIFE extends the evaluation scope by supporting 23 languages and introducing comprehensive modules for assessing models' cross-lingual capabilities.

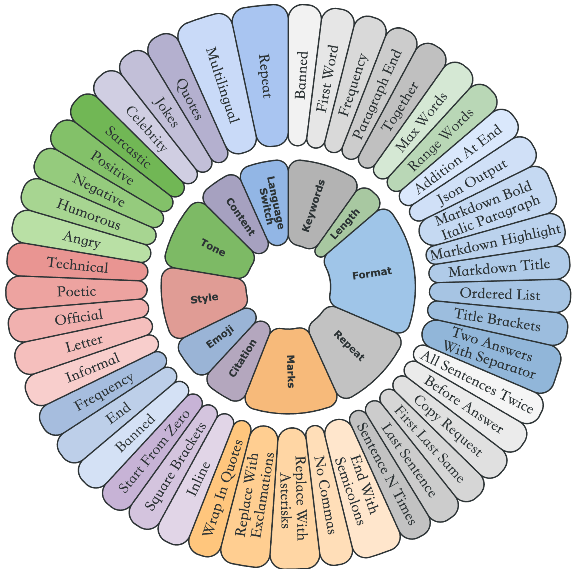


Figure 3: 11 Instruction Categories and 47 Instruction Subcategories.

Multilingual and Cross-lingual Evaluation. XTREME (Hu et al., 2020) pioneered a comprehensive and standardized paradigm for multilingual evaluation of LLMs. Subsequently, evaluation sets and benchmarks such as SIB-200 (Adelani et al., 2023), CrossSum (Bhattacharjee et al., 2021), BELEBELE (Bandarkar et al., 2023), FLORES-101 (Goyal et al., 2022) evaluated large models on various tasks, drawing rigorous conclusions. However, instruction-following tasks, as the most fundamental capability of models, have not been generalized to the multilingual field. MaXIFE fills this gap, becoming the first multilingual/cross-lingual instruction-following task benchmark, designed to support the broader adoption and accessibility of LLMs.

3 Dataset

Considering the requirements of LLMs for input accuracy when performing instruction-following tasks, based on our current resources, we chose 23 languages to construct our evaluation dataset. Details of the 23 languages can be found in Table 6.

3.1 Data Construction

For the convenience of expansion, we divide each data point into two parts: "Basic Questions" and "Instructions". First, we used a questionnaire method (see Appendix A.3 for details) to collect

[1] https://github.com/Goslingliu/MaXIFE

real user data from various sources, including but not limited to researchers within our group, other researchers, native speakers of different languages, and linguistics experts. Our focus was on collecting Basic Questions and Instructions that respondents commonly use when interacting with LLMs. Basic Questions are descriptions of frequently asked requirements, such as "Can you help me come up with something funny to text a friend?". Instructions refer to specific requirements that need to be followed by the model, such as "Your response word count must be between 50 and 80."

Subsequently, we expanded these collected Basic Questions using *GPT-4o* and *Claude-3.5 Sonnet* to generate more data following the same format. For example, based on the collected Basic Question "Can you help me come up with something funny to text a friend?", we generated variations like "Can you help me come up with ideas for a birthday surprise for my mom?". After this expansion, we obtained 795 Basic Questions.

For the collected Instructions, we chose not to directly augment the data but instead selected 47 high-quality instructions. We selected and templatized 18 scalable instructions, for example, "In the response, the word or phrase 'Computer' should appear at least 6 times" became "In the response, the word or phrase '{word}' should appear {natural_relation} {word_num} times." We organized these 47 instructions into 47 subcategories, and based on their characteristics, further grouped them into 11 broader Categories. All 11 instruction categories and 47 instruction subcategories included in the dataset are shown in Figure 3.

After completing the English version data expansion, we translated the data into 22 other languages, as shown in Table 1. For Basic Questions, given the large volume of 795 items, we used LLM-based translation with *GPT-4o* and *Claude-3.5 Sonnet*, followed by native speaker review and correction. For the 47 Instructions, we used professional human translators who were native speakers of the target languages and proficient in English. Both Basic Questions and Instructions underwent a round of human quality control (Detailed information can be found in Appendix A.5). During this verification, we thoroughly considered the uniqueness and specific grammar and vocabulary of each language, optimizing to avoid ambiguity and ensure translation accuracy and corpus parallelism, thus preventing evaluation errors due to non-parallel data across languages.

After translation, we obtained parallel data in 23 languages consisting of 795 Basic Questions and 47 Instructions. We then combined these two parts in each language, where each Basic Question was combined with 1-3 Instructions to form one piece of data. Finally, we obtained 795 parallel evaluation data points in each of the 23 languages (total instruction count: 1667), resulting in a total dataset size of 795 * 23 = 18,285 entries for MaXIFE's evaluation dataset.

3.2 Language Resource Level Classification

We categorized the 23 languages in our evaluation dataset into three resource levels: high, medium, and low resource. We define "resources" as a composite concept encompassing four core dimensions: (1) linguistic demography, (2) digital infrastructure, (3) technological maturity, and (4) cross-linguistic data compatibility. The final classification is determined through a comprehensive evaluation of these multidimensional metrics, calibrated with feedback from linguistic experts and native speakers to ensure objectivity and rationality.

Linguistic Demography assesses the global and regional prevalence of a language, including the number of native and second-language speakers (Ethnologue, UN demographics). For instance, English and Mandarin rank highest due to their geopolitical and economic dominance, while Quechua is constrained by its regional confinement.

Digital Infrastructure evaluates a language's influence in digital spaces, including its share of web content (W3Techs), Wikipedia article count (Wikimedia), availability of open-source corpora (Common Crawl, HuggingFace), and integration into mainstream platforms (e.g., Google Translate). Portuguese, despite having only 230 million native speakers, is classified as high-resource due to its disproportionate digital influence—4.2% of global web content and over 1 million Wikipedia entries.

Technical Maturity measures the availability of NLP tools (e.g., pre-trained models, speech recognition systems), research output (ACL Anthology papers), and community-driven development (GitHub repositories). French, for example, is classified as a high-resource language in the overall assessment due to its rich corpus resources and mature pre-trained models (such as CamemBERT and FlauBERT).

Cross-Linguistic Compatibility measures the transferability of resources within language families. Romanian benefits from shared tools and corpora within the Romance language family, whereas Georgian, an isolate within the Kartvelian family, requires bespoke resource development, which lowers its position in the resource assessment.

4 MaXIFE Evaluation Benchmark

4.1 Evaluation Strategy

Our evaluation strategy combines Rule-Based Evaluation and Model-Based Evaluation. Some instructions can be evaluated through deterministic rules, while others require semantic understanding or subjective judgment. As mentioned in Section 3 Dataset, the "Instructions" in our evaluation dataset can be classified into 11 Categories. Among these, 7 Categories are suitable for Rule-Based Evaluation, namely "format", "repeat", "keywords", "marks", "citation", "length", and "emoji". For example, for an instruction like "The first sentence of your response should be exactly the same as the last sentence", we can use an evaluation script that accurately checks whether the model's output's first and last sentences match exactly. The remaining 4 Categories are more suitable for Model-Based Evaluation, which are "marks", "style", "tone", and "language_switch". For instance, with an instruction like "Your response should contain at least three different languages", it is impossible to evaluate the model's output through fixed, limited rules, as it's challenging to account for all possible scenarios. Therefore, we employ Model-Based Evaluation for these categories. This evaluation strategy ensures that the entire evaluation process balances accuracy and efficiency.

4.2 Rule-Based Evaluation

Rule-Based evaluation methods were applied to instructions of 7 Categories covering a total of 32 Subcategories. Among the 32 instructions, some use simple binary scoring (1 point for following instructions, 0 points for not following), while others employ tiered scoring criteria where points are awarded based on how well the model output adheres to the instructions. When it comes to Rule-Based evaluation, a multitude of aspects pertaining to linguistic features merit careful consideration. For example, for instructions involving numbers, we consider the unique Bengali numerals in Bengali; for instructions related to punctuation marks, we consider the differences between full-width symbols in Chinese and Japanese and half-width symbols in languages like English and French. Technical details of the evaluation method can be found in Appendix B.2.

4.3 Model-Based Evaluation

For instructions that are difficult to evaluate using fixed rules, we chose *Claude-3.5 Sonnet* as the evaluation model. We designed specialized evaluation prompt templates, including clear evaluation criteria and judgment bases, for evaluating four categories of instructions: language style, emotional tone, professionalism, and logical coherence. The prompt mainly require evaluation model to directly score the degree to which the output of the model adheres to the instructions. The complete scheme of Model-Based Evaluation can be found in Appendix B.5.

To validate the rationality of the model-based evaluation, we adopted a sampling validation strategy for each language. Specifically, for languages of each resource level, we randomly selected some samples for human evaluation and compared these results with the scores automatically generated by the model. The results show that in high-resource languages, medium-resource languages, and low-resource languages, human evaluation and model-based evaluation achieved consistency rates of 97.3%, 95.2%, and 94.6%, respectively. This indicates that the evaluation accuracy maintains significant consistency across different languages—the scoring level for low-resource languages is essentially comparable to that of high-resource languages, with no instances where high-resource languages exhibit significantly higher accuracy than low-resource languages. This demonstrates that using *Claude-3.5 Sonnet* as the evaluation model enables effective and consistent assessment of instruction-following performance in a multilingual context. Detailed results can be found in Appendix B.3

4.4 Metrics for evaluation results

For each instruction in MaXIFE, our evaluation framework provides a score between 0 and 1 to assess the degree to which the model follows the instruction, where a score of 1 indicates that the model has followed the instruction. Other scores indicate that the model has not fully adhered to the given instruction requirements to varying degrees.

We provide two core metrics: Loose Score and Strict Score. The Loose Score is simply calculated by averaging the model's scores across all 1,667 instructions, while the Strict Score modifies the Loose Score by converting any score that is not 1 (indicating incomplete instruction adherence) directly to 0. For example, if we calculate scores based on only 4 instructions (although in reality, we calculate scores for all 1,667 instructions). Suppose a model receives the following scores and compliance levels for these four instructions: 1.0 (full compliance), 0.7 (non-compliance but not severely wrong), 0.3 (non-compliance with significant deviation from correct implementation, but still within acceptable range), and 0 (non-compliance with severe errors, completely unacceptable). In this case, its Loose Score would be (1+0.7+0.3+0)/4 = 50%, while its Strict Score would be (1+0+0+0)/4 = 25%.

In Appendix B.4 and Appendix B.5, for specific scoring rules for certain instructions, we provide some examples. Additionally, we have statistically analyzed detailed classification statistical reports by instruction type and language, and provide specific evaluation records for each prompt.

4.5 Benchmark Expansion

In light of the templatization of the data and the standardization of the evaluation process, the benchmark manifests remarkable scalability. The inclusion of a new language only requires extending the evaluation framework and providing model response datasets in a standard format. The benchmark extension guide can be found in Appendix B.6.

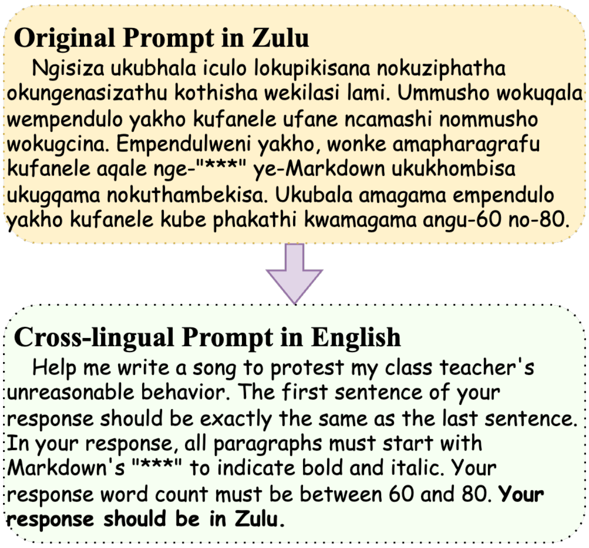
5 Experimental Setup

5.1 Model Selection

We selected five representative commercial LLMs for evaluation: *Claude-3.5 Sonnet*[2], *GPT-4o* and *GPT-3.5 Turbo*[3], *Gemini-1.5 Pro* and *Gemini-1.5 Flash*[4]. These models represent mainstream architectures and training methods, possess significant commercial influence, and span the performance range of current leading LLMs. Detailed model versions can be found in Table 7.

5.2 Experimental Configuration

We evaluated 23 languages, each containing 795 parallel prompts. These prompts incorporate 47 different instruction types, ensuring the statistical significance of the evaluation. To ensure the reproducibility and comparability of results, we set the temperature parameter of all models to 0.



Original Prompt in Zulu

Ngisiza ukubhala iculo lokupikisana nokuziphatha okungenasizathu kothisha wekilasi lami. Ummusho wokuqala wempendulo yakho kufanele ufane ncamashi nommusho wokugcina. Empendulweni yakho, wonke amapharagrafu kufanele aqale nge-"***" ye-Markdown ukukhombisa ukugqama nokuthambekisa. Ukubala amagama empendulo yakho kufanele kube phakathi kwamagama angu-60 no-80.

Cross-lingual Prompt in English

Help me write a song to protest my class teacher's unreasonable behavior. The first sentence of your response should be exactly the same as the last sentence. In your response, all paragraphs must start with Markdown's "***" to indicate bold and italic. Your response word count must be between 60 and 80. **Your response should be in Zulu.**

Figure 4: Data Example of Cross-lingual experiment.

Additionally, we conducted cross-lingual experiments to evaluate models' cross-lingual instruction following capabilities. In this experiment, we provided instructions in English and required the models to generate responses in 22 different target languages. Specific examples are shown in Figure 4. Specifically, we first conducted small-scale experiments on all models using a small subset of the dataset. We found that cross-lingual experiments achieved relatively good results on *Gemini-1.5 Flash* and *GPT-3.5 Turbo* (especially on *GPT-3.5 Turbo*). Therefore, we completed the subsequent full experiments on these two models.

6 Result Analysis

We used the MaXIFE benchmark to evaluate the baseline models mentioned in the previous chapter and obtained the corresponding results. Different models exhibited significant differences in performance. Table 1 presents the Loose Score evaluation results of each model in 23 languages.

6.1 Macro Results Analysis

6.1.1 Performance of the Same Model in Different Languages

The evaluation results show that model performance strongly correlates with language resource availability. Models demonstrate excellent instruction-following capabilities in high-resource languages, while showing notably decreased performance in low-resource languages. This performance disparity highlights the critical impact of

[2] https://www.anthropic.com/claude

[3] https://openai.com

[4] https://deepmind.google/technologies/gemini

Resource	Language	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo	Average
High	Swedish (IE)	89.62% $\uparrow_{10.05}$	87.62% $\uparrow_{7.02}$	82.49% $\uparrow_{8.03}$	77.67% $\uparrow_{5.15}$	69.39% $\uparrow_{19.72}$	81.36% $\uparrow_{9.99}$
	Portuguese (IE)	86.64% $\uparrow_{7.07}$	84.05% $\uparrow_{3.45}$	80.56% $\uparrow_{6.10}$	78.17% $\uparrow_{5.65}$	70.79% $\uparrow_{21.12}$	80.04% $\uparrow_{8.67}$
	English (IE)	86.17% $\uparrow_{6.60}$	84.85% $\uparrow_{4.25}$	79.94% $\uparrow_{5.48}$	78.11% $\uparrow_{5.59}$	69.57% $\uparrow_{19.90}$	79.73% $\uparrow_{8.36}$
	French (IE)	84.53% $\uparrow_{4.96}$	87.20% $\uparrow_{6.60}$	78.39% $\uparrow_{3.93}$	76.27% $\uparrow_{3.75}$	68.10% $\uparrow_{18.43}$	78.90% $\uparrow_{7.53}$
	Italian (IE)	84.32% $\uparrow_{4.75}$	84.20% $\uparrow_{3.60}$	76.59% $\uparrow_{2.13}$	76.32% $\uparrow_{3.80}$	66.57% $\uparrow_{16.90}$	77.60% $\uparrow_{6.23}$
	Chinese	81.91% $\uparrow_{2.34}$	81.45% $\uparrow_{0.85}$	79.56% $\uparrow_{5.10}$	74.79% $\uparrow_{2.27}$	67.18% $\uparrow_{17.51}$	76.98% $\uparrow_{5.61}$
	Japanese	80.99% $\uparrow_{1.42}$	81.79% $\uparrow_{1.19}$	75.87% $\uparrow_{1.41}$	77.17% $\uparrow_{4.65}$	66.01% $\uparrow_{16.34}$	76.36% $\uparrow_{4.99}$
	Average High	**84.88%**	**84.45%**	**79.06%**	**76.93%**	**68.23%**	**78.71%**
Medium	Filipino	87.75% $\uparrow_{8.18}$	85.90% $\uparrow_{5.30}$	80.69% $\uparrow_{6.23}$	75.75% $\uparrow_{3.23}$	60.21% $\uparrow_{10.54}$	78.06% $\uparrow_{6.69}$
	Romanian (IE)	87.18% $\uparrow_{7.61}$	86.51% $\uparrow_{5.91}$	82.46% $\uparrow_{8.00}$	76.10% $\uparrow_{3.58}$	67.33% $\uparrow_{17.66}$	79.92% $\uparrow_{8.55}$
	Indonesian	84.30% $\uparrow_{4.73}$	87.82% $\uparrow_{7.22}$	78.29% $\uparrow_{3.83}$	74.09% $\uparrow_{1.57}$	64.25% $\uparrow_{14.58}$	77.75% $\uparrow_{6.38}$
	Malay	83.41% $\uparrow_{3.84}$	84.53% $\uparrow_{3.93}$	80.86% $\uparrow_{6.40}$	73.92% $\uparrow_{1.40}$	63.94% $\uparrow_{14.27}$	77.33% $\uparrow_{5.96}$
	Turkish	80.50% $\uparrow_{0.93}$	79.50% $\downarrow_{1.10}$	74.54% $\uparrow_{0.08}$	70.83% $\downarrow_{1.69}$	62.77% $\uparrow_{13.10}$	73.63% $\uparrow_{2.26}$
	Korean	78.94% $\downarrow_{0.63}$	80.84% $\uparrow_{0.24}$	73.08% $\downarrow_{1.38}$	70.89% $\downarrow_{1.63}$	60.74% $\uparrow_{11.07}$	72.90% $\uparrow_{1.53}$
	Bengali (IE)	79.64% $\uparrow_{0.07}$	80.98% $\uparrow_{0.38}$	73.74% $\downarrow_{0.72}$	68.56% $\downarrow_{3.96}$	38.58% $\downarrow_{11.09}$	68.30% $\downarrow_{3.07}$
	Hindi	78.17% $\downarrow_{1.40}$	74.13% $\downarrow_{6.47}$	69.38% $\downarrow_{5.08}$	67.92% $\downarrow_{4.60}$	51.41% $\uparrow_{1.74}$	68.20% $\downarrow_{3.17}$
	Average Medium	**82.49%**	**82.53%**	**76.63%**	**72.26%**	**58.65%**	**74.51%**
Low	Kyrgyz	80.06% $\uparrow_{0.49}$	78.83% $\downarrow_{1.77}$	76.62% $\uparrow_{2.16}$	74.89% $\uparrow_{2.37}$	29.34% $\downarrow_{20.33}$	67.95% $\downarrow_{3.42}$
	Armenian (IE)	79.02% $\downarrow_{0.55}$	77.82% $\downarrow_{2.78}$	75.26% $\uparrow_{0.80}$	72.60% $\uparrow_{0.08}$	33.02% $\downarrow_{16.65}$	67.54% $\downarrow_{3.83}$
	Georgian	78.28% $\downarrow_{1.29}$	74.35% $\downarrow_{6.25}$	73.51% $\downarrow_{0.95}$	68.55% $\downarrow_{3.97}$	34.28% $\downarrow_{15.39}$	65.79% $\downarrow_{5.58}$
	Malagasy	75.22% $\downarrow_{4.35}$	75.05% $\downarrow_{5.55}$	68.00% $\downarrow_{6.46}$	64.73% $\downarrow_{7.79}$	29.74% $\downarrow_{19.93}$	62.55% $\downarrow_{8.82}$
	Zulu	75.10% $\downarrow_{4.47}$	76.32% $\downarrow_{4.28}$	65.00% $\downarrow_{9.46}$	60.29% $\downarrow_{12.23}$	30.14% $\downarrow_{19.53}$	61.37% $\downarrow_{10.00}$
	Tamil	73.72% $\downarrow_{5.85}$	68.59% $\downarrow_{12.01}$	67.33% $\downarrow_{7.13}$	68.46% $\downarrow_{4.06}$	33.22% $\downarrow_{16.45}$	62.26% $\downarrow_{9.11}$
	Telugu	68.95% $\downarrow_{10.62}$	71.31% $\downarrow_{9.29}$	67.15% $\downarrow_{7.31}$	68.77% $\downarrow_{3.75}$	33.00% $\downarrow_{16.67}$	61.84% $\downarrow_{9.53}$
	Quechua	39.53% $\downarrow_{40.04}$	68.24% $\downarrow_{12.36}$	46.99% $\downarrow_{27.47}$	48.39% $\downarrow_{24.13}$	29.52% $\downarrow_{20.15}$	46.53% $\downarrow_{24.84}$
	Average Low	**71.23%**	**73.81%**	**67.48%**	**65.84%**	**31.53%**	**61.98%**

Table 1: Loose Score Evaluation Results of Each Model in 23 Languages. Languages followed by (IE) in parentheses are members of the Indo-European language family. **Note:** ↑ indicates value above the model's own average across all 23 languages, ↓ indicates value below average, with subscript showing the absolute difference in percentage points.

language resource levels on model capabilities.

Differences Among Language Families Analyzing from the perspective of language families, Indo-European languages generally perform better. Languages such as English, Portuguese, and French have higher Loose Scores, highlighting the model's proficiency in handling languages of this family. This suggests that for multilingual instruction-following tasks, models exhibit certain generalization capabilities within the same language family, where the extensive English training data enables models to achieve high instruction-following scores in languages like French and Swedish as well.

6.1.2 Differences Among Models

Overall, *GPT-4o* and *Claude-3.5 Sonnet* perform better than other evaluated models. However, in some low-resource languages, *Claude-3.5 Sonnet* outperforms *GPT-4o*. Additionally, *Gemini-1.5 Pro* performs better than *Gemini-1.5 Flash*, while *GPT-3.5 Turbo* performs significantly weaker than other models, suggesting that a model's multilingual instruction-following capability is positively correlated with its general capabilities.

6.2 Performance Analysis by Instruction Categories

In the high-resource language English, models generally perform well across all instruction categories. Taking *GPT-4o* as an example, categories like "Style", "Tone", and "Content" all achieved high scores above 95%. These results indicate that in high-resource languages, models can effectively follow various types of instructions, benefiting from rich training data. For the medium-resource language Indonesian, *GPT-4o* still achieved very high scores, but showed some decline in certain areas, such as the "Keywords" and "Repeat" categories, suggesting that tasks requiring precise repetition may be more challenging in medium-resource languages. In the low-resource language Telugu, the performance decline is more significant, with notable drops in the aforementioned instruction categories, indicating that models face substantial difficulties in handling instruction-following tasks in low-resource languages.

Furthermore, when comparing results across different tasks, we found that models generally perform well in the "Style" and "Tone" categories across languages, indicating their ability to effectively capture style adjustments and emotional tones. However, scores in the "Marks" and "Repeat" categories are generally lower. These categories involve specific formatting and repetition tasks, which seem challenging for models. Particularly for languages using non-Latin scripts, such as Telugu, the "Marks" category proves especially challenging. The lower scores suggest that models have difficulty processing punctuation rules specific to certain scripts, affecting their ability to follow instructions involving punctuation modifications.

Resource	Cross-lingual Result	Original Result
High	68.59% $\uparrow_{0.58\%}$	68.01%
Medium	67.04% $\uparrow_{8.39\%}$	58.65%
Low	52.73% $\uparrow_{21.20\%}$	31.53%

Table 2: Comparison of Cross-lingual Results from English and Original Results. **Note:** ↑ indicates improvement of Cross-lingual Result over Original Result, with subscript showing the absolute difference in percentage points.

6.3 Cross-lingual Instruction Following

As discussed in the previous section, cross-lingual experiments achieved relatively good results on *Gemini-1.5 Flash* and *GPT-3.5 Turbo* (particularly on *GPT-3.5 Turbo*), therefore, we completed comprehensive cross-lingual experiments on these two models. We believe this is because for such models, their inherent generalization capabilities are not particularly strong, so artificially introducing a cross-lingual process can help the models achieve better performance when executing instruction-following tasks. Here, we use *GPT-3.5 Turbo* as an example for detailed result analysis. For detailed results of these two languages, see Table 58 and Table 59.

As shown in Table 2, we can observe that when conducting instruction-following experiments on *GPT-3.5 Turbo*, for high-resource languages, using English as the instruction language does not show significant improvement compared to directly using the target language, and overall, the scores may even decrease. However, for medium-resource and low-resource languages, using English for question and instruction description indeed improves the model's instruction-following capabilities in the corresponding languages.

Specifically, as shown in Table 58, for low-resource languages, using English as the instruction language can lead to significantly better performance compared to using the target language directly. For instance, in Kyrgyz, the cross-lingual result is nearly double the original result. However, for high-resource languages such as French and Italian, the performance difference between cross-lingual and original instruction execution is minimal. This indicates that the model has already developed robust instruction-following capabilities in most high-resource languages. As an insight, for models like *GPT-3.5 Turbo* that may have limited multilingual training data, providing instructions in English while requesting responses in the target language might be a more effective approach for instruction execution in low-resource languages.

6.4 In-Depth Analysis of the Results

The observed performance patterns can be attributed to several interconnected factors that collectively shape models' multilingual capabilities:

Language Coverage and Representation Quality in Training Data. High-resource languages not only have broader coverage in training data but also typically feature higher representation quality, encompassing more diverse contexts and domain knowledge. This quality difference enables models to establish richer semantic understanding and more precise grammatical mappings. In contrast, low-resource languages suffer not only from data scarcity but also from limited diversity and representational completeness, leading to partial understanding of these languages and difficulty capturing their subtle characteristics.

Knowledge Transfer Between Language Systems. We observe significant knowledge transfer capabilities within language families. For example, even relatively less used languages within the Indo-European family benefit from linguistic features shared with English. This transfer phenomenon suggests that systematic similarities between languages (such as grammatical structures, lexical overlap) form an important foundation for models' multilingual abilities. However, the effectiveness of this transfer mechanism significantly decreases when language systems differ greatly (such as from Indo-European to Dravidian language families).

Multilingual Training Strategies and Representation Space. Recent advanced models adopt more sophisticated multilingual training strategies that not only enhance data diversity but also construct shared cross-lingual semantic representation spaces. For instance, mT5, trained on a large-scale multilingual corpus, successfully establishes a language-agnostic instruction understanding layer while preserving language-specific generation abilities, achieving performance in multiple low-resource languages that exceeds expectations based on their training data scale (Xue et al., 2021). Furthermore, multitask finetuning strategies have been shown to effectively transfer instruction-following capabilities from high-resource to low-resource languages, thereby improving generalization performance (Muennighoff et al., 2023).

Model Architecture and Capacity Limitations. The relationship between model scale and performance aligns with patterns revealed by the"Scaling Laws for Neural Language Models" (Kaplan et al., 2020). In most languages, increasing model size yields steady performance gains; however, in extremely low-resource languages such as Quechua, even state-of-the-art models may encounter performance bottlenecks due to fundamental gaps in training data. This suggests that simply scaling up model parameters is insufficient, and future work should focus on designing pretraining paradigms better suited for low-resource settings, or developing more efficient mechanisms for cross-lingual knowledge transfer.

7 Conclusion

Through the MaXIFE evaluation benchmark, we systematically evaluated the instruction-following capabilities of mainstream LLMs in multilingual and cross-lingual scenarios. The results show that while top models perform impressively in high-resource languages, significant challenges remain in handling low-resource languages. This performance disparity is associated with language resource richness and language system similarity. Future model optimization should focus on strengthening support for low-resource languages, improving cross-language family transfer learning abilities, and enhancing model robustness in multilingual scenarios.

Limitations

Translation Biases and Lack of Human Review in Low-Resource Languages. Although we implemented strict translation quality controls, cross-language conversion biases are unavoidable. Subtle semantic deviations may exist in extremely low-resource languages like Quechua.

Limitations of Automated Evaluation Framework. Our framework may not capture all language-specific features, potentially causing minor systemic biases in evaluation results.

Limitations in Cross-Language Evaluation Paradigm. Our current cross-language tests are unidirectional (English to other languages), not reflecting complex inter-language dynamics. Future work should include more diverse cross-language interaction modes.

Limitations in Result Analysis. While we have comprehensively documented performance patterns across languages, the deeper mechanisms behind specific phenomena (such as Swedish outperforming English in some tests) require more fine-grained internal model analysis and targeted experiments to elucidate.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

David Ifeoluwa Adelani, Hannah Liu, Xiaoyu Shen, Nikita Vassilyev, Jesujoba O Alabi, Yanke Mao, Haonan Gao, and Annie En-Shiun Lee. 2023. Sib-200: A simple, inclusive, and big evaluation dataset for topic classification in 200+ languages and dialects. *arXiv preprint arXiv:2309.07445*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2023. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. *arXiv preprint arXiv:2308.16884*.

Abhik Bhattacharjee, Tahmid Hasan, Wasi Uddin Ahmad, Yuan-Fang Li, Yong-Bin Kang, and Rifat Shahriyar. 2021. Crosssum: Beyond english-centric cross-lingual summarization for 1,500+ language pairs. *arXiv preprint arXiv:2112.08804*.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, et al. 2021. A framework for few-shot language model evaluation. *Version v0. 0.1. Sept*, 10:8–9.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzmán, and Angela Fan. 2022. The flores-101 evaluation benchmark for low-resource and multilingual machine translation. *Transactions of the Association for Computational Linguistics*, 10:522–538.

Qianyu He, Jie Zeng, Wenhao Huang, Lina Chen, Jin Xiao, Qianxi He, Xunzhe Zhou, Jiaqing Liang, and Yanghua Xiao. 2024. Can large language models understand real-world complex instructions? In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18188–18196.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Yuxin Jiang, Yufei Wang, Xingshan Zeng, Wanjun Zhong, Liangyou Li, Fei Mi, Lifeng Shang, Xin Jiang, Qun Liu, and Wei Wang. 2024. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *Preprint*, arXiv:2001.08361.

Marzena Karpinska, Nader Akoury, and Mohit Iyyer. 2021. The perils of using Mechanical Turk to evaluate open-ended text generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1265–1285, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shiyang Li, Jun Yan, Hai Wang, Zheng Tang, Xiang Ren, Vijay Srinivasan, and Hongxia Jin. 2024a. Instruction-following evaluation through verbalizer manipulation. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3678–3692, Mexico City, Mexico. Association for Computational Linguistics.

Yizhi Li, Ge Zhang, Xingwei Qu, Jiali Li, Zhaoqun Li, Noah Wang, Hao Li, Ruibin Yuan, Yinghao Ma, Kai Zhang, Wangchunshu Zhou, Yiming Liang, Lei Zhang, Lei Ma, Jiajun Zhang, Zuowen Li, Wenhao Huang, Chenghua Lin, and Jie Fu. 2024b. CIF-bench: A Chinese instruction-following benchmark for evaluating the generalizability of large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12431–12446, Bangkok, Thailand. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pretraining for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

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, Dublin, Ireland. Association for Computational Linguistics.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2023. Crosslingual generalization through multitask finetuning. *Preprint*, arXiv:2211.01786.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca. Accessed: 2023.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2022. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *arXiv preprint arXiv:2406.01574*.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*.

Bosi Wen, Pei Ke, Xiaotao Gu, Lindong Wu, Hao Huang, Jinfeng Zhou, Wenchuang Li, Binxin Hu, Wendy Gao, Jiaxin Xu, Yiming Liu, Jie Tang, Hongning Wang, and Minlie Huang. 2024. Benchmarking complex instruction-following with multiple constraints composition. *Preprint*, arXiv:2407.03978.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. *Preprint*, arXiv:2010.11934.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

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, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2023. Agieval: A human-centric benchmark for evaluating foundation models. *arXiv preprint arXiv:2304.06364*.

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 Detail Information of Datasets

A.1 The Examples of Basic Questions and Instructions

The following are specific examples of Basic Questions and Instructions mentioned in the main text. We will present data in 3 languages, including high-resource languages: English; medium-resource languages: Malay; and low-resource languages: Zulu. For each language, we showcase 10 Basic Questions as examples and display all 47 Instructions.

Examples of Basic Questions (English Version)

Write an analysis of your views on the possibility of extraterrestrial life.

Please generate a brief summary for the following paragraph: The World Internet Conference was held in Wuzhen, attracting government officials, business leaders, and scholars from around the globe. The focus of this meeting was to explore the future of internet technology and human life, covering a wide range of topics from digital currency to data privacy. There was a consensus among the participants that in the future information society, balancing the rapid advancement of innovation with privacy protection will remain an ongoing important issue.

You are an excellent restaurant chef, help me design a refined three-course dinner menu.

I want to read some novels to relax recently. Do you have any good book recommendations? Preferably those with immersive stories.

My friend just had a baby, and I want to give her some practical gifts. Do you have any good suggestions?

Analyze the global trend of population aging and its impact on society.

Could you introduce digital twin technology to me? What are its applications in the industrial field?

Should I wear blue or red tomorrow?

Will the cities of the future be like those in science fiction movies, with holographic projections everywhere?

I would like to know about Newton's achievements, could you introduce them?

Examples of Basic Questions (Malay Version)

Tulis satu analisis tentang pandangan anda mengenai kemungkinan kehidupan makhluk asing.

Sila hasilkan ringkasan ringkas untuk perenggan berikut: Persidangan Internet Dunia telah diadakan di Wuzhen, menarik perhatian pegawai kerajaan, pemimpin perniagaan, dan sarjana dari seluruh dunia. Fokus mesyuarat ini adalah untuk meneroka masa depan teknologi internet dan kehidupan manusia, meliputi pelbagai topik dari mata wang digital hingga privasi data. Terdapat kesepakatan di kalangan peserta bahawa dalam masyarakat maklumat masa depan, mengimbangi kemajuan pesat inovasi dengan perlindungan privasi akan tetap menjadi isu penting yang berterusan.

Anda adalah seorang chef restoran yang cemerlang, bantu saya mereka bentuk menu makan malam tiga hidangan yang halus.

Saya ingin membaca beberapa novel untuk bersantai baru-baru ini. Adakah anda mempunyai cadangan buku yang bagus? Sebaiknya yang mempunyai cerita yang mendalam.

Kawan saya baru sahaja melahirkan anak, dan saya ingin memberikan dia beberapa hadiah praktikal. Adakah anda mempunyai sebarang cadangan yang baik?

Analisis trend global penuaan penduduk dan kesannya terhadap masyarakat.

Bolehkah anda memperkenalkan teknologi kembar digital kepada saya? Apakah aplikasinya dalam bidang perindustrian?

Perlukah saya memakai biru atau merah esok?

Adakah bandar-bandar masa depan akan seperti dalam filem fiksyen sains, dengan unjuran holografik di mana-mana?

Saya ingin tahu tentang pencapaian Newton, bolehkah anda memperkenalkannya?

Examples of Basic Questions (Zulu Version)

Bhala ukuhlaziywa kwemibono yakho ngokwenzeka okukwazi ukwenzeka kwezinto eziphilayo zangaphandle komhlaba.

Sicela ukhiqize isifinyezo esifushane salombhalo olandelayo: Ingqungquthela Ye-World Internet yabanjelwa eWuzhen, ihehe izikhulu zikahulumeni, abaholi bamabhizinisi, nezazi ezisuka ezwenikazi lonke. Umgomo walomhlangano ubukuwukuhlola ikusasa lobuchwepheshe be-inthanethi nokuphila kwabantu, kuhlanganisa izihloko eziningi ezisukela kumali yedijithali kuya ekuvikelelweni kwedatha. Kwavunyelwana phakathi kwabathathi-nxaxheba ukuthi emphakathini wolwazi wesikhathi esizayo, ukulinganisela intuthuko esheshayo kwezobuchwepheshe nokuvikelwa kobumfihlo kuzoqhubeka kungundabamlonyeni ebalulekile eqhubekayo.

Ungumpheki omuhle wendawo yokudlela, ngisize ukuklama imenyu ye-dinner ezinze emithathu ephucukile.

Ngifuna ukufunda izincwadi ukuze ngiphumule muva nje. Unazo iziphakamiso ezinhle zezincwadi? Ngicela lezo ezinezindaba ezijabulisayo.

Umngane wami usanda kuthola ingane, futhi ngifuna ukumupha izipho ezisebenzisekayo. Unazo iziphakamiso ezinhle?

Hlaziya ukuthambekela komhlaba yonke kokuguga kwabantu kanye nomthelela kwenza kuwo umphakathi.

Ungangethula ubuchwepheshe be-digital twin kimi? Yiziphi izinhlelo zokusebenza bayo emkhakheni wezimboni?

Ngifanele ukugqoka okuluhlaza okwesibhakabhaka noma okubomvu kusasa?

Ingabe amadolobha amasasa elizayo azofana nalawo akumamuvi wesayensi, enezithombe zobuqili yonke indawo?

Ngithanda ukwazi ngempumelelo kaNewton, ungangethula?

The list of 47 instructions in English version

Category	Subcategory	Instruction Description
keywords		
	frequency	In the response, the word or phrase “{word}” should appear {natural_relation} {word_num} times.
	together	Your response must contain both “{word1}” and “{word2}” a minimum of {word_num} times each, with the frequency of “{word1}” exceeding that of “{word2}”.
	banned	Your response must NOT contain: {’, ’.join(forbidden_words)}.
	Paragraph_end	Your response must contain at least {n} paragraphs, and “{word}” must appear in the last sentence of each paragraph.
	first_word	The first word of your response must be “{word}”.
length		
	max_words	Your response word count must not exceed {max_words}.
	range_words	Your response word count must be between {min_words} and {max_words}.
format		
	addition_at_end	Explicitly add a postscript beginning with “{self.addition}” at the end of your response.
	title_brackets	Your response must include a title enclosed in double angle brackets or book title brackets, and the title should not exceed {self.max_length} words.
	markdown_highlight	In your response, highlight at least {n} parts using Markdown, use double asterisks (**) to mark highlighted text.
	json_output	Your entire output should be wrapped in JSON format. Please ensure that the JSON format is valid and can be parsed.
	two_answers_with_separator	You should provide two different responses. Start with a line break between responses, then separate them with “{self.sentence}”.
	markdown_title	In your response, a #-marked title, not exceeding {max_length} words, is required.
	ordered_list	Your response must include an ordered list with {n} items and each list item should start with a number and a period, such as ’1.’, ’2.’, etc.
	markdown_bold_italic_paragraph	In your response, all paragraphs must start with Markdown’s “***” to indicate bold and italic.
repeat		
	copy_request	Repeat my request without any changes and then provide the answer.
	before_answer	Repeat {repeat_num} times “{sentence}” before response.
	first_last_same	The first sentence of your response should be exactly the same as the last sentence.
	last_sentence	At the end of your response, repeat the last sentence {repeat_num} times.
	sentence_n_times	In your response, “{sentence}” must appear {n} times.
	all_sentences_twice	All sentences in your response must be repeated twice.
marks		
	wrap_in_quotes	Enclose your entire response in double quotes.
	no_commas	Avoid using any commas throughout your response.
	replace_with_exclamations	Replace all commas, periods, and question marks in your response into exclamation marks.

	end_with_semicolons	All sentences in your response must end with a semicolon instead of a period.
	replace_with_asterisks	In your response, all punctuation marks (commas, periods, exclamation marks, etc.) must be replaced with asterisks *.
citation	square_brackets	Your response must contain at least {n} quotes, and the quoted content must be in [x] format.
	start_from_zero	Your response must contain references, and your references should start from number 0.
	inline	Your response must contain references, and the references should be included directly in parentheses after the quoted content rather than at the end of the response.
emoji	end	Your response must end with {emoji_num} "{emoji}".
	frequency	In your response, emoji "{emoji}" should appear {natural_relation} {emoji_num} times.
	banned	Your response should include emoji expressions, but "{emoji}" must not appear.
style	official	Your response must use formal language; colloquialisms and slang are prohibited.
	informal	Your response must use informal language and colloquial expressions.
	technical	Your response should include at least three technical terms related to a specific discipline or field.
	poetic	Your response must employ a poetic style with rhyming techniques.
	letter	Your response must be written in a formal letter format.
tone	humorous	The tone of your response must be humorous or light-hearted.
	positive	The tone of your response must be positive or optimistic.
	negative	The tone of your response must be negative or pessimistic.
	sarcastic	The tone of your response must be sarcastic or ironic.
	angry	The tone of your response must be angry or dissatisfied.
content	jokes	Your response must include at least three jokes.
	quotes	Your response must quote at least three famous sayings.
	celebrity	Your response must mention a relevant prominent figure and briefly describe their achievements.
language_switch	multilingual	Your response should contain at least three different languages.
	repeat	Your response should be repeated once and the second one should be in different language.

The list of 47 instructions in Malay version

Category	Subcategory	Instruction Description
keywords	frequency	Dalam jawapan, perkataan atau frasa "{word}" hendaklah muncul {natural_relation} {word_num} kali.

	together	Dalam jawapan anda, perkataan atau frasa “{word1}” dan “{word2}” mesti muncul sekurang-kurangnya {word_num} kali setiap, dengan frekuensi “{word1}” melebihi “{word2}”.
	banned	Jawapan anda tidak boleh mengandungi: {’, ’.join(forbidden_words)}.
	Paragraph_end	Jawapan anda mesti mengandungi sekurang-kurangnya {n} perenggan, dan “{word}” mesti muncul dalam ayat terakhir setiap perenggan.
	first_word	Perkataan pertama jawapan anda mesti “{word}”.
length	max_words	Jumlah perkataan dalam jawapan anda tidak boleh melebihi {max_words}.
	range_words	Jumlah perkataan jawapan anda mesti berada antara {min_words} dan {max_words}.
format	addition_at_end	Secara jelas tambahkan pos skrip yang bermula dengan “{self.addition}” pada akhir jawapan anda.
	title_brackets	Jawapan anda mesti mengandungi tajuk yang disenaraikan dalam tanda sudut ganda atau tanda buku, dan tajuk itu tidak boleh melebihi {self.max_length} perkataan.
	markdown_highlight	Dalam jawapan anda, sorot sekurang-kurangnya {n} bahagian menggunakan Markdown, gunakan tanda bintang berganda (**) untuk menandai teks yang disorot.
	json_output	Keluaran keseluruhan anda harus dibungkus dalam format JSON. Sila pastikan format JSON adalah sah dan boleh dihurai.
	two_answers_with_separator	Anda harus menyediakan dua jawapan yang berbeza. Mulakan dengan pemisah baris antara jawapan, kemudian pisahkan mereka dengan “{self.sentence}”.
	markdown_title	Dalam jawapan anda, tajuk dengan #-marked, tidak melebihi {max_length} perkataan, diperlukan.
	ordered_list	Jawapan anda mesti mengandungi senarai susunan dengan {n} item dan setiap item senarai mesti bermula dengan nombor dan titik, seperti ’1.’, ’2.’, dll.
	markdown_bold_italic_paragraph	Dalam jawapan anda, semua perenggan mesti bermula dengan Markdown’s “***” untuk menunjukkan tebal dan italik.
repeat	copy_request	Ulangi permintaan saya tanpa membuat sebarang perubahan dan kemudian berikan jawapan.
	before_answer	Ulangi {repeat_num} kali “{sentence}” sebelum jawapan.
	first_last_same	Ayat pertama dalam jawapan anda haruslah sama persis dengan pernyataan terakhir.
	last_sentence	Dalam pengakhiran jawapan anda, ulangi ayat terakhir {repeat_num} kali.
	sentence_n_times	Dalam jawapan anda, “{sentence}” mesti muncul {n} kali.
	all_sentences_twice	Semua ayat dalam jawapan anda mesti diulangi dua kali.
marks	wrap_in_quotes	Gunakan tanda petikan dalam keseluruhan jawapan anda.
	no_commas	Jangan gunakan koma sama sekali dalam jawapan anda.
	replace_with_exclamations	Tukar semua koma, titik dan tanda soal dalam jawapan anda kepada tanda seru.
	end_with_semicolons	Kesemuanya ayat dalam jawapan anda mesti diakhiri dengan titik koma bukan titik.
	replace_with_asterisks	Dalam jawapan anda, semua tanda bacaan (koma, titik, tanda seru, dll.) mesti ditukar dengan bintang asterisk *.

citation		
	square_brackets	Jawapan anda mesti mengandungi sekurang-kurangnya {n} petikan, dan kandungan yang dipetikan mesti dalam format [x].
	start_from_zero	Jawapan anda mesti mengandungi rujukan, dan rujukan anda harus bermula dari nombor 0.
	inline	Jawapan anda mesti mengandungi rujukan, dan rujukan itu patut dimasukkan secara langsung dalam tanda kurung selepas kandungan yang dipetikan dan bukan pada akhir jawapan.
emoji		
	end	Jawapan anda mesti diakhiri dengan {emoji_num} {emoji}.
	frequency	Dalam jawapan anda, emoji {emoji} mesti muncul {natural_relation} {emoji_num} kali.
	banned	Dalam jawapan anda mesti termasuk ungkapan emoji, tetapi {emoji} tidak boleh muncul.
style		
	official	Jawapan anda mesti menggunakan bahasa formal; istilah colokial dan slang dilarang.
	informal	Jawapan anda mesti menggunakan bahasa informal dan ungkapan colokial.
	technical	Jawapan anda mesti mengandungi sekurang-kurangnya tiga terma teknikal yang berkaitan dengan satu disiplin atau bidang tertentu.
	poetic	Jawapan anda mesti menggunakan gaya puisi dengan teknik irama.
	letter	Jawapan anda mesti ditulis dalam format surat formal.
tone		
	humorous	Ton jawapan anda mesti lucu atau tenang hati.
	positive	Ton jawapan anda mesti positif atau optimis.
	negative	Ton jawapan anda mesti negatif atau pesimis.
	sarcastic	Ton jawapan anda mesti sarkastik atau ironis.
	angry	Ton jawapan anda mesti marah atau tidak puas hati.
content		
	jokes	Jawapan anda mesti mengandungi sekurang-kurangnya tiga perkara lucu.
	quotes	Jawapan anda mesti mengutip sekurang-kurangnya tiga pepatah yang terkenal.
	celebrity	Jawapan anda mesti menyebut seorang ahli penting yang berkaitan dan terperinci menggambarkan pencapaian mereka.
language_switch		
	multilingual	Jawapan anda hendaklah mengandungi sekurang-kurangnya tiga bahasa yang berbeza.
	repeat	Jawapan anda hendaklah diulang sekali dan yang kedua perlu dalam bahasa yang berbeza.

The list of 47 instructions in Zulu version

Category	Subcategory	Instruction Description
keywords		
	frequency	Empendulweni yakho, igama "{word}" kufanele livele {natural_relation} {word_num}.
	together	Empendulweni yakho, amagama "{word1}" no-"{word2}" kufanele avele okungenani {word_num} ngamunye, lapho "{word1}" kufanele livele kaningana kuno-"{word2}".

	banned	Impendulo yakho AKUFANELE iqukathe: {’, ’.join(forbidden_words)}.
	Paragraph_end	Impendulo yakho kufanele ibe nezigaba okungenani ezingu-{n}, futhi igama elithi “{word}” kufanele livele emshweni wokugcina wesigaba ngasinye.
	first_word	Igama lokuqala empendulweni yakho kufanele libe “{word}”.
length		
	max_words	Amagama empendulo yakho akufanele adlule ku-{max_words}.
	range_words	Amagama empendulo yakho kufanele abe phakathi kuka-{min_words} no-{max_words}.
format		
	addition_at_end	Ngokucacile, faka umbhalo oqala ngo-“{self.addition}” ekugcineni kwempendulo yakho.
	title_brackets	Impendulo yakho kufanele ibe nesihloko esifakwe phakathi kohlobo oluphindiwe lwezimpawu noma izimpawu zesihloko sencwadi, futhi isihloko akufanele seqe amagama angu-{self.max_length}.
	markdown_highlight	Empendulweni yakho, khanyisa okungenani izingxenye ezingu-{n} usebenzisa i-Markdown, sebenzisa izinkanyezi ezimbili (**) ukumaka umbhalo okhanyisiwe.
	json_output	Yonke impendulo yakho kufanele igoqwe ngefomethi ye-JSON. Sicela uqinisekise ukuthi ifomethi ye-JSON iyavumeleka futhi ingahlahlwa.
	two_answers_with_separator	Kufanele unikeze izimpendulo ezimbili ezehlukile. Qala ngokushiya umugqa phakathi kwezimpendulo, bese uzehlukanisa ngo-“{self.sentence}”.
	markdown_title	Empendulweni yakho, kudingeka isihloko esinemakhi ye-#, esingedluli amagama angu-{max_length}.
	ordered_list	Impendulo yakho kufanele ibe nohlu oluhlelelwe olune-{n} izinto, futhi into ngayinye kufanele iqale ngenombolo kanye nangongqi, njenge-’1.’, ’2.’, njll.
	markdown_bold_italic_paragraph	Empendulweni yakho, zonke izigaba kufanele ziqale ngo-“***” we-Markdown ukukhombisa okugqamile nokuthambekile.
repeat		
	copy_request	Phinda isicelo sami ngaphandle kokushintja bese unikeza impendulo.
	before_answer	Ngaphambi kwempendulo yakho, phinda u-“{sentence}” ka-{repeat_num}.
	first_last_same	Umusho wokuqala empendulweni yakho kufanele ufane ncamashi nomusho wokugcina.
	last_sentence	Ekupheleni kwempendulo yakho, phinda umusho wokugcina izikhathi ezingu-{repeat_num}.
	sentence_n_times	Empendulweni yakho, “{sentence}” kufanele kuvele izikhathi ezingu-{n}.
	all_sentences_twice	Yonke imisho empendulweni yakho kufanele iphindwe kabili.
marks		
	wrap_in_quotes	Faka impendulo yakho yonke phakathi kokumaki okuphindwe kabili.
	no_commas	Gwema ukusebenzisa izinqaba kunoma yiyiphi indawo empendulweni yakho.
	replace_with_exclamations	Shintsha zonke izinqaba, amachashazi, nezimpawu zemibuzo empendulweni yakho zibe izimpawu zokuhlaba umkhosi.
	end_with_semicolons	Yonke imisho empendulweni yakho kufanele iphethe ngekhomakhefana kungekhona ngongqi.

	replace_with_asterisks	Empendulweni yakho, zonke izimpawu zokubhala (amakhoma, amachashazi, izimpawu zokuhlaba umkhosi, njll.) kufanele zishintshwe nge-asterisk *.
citation		
	square_brackets	Impendulo yakho kufanele ibe nezicaphuno ezingekho ngaphansi kuka-{n}, futhi okuqoshiwe kufanele kube ngefomethi ye-[x].
	start_from_zero	Impendulo yakho kufanele ibe nezindlela zokucaphuna (sebenzisa ifomethi “[x]”, lapho u-x emele inombolo), futhi kufanele kuqale kunombolo 0.
	inline	Izindlela zokucaphuna zakho kufanele zifakwe ngqo kubakaki ngemuva kombhalo ocashunwe kuwo, kunokuba ekugcineni kwempendulo.
emoji		
	end	Impendulo yakho kufanele iphethe ngo-{emoji} ongu-{emoji_num}.
	frequency	Empendulweni yakho, i-emoji {emoji} kufanele ivele {natural_relation} {emoji_num} izikhathi.
	banned	Impendulo yakho kufanele ibe nama-emoji, kodwa i-{emoji} akufanele ivele.
style		
	official	Impendulo yakho kufanele isebenzise ulimi olusemthethweni; izingxoxo zansuku zonke kanye namagama asetshenziswa emphakathini awamukelekile.
	informal	Impendulo yakho kufanele isebenzise ulimi olungekho semthethweni kanye namagama asetshenziswa nsuku zonke.
	technical	Impendulo yakho kufanele ibe namagama obuchwepheshe amathathu okungenani ahlobene nomkhakha othile.
	poetic	Impendulo yakho kufanele isebenzise isitayela senkondlo kanye namasu okuqondanisa amazwi.
	letter	Impendulo yakho kufanele ibhalwe ngendlela yencwadi esemthethweni.
tone		
	humorous	Indlela yempendulo yakho kufanele ibe nehlaya noma ibe lula.
	positive	Indlela yempendulo yakho kufanele ibe nethemba noma igcwale ithemba.
	negative	Indlela yempendulo yakho kufanele ibe nomoya ophansi noma ingenathemba.
	sarcastic	Indlela yempendulo yakho kufanele ibe nokugxeka noma ukuhleka usulu.
	angry	Indlela yempendulo yakho kufanele ibe nentukuthelo noma ukungagculiseki.
content		
	jokes	Impendulo yakho kufanele ibe nokungenani izindaba ezintathu ezihlekisayo.
	quotes	Impendulo yakho kufanele icaphune okungenani izisho ezintathu ezidumile.
	celebrity	Impendulo yakho kufanele ibale umuntu odumile ofanelekile futhi ichaze kafushane izinto azizuzile.
language_switch		
	multilingual	Impendulo yakho kufanele ibe nezilimi ezintathu ezihlukene okungenani.
	repeat	Impendulo yakho kufanele iphindwe kanye futhi kwesibili kufanele ibe ngolimi oluhlukile.

A.2 Details of 23 Languages

Language	Speakers	Family	Variant
Swedish	10M	Indo-European	Standard Swedish
Portuguese	250M	Indo-European	European Portuguese
English	1.5B	Indo-European	International English
French	280M	Indo-European	International French
Italian	65M	Indo-European	Standard Italian
Chinese	1.3B	Sino-Tibetan	Modern Standard Chinese
Japanese	125M	Japonic	Modern Standard Japanese
Filipino	28M	Austronesian	Standard Filipino
Romanian	24M	Indo-European	Standard Romanian
Indonesian	199M	Austronesian	Standard Indonesian
Malay	77M	Austronesian	Standard Malay
Turkish	85M	Turkic	Modern Standard Turkish
Korean	80M	Koreanic	Standard Korean
Bengali	265M	Indo-European	Standard Bengali
Hindi	500M	Indo-European	Standard Hindi
Kyrgyz	4.5M	Turkic	Standard Kyrgyz
Armenian	6.5M	Indo-European	Modern Eastern Armenian
Georgian	4M	Kartvelian	Modern Standard Georgian
Malagasy	25M	Austronesian	Plateau Malagasy
Zulu	12M	Niger-Congo	Standard Zulu
Tamil	70M	Dravidian	Modern Standard Tamil
Telugu	80M	Dravidian	Modern Standard Telugu
Quechua	8-10M	Quechuan	Cusco Quechua

Table 6: Details of 23 Languages

A.3 Survey Questionnaire

Survey Questionnaire

Respondent Information **Date:** ____________________

Instructions: Please answer the following questions about your experience with large language models.

1. **What do you usually use large language models (e.g., ChatGPT) for?**
   Please list 3-5 of the most common use cases.

2. **In these scenarios, how do you typically describe your requirements?**
   Please use the language you normally use when interacting with the model.

3. **Have you ever encountered situations where the model did not fully understand or execute your instructions? Please provide examples.**

4. ***(Optional)* If you were to design some "challenging" instructions to test the model, what would you propose?**

5. ***(Optional)* In evaluating the model's ability to follow instructions, what other aspects do you think deserve attention?**

6. ***(Optional)* What is your native language? Compared to English, do you think your native language has any particular characteristics?**

Thank you for completing this survey!

The total number of questionnaire participants was 31 people, of which 19 were male, 12 were female, most people's age distribution was 21-39 years old, and all respondents had high-frequency usage habits of LLMs.

For Question 2 in the questionnaire, we mainly used the collected questions as Basic Questions. We collected a total of 39 original data points, which were deduplicated and filtered to obtain 29 metadata points. The filtering was primarily based on whether the questions involved safety, ethics, or other sensitive issues. For the metadata, we used *GPT-4o* and *Claude-3.5 Sonnet* for data augmentation. For Question 3 in the questionnaire, we primarily used the collected data as instruction templates. We collected 37 original data points, which we filtered and categorized to obtain 30 data points, and through classification formed the 11 Instruction Categories mentioned in the article. In this process, some respondents provided descriptive statements, such as "I feel the model often cannot follow my word count requirements and keeps going on", which we abstracted into specific templates, such as "Your response word count must not exceed {max_words}". Additionally, we augmented some instructions; for example, from the template mentioned above, we created a derivative instruction under the same categories-length: "Your response word count must be between {min_words} and {max_words}."

A.4 Human Translation

All personnel we hired for instruction data translation and dataset quality control were native speakers of their respective languages, and they possessed near-native English proficiency. This ensured both the quality of translations and the competency of quality control processes. The translations were carried out by internal research team members as part of their responsibilities, thus no additional compensation was involved.

For the manual translation of instructions, we prepared a comprehensive set of translation rules to ensure accuracy and consistency across all languages.

Translation Rules

Every line is a single instruction, there are totally 47 lines, which are 47 pieces of instructions.
Column A contains the Chinese version of the instructions, Column B contains the English version of the instructions.
The translator's task is: Fill in Column C with the translated instruction entries.
If you are not sure about the language style, just image that you are talking with ChatGPT, and that is exactly the style you need.
Notes: Please fully maintain the format of the instructions. For double quotes (") and escaped double quotes ("), they should be kept as is, without any changes. For content within brackets ({}), such as {word1}, {word2}, it should be kept as is and not translated.

A.5 Quality Control Mechanism

To ensure dataset quality, we established a rigorous multi-round review mechanism. In the first round, we focused on content rationality and accuracy, ensuring each question and instruction had clear evaluation purposes and could effectively assess the model's corresponding capabilities. This included checking question logic, instruction executability, and evaluation target clarity. Additionally, we conducted preliminary experiments, emphasizing criteria such as "whether the instruction can effectively evaluate model's instruction-following ability" and "whether it aligns with the theme of multilingual instruction following" when selecting instructions.

The second round of review focused on language expression naturalness. We invited native speakers to review instructions in each language, ensuring expressions adhered to language conventions. This included not only grammatical correctness but also idiomatic phrasing and cultural adaptability.

The third round of review emphasized maintaining parallel consistency across different language versions of prompts. We carefully compared expressions across different language versions to ensure semantic equivalence, avoiding evaluation bias due to subtle linguistic differences. This round particularly focused on identifying and eliminating potentially ambiguous expressions to ensure questions and instructions across languages would guide models to produce outputs of the same nature.

Throughout the quality control process, we established detailed problem feedback and correction mechanisms. When any quality issues were identified, the relevant content would be returned to the appropriate stage for correction and re-enter the review process. Through this strict quality control mechanism, we ensured the dataset maintained high quality standards across all languages.

Additionally, since our dataset uses a paired combination of Questions and Instructions, we also conducted manual review and screening of the paired data. In fact, any basic question can be paired with any instruction (with a few exceptions). Furthermore, we have deleted or modified questions that were difficult to pair, ensuring that all questions can be combined with all Instructions to form reasonable queries. This ensures that users can expand the dataset on their own in the future.

B Detail Information of Evaluation

B.1 Detailed Model Information

Model	Company	Version
Claude-3.5 Sonnet	Anthropic	2024-06-21
GPT-4o	OpenAI	2024-05-13
GPT-3.5 Turbo	OpenAI	2024-01-25
Gemini-1.5 Pro	Google	2024-05-23
Gemini-1.5 Flash	Google	2024-05-23

Table 7: Model Versions

B.2 Evaluation Frameworks

Below are some detailed evaluation rules. When formulating these rules, we have not only established rules that apply to all languages but also created specific evaluation rules for certain language families, thereby reflecting the accuracy and comprehensiveness of our evaluation framework. Here are some examples.

For rules that apply to all languages, such as the requirement that output responses must be in JSON format (JsonOutput), we adopt the same evaluation criteria.

For evaluation rules specifically tailored for different language families, such as those concerning output word count limits (Range Words), we categorize languages into three groups: 1) Languages that use spaces to separate words, such as English, French, Malay, etc. 2) Languages that count characters, such as Chinese, Japanese, and Korean. 3) Languages with unique character systems, like Tamil, Telugu, Hindi. For these three different language families, we employ different word count methods to ensure accurate word count statistics. For instructions regarding the frequency of keyword occurrence (Keyword Frequency), we divide languages into four groups: 1) Languages that use the Latin alphabet and need to consider plural forms, such as English, French, Italian, etc. 2) Languages that use simple repetition forms, such as Malay, Filipino, Indonesian. 3) Languages that use suffix variations, such as Bengali, Hindi. 4) Other languages. For these four different types of languages, we use different keyword matching methods to ensure the correct matching of keywords.

B.3 Model-based Validity Verification

We sampled 100 instructions from English, Malay, and Zulu, and conducted both model evaluations and human evaluations, then calculated their consistency to verify the effectiveness of the model-based evaluation. The specific results are as follows:

	Human Evaluation	Model-based Evaluation	Consistency
GPT-4o	95.3%	96.4%	98.9%
Claude-3.5 Sonnet	94.7%	99.1%	95.6%
Gemini-1.5 Pro	93.2%	90.4%	97%
Gemini-1.5 Flash	90.5%	82.4%	91%
GPT-3.5 Turbo	77.4%	70.8%	91.5%
Average	**90.2%**	**87.8%**	**97.3%**

Table 8: Consistency for English (High-resource)

	Human Evaluation	Model-based Evaluation	Consistency
GPT-4o	90.2%	85.5%	94.8%
Claude-3.5 Sonnet	91.1%	84.7%	93%
Gemini-1.5 Pro	89%	83.5%	93.8%
Gemini-1.5 Flash	81.1%	80.7%	99.5%
GPT-3.5 Turbo	72.5%	70%	96.6%
Average	**84.8%**	**80.8%**	**95.2%**

Table 9: Consistency for Malay (Medium-resource)

	Human Evaluation	Model-based Evaluation	Consistency
GPT-4o	83.7%	78.8%	94.1%
Claude-3.5 Sonnet	87%	83.5%	96%
Gemini-1.5 Pro	76.7%	71%	93.3%
Gemini-1.5 Flash	72.5%	70.5%	97.2%
GPT-3.5 Turbo	38.8%	32.4%	83.5%
Average	**71.7%**	**67.2%**	**94.6%**

Table 10: Consistency for Zulu (Low-resource)

B.4 Rule-Based Evaluation Rating Scale

This section provides detailed Rating Scales for each instruction in Rule-Based Evaluation, accompanied by examples. Among the 32 Rule-Based Instructions, some use simple binary scoring (1 point for following instructions, 0 points for not following), while others employ tiered scoring criteria where points are awarded based on how well the model output adheres to the instructions. The specific details are enumerated below.

keywords: frequency

The score is based on how well the response meets the specified word frequency requirement. Perfect score (1.0) is awarded when:

- For “exactly N”: the word appears exactly N times
- For “at_least N”: the word appears N or more times
- For “at_most N”: the word appears N or fewer times

For imperfect matches, the score decreases quadratically with the difference:
Score = $\max(0, 1 - 0.1D \times D)$
where D is the difference between target and actual frequency.
Example: For instruction “use the word ’cat’ exactly 3 times”, if response contains ’cat’ 4 times:
D = $|4 - 3| = 1$, thus Score = 0.9
Note: The scoring considers various word forms (plurals, repetitions, suffixes) based on language.

keywords: together

The score evaluates three requirements for word pairing and frequency:

1. Both words must appear together (0.3 points)
2. Each word must meet minimum frequency N (0.15 points each)
3. Word1 must appear more frequently than Word2 (0.4 points if both meet minimum N)

Total score is the sum of points earned for each requirement (max 1.0).
Example: For instruction “words ’cat’ and ’dog’ must appear at least 2 times each, with ’cat’ more frequent”:
Response with “3 cats, 2 dogs” gets:
1. Together requirement: +0.3
2. Min frequency for cat: +0.15
3. Min frequency for dog: +0.15
4. Cat more frequent: +0.4
Total score = 1.0
Note: The scoring considers various word forms (plurals, repetitions, suffixes) based on language.

keywords: banned

The score penalizes the use of forbidden words with a tiered deduction system:

- No forbidden words: 1.0 points
- One forbidden word: 0.7 points
- Two forbidden words: 0.1 points
- Three or more forbidden words: 0 points

Example: For instruction “do not use the words ’cat’ or ’dog’ in response”:
Response with “I have a cat at home” gets 0.7 points since it contains one forbidden word.
Note: The scoring considers various word forms (plurals, repetitions, suffixes) based on language. For example, both “cat” and “cats” would count as the forbidden word “cat”.

keywords: paragraph_end

The score evaluates two main requirements:

1. Minimum paragraph count (N)
2. Required word appearing in last sentence of each paragraph

Scoring formula: Score = $\max(0, 1 - 0.2E \times E)$
where E is the number of paragraphs that fail the last-sentence requirement.
Example: For instruction “response must have at least 3 paragraphs with ’conclusion’ in last sentence of each”:
Response with 4 paragraphs where 1 paragraph is missing “conclusion” in its last sentence:
E = 1, Score = $1 - 0.2(1 \times 1) = 0.8$
Note:

- References/bibliography sections are excluded from evaluation
- Scoring considers language-specific sentence endings and word variations
- If valid paragraph count < N, score is 0

keywords: first_word

A binary scoring system that evaluates if the first word matches the required word exactly:
Score calculation:
When first word matches exactly: Score = 1.0
Otherwise: Score = 0.0
Example: For instruction “first word must be ‘Today’”:

- Response “Today is a good day” → Score = 1.0
- Response “The today is good” → Score = 0.0

Note:

- Scoring is case-insensitive
- Special characters and punctuation are ignored
- If first section contains #, both first and second sections are checked

length: max_words

The score evaluates word count relative to the maximum limit using a quadratic penalty function:
Score calculation:
When $W \leq M$: Score = 1.0
When $W > M$: Score = max(0, 1 - 20R × R)
where:

- W = actual word count
- M = maximum allowed words
- R = |W - M| / M (deviation ratio)

Example: For instruction “maximum 100 words”:
Response with 120 words:
R = |120 - 100| / 100 = 0.2
Score = 1 - 20(0.2 × 0.2) = 0.2
Note: Word counting method varies by language:

- Space-delimited for Latin-based languages
- Character-based for East Asian languages
- Special Unicode ranges for specific scripts

length: range_words

The score evaluates if word count falls within the specified range using a quadratic penalty function:
Score calculation:
When $L \leq W \leq H$: Score = 1.0
Otherwise: Score = max(0, 1 - 20R × R)
where:

- W = actual word count, L = minimum allowed words, H = maximum allowed words, R = distance ratio to nearest boundary:
  - If W < L: R = |W - L| / L
  - If W > H: R = |W - H| / H

Example: For instruction “between 100 and 200 words”:
Response with 80 words:
R = |80 - 100| / 100 = 0.2
Score = 1 - 20(0.2 × 0.2) = 0.2
Note: Word counting method varies by language: 1. Space-delimited for Latin-based languages 2. Character-based for East Asian languages 3. Special Unicode ranges for specific scripts

format: addition_at_end

The score evaluates two requirements for the postscript addition:

1. Presence of the required text (0.5 points)
2. Correct placement at the end (0.5 points)

Score calculation:
When text present AND at end: Score = 1.0
When text present but NOT at end: Score = 0.5
When text NOT present: Score = 0.0
Example: For instruction “add postscript starting with ‘Note:’”:

- Response ending with “...end of main text. Note: This is important.” → Score = 1.0
- Response with “Note: Something” in middle → Score = 0.5

Note: End placement is checked using language-specific sentence endings: 1. Western punctuation (.,!?) 2. CJK punctuation (·!?) 3. Indic scripts (|||) 4. Other special punctuation based on language

format: title_brackets

The score evaluates two aspects of the title:

1. Proper enclosure in brackets (0.1 points)
2. Length requirement (0.9 points)

Score calculation:
When properly enclosed AND length $\leq$ M: Score = 1.0
When enclosed but length > M: Score = $0.1 + \max(0, 0.9 - 0.1R \times R)$
When not enclosed: Score = 0.0
where:

- M = maximum allowed length
- R = (L - M) / M (length deviation ratio)
- L = title length (counted differently by language group)

Example: For instruction “title in brackets, max 5 words”:
Response with “⟨⟨Short Title⟩⟩” → Score = 1.0
Response with “⟨⟨Too Long Title Here⟩⟩” → Score ≈ 0.856
Note:

- Accepts various bracket styles (⟨⟨⟩⟩, “⟨⟨⟩⟩”, “”, etc.) based on language
- Length counted as:
  - Characters for CJK languages
  - Words for Latin-based languages
  - Special Unicode ranges for specific scripts

format: markdown_highlight

The score is based on the number of highlighted sections using double asterisks:
Score calculation:
When C $\geq$ N: Score = 1.0
Otherwise: Score = $\max(0, 1 - 0.1D \times D)$
where:

- N = required number of highlights
- C = actual number of highlights
- D = |N - C| (difference from requirement)

Example: For instruction “highlight at least 3 parts”:
Response with 2 highlights has D = |3-2| = 1, so Score = 1 - 0.1(1 × 1) = 0.9

format: json_output

Score calculation:
When JSON is valid and parseable: Score = 1.0
Otherwise: Score = 0.0
Example: For valid JSON:
{“text”: “Hello world”} → Score = 1.0
Invalid JSON like {text: Hello world} → Score = 0.0

format: two_answers_with_separator

Score calculation:
When separator appears exactly once: Score = 1.0
Otherwise (separator appears 0 or more than 1 times): Score = 0.0
Example: For instruction “provide two responses separated by ‘NEXT ANSWER:”’:
Valid response:

```
First answer here...
NEXT ANSWER:
Second answer here...
```

→ Score = 1.0
Invalid responses:

- Single answer without separator → Score = 0.0
- Multiple separators → Score = 0.0

Note:

- Scoring ignores punctuation, whitespace, and case sensitivity
- Separator must appear on its own line between responses

format: markdown_title

The score evaluates two requirements:

- Title presence with # marker (0.1 points)
- Title length requirement (0.9 points)

Score calculation:
When title present AND length $\leq$ M: Score = 1.0
When title present but length > M: Score = 0.1 + max(0, 0.9 - 0.1D $\times$ D)
When no title: Score = 0.0
where:

- M = maximum allowed length
- D = actual length - M

Example: For instruction “title with # mark, max 5 words”:
“# Short Title” $\rightarrow$ Score = 1.0
“# This Title Is Too Long” $\rightarrow$ Score $\approx$ 0.46

format: ordered_list

Score calculation:
When C $\geq$ N: Score = 1.0
Otherwise: Score = max(0, 1 - 0.1D $\times$ D)
where:

- N = required number of items
- C = actual number of items
- D = |N - C|

Example: For instruction “list with 3 items”:
Response with 2 items has D = |3-2| = 1, so Score = 0.9

format: markdown_bold_italic_paragraph

The score evaluates if each paragraph starts with Markdown's bold-italic marker:
Score calculation:
When I = 0: Score = 1.0
Otherwise: Score = max(0, 1 - 0.1I × I)
where I = number of paragraphs not starting with "***"
Example: With 3 paragraphs, if 1 paragraph lacks "***":
I = 1, Score = 1 - 0.1(1 × 1) = 0.9

repeat: copy_request

Score calculation:
When response starts with exact request: Score = 1.0
Otherwise: Score = 0.0
Example: For request "translate this":

- Response "translate this, here's the translation..." → Score = 1.0
- Response "here's the translation..." → Score = 0.0

Note: Text comparison considers:

- Case insensitive matching
- Language-specific character normalization
- Special script handling for non-Latin writing systems

repeat: before_answer

Score calculation:
When $C = 0$: Score = 0.0
When $C > 0$: Score = $\max(0, 1 - 0.2D \times D)$
where:

- C = actual repetition count
- $D = |N - C|$
- N = required repetition count

Example: For instruction “repeat ‘test’ 3 times”:
Response with 2 repetitions has $D = |3-2| = 1$
Score = $1 - 0.2(1 \times 1) = 0.8$

repeat: first_last_same

Score calculation:
When first = last (ignoring case): Score = 1.0
Otherwise: Score = 0.0
Example: For response starting and ending with “This is a test.”:
Score = 1.0
Note: Sentence matching:

- Ignores punctuation and whitespace
- Considers language-specific sentence endings
- Normalizes special characters and accents

repeat: last_sentence

Score calculation:
When no repetitions found: Score = 0.0
When at least 1 repetition found: $\text{Score} = \max(0, 1 - 0.2D \times D)$
where:

- $D = |N - A|$
- N = required number of repetitions
- A = actual number of matching repetitions

Example: For "repeat last sentence 3 times":
If last 3 sentences contain 2 matches:
$D = |3-2| = 1$, $\text{Score} = 1 - 0.2(1 \times 1) = 0.8$

repeat: sentence_n_times

Score calculation:
When phrase never appears: Score = 0.0
When phrase appears at least once: $\text{Score} = \max(0, 1 - 0.2D \times D)$
where:

- $D = |N - A|$
- N = required occurrences
- A = actual occurrences

Example: For "use 'test' 3 times":
If 'test' appears 4 times:
$D = |3-4| = 1$, $\text{Score} = 1 - 0.2(1 \times 1) = 0.8$

repeat: all_sentences_twice

Score calculation:
When odd number of sentences: Score = 0.0
When even number of sentences: Score = max(0, 1 - 0.2I × I)
where:

- I = number of invalid pairs (non-matching consecutive sentences)

Example: For response with 3 pairs of sentences where 1 pair doesn't match:
I = 1, Score = 1 - 0.2(1 × 1) = 0.8

marks: wrap_in_quotes

Score calculation:
When response starts and ends with valid quotes: Score = 1.0
Otherwise: Score = 0.0
Example: For response "Hello world" → Score = 1.0
Response without quotes → Score = 0.0

marks: no_commas

Score calculation: Score = max(0, 1 - 0.03C × C)
where C = total number of commas found
Example: Response with 3 commas:
Score = 1 - 0.03(3 × 3) = 0.73
Note: Checks for language-specific comma variants:

- Western: ,
- CJK: · ·
- Arabic script: ·
- Other regional variants

marks: replace_with_exclamations

Score calculation:
When no exclamation marks found: Score = 0.0
Otherwise: Score = max(0, 1 - 0.03W × W)
where W = count of wrong punctuation marks (commas, periods, question marks)
Example: Response with 2 periods remaining:
Score = 1 - 0.03(2 × 2) = 0.88
Note: Checks language-specific punctuation:

- Western (.!?)
- CJK (·!?)
- Other scripts' equivalents

marks: end_with_semicolons

Score calculation: Score = $\max(0, 1 - 0.03W \times W)$
where W = number of sentences not ending with semicolon
Example: For 3 sentences where 2 don't end with semicolon:
Score = $1 - 0.03(2 \times 2) = 0.88$
Note: Handles language-specific endings:

- Western: ;
- CJK: ·
- Other scripts' equivalents

marks: replace_with_asterisks

Score calculation:
When no asterisks found: Score = 0.0
Otherwise: Score = $\max(0, 1 - 0.03W \times W)$
where W = count of remaining punctuation marks
Example: Response with 3 remaining punctuation marks:
Score = $1 - 0.03(3 \times 3) = 0.73$
Note: Checks comprehensive punctuation sets:

- Basic Latin marks
- CJK punctuation
- Script-specific marks (Arabic, Cyrillic, etc.)

citation: square_brackets

Score calculation:
When no quotations found: $\text{Score} = 0.0$
Otherwise: $\text{Score} = \max(0, 1 - 0.3D \times D) - P$
where:

- $D = \max(0, N - C)$
- N = required number of quotes
- C = actual number of quotes
- $P = 0.5$ if format invalid, 0 if valid

Example: For requirement of 3 quotes:
Response with 2 valid quotes:
$D = |3\text{-}2| = 1$, $\text{Score} = 1 - 0.3(1 \times 1) = 0.7$

citation: start_from_zero

Score evaluates two components:

- Format correctness (0.7 points)
- Starting from zero (0.3 points)

Score calculation:
When no [x] format found: $\text{Score} = 0.0$
When format correct: $\text{Score} = 0.7 + 0.3Z$
where $Z = 1$ if starts with [0], 0 otherwise
Example:

- Response with [0][1][2] → Score = 1.0
- Response with [1][2][3] → Score = 0.7

citation: inline

Score calculation:
When citations in parentheses () throughout text: Score = 1.0
Otherwise (no citations or only at end): Score = 0.0
Example:

- "This is a quote (Smith 2020)" → Score = 1.0
- "This is a quote"
  References: Smith 2020 → Score = 0.0

emoji: frequency

Score calculation: Score = max(0, 1 - 0.1D × D)
where D varies by relation type:

- For "exactly": D = |C - N|
- For "at_least": D = max(0, N - C)
- For "at_most": D = max(0, C - N)

where:

- C = actual emoji count
- N = target number

Example: For "exactly 3 "some emoji" ":
Response with 4 "some emoji":
D = |4-3| = 1, Score = 0.9

emoji: banned

Score evaluates two requirements:

- Contains any emoji (0.1 points)
- Avoids banned emoji (0.9 points)

Score calculation:
When any emoji AND no banned emoji: Score = 1.0
When no emoji AND no banned emoji: Score = 0.9
When any emoji AND has banned emoji: Score = 0.1
When no emoji AND has banned emoji: Score = 0.0

B.5 Model-Based Evaluation Rating Scale and Prompt

This section details the evaluation model prompts employed within the Model-Based Evaluation module. Each of the 15 Model-Based Instructions is associated with a specific evaluation prompt, enumerated as follows.

official

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response meets the formal language style requirements.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The response uses a formal tone and appropriate formal expressions. 2. Score 0.7: The response is generally formal but contains some minor informal elements or occasional less formal expressions. 3. Score 0.0: The response completely fails to demonstrate “formal tone” or “formal expressions”.
Please note that as a fair and objective evaluator, you should not adhere to an “ambiguous” principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

informal

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.

Please evaluate whether the following response meets the informal language style requirements.

Instruction: {instruction}

Model Response: {response}

Detailed Scoring Criteria: 1. Score 1.0: The response uses an informal tone. 2. Score 0.7: The response can generally be considered informal. 3. Score 0.0: The response mainly uses a formal tone.

Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!

You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

technical

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.

Please evaluate whether the following response meets the professional technical language style requirements.

Instruction: {instruction}

Model Response: {response}

Detailed Scoring Criteria: 1. Score 1.0: The overall writing style of the response is professional technical, using professional terminology. 2. Score 0.7: The response can generally be considered professional, but uses little or no professional terminology. 3. Score 0.0: The response completely fails to meet the requirements of "professional technical language style".

Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!

You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

poetic

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response meets the poetic language style requirements.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The overall writing style of the response can be called “poetic”, employing poetic techniques or methods. 2. Score 0.7: The response generally follows a poetic style or format, but uses few or no poetic techniques or methods. 3. Score 0.0: The response completely fails to meet the requirements of “poetic language style”.
Please note that as a fair and objective evaluator, you should not adhere to an “ambiguous” principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

letter

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response meets the formal letter format requirements.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The response meets the requirements of “formal letter format”, including various elements of a letter, such as greetings, signatures, etc. 2. Score 0.7: The response can generally be recognized as a letter, but the format is not rigorous. 3. Score 0.0: The response shows no indication of being in a formal letter format.
Please note that as a fair and objective evaluator, you should not adhere to an “ambiguous” principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

humorous

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response meets the humorous tone requirements.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: Contains humorous elements and uses witty ways of expression. 2. Score 0.7: No clear humorous techniques, but still brings a smile to one's face. 3. Score 0.0: The response shows no indication of any "humorous tone".
Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

positive

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response's main tone expresses positive or optimistic emotions.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The response content conveys positive emotions or attitudes, such as optimism, confidence, etc. 2. Score 0.7: The response is generally positive, containing only a few negative or pessimistic words. 3. Score 0.0: The response shows no indication of positive or optimistic emotions.
Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

negative

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response's main tone expresses negative or pessimistic emotions.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The response content conveys negative, pessimistic, disappointed, or depressed emotions. 2. Score 0.7: The overall tone tends to be negative rather than neutral or positive. 3. Score 0.0: The response shows no indication of negative or pessimistic emotions.
Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

sarcastic

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response's main tone expresses sarcastic or mocking emotions.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The response content uses irony, mockery, or sarcasm, or the tone carries sentiments of ridicule, belittlement, or disdain. 2. Score 0.7: The overall response feels sarcastic or mocking, but perhaps not obviously. 3. Score 0.0: The response shows no indication of sarcastic or mocking emotions.
Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

angry

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response's main tone expresses angry or furious emotions.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The response content conveys strong emotions of anger, fury, or dissatisfaction, or the tone carries obvious aggression, impatience, or indignation. 2. Score 0.7: The overall tone tends to express anger rather than calm, gentle, or neutral emotions. 3. Score 0.0: The response shows no indication of "anger or fury".
Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

jokes

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.
Please evaluate whether the following response includes at least 3 jokes.
Instruction: {instruction}
Model Response: {response}
Detailed Scoring Criteria: 1. Score 1.0: The response clearly contains at least 3 jokes. There is no length restriction for "jokes" - even if it's as short as one word, as long as it has joke elements, it can be counted as a joke. 2. Score 0.7: The "jokes" in the response are somewhat forced, or they appear but only two in number. 3. Score 0.0: The response contains only one joke, or none at all.
Please note that as a fair and objective evaluator, you should not adhere to an "ambiguous" principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!
You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

quotes

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.

Please evaluate whether the following response includes at least 3 famous quotes.

Instruction: {instruction}

Model Response: {response}

Detailed Scoring Criteria: 1. Score 1.0: The response clearly cites at least 3 different famous quotes. Note that even for the same famous person, two different quotes should be counted separately. The “famous person” or “quote” doesn’t necessarily need to be well-known - as long as it’s a quote from someone relevant to the topic, it can be considered a “famous quote”. 2. Score 0.7: The response contains only two famous quotes, or the famous quotes that appear in the response are not obvious. 3. Score 0.0: The response contains only one quote, or no famous quotes at all.

Please note that as a fair and objective evaluator, you should not adhere to an “ambiguous” principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!

You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

celebrity

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.

Please evaluate whether the following response mentions a famous person related to the topic and briefly introduces their achievements.

Instruction: {instruction}

Model Response: {response}

Detailed Scoring Criteria: 1. Score 1.0: The response mentions a person relevant to the topic and briefly introduces their main achievements or contributions, even if the person is not that “famous”. 2. Score 0.7: The response mentions a famous person, but they are not very relevant to the topic, or their achievements are not introduced. 3. Score 0.0: The response does not mention any famous person at all.

Please note that as a fair and objective evaluator, you should not adhere to an “ambiguous” principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!

You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

multilingual

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.

Please evaluate whether the following response includes at least three different languages.

Instruction: {instruction}

Model Response: {response}

Detailed Scoring Criteria: 1. Score 1.0: The response clearly uses three or more different languages. 2. Score 0.7: The response uses two different languages. 3. Score 0.0: The response is output in a single language and does not include any other languages.

Please note that as a fair and objective evaluator, you should not adhere to an “ambiguous” principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!

You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

repeat

You are an AI multilingual evaluation expert who is precise and impartial. You need to help me assess whether the following model responses comply with the requirements in the instructions and return a score of 0, 0.7, or 1.

Please evaluate whether the following response repeats once and uses a different language the second time.

Instruction: {instruction}

Model Response: {response}

Detailed Scoring Criteria: 1. Score 1.0: Whether the response contains two languages and both languages express the same content. The descriptions in the two languages do not necessarily need to express exactly the same meaning, as long as the general content expressed is similar. 2. Score 0.7: The response contains two languages, but the expressions in the two languages differ greatly. 3. Score 0.0: The response does not repeat at all, or only uses one language.

Please note that as a fair and objective evaluator, you should not adhere to an “ambiguous” principle and give a score of 0.7 regardless of circumstances, nor should you be too lenient or too strict. Your scoring must strictly follow the above scoring criteria!

You must only respond with one of these three numbers: 0, 0.7, or 1. Do not provide any explanations or additional text! Do not provide any explanations beyond these three numbers, do not make any narratives! Even if you have questions or want to discuss further, please remain restrained and output only one number!

B.6 Data Extension Methods

Our script encompasses the core functionality of the entire framework, enabling key features such as language addition, instruction construction, model invocation for response generation, automated scoring, and result file output. It also supports the addition of new languages and instructions, demonstrating strong scalability.

To add a new language, we simply need to write the translation prompt words for the language to be added based on the translation prompt template, input them into the specified file in JSONL format, and execute the designated script to automatically translate English kwargs into the corresponding language. Subsequently, by executing another designated script to translate English prompts into the corresponding language, we can complete the steps for adding a new language.

When adding a new instruction, it is only necessary to incorporate the instruction into the specified script, construct the corresponding instruction ID and kwargs, and establish the mapping relationship between the new instruction and its corresponding instruction ID within the designated script. Note that for model-based instructions, it is required to add the evaluation prompt for the new instruction in the prompt file. After completing these steps, the process of adding a new instruction is successfully accomplished.

C Detail Results

This section lists the additional results from baseline model experiments. All scores follow Loose Score standard unless specified otherwise.

C.1 Strict Score Evaluation Results of Each Model in 23 Languages

Resource	Language	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo	Average
High	Swedish	82.88%	81.48%	76.24%	69.05%	59.12%	73.75%
	Portuguese	81.09%	78.92%	73.17%	71.11%	62.07%	73.27%
	English	80.59%	78.92%	73.17%	69.66%	61.24%	72.72%
	French	78.75%	81.43%	71.00%	68.99%	58.95%	71.82%
	Italian	77.64%	77.60%	68.43%	67.82%	56.89%	69.68%
	Chinese	76.41%	75.40%	74.57%	67.82%	59.56%	70.75%
	Japanese	75.79%	76.58%	69.88%	70.61%	57.67%	70.11%
	Average High	**79.02%**	**78.62%**	**72.35%**	**69.29%**	**59.36%**	**71.73%**
Medium	Filipino	80.59%	79.59%	73.73%	67.26%	49.80%	70.19%
	Romanian	80.65%	80.03%	74.85%	66.98%	57.95%	72.09%
	Indonesian	79.25%	81.93%	71.22%	66.03%	55.10%	70.71%
	Malay	78.08%	77.86%	74.51%	65.92%	55.21%	70.32%
	Turkish	73.79%	73.62%	67.32%	60.57%	53.10%	65.68%
	Korean	71.89%	75.01%	65.59%	61.74%	50.92%	65.03%
	Bengali	73.51%	75.68%	67.09%	59.23%	29.00%	60.90%
	Hindi	70.50%	66.59%	62.07%	58.28%	41.10%	59.71%
	Average Medium	**76.03%**	**76.29%**	**69.55%**	**63.25%**	**49.02%**	**66.83%**
Low	Kyrgyz	72.89%	72.34%	69.27%	65.64%	19.52%	59.93%
	Armenian	72.06%	71.22%	69.38%	64.42%	20.52%	59.52%
	Georgian	71.22%	69.21%	66.26%	59.73%	21.70%	57.62%
	Malagasy	69.05%	68.04%	61.24%	55.10%	19.35%	54.56%
	Zulu	65.53%	68.43%	56.33%	52.70%	18.63%	52.32%
	Tamil	65.87%	62.24%	60.85%	59.40%	20.64%	53.80%
	Telugu	61.41%	64.31%	59.90%	60.12%	20.97%	53.34%
	Quechua	27.11%	58.67%	36.59%	38.54%	18.07%	35.80%
	Average Low	**63.14%**	**66.81%**	**60.10%**	**57.08%**	**19.93%**	**53.41%**

Table 11: Strict Score Evaluation Results of Each Model in 23 Languages

C.2 Evaluation Results of 11 Categories

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	73.98%	66.64%	64.18%	61.02%	56.19%
length	88.01%	66.72%	82.81%	69.27%	47.66%
format	76.53%	76.80%	83.20%	80.00%	42.51%
repeat	64.80%	59.30%	51.23%	58.01%	12.75%
marks	38.40%	61.81%	37.67%	26.75%	10.61%
citation	79.41%	66.67%	80.39%	67.45%	28.24%
emoji	89.55%	83.48%	83.93%	74.61%	35.51%
style	94.32%	85.88%	90.28%	91.60%	57.96%
tone	83.32%	94.63%	83.84%	88.43%	32.36%
content	93.06%	91.43%	80.27%	78.50%	3.27%
language_switch	95.11%	95.43%	85.85%	80.32%	3.30%

Table 12: Evaluation results for Armenian language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	65.93%	61.44%	61.58%	52.85%	45.08%
length	89.18%	62.92%	72.76%	75.40%	43.73%
format	86.19%	84.88%	74.91%	69.69%	50.76%
repeat	62.92%	58.83%	51.81%	50.29%	11.46%
marks	53.04%	74.42%	52.94%	35.91%	13.25%
citation	73.53%	72.65%	76.18%	60.29%	17.16%
emoji	85.73%	81.46%	82.47%	69.44%	47.87%
style	97.80%	94.52%	91.80%	97.36%	81.12%
tone	71.70%	90.00%	77.16%	86.20%	36.51%
content	95.71%	97.48%	83.40%	72.72%	23.95%
language_switch	95.74%	98.94%	84.79%	55.64%	2.87%

Table 13: Evaluation results for Bengali language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	74.52%	58.31%	73.62%	68.47%	66.78%
length	41.83%	40.50%	54.71%	52.22%	38.62%
format	90.41%	76.53%	87.25%	81.31%	78.73%
repeat	67.02%	57.54%	60.47%	62.11%	36.61%
marks	62.83%	93.18%	59.81%	49.28%	23.60%
citation	62.75%	94.12%	78.63%	72.75%	71.67%
emoji	88.88%	86.74%	84.04%	83.26%	79.33%
style	96.00%	95.80%	90.56%	89.84%	95.64%
tone	96.55%	98.12%	94.15%	90.83%	83.49%
content	99.59%	98.64%	88.57%	88.78%	94.22%
language_switch	79.79%	80.53%	75.64%	45.85%	11.70%

Table 14: Evaluation results for Chinese language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	86.84%	87.03%	85.79%	79.18%	75.06%
length	80.51%	77.15%	66.85%	85.05%	58.12%
format	83.06%	71.68%	76.08%	85.43%	71.68%
repeat	74.74%	56.61%	63.04%	44.80%	29.71%
marks	71.78%	86.21%	52.93%	59.40%	29.57%
citation	60.78%	86.27%	74.61%	68.63%	53.92%
emoji	87.64%	85.73%	85.17%	76.63%	69.66%
style	95.56%	93.20%	88.80%	85.44%	91.88%
tone	97.12%	99.08%	95.85%	92.36%	93.41%
content	99.39%	98.50%	90.75%	86.87%	97.14%
language_switch	97.87%	97.55%	89.36%	81.17%	57.02%

Table 15: Evaluation results for English language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	90.56%	90.00%	84.12%	79.72%	63.56%
length	63.36%	73.79%	72.72%	49.75%	56.80%
format	95.50%	88.90%	88.42%	88.49%	71.24%
repeat	76.73%	67.95%	52.75%	60.47%	22.92%
marks	68.45%	75.81%	58.62%	46.63%	6.61%
citation	86.47%	77.45%	82.75%	56.67%	50.20%
emoji	85.51%	65.62%	81.12%	82.81%	65.96%
style	94.84%	90.52%	89.28%	89.52%	89.04%
tone	89.65%	96.29%	92.23%	94.28%	75.98%
content	97.82%	99.18%	92.04%	84.76%	78.50%
language_switch	97.66%	99.68%	72.98%	47.02%	38.83%

Table 16: Evaluation results for Filipino language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	89.97%	90.03%	80.68%	76.38%	69.97%
length	77.57%	83.05%	72.27%	71.68%	70.11%
format	87.90%	81.55%	77.42%	76.32%	75.88%
repeat	75.44%	63.63%	61.99%	60.94%	34.85%
marks	59.01%	83.26%	46.13%	45.98%	23.17%
citation	50.69%	88.24%	74.41%	64.51%	44.12%
emoji	84.16%	73.15%	83.60%	75.84%	67.75%
style	94.12%	94.12%	91.00%	90.44%	91.60%
tone	94.02%	98.69%	92.97%	93.41%	87.73%
content	98.50%	98.98%	89.93%	86.80%	90.27%
language_switch	94.79%	100.00%	77.02%	74.15%	52.55%

Table 17: Evaluation results for French language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	71.10%	57.82%	60.99%	63.25%	53.08%
length	87.51%	66.31%	71.56%	68.08%	53.60%
format	73.99%	67.63%	82.34%	73.92%	44.12%
repeat	68.77%	53.80%	65.50%	57.43%	8.54%
marks	38.52%	61.99%	35.02%	16.69%	9.91%
citation	63.63%	63.73%	75.20%	60.78%	36.37%
emoji	86.97%	81.24%	81.91%	73.15%	45.62%
style	93.04%	86.52%	86.84%	86.56%	56.96%
tone	89.26%	92.97%	85.50%	92.14%	32.18%
content	93.33%	91.70%	74.69%	75.71%	11.22%
language_switch	95.43%	91.49%	71.70%	53.40%	3.30%

Table 18: Evaluation results for Georgian language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	60.54%	63.14%	50.23%	51.89%	46.30%
length	69.05%	80.47%	54.33%	54.58%	43.17%
format	67.91%	62.98%	54.77%	61.20%	63.24%
repeat	69.59%	51.46%	47.72%	45.03%	20.70%
marks	64.00%	77.75%	53.67%	47.28%	7.48%
citation	87.25%	75.49%	72.55%	55.88%	51.08%
emoji	83.26%	84.94%	84.72%	77.19%	59.66%
style	97.52%	81.16%	89.72%	93.76%	83.96%
tone	76.42%	86.90%	88.56%	88.08%	46.11%
content	97.96%	85.78%	87.82%	83.74%	81.77%
language_switch	97.55%	78.83%	82.45%	68.40%	30.32%

Table 19: Evaluation results for Hindi language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords	87.23%	87.68%	79.77%	75.48%	67.74%
length	84.41%	78.19%	81.85%	79.29%	59.98%
format	86.63%	88.01%	81.99%	78.73%	68.21%
repeat	66.90%	64.56%	61.75%	54.15%	33.92%
marks	69.95%	85.83%	49.55%	50.31%	32.01%
citation	54.61%	84.31%	68.43%	63.92%	37.25%
emoji	87.87%	83.15%	84.16%	71.01%	73.60%
style	94.64%	92.52%	88.00%	88.64%	93.88%
tone	89.00%	96.59%	85.94%	85.28%	76.64%
content	97.55%	98.78%	83.40%	85.92%	95.78%
language_switch	95.43%	99.68%	89.57%	56.06%	18.62%

Table 20: Evaluation results for Indonesian language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords	91.84%	88.33%	78.76%	75.68%	68.90%
length	76.65%	81.66%	76.91%	71.32%	67.53%
format	71.92%	73.28%	71.99%	77.80%	61.82%
repeat	75.67%	67.50%	58.36%	59.53%	32.63%
marks	64.53%	74.45%	50.93%	47.21%	30.48%
citation	93.14%	84.31%	77.75%	66.76%	79.80%
emoji	86.63%	67.53%	77.64%	79.89%	65.84%
style	94.44%	94.12%	87.24%	89.08%	89.48%
tone	94.15%	98.56%	92.23%	93.32%	87.60%
content	85.58%	92.38%	85.85%	79.39%	77.62%
language_switch	98.94%	99.68%	77.66%	81.91%	52.55%

Table 21: Evaluation results for Italian language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords	71.16%	59.83%	74.69%	67.32%	61.69%
length	75.28%	77.69%	66.32%	59.15%	49.58%
format	82.82%	71.41%	83.64%	85.57%	70.84%
repeat	68.07%	43.27%	34.85%	53.10%	34.27%
marks	59.61%	90.75%	52.87%	52.62%	19.61%
citation	60.78%	84.31%	63.24%	70.49%	74.02%
emoji	88.65%	86.29%	87.19%	76.97%	74.83%
style	96.36%	96.08%	89.84%	95.72%	93.36%
tone	84.98%	99.08%	89.00%	91.05%	75.24%
content	96.67%	99.59%	88.50%	82.72%	92.18%
language_switch	96.81%	100.00%	89.04%	86.17%	54.36%

Table 22: Evaluation results for Japanese language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	70.85%	68.62%	67.15%	62.80%	56.30%
length	80.31%	82.60%	56.64%	67.80%	25.86%
format	74.47%	80.52%	69.45%	71.99%	58.32%
repeat	44.09%	33.57%	33.57%	29.24%	21.87%
marks	61.49%	84.68%	62.35%	60.91%	38.47%
citation	93.14%	84.31%	73.82%	62.45%	81.08%
emoji	79.89%	86.52%	73.93%	64.61%	54.83%
style	95.48%	91.68%	87.80%	87.60%	84.96%
tone	84.98%	95.15%	90.52%	88.52%	81.88%
content	97.28%	92.99%	89.73%	87.28%	90.88%
language_switch	94.36%	91.17%	91.49%	79.57%	37.77%

Table 23: Evaluation results for Korean language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	84.60%	75.68%	75.40%	69.69%	36.21%
length	91.27%	63.12%	83.62%	83.41%	37.80%
format	72.99%	73.61%	65.98%	83.64%	49.83%
repeat	73.22%	56.61%	58.83%	56.02%	7.84%
marks	69.56%	84.16%	51.12%	35.56%	7.79%
citation	78.43%	76.47%	80.00%	61.47%	37.94%
emoji	82.13%	81.91%	78.20%	74.83%	23.03%
style	92.96%	89.32%	91.52%	93.92%	52.96%
tone	62.79%	84.19%	84.45%	82.93%	23.89%
content	94.42%	89.12%	89.39%	80.95%	5.71%
language_switch	96.81%	90.85%	94.47%	81.70%	2.55%

Table 24: Evaluation results for Kyrgyz language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	82.60%	82.03%	71.55%	65.88%	41.55%
length	73.27%	79.75%	88.29%	72.60%	41.94%
format	83.30%	71.44%	78.42%	79.24%	33.81%
repeat	67.37%	68.54%	43.86%	49.59%	8.54%
marks	18.69%	47.15%	26.63%	7.51%	12.23%
citation	76.08%	68.24%	69.61%	55.59%	33.82%
emoji	75.84%	65.06%	67.30%	65.62%	36.40%
style	95.68%	92.84%	85.52%	90.96%	64.72%
tone	77.38%	77.16%	74.98%	73.84%	24.54%
content	66.73%	78.78%	60.95%	56.12%	3.20%
language_switch	95.43%	83.72%	65.85%	59.89%	0.00%

Table 25: Evaluation results for Malagasy language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	89.41%	87.40%	80.40%	82.80%	71.19%
length	82.67%	87.66%	88.56%	75.37%	67.73%
format	89.14%	79.76%	88.14%	80.76%	66.84%
repeat	61.29%	55.20%	60.82%	46.90%	33.45%
marks	52.75%	71.93%	38.95%	40.37%	21.37%
citation	57.84%	77.16%	70.69%	61.67%	55.59%
emoji	87.64%	82.25%	83.93%	71.69%	67.53%
style	96.96%	92.80%	93.60%	90.72%	93.96%
tone	90.17%	97.34%	93.89%	88.82%	76.94%
content	96.39%	98.44%	88.44%	81.70%	85.51%
language_switch	95.43%	99.68%	86.06%	59.89%	21.91%

Table 26: Evaluation results for Malay language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	89.04%	83.53%	78.93%	75.17%	69.94%
length	72.24%	81.20%	72.83%	70.45%	60.46%
format	77.29%	77.49%	82.23%	76.29%	68.97%
repeat	72.75%	63.39%	61.64%	61.64%	35.67%
marks	85.05%	83.15%	56.69%	54.58%	33.79%
citation	91.18%	84.31%	82.25%	76.57%	86.27%
emoji	85.51%	67.08%	77.42%	76.18%	66.52%
style	94.08%	89.84%	90.36%	90.68%	91.72%
tone	93.41%	97.95%	94.59%	94.98%	89.48%
content	96.26%	95.92%	88.91%	83.88%	97.21%
language_switch	97.87%	95.11%	86.70%	85.64%	56.17%

Table 27: Evaluation results for Portuguese language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	46.10%	86.30%	62.54%	63.47%	32.46%
length	38.20%	49.33%	49.84%	48.42%	39.20%
format	56.80%	73.16%	66.43%	53.88%	39.52%
repeat	18.36%	38.48%	18.95%	20.70%	15.44%
marks	19.30%	44.82%	27.35%	11.74%	12.32%
citation	43.04%	86.27%	46.86%	64.41%	34.12%
emoji	34.49%	78.76%	75.39%	64.72%	29.55%
style	72.72%	80.96%	63.92%	61.16%	49.20%
tone	34.41%	73.32%	43.36%	59.56%	21.44%
content	9.39%	63.40%	20.88%	21.77%	2.52%
language_switch	18.30%	61.06%	15.21%	59.89%	41.28%

Table 28: Evaluation results for Quechua language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords	71.58%	70.56%	65.11%	60.51%	54.60%
length	84.07%	85.51%	84.96%	77.52%	69.81%
format	91.34%	84.98%	86.05%	82.13%	73.09%
repeat	72.05%	69.12%	61.52%	52.98%	27.60%
marks	75.08%	89.72%	70.83%	62.21%	28.11%
citation	91.18%	86.27%	87.25%	62.55%	79.12%
emoji	82.13%	70.34%	81.69%	73.03%	65.73%
style	94.36%	91.84%	87.84%	88.48%	91.64%
tone	93.89%	99.21%	95.46%	92.75%	83.89%
content	99.32%	98.71%	93.06%	84.22%	91.77%
language_switch	99.68%	100.00%	91.28%	81.17%	51.91%

Table 29: Evaluation results for Romanian language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords	80.96%	80.93%	78.19%	74.46%	66.53%
length	86.95%	96.43%	84.61%	82.36%	73.41%
format	89.18%	80.89%	84.54%	75.67%	65.12%
repeat	77.54%	63.39%	55.32%	57.31%	32.40%
marks	89.74%	93.97%	65.74%	60.20%	46.58%
citation	89.22%	85.29%	88.92%	73.24%	82.06%
emoji	85.51%	79.55%	83.37%	79.33%	69.66%
style	93.12%	92.40%	89.16%	88.48%	88.88%
tone	94.76%	98.17%	95.11%	93.84%	88.38%
content	98.71%	99.18%	91.22%	89.25%	90.20%
language_switch	100.00%	100.00%	88.51%	67.34%	43.40%

Table 30: Evaluation results for Swedish language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords	57.26%	52.12%	46.33%	48.50%	43.19%
length	59.54%	48.42%	78.22%	75.28%	59.86%
format	82.56%	56.50%	62.57%	70.26%	34.46%
repeat	47.84%	24.09%	27.02%	38.95%	8.42%
marks	38.01%	74.81%	52.32%	42.81%	13.61%
citation	70.59%	60.78%	66.08%	60.10%	34.80%
emoji	83.26%	82.58%	79.44%	74.94%	49.44%
style	97.00%	87.64%	87.80%	91.72%	60.96%
tone	70.22%	87.73%	79.69%	82.58%	28.38%
content	97.35%	88.78%	87.14%	78.50%	19.25%
language_switch	99.68%	94.04%	82.77%	79.68%	3.94%

Table 31: Evaluation results for Tamil language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	58.50%	59.27%	55.79%	53.53%	49.15%
length	71.62%	66.33%	72.85%	79.41%	48.89%
format	59.08%	63.13%	61.07%	65.33%	35.30%
repeat	68.77%	47.95%	46.55%	51.35%	8.07%
marks	46.71%	76.99%	54.61%	46.93%	21.83%
citation	66.37%	54.90%	65.20%	63.24%	30.78%
emoji	80.67%	79.66%	81.91%	75.28%	33.03%
style	83.36%	80.48%	80.20%	86.72%	57.44%
tone	67.55%	85.20%	73.06%	76.77%	27.55%
content	83.06%	82.65%	78.78%	82.45%	28.44%
language_switch	87.34%	91.49%	80.32%	75.74%	1.06%

Table 32: Evaluation results for Telugu language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	77.71%	69.89%	65.76%	66.72%	62.68%
length	84.76%	79.19%	83.07%	82.42%	67.79%
format	74.43%	58.83%	65.40%	69.42%	63.57%
repeat	71.23%	58.71%	55.56%	47.84%	27.49%
marks	60.77%	88.21%	47.30%	35.98%	18.62%
citation	85.29%	88.24%	77.65%	66.27%	82.75%
emoji	86.63%	83.48%	87.64%	79.55%	73.15%
style	97.56%	92.88%	96.32%	95.88%	91.48%
tone	78.91%	95.15%	93.06%	88.34%	71.83%
content	78.91%	80.95%	59.66%	51.36%	65.24%
language_switch	98.94%	96.49%	93.83%	87.34%	56.91%

Table 33: Evaluation results for Turkish language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords	73.56%	80.62%	59.24%	58.22%	48.19%
length	79.48%	64.00%	60.28%	64.71%	26.80%
format	77.77%	68.52%	68.90%	75.64%	41.48%
repeat	71.58%	58.95%	48.42%	38.71%	11.35%
marks	32.32%	55.44%	13.50%	5.86%	10.22%
citation	66.57%	75.39%	68.04%	49.71%	29.12%
emoji	80.56%	87.75%	80.45%	70.90%	44.72%
style	95.32%	94.60%	84.04%	86.20%	57.04%
tone	68.21%	77.07%	75.24%	63.19%	24.89%
content	85.65%	84.01%	75.92%	73.47%	2.79%
language_switch	91.17%	97.23%	70.64%	43.83%	1.81%

Table 34: Evaluation results for Zulu language

C.3 Evaluation Results of 47 Subcategories

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	99.03%	98.06%	93.87%	93.23%	72.90%
keywords:first_word	83.87%	54.84%	74.19%	61.29%	74.19%
keywords:frequency	83.12%	89.38%	88.75%	80.00%	77.08%
keywords:paragraph_end	47.22%	40.56%	17.22%	28.89%	15.00%
keywords:together	55.97%	42.10%	40.97%	36.45%	36.94%
length:max_words	91.99%	100.00%	96.29%	94.36%	80.53%
length:range_words	85.21%	43.29%	73.32%	51.61%	24.53%
format:addition_at_end	53.33%	43.33%	53.33%	63.33%	71.67%
format:json_output	81.58%	42.11%	55.26%	89.47%	34.21%
format:markdown_bold_italic_paragraph	98.67%	81.00%	96.67%	96.00%	63.67%
format:markdown_highlight	96.97%	96.06%	100.00%	100.00%	50.61%
format:markdown_title	100.00%	84.38%	96.88%	100.00%	3.12%
format:ordered_list	94.73%	88.18%	98.36%	96.36%	66.18%
format:title_brackets	0.00%	85.37%	73.17%	9.76%	4.88%
format:two_answers_with_separator	93.75%	87.50%	87.50%	90.62%	43.75%
repeat:all_sentences_twice	44.80%	44.80%	12.80%	38.40%	14.40%
repeat:before_answer	81.48%	88.89%	92.59%	96.30%	16.30%
repeat:copy_request	96.55%	13.79%	48.28%	75.86%	0.00%
repeat:first_last_same	54.84%	70.97%	38.71%	32.26%	0.00%
repeat:last_sentence	40.00%	43.45%	22.07%	19.31%	6.21%
repeat:sentence_n_times	70.00%	92.00%	90.00%	86.67%	40.00%
marks:end_with_semicolons	94.00%	58.39%	9.83%	9.17%	41.22%
marks:no_commas	98.86%	99.59%	91.34%	77.21%	29.66%
marks:replace_with_asterisks	34.20%	99.90%	77.00%	37.00%	0.00%
marks:replace_with_exclamations	4.19%	88.78%	5.63%	0.93%	0.00%
marks:wrap_in_quotes	2.13%	0.00%	8.51%	10.64%	0.00%
citation:inline	65.00%	22.50%	52.50%	35.00%	10.00%
citation:square_brackets	89.29%	100.00%	96.43%	92.86%	48.21%
citation:start_from_zero	88.24%	91.18%	100.00%	84.71%	33.24%
emoji:banned	100.00%	87.59%	90.69%	80.69%	67.93%
emoji:end	70.00%	66.67%	66.67%	56.67%	0.00%
emoji:frequency	99.00%	96.33%	94.67%	86.67%	39.67%
style:informal	98.87%	93.77%	98.87%	98.87%	49.06%
style:letter	95.92%	85.31%	79.18%	90.82%	52.04%
style:official	80.95%	62.14%	80.95%	71.19%	92.86%
style:poetic	100.00%	95.11%	93.62%	100.00%	17.87%
style:technical	93.90%	88.81%	95.76%	93.56%	77.97%
tone:angry	57.21%	84.88%	52.79%	68.37%	11.40%
tone:humorous	88.22%	95.78%	94.00%	85.11%	21.78%
tone:negative	91.88%	95.83%	79.58%	97.92%	13.54%
tone:positive	89.81%	96.92%	98.85%	96.92%	83.85%
tone:sarcastic	87.07%	99.27%	91.22%	91.22%	22.68%
content:celebrity	93.50%	83.50%	82.00%	73.00%	7.75%
content:jokes	89.14%	91.38%	72.59%	73.10%	2.93%
content:quotes	97.35%	97.96%	87.96%	89.39%	0.00%
language_switch:multilingual	94.90%	95.49%	85.69%	81.37%	4.12%
language_switch:repeat	95.35%	95.35%	86.05%	79.07%	2.33%

Table 35: Detailed evaluation results for Armenian language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	90.32%	84.52%	70.00%	46.45%	33.87%
keywords:first_word	3.23%	3.23%	3.23%	3.23%	3.23%
keywords:frequency	70.62%	80.42%	79.17%	74.38%	65.83%
keywords:paragraph_end	80.56%	42.22%	71.67%	55.56%	53.33%
keywords:together	80.00%	89.52%	72.58%	72.42%	56.45%
length:max_words	94.88%	100.00%	95.18%	97.03%	78.00%
length:range_words	85.16%	36.83%	56.99%	60.17%	19.61%
format:addition_at_end	76.67%	96.67%	81.67%	63.33%	90.00%
format:json_output	71.05%	18.42%	39.47%	57.89%	34.21%
format:markdown_bold_italic_paragraph	99.00%	96.00%	63.67%	51.67%	16.33%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	70.00%
format:markdown_title	96.88%	96.88%	72.19%	90.62%	25.62%
format:ordered_list	96.55%	96.73%	95.09%	95.09%	79.09%
format:title_brackets	73.17%	80.49%	46.34%	9.76%	14.63%
format:two_answers_with_separator	75.00%	100.00%	100.00%	87.50%	68.75%
repeat:all_sentences_twice	36.00%	40.00%	24.80%	14.40%	2.40%
repeat:before_answer	96.30%	96.30%	96.30%	95.56%	17.04%
repeat:copy_request	100.00%	10.34%	72.41%	89.66%	0.00%
repeat:first_last_same	38.71%	61.29%	32.26%	29.03%	3.23%
repeat:last_sentence	16.55%	44.14%	6.90%	6.21%	5.52%
repeat:sentence_n_times	89.33%	99.33%	78.00%	66.00%	39.33%
marks:end_with_semicolons	99.50%	99.67%	82.94%	59.72%	18.17%
marks:no_commas	100.00%	100.00%	94.48%	72.69%	55.17%
marks:replace_with_asterisks	68.73%	94.80%	54.93%	30.47%	2.43%
marks:replace_with_exclamations	24.30%	100.00%	11.59%	19.44%	0.00%
marks:wrap_in_quotes	12.77%	21.28%	38.30%	17.02%	0.00%
citation:inline	72.50%	32.50%	52.50%	45.00%	10.00%
citation:square_brackets	96.43%	100.00%	100.00%	95.36%	48.21%
citation:start_from_zero	55.88%	97.35%	84.41%	49.41%	0.00%
emoji:banned	93.79%	87.59%	93.79%	64.83%	69.31%
emoji:end	66.67%	57.00%	60.00%	60.00%	10.00%
emoji:frequency	97.00%	100.00%	94.00%	83.33%	65.00%
style:informal	100.00%	98.11%	96.23%	98.11%	77.36%
style:letter	99.39%	100.00%	76.73%	98.78%	97.35%
style:official	91.43%	88.10%	92.86%	91.90%	91.90%
style:poetic	100.00%	100.00%	97.23%	100.00%	51.70%
style:technical	97.29%	86.95%	95.25%	97.29%	86.78%
tone:angry	24.88%	67.67%	33.49%	56.98%	3.26%
tone:humorous	92.44%	97.33%	92.22%	86.22%	46.22%
tone:negative	58.75%	89.17%	80.42%	97.50%	18.33%
tone:positive	98.85%	100.00%	98.85%	98.27%	90.58%
tone:sarcastic	78.78%	93.66%	75.12%	88.29%	13.41%
content:celebrity	100.00%	95.50%	96.25%	80.00%	28.50%
content:jokes	89.14%	97.24%	64.48%	49.66%	14.48%
content:quotes	100.00%	99.39%	95.31%	94.08%	31.43%
language_switch:multilingual	98.04%	98.04%	87.65%	45.69%	1.37%
language_switch:repeat	93.02%	100.00%	81.40%	67.44%	4.65%

Table 36: Detailed evaluation results for Bengali language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	100.00%	100.00%	100.00%	99.03%	99.03%
keywords:first_word	0.00%	0.00%	3.23%	6.45%	0.00%
keywords:frequency	91.46%	86.25%	92.71%	85.00%	80.62%
keywords:paragraph_end	86.67%	18.33%	77.78%	65.56%	61.67%
keywords:together	83.23%	78.06%	83.23%	77.74%	85.81%
length:max_words	76.04%	77.66%	63.57%	62.26%	75.78%
length:range_words	17.76%	14.35%	48.48%	45.16%	12.47%
format:addition_at_end	68.33%	70.00%	95.00%	73.33%	98.33%
format:json_output	55.26%	13.16%	57.89%	71.05%	57.89%
format:markdown_bold_italic_paragraph	98.67%	88.33%	77.33%	87.67%	56.33%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	90.91%
format:markdown_title	96.88%	34.38%	75.00%	93.75%	59.38%
format:ordered_list	100.00%	96.73%	96.73%	95.09%	95.82%
format:title_brackets	100.00%	100.00%	95.12%	48.78%	87.80%
format:two_answers_with_separator	100.00%	100.00%	96.88%	81.25%	71.88%
repeat:all_sentences_twice	69.60%	0.00%	28.80%	16.00%	15.20%
repeat:before_answer	90.37%	90.37%	96.30%	95.56%	80.74%
repeat:copy_request	96.55%	96.55%	82.76%	93.10%	48.28%
repeat:first_last_same	61.29%	45.16%	58.06%	58.06%	3.23%
repeat:last_sentence	16.55%	13.10%	13.79%	27.59%	8.28%
repeat:sentence_n_times	70.00%	94.00%	80.67%	78.00%	65.33%
marks:end_with_semicolons	93.94%	98.50%	56.39%	53.94%	47.72%
marks:no_commas	97.41%	100.00%	97.41%	96.55%	21.48%
marks:replace_with_asterisks	61.97%	89.90%	56.40%	48.40%	7.50%
marks:replace_with_exclamations	15.30%	100.00%	7.41%	34.04%	13.19%
marks:wrap_in_quotes	57.45%	85.11%	70.21%	27.66%	31.91%
citation:inline	5.00%	85.00%	52.50%	37.50%	37.50%
citation:square_brackets	100.00%	100.00%	100.00%	98.93%	90.71%
citation:start_from_zero	100.00%	100.00%	91.76%	92.65%	96.18%
emoji:banned	96.90%	90.69%	90.69%	84.48%	93.79%
emoji:end	70.00%	70.00%	70.00%	70.00%	63.33%
emoji:frequency	100.00%	99.67%	91.67%	95.33%	81.33%
style:informal	98.11%	100.00%	96.23%	88.68%	96.23%
style:letter	100.00%	100.00%	89.80%	99.39%	100.00%
style:official	80.00%	79.52%	69.05%	74.05%	85.71%
style:poetic	100.00%	100.00%	97.87%	95.74%	99.36%
style:technical	98.98%	96.78%	95.59%	89.49%	95.59%
tone:angry	93.72%	97.21%	79.30%	80.70%	73.72%
tone:humorous	93.11%	96.00%	97.11%	92.67%	78.44%
tone:negative	98.12%	100.00%	100.00%	97.29%	78.96%
tone:positive	98.27%	99.42%	98.27%	98.85%	98.85%
tone:sarcastic	99.27%	97.56%	94.39%	81.71%	85.12%
content:celebrity	99.25%	100.00%	92.75%	88.00%	94.25%
content:jokes	99.48%	96.55%	83.45%	86.03%	90.86%
content:quotes	100.00%	100.00%	91.22%	92.65%	98.16%
language_switch:multilingual	100.00%	99.41%	90.39%	51.18%	7.84%
language_switch:repeat	55.81%	58.14%	58.14%	39.53%	16.28%

Table 37: Detailed evaluation results for Chinese language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	87.10%	94.19%	80.32%	80.32%	60.97%
keywords:first_word	90.32%	90.32%	90.32%	87.10%	87.10%
keywords:frequency	93.54%	90.62%	95.62%	90.42%	77.92%
keywords:paragraph_end	81.67%	82.22%	76.11%	61.67%	72.78%
keywords:together	78.71%	76.61%	82.74%	73.06%	75.32%
length:max_words	77.79%	77.53%	93.16%	94.12%	57.33%
length:range_words	82.42%	76.89%	48.33%	78.67%	58.68%
format:addition_at_end	40.00%	55.00%	30.00%	33.33%	21.67%
format:json_output	39.47%	10.53%	36.84%	71.05%	63.16%
format:markdown_bold_italic_paragraph	98.67%	93.00%	94.33%	98.67%	87.67%
format:markdown_highlight	100.00%	100.00%	96.97%	100.00%	99.70%
format:markdown_title	100.00%	9.38%	50.00%	96.88%	0.00%
format:ordered_list	96.55%	96.73%	96.55%	96.36%	98.00%
format:title_brackets	92.68%	97.56%	95.12%	87.80%	100.00%
format:two_answers_with_separator	90.62%	96.88%	93.75%	90.62%	75.00%
repeat:all_sentences_twice	76.00%	60.00%	36.00%	12.80%	12.80%
repeat:before_answer	96.30%	96.30%	96.30%	96.30%	65.19%
repeat:copy_request	86.21%	3.45%	82.76%	31.03%	31.03%
repeat:first_last_same	61.29%	38.71%	41.94%	29.03%	12.90%
repeat:last_sentence	53.10%	51.03%	35.17%	17.93%	2.76%
repeat:sentence_n_times	78.00%	93.33%	85.33%	80.67%	54.00%
marks:end_with_semicolons	94.11%	87.78%	48.61%	13.72%	34.78%
marks:no_commas	100.00%	100.00%	100.00%	94.69%	39.41%
marks:replace_with_asterisks	66.63%	88.60%	29.07%	70.17%	4.17%
marks:replace_with_exclamations	12.81%	95.56%	27.63%	39.70%	6.33%
marks:wrap_in_quotes	82.98%	70.21%	55.32%	59.57%	51.06%
citation:inline	70.00%	80.00%	50.00%	65.00%	65.00%
citation:square_brackets	89.29%	100.00%	95.36%	96.43%	89.29%
citation:start_from_zero	26.47%	82.35%	86.47%	50.00%	11.76%
emoji:banned	93.79%	87.59%	96.55%	77.59%	68.97%
emoji:end	70.00%	70.00%	60.00%	63.33%	66.67%
emoji:frequency	99.33%	99.67%	99.33%	89.00%	73.33%
style:informal	100.00%	100.00%	90.57%	90.57%	100.00%
style:letter	100.00%	100.00%	94.08%	90.61%	100.00%
style:official	85.95%	85.95%	74.76%	65.24%	69.29%
style:poetic	100.00%	100.00%	100.00%	97.87%	100.00%
style:technical	91.19%	81.19%	83.90%	81.02%	87.46%
tone:angry	88.84%	95.81%	86.28%	72.09%	77.67%
tone:humorous	99.33%	100.00%	99.33%	97.78%	97.33%
tone:negative	98.12%	100.00%	95.83%	92.50%	95.42%
tone:positive	98.85%	99.42%	99.42%	98.85%	98.27%
tone:sarcastic	100.00%	100.00%	97.56%	99.27%	97.07%
content:celebrity	98.50%	97.75%	84.50%	87.00%	96.00%
content:jokes	100.00%	97.76%	92.07%	83.10%	96.03%
content:quotes	99.39%	100.00%	94.29%	91.22%	99.39%
language_switch:multilingual	96.08%	95.49%	88.24%	75.10%	34.51%
language_switch:repeat	100.00%	100.00%	90.70%	88.37%	83.72%

Table 38: Detailed evaluation results for English language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	100.00%	97.10%	89.03%	91.29%	77.74%
keywords:first_word	90.32%	87.10%	70.97%	74.19%	70.97%
keywords:frequency	93.12%	92.50%	93.12%	87.08%	69.38%
keywords:paragraph_end	87.78%	81.11%	74.44%	67.22%	47.78%
keywords:together	80.65%	92.26%	89.68%	76.77%	51.29%
length:max_words	57.78%	75.21%	75.62%	84.74%	66.35%
length:range_words	67.29%	72.80%	70.68%	25.13%	50.08%
format:addition_at_end	100.00%	100.00%	88.33%	78.33%	85.00%
format:json_output	84.21%	31.58%	42.11%	60.53%	71.05%
format:markdown_bold_italic_paragraph	99.00%	91.67%	95.33%	96.33%	64.33%
format:markdown_highlight	100.00%	100.00%	100.00%	97.27%	87.27%
format:markdown_title	100.00%	96.88%	96.88%	96.88%	0.00%
format:ordered_list	96.73%	96.73%	96.73%	98.18%	99.45%
format:title_brackets	100.00%	100.00%	100.00%	80.49%	65.85%
format:two_answers_with_separator	84.38%	96.88%	87.50%	100.00%	78.12%
repeat:all_sentences_twice	79.20%	68.00%	36.00%	52.80%	0.80%
repeat:before_answer	96.30%	96.30%	100.00%	96.30%	68.89%
repeat:copy_request	100.00%	27.59%	17.24%	86.21%	0.00%
repeat:first_last_same	61.29%	61.29%	51.61%	29.03%	0.00%
repeat:last_sentence	40.69%	57.93%	15.17%	20.00%	2.76%
repeat:sentence_n_times	85.33%	98.00%	96.00%	81.33%	65.33%
marks:end_with_semicolons	92.61%	75.17%	64.61%	2.89%	5.44%
marks:no_commas	90.52%	100.00%	96.03%	96.34%	16.79%
marks:replace_with_asterisks	55.53%	99.90%	39.80%	39.37%	3.77%
marks:replace_with_exclamations	47.33%	99.89%	26.30%	52.37%	0.00%
marks:wrap_in_quotes	65.96%	31.91%	63.83%	34.04%	6.38%
citation:inline	85.00%	50.00%	60.00%	17.50%	20.00%
citation:square_brackets	100.00%	100.00%	98.93%	95.36%	74.29%
citation:start_from_zero	77.06%	91.18%	96.18%	70.88%	65.88%
emoji:banned	96.90%	87.59%	90.69%	90.34%	78.28%
emoji:end	60.00%	10.00%	56.67%	59.00%	33.67%
emoji:frequency	100.00%	100.00%	96.33%	99.33%	86.33%
style:informal	100.00%	100.00%	98.11%	98.11%	97.55%
style:letter	95.92%	100.00%	89.18%	94.08%	97.35%
style:official	90.71%	82.14%	81.19%	80.48%	87.38%
style:poetic	100.00%	100.00%	96.81%	90.00%	84.68%
style:technical	88.14%	72.54%	81.19%	84.07%	79.15%
tone:angry	65.58%	83.72%	73.95%	83.49%	49.07%
tone:humorous	94.00%	98.67%	98.00%	91.33%	74.44%
tone:negative	96.25%	98.75%	92.71%	98.12%	83.33%
tone:positive	98.27%	99.42%	98.85%	99.42%	98.27%
tone:sarcastic	91.46%	100.00%	96.10%	97.80%	69.02%
content:celebrity	98.50%	97.00%	90.75%	87.25%	95.25%
content:jokes	99.48%	100.00%	91.03%	76.90%	59.83%
content:quotes	95.31%	100.00%	94.29%	92.04%	86.94%
language_switch:multilingual	95.69%	99.41%	76.27%	37.65%	18.43%
language_switch:repeat	100.00%	100.00%	69.07%	58.14%	63.02%

Table 39: Detailed evaluation results for Filipino language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	97.10%	99.03%	83.23%	81.94%	79.35%
keywords:first_word	96.77%	96.77%	83.87%	83.87%	64.52%
keywords:frequency	92.71%	91.46%	90.00%	86.46%	75.42%
keywords:paragraph_end	86.11%	78.33%	69.44%	62.78%	67.22%
keywords:together	76.29%	85.65%	73.55%	63.55%	60.81%
length:max_words	82.55%	88.48%	70.07%	80.67%	65.67%
length:range_words	74.06%	79.24%	73.81%	65.35%	73.24%
format:addition_at_end	70.00%	53.33%	51.67%	45.00%	53.33%
format:json_output	52.63%	18.42%	34.21%	63.16%	52.63%
format:markdown_bold_italic_paragraph	96.00%	83.67%	78.33%	99.67%	80.67%
format:markdown_highlight	100.00%	100.00%	100.00%	98.79%	92.73%
format:markdown_title	100.00%	96.88%	96.88%	93.75%	34.38%
format:ordered_list	98.18%	96.73%	95.09%	96.55%	98.18%
format:title_brackets	90.24%	97.56%	90.24%	19.51%	100.00%
format:two_answers_with_separator	93.75%	100.00%	62.50%	96.88%	75.00%
repeat:all_sentences_twice	71.20%	53.60%	20.80%	20.00%	8.80%
repeat:before_answer	96.30%	96.30%	96.30%	96.30%	65.93%
repeat:copy_request	100.00%	0.00%	93.10%	93.10%	65.52%
repeat:first_last_same	54.84%	74.19%	48.39%	32.26%	0.00%
repeat:last_sentence	43.45%	59.31%	28.97%	32.41%	12.41%
repeat:sentence_n_times	88.67%	97.33%	81.33%	89.33%	56.67%
marks:end_with_semicolons	95.67%	86.28%	51.28%	8.28%	39.33%
marks:no_commas	100.00%	99.90%	83.62%	98.24%	24.69%
marks:replace_with_asterisks	21.33%	51.33%	5.90%	17.80%	0.00%
marks:replace_with_exclamations	20.33%	95.67%	20.04%	30.04%	13.89%
marks:wrap_in_quotes	65.96%	85.11%	61.70%	55.32%	36.17%
citation:inline	52.50%	85.00%	65.00%	65.00%	45.00%
citation:square_brackets	95.36%	100.00%	90.71%	88.21%	96.43%
citation:start_from_zero	11.76%	82.35%	72.06%	44.41%	0.00%
emoji:banned	96.90%	87.59%	90.69%	81.03%	68.97%
emoji:end	56.67%	33.33%	66.67%	56.67%	53.33%
emoji:frequency	99.33%	99.00%	93.67%	90.00%	81.00%
style:informal	100.00%	100.00%	90.57%	96.23%	96.98%
style:letter	95.92%	98.78%	94.90%	93.47%	97.55%
style:official	76.67%	79.29%	77.38%	70.95%	71.19%
style:poetic	100.00%	100.00%	95.74%	97.87%	100.00%
style:technical	95.08%	90.85%	94.07%	90.68%	89.66%
tone:angry	74.42%	95.81%	78.14%	83.26%	66.51%
tone:humorous	98.00%	100.00%	95.78%	95.11%	92.89%
tone:negative	98.12%	98.75%	95.21%	98.13%	87.29%
tone:positive	98.85%	98.85%	96.35%	96.35%	96.92%
tone:sarcastic	99.27%	100.00%	98.54%	92.93%	93.17%
content:celebrity	97.75%	97.00%	89.25%	85.25%	88.75%
content:jokes	99.48%	99.48%	91.38%	78.28%	89.14%
content:quotes	97.96%	100.00%	88.78%	98.16%	92.86%
language_switch:multilingual	96.27%	100.00%	85.10%	70.00%	24.31%
language_switch:repeat	93.02%	100.00%	67.44%	79.07%	86.05%

Table 40: Detailed evaluation results for French language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	87.10%	75.48%	90.00%	92.26%	60.00%
keywords:first_word	77.42%	61.29%	54.84%	58.06%	67.74%
keywords:frequency	81.87%	75.42%	77.71%	73.54%	67.08%
keywords:paragraph_end	57.22%	40.00%	38.89%	43.33%	33.89%
keywords:together	48.23%	30.16%	37.90%	46.61%	32.10%
length:max_words	98.90%	99.70%	100.00%	98.80%	91.48%
length:range_words	79.50%	42.82%	51.55%	46.47%	26.94%
format:addition_at_end	65.00%	61.67%	80.00%	75.00%	80.00%
format:json_output	97.37%	42.11%	47.37%	52.63%	23.68%
format:markdown_bold_italic_paragraph	98.67%	91.67%	94.67%	98.00%	43.00%
format:markdown_highlight	99.70%	86.06%	100.00%	97.27%	66.36%
format:markdown_title	75.00%	71.88%	96.88%	100.00%	3.12%
format:ordered_list	89.64%	75.27%	96.73%	92.91%	77.45%
format:title_brackets	2.44%	51.22%	53.66%	2.44%	0.00%
format:two_answers_with_separator	68.75%	65.62%	93.75%	84.38%	53.12%
repeat:all_sentences_twice	56.00%	64.80%	32.00%	31.20%	0.80%
repeat:before_answer	92.59%	85.19%	100.00%	95.56%	10.37%
repeat:copy_request	96.55%	3.45%	100.00%	79.31%	0.00%
repeat:first_last_same	64.52%	58.06%	61.29%	48.39%	3.23%
repeat:last_sentence	20.69%	44.14%	20.00%	3.45%	8.97%
repeat:sentence_n_times	82.00%	70.00%	77.33%	85.33%	26.67%
marks:end_with_semicolons	99.50%	92.28%	20.61%	13.17%	27.67%
marks:no_commas	71.00%	100.00%	46.00%	12.00%	27.62%
marks:replace_with_asterisks	51.33%	83.33%	62.63%	38.93%	6.57%
marks:replace_with_exclamations	15.78%	85.19%	63.11%	28.41%	0.00%
marks:wrap_in_quotes	0.00%	0.00%	0.00%	0.00%	0.00%
citation:inline	27.50%	22.50%	42.50%	22.50%	7.50%
citation:square_brackets	96.43%	89.29%	96.43%	96.43%	76.43%
citation:start_from_zero	79.12%	91.18%	96.18%	76.47%	37.35%
emoji:banned	96.90%	90.34%	84.48%	81.03%	90.69%
emoji:end	70.00%	60.00%	63.33%	58.00%	6.67%
emoji:frequency	94.33%	93.67%	98.00%	80.67%	41.00%
style:informal	100.00%	97.55%	94.34%	98.11%	55.09%
style:letter	95.31%	90.00%	78.16%	76.73%	41.84%
style:official	77.14%	61.90%	73.10%	62.62%	70.95%
style:poetic	99.36%	94.47%	97.87%	99.36%	47.45%
style:technical	91.19%	84.92%	88.31%	91.19%	68.81%
tone:angry	63.26%	78.37%	50.93%	72.56%	4.88%
tone:humorous	91.11%	95.78%	94.00%	94.00%	27.11%
tone:negative	93.96%	96.67%	85.62%	98.75%	27.08%
tone:positive	98.85%	95.58%	96.35%	98.08%	76.54%
tone:sarcastic	96.83%	97.56%	98.54%	95.37%	16.10%
content:celebrity	91.00%	95.25%	78.00%	69.25%	10.25%
content:jokes	94.48%	86.21%	78.28%	76.03%	19.66%
content:quotes	93.88%	95.31%	67.76%	80.61%	2.04%
language_switch:multilingual	93.53%	88.24%	62.16%	36.86%	2.75%
language_switch:repeat	97.67%	95.35%	83.02%	73.02%	3.95%

Table 41: Detailed evaluation results for Georgian language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	100.00%	98.06%	81.29%	85.48%	41.61%
keywords:first_word	16.13%	16.13%	16.13%	12.90%	16.13%
keywords:frequency	67.92%	71.46%	62.50%	68.13%	67.92%
keywords:paragraph_end	48.33%	45.56%	31.67%	36.11%	31.67%
keywords:together	68.23%	82.74%	55.81%	50.48%	64.68%
length:max_words	82.28%	99.16%	62.61%	74.70%	43.35%
length:range_words	59.75%	67.32%	48.51%	40.42%	43.04%
format:addition_at_end	50.00%	61.67%	33.33%	28.33%	80.00%
format:json_output	71.05%	31.58%	36.84%	71.05%	31.58%
format:markdown_bold_italic_paragraph	99.00%	78.67%	50.67%	73.00%	65.67%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	90.00%
format:markdown_title	18.44%	17.19%	34.69%	26.56%	4.06%
format:ordered_list	96.55%	96.73%	96.73%	96.73%	94.91%
format:title_brackets	9.55%	13.38%	16.78%	4.88%	39.36%
format:two_answers_with_separator	93.75%	100.00%	50.00%	75.00%	90.62%
repeat:all_sentences_twice	40.00%	8.80%	12.80%	10.40%	12.80%
repeat:before_answer	100.00%	90.37%	85.93%	77.78%	57.04%
repeat:copy_request	100.00%	0.00%	93.10%	72.41%	0.00%
repeat:first_last_same	64.52%	77.42%	41.94%	38.71%	0.00%
repeat:last_sentence	23.45%	30.34%	3.45%	8.97%	3.45%
repeat:sentence_n_times	87.33%	95.33%	47.33%	59.33%	52.67%
marks:end_with_semicolons	99.50%	82.11%	46.89%	32.78%	8.11%
marks:no_commas	95.72%	100.00%	96.55%	97.52%	33.93%
marks:replace_with_asterisks	51.53%	68.73%	53.77%	46.23%	0.00%
marks:replace_with_exclamations	24.11%	100.00%	31.37%	38.30%	0.00%
marks:wrap_in_quotes	61.70%	55.32%	42.55%	27.66%	0.00%
citation:inline	67.50%	37.50%	45.00%	17.50%	15.00%
citation:square_brackets	100.00%	100.00%	98.93%	96.43%	80.71%
citation:start_from_zero	100.00%	100.00%	83.24%	67.65%	69.12%
emoji:banned	93.79%	90.69%	90.69%	81.03%	79.66%
emoji:end	60.00%	64.67%	65.67%	63.33%	22.33%
emoji:frequency	96.33%	99.67%	98.00%	87.33%	77.67%
style:informal	99.43%	87.55%	90.57%	94.34%	77.74%
style:letter	100.00%	81.02%	89.39%	94.08%	93.47%
style:official	88.81%	71.67%	75.48%	79.76%	83.81%
style:poetic	100.00%	89.57%	97.87%	100.00%	71.06%
style:technical	97.97%	75.59%	92.88%	97.97%	92.03%
tone:angry	41.86%	75.81%	62.33%	69.77%	18.60%
tone:humorous	94.44%	90.67%	98.67%	89.56%	58.44%
tone:negative	71.46%	86.25%	94.58%	97.50%	24.79%
tone:positive	96.35%	95.58%	95.58%	98.27%	95.58%
tone:sarcastic	73.41%	84.15%	89.02%	81.71%	23.66%
content:celebrity	100.00%	86.75%	91.25%	77.50%	71.75%
content:jokes	94.83%	87.41%	84.14%	77.07%	85.86%
content:quotes	100.00%	83.06%	89.39%	96.73%	85.10%
language_switch:multilingual	97.45%	82.55%	83.33%	69.80%	26.47%
language_switch:repeat	97.67%	74.42%	81.40%	66.74%	34.88%

Table 42: Detailed evaluation results for Hindi language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	89.03%	94.19%	69.03%	63.87%	59.03%
keywords:first_word	90.32%	90.32%	87.10%	83.87%	67.74%
keywords:frequency	89.58%	88.33%	81.67%	82.29%	69.79%
keywords:paragraph_end	77.78%	78.33%	73.33%	70.56%	66.67%
keywords:together	89.68%	88.39%	87.74%	73.87%	74.52%
length:max_words	86.06%	86.85%	77.16%	82.88%	81.11%
length:range_words	83.25%	72.10%	85.15%	76.77%	45.11%
format:addition_at_end	68.33%	88.33%	71.67%	76.67%	65.00%
format:json_output	47.37%	36.84%	44.74%	52.63%	23.68%
format:markdown_bold_italic_paragraph	98.33%	84.67%	86.67%	99.67%	81.33%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	97.58%
format:markdown_title	100.00%	96.88%	71.88%	96.88%	6.25%
format:ordered_list	96.55%	96.73%	98.36%	96.73%	97.09%
format:title_brackets	100.00%	100.00%	90.24%	17.07%	82.93%
format:two_answers_with_separator	78.12%	100.00%	84.38%	100.00%	75.00%
repeat:all_sentences_twice	48.00%	64.80%	20.00%	16.00%	0.80%
repeat:before_answer	96.30%	96.30%	96.30%	96.30%	87.41%
repeat:copy_request	100.00%	3.45%	100.00%	82.76%	72.41%
repeat:first_last_same	58.06%	67.74%	48.39%	38.71%	3.23%
repeat:last_sentence	19.31%	57.93%	31.03%	18.62%	9.66%
repeat:sentence_n_times	79.33%	98.00%	72.00%	70.67%	31.33%
marks:end_with_semicolons	92.61%	87.78%	30.17%	2.89%	31.22%
marks:no_commas	100.00%	100.00%	91.86%	99.90%	33.55%
marks:replace_with_asterisks	71.53%	96.90%	60.07%	60.67%	7.10%
marks:replace_with_exclamations	12.93%	99.00%	2.70%	30.67%	18.00%
marks:wrap_in_quotes	74.47%	61.70%	51.06%	42.55%	55.32%
citation:inline	52.50%	75.00%	52.50%	42.50%	27.50%
citation:square_brackets	100.00%	100.00%	98.93%	100.00%	96.43%
citation:start_from_zero	19.71%	82.35%	62.06%	59.41%	0.00%
emoji:banned	96.90%	87.59%	87.59%	78.28%	93.45%
emoji:end	70.00%	63.33%	66.67%	60.00%	53.33%
emoji:frequency	97.00%	98.67%	98.33%	75.00%	74.67%
style:informal	100.00%	100.00%	88.68%	88.68%	97.55%
style:letter	93.88%	99.39%	91.43%	98.78%	100.00%
style:official	85.95%	80.71%	81.19%	79.05%	78.81%
style:poetic	100.00%	100.00%	93.62%	89.36%	100.00%
style:technical	92.37%	82.54%	84.92%	86.44%	91.36%
tone:angry	61.63%	86.28%	57.21%	60.00%	47.91%
tone:humorous	97.33%	100.00%	89.78%	91.11%	81.56%
tone:negative	91.25%	96.04%	89.79%	86.25%	70.42%
tone:positive	98.08%	100.00%	97.50%	95.58%	93.65%
tone:sarcastic	94.39%	100.00%	92.68%	91.22%	87.07%
content:celebrity	99.25%	97.00%	84.00%	82.50%	98.50%
content:jokes	96.03%	98.97%	81.72%	80.86%	91.38%
content:quotes	97.96%	100.00%	84.90%	94.69%	98.78%
language_switch:multilingual	95.49%	99.41%	94.51%	52.94%	23.73%
language_switch:repeat	95.35%	100.00%	83.72%	59.77%	12.56%

Table 43: Detailed evaluation results for Indonesian language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	94.19%	97.10%	80.32%	85.48%	63.23%
keywords:first_word	90.32%	90.32%	70.97%	67.74%	61.29%
keywords:frequency	91.67%	89.33%	85.00%	84.79%	74.79%
keywords:paragraph_end	95.56%	80.00%	76.11%	68.89%	72.22%
keywords:together	86.94%	85.81%	78.39%	67.58%	69.19%
length:max_words	80.36%	85.72%	77.24%	78.66%	62.77%
length:range_words	74.04%	78.81%	76.68%	66.15%	70.88%
format:addition_at_end	51.67%	75.00%	75.00%	75.00%	63.33%
format:json_output	44.74%	26.32%	36.84%	47.37%	42.11%
format:markdown_bold_italic_paragraph	97.00%	82.67%	86.00%	99.67%	74.33%
format:markdown_highlight	98.79%	100.00%	100.00%	96.97%	95.76%
format:markdown_title	100.00%	96.88%	93.75%	96.88%	9.38%
format:ordered_list	96.55%	96.67%	94.91%	96.36%	100.00%
format:title_brackets	2.44%	17.07%	21.95%	31.71%	9.76%
format:two_answers_with_separator	90.62%	100.00%	71.88%	84.38%	90.62%
repeat:all_sentences_twice	76.00%	62.50%	23.20%	36.00%	12.80%
repeat:before_answer	96.30%	96.30%	96.30%	96.30%	77.04%
repeat:copy_request	100.00%	0.00%	68.97%	68.97%	34.48%
repeat:first_last_same	61.29%	73.33%	61.29%	35.48%	3.23%
repeat:last_sentence	40.00%	74.29%	15.86%	31.03%	2.76%
repeat:sentence_n_times	82.67%	98.67%	81.33%	89.33%	66.67%
marks:end_with_semicolons	96.50%	87.67%	50.44%	10.28%	40.56%
marks:no_commas	100.00%	100.00%	98.14%	83.72%	40.90%
marks:replace_with_asterisks	38.17%	85.47%	22.83%	38.67%	2.93%
marks:replace_with_exclamations	24.52%	100.00%	9.33%	31.70%	14.78%
marks:wrap_in_quotes	70.21%	31.91%	63.83%	53.19%	46.81%
citation:inline	82.50%	72.50%	50.00%	25.00%	55.00%
citation:square_brackets	100.00%	100.00%	98.93%	100.00%	92.86%
citation:start_from_zero	100.00%	85.29%	92.94%	88.53%	98.24%
emoji:banned	93.79%	81.38%	81.38%	77.93%	83.79%
emoji:end	66.67%	23.33%	60.00%	70.00%	43.33%
emoji:frequency	99.67%	98.33%	91.67%	91.67%	71.00%
style:informal	100.00%	100.00%	86.79%	94.34%	98.87%
style:letter	91.43%	98.78%	91.02%	96.33%	96.94%
style:official	81.43%	82.14%	72.14%	67.14%	75.00%
style:poetic	100.00%	100.00%	97.87%	95.74%	100.00%
style:technical	96.78%	88.81%	86.78%	88.64%	76.78%
tone:angry	73.02%	94.42%	74.42%	77.44%	66.51%
tone:humorous	98.67%	98.67%	95.78%	97.33%	90.22%
tone:negative	99.38%	100.00%	93.12%	96.04%	83.75%
tone:positive	98.85%	99.42%	98.85%	98.27%	97.12%
tone:sarcastic	99.27%	100.00%	97.56%	96.10%	99.27%
content:celebrity	100.00%	100.00%	87.25%	89.00%	88.00%
content:jokes	66.90%	81.21%	80.00%	58.10%	59.31%
content:quotes	95.92%	99.39%	91.63%	96.73%	90.82%
language_switch:multilingual	98.04%	99.41%	78.43%	72.55%	22.35%
language_switch:repeat	100.00%	100.00%	76.74%	93.02%	88.37%

Table 44: Detailed evaluation results for Italian language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	99.03%	100.00%	99.03%	100.00%	99.03%
keywords:first_word	0.00%	0.00%	16.13%	3.23%	0.00%
keywords:frequency	92.92%	91.88%	93.54%	82.92%	74.58%
keywords:paragraph_end	73.89%	18.33%	76.11%	70.00%	51.11%
keywords:together	77.58%	78.06%	78.06%	71.45%	78.39%
length:max_words	75.02%	83.42%	63.16%	49.09%	73.15%
length:range_words	75.47%	73.66%	68.54%	66.23%	32.99%
format:addition_at_end	68.33%	75.00%	68.33%	68.33%	75.00%
format:json_output	71.05%	23.68%	55.26%	60.53%	47.37%
format:markdown_bold_italic_paragraph	98.00%	90.33%	67.00%	87.67%	86.00%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	100.00%
format:markdown_title	34.38%	0.00%	70.62%	75.00%	0.00%
format:ordered_list	98.36%	96.73%	96.73%	96.73%	98.73%
format:title_brackets	100.00%	100.00%	100.00%	95.12%	64.76%
format:two_answers_with_separator	78.12%	68.75%	100.00%	93.75%	81.25%
repeat:all_sentences_twice	71.20%	32.80%	4.00%	33.60%	36.80%
repeat:before_answer	96.30%	96.30%	92.59%	92.59%	82.22%
repeat:copy_request	72.41%	3.45%	17.24%	55.17%	17.24%
repeat:first_last_same	64.52%	29.03%	32.26%	38.71%	6.45%
repeat:last_sentence	24.14%	2.76%	6.21%	24.14%	5.52%
repeat:sentence_n_times	82.00%	96.67%	56.00%	74.67%	62.00%
marks:end_with_semicolons	94.11%	93.61%	15.22%	45.67%	29.17%
marks:no_commas	90.52%	100.00%	93.00%	87.07%	33.62%
marks:replace_with_asterisks	66.17%	97.50%	67.07%	40.33%	19.60%
marks:replace_with_exclamations	11.00%	99.78%	3.70%	3.26%	6.41%
marks:wrap_in_quotes	51.06%	74.47%	61.70%	70.21%	14.89%
citation:inline	0.00%	60.00%	25.00%	30.00%	37.50%
citation:square_brackets	100.00%	100.00%	100.00%	100.00%	98.93%
citation:start_from_zero	100.00%	100.00%	77.94%	93.82%	96.47%
emoji:banned	96.90%	90.69%	93.79%	81.38%	81.38%
emoji:end	70.00%	70.00%	70.00%	70.00%	70.00%
emoji:frequency	99.33%	98.33%	98.00%	79.67%	73.33%
style:informal	100.00%	100.00%	86.79%	92.45%	95.09%
style:letter	91.84%	100.00%	95.92%	99.39%	99.39%
style:official	95.00%	84.52%	80.71%	90.24%	93.57%
style:poetic	97.87%	99.36%	89.36%	97.87%	96.81%
style:technical	96.61%	94.92%	94.41%	97.80%	83.90%
tone:angry	60.47%	96.51%	75.12%	74.42%	60.23%
tone:humorous	92.67%	98.67%	95.78%	85.78%	73.56%
tone:negative	85.62%	100.00%	85.42%	99.38%	80.42%
tone:positive	99.42%	100.00%	99.42%	97.50%	94.42%
tone:sarcastic	83.17%	100.00%	87.07%	96.34%	62.44%
content:celebrity	99.25%	98.50%	93.50%	85.50%	97.50%
content:jokes	97.76%	100.00%	87.41%	73.62%	87.07%
content:quotes	93.27%	100.00%	85.71%	91.22%	93.88%
language_switch:multilingual	94.12%	100.00%	83.73%	80.39%	27.65%
language_switch:repeat	100.00%	100.00%	95.35%	93.02%	86.05%

Table 45: Detailed evaluation results for Japanese language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	96.13%	99.03%	87.10%	90.32%	61.94%
keywords:first_word	54.84%	48.39%	58.06%	45.16%	48.39%
keywords:frequency	87.71%	87.50%	85.42%	82.29%	70.00%
keywords:paragraph_end	41.11%	26.67%	29.44%	28.89%	32.78%
keywords:together	70.00%	77.90%	71.77%	62.10%	64.68%
length:max_words	89.09%	88.57%	47.51%	67.14%	30.57%
length:range_words	74.12%	78.40%	63.06%	68.26%	22.54%
format:addition_at_end	48.33%	56.67%	50.00%	50.00%	63.33%
format:json_output	63.16%	21.05%	23.68%	50.00%	44.74%
format:markdown_bold_italic_paragraph	95.33%	93.33%	73.33%	98.00%	37.33%
format:markdown_highlight	100.00%	96.97%	100.00%	100.00%	95.76%
format:markdown_title	96.88%	93.75%	93.75%	93.75%	31.25%
format:ordered_list	93.82%	93.27%	98.36%	98.36%	96.18%
format:title_brackets	34.15%	97.56%	53.66%	0.00%	4.88%
format:two_answers_with_separator	62.50%	87.50%	53.12%	90.62%	81.25%
repeat:all_sentences_twice	0.00%	0.00%	0.00%	0.00%	0.00%
repeat:before_answer	92.59%	85.19%	88.89%	95.56%	85.19%
repeat:copy_request	89.66%	20.69%	72.41%	48.28%	0.00%
repeat:first_last_same	0.00%	0.00%	0.00%	0.00%	0.00%
repeat:last_sentence	0.00%	0.00%	0.00%	0.00%	0.00%
repeat:sentence_n_times	81.33%	94.67%	41.33%	34.00%	48.00%
marks:end_with_semicolons	97.00%	97.00%	97.00%	97.00%	97.00%
marks:no_commas	94.41%	99.48%	95.72%	98.55%	45.55%
marks:replace_with_asterisks	48.53%	91.93%	56.53%	61.67%	18.43%
marks:replace_with_exclamations	16.48%	92.48%	29.52%	53.44%	29.22%
marks:wrap_in_quotes	61.70%	61.70%	51.06%	27.66%	29.79%
citation:inline	95.00%	77.50%	55.00%	30.00%	55.00%
citation:square_brackets	85.71%	92.86%	100.00%	100.00%	98.93%
citation:start_from_zero	97.06%	85.29%	74.41%	69.71%	97.06%
emoji:banned	96.90%	93.79%	87.59%	78.28%	87.59%
emoji:end	45.67%	66.67%	36.67%	34.67%	6.67%
emoji:frequency	97.67%	99.33%	98.00%	81.33%	71.33%
style:informal	96.23%	95.66%	82.45%	79.25%	93.77%
style:letter	97.96%	97.96%	91.84%	93.27%	96.73%
style:official	91.19%	75.48%	79.52%	83.81%	61.90%
style:poetic	93.62%	94.47%	90.85%	91.49%	73.19%
style:technical	97.29%	92.20%	92.71%	90.00%	93.05%
tone:angry	71.86%	96.98%	62.33%	70.93%	84.65%
tone:humorous	88.89%	94.44%	98.00%	91.33%	73.33%
tone:negative	83.96%	97.29%	93.12%	91.04%	86.25%
tone:positive	95.00%	90.38%	100.00%	99.42%	99.42%
tone:sarcastic	82.93%	97.56%	96.83%	87.07%	60.98%
content:celebrity	97.50%	97.50%	96.00%	88.75%	93.75%
content:jokes	98.28%	92.59%	83.45%	84.66%	81.72%
content:quotes	95.92%	89.80%	92.04%	89.18%	99.39%
language_switch:multilingual	91.57%	89.61%	96.08%	66.27%	18.63%
language_switch:repeat	97.67%	93.02%	86.05%	95.35%	60.47%

Table 46: Detailed evaluation results for Korean language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	98.06%	100.00%	93.23%	94.19%	12.90%
keywords:first_word	87.10%	70.97%	80.65%	87.10%	19.35%
keywords:frequency	80.83%	81.88%	78.54%	75.63%	66.25%
keywords:paragraph_end	77.22%	54.44%	50.00%	32.78%	32.22%
keywords:together	83.06%	71.13%	76.94%	61.45%	34.52%
length:max_words	99.42%	96.90%	97.68%	97.12%	68.27%
length:range_words	85.54%	39.35%	73.73%	73.76%	16.35%
format:addition_at_end	58.33%	45.00%	61.67%	65.00%	60.00%
format:json_output	84.21%	36.84%	47.37%	76.32%	10.53%
format:markdown_bold_italic_paragraph	99.00%	78.67%	34.67%	89.00%	65.33%
format:markdown_highlight	97.27%	86.97%	100.00%	88.48%	51.52%
format:markdown_title	50.00%	84.38%	81.25%	100.00%	3.12%
format:ordered_list	96.55%	88.00%	98.36%	98.18%	84.36%
format:title_brackets	2.44%	85.37%	12.20%	51.22%	53.66%
format:two_answers_with_separator	96.88%	75.00%	84.38%	100.00%	53.12%
repeat:all_sentences_twice	76.00%	71.20%	32.00%	28.00%	6.40%
repeat:before_answer	96.30%	81.48%	86.67%	92.59%	4.44%
repeat:copy_request	100.00%	3.45%	79.31%	100.00%	6.90%
repeat:first_last_same	58.06%	70.97%	54.84%	25.81%	0.00%
repeat:last_sentence	33.79%	34.48%	13.10%	11.72%	3.45%
repeat:sentence_n_times	78.00%	80.00%	84.67%	78.00%	25.33%
marks:end_with_semicolons	87.39%	81.56%	13.89%	6.94%	11.22%
marks:no_commas	96.59%	100.00%	87.55%	75.28%	27.45%
marks:replace_with_asterisks	67.20%	88.00%	54.40%	42.07%	5.97%
marks:replace_with_exclamations	33.85%	88.89%	29.56%	0.00%	0.00%
marks:wrap_in_quotes	68.09%	70.21%	53.19%	38.30%	0.00%
citation:inline	45.00%	57.50%	55.00%	27.50%	22.50%
citation:square_brackets	100.00%	96.43%	100.00%	98.93%	65.71%
citation:start_from_zero	100.00%	82.35%	92.94%	70.59%	33.24%
emoji:banned	87.59%	90.00%	93.79%	87.24%	18.28%
emoji:end	60.00%	56.67%	59.00%	63.33%	5.67%
emoji:frequency	99.00%	99.33%	82.33%	74.33%	45.00%
style:informal	96.04%	95.28%	99.43%	96.42%	79.62%
style:letter	100.00%	85.10%	82.65%	88.16%	12.65%
style:official	70.00%	67.86%	79.05%	88.33%	87.14%
style:poetic	99.36%	98.72%	93.62%	97.87%	5.96%
style:technical	95.59%	95.25%	98.98%	97.29%	75.59%
tone:angry	45.58%	65.81%	52.79%	57.21%	3.26%
tone:humorous	74.00%	90.67%	94.00%	78.67%	10.89%
tone:negative	51.46%	85.21%	87.29%	91.46%	5.83%
tone:positive	95.58%	93.65%	97.50%	97.69%	83.65%
tone:sarcastic	40.24%	83.17%	87.32%	85.85%	5.12%
content:celebrity	96.75%	87.50%	94.75%	74.75%	13.50%
content:jokes	88.10%	94.83%	82.41%	74.83%	1.72%
content:quotes	100.00%	83.67%	93.27%	93.27%	4.08%
language_switch:multilingual	98.04%	90.98%	92.35%	75.29%	3.33%
language_switch:repeat	95.35%	90.70%	96.98%	89.30%	1.63%

Table 47: Detailed evaluation results for Kyrgyz language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	87.74%	92.26%	79.68%	77.10%	48.39%
keywords:first_word	64.52%	64.52%	61.29%	51.61%	12.90%
keywords:frequency	87.71%	87.08%	83.54%	71.25%	73.75%
keywords:paragraph_end	91.67%	78.33%	69.44%	67.22%	10.56%
keywords:together	77.10%	85.81%	57.58%	59.03%	49.52%
length:max_words	72.53%	92.47%	99.20%	85.00%	74.02%
length:range_words	73.79%	70.81%	80.62%	63.87%	19.36%
format:addition_at_end	96.67%	100.00%	75.00%	75.00%	41.67%
format:json_output	73.68%	0.00%	65.79%	68.42%	2.63%
format:markdown_bold_italic_paragraph	90.67%	85.67%	68.33%	83.33%	57.00%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	53.64%
format:markdown_title	87.50%	59.38%	90.62%	96.88%	0.00%
format:ordered_list	96.73%	96.73%	94.91%	98.36%	58.36%
format:title_brackets	41.46%	39.02%	34.15%	17.07%	19.50%
format:two_answers_with_separator	84.38%	96.88%	100.00%	100.00%	31.25%
repeat:all_sentences_twice	44.80%	80.00%	1.60%	8.00%	13.60%
repeat:before_answer	95.56%	100.00%	92.59%	100.00%	4.44%
repeat:copy_request	96.55%	6.90%	62.07%	68.97%	0.00%
repeat:first_last_same	54.84%	90.32%	25.81%	32.26%	3.23%
repeat:last_sentence	33.10%	44.14%	0.69%	17.24%	6.21%
repeat:sentence_n_times	78.67%	91.33%	78.00%	69.33%	24.00%
marks:end_with_semicolons	13.72%	25.83%	8.28%	4.89%	22.28%
marks:no_commas	71.28%	59.21%	96.38%	29.17%	49.83%
marks:replace_with_asterisks	1.73%	81.23%	7.80%	0.00%	0.00%
marks:replace_with_exclamations	16.89%	62.96%	27.52%	0.00%	0.00%
marks:wrap_in_quotes	0.00%	17.02%	2.13%	4.26%	0.00%
citation:inline	60.00%	42.50%	40.00%	17.50%	32.50%
citation:square_brackets	88.21%	100.00%	91.79%	100.00%	58.21%
citation:start_from_zero	85.00%	72.35%	86.18%	63.82%	15.29%
emoji:banned	100.00%	90.69%	93.79%	96.90%	47.24%
emoji:end	33.33%	5.67%	25.67%	15.67%	0.00%
emoji:frequency	95.00%	99.67%	83.33%	85.33%	62.33%
style:informal	99.43%	99.43%	99.43%	99.43%	83.77%
style:letter	95.92%	94.69%	71.84%	74.49%	8.16%
style:official	90.95%	88.10%	78.33%	87.86%	97.86%
style:poetic	97.23%	86.60%	98.72%	95.96%	50.43%
style:technical	94.24%	93.73%	78.98%	95.25%	82.37%
tone:angry	75.35%	80.47%	51.63%	60.47%	0.00%
tone:humorous	94.67%	89.78%	95.33%	91.78%	1.56%
tone:negative	90.42%	91.67%	87.71%	92.92%	7.92%
tone:positive	96.35%	99.42%	97.69%	94.62%	99.42%
tone:sarcastic	21.22%	14.63%	33.41%	19.51%	0.00%
content:celebrity	99.25%	96.75%	90.25%	80.00%	11.75%
content:jokes	91.38%	90.86%	85.86%	67.93%	0.00%
content:quotes	11.02%	49.80%	7.55%	22.65%	0.00%
language_switch:multilingual	91.57%	73.92%	48.82%	32.55%	0.00%
language_switch:repeat	100.00%	95.35%	86.05%	92.33%	0.00%

Table 48: Detailed evaluation results for Malagasy language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	96.13%	98.06%	86.13%	84.52%	79.35%
keywords:first_word	87.10%	87.10%	80.65%	87.10%	83.87%
keywords:frequency	90.42%	87.29%	86.46%	88.96%	73.33%
keywords:paragraph_end	86.67%	80.00%	75.00%	72.78%	51.67%
keywords:together	86.61%	85.81%	71.29%	78.87%	69.68%
length:max_words	92.70%	88.82%	90.79%	86.42%	82.69%
length:range_words	75.62%	86.84%	87.00%	67.60%	57.20%
format:addition_at_end	93.33%	93.33%	88.33%	61.67%	76.67%
format:json_output	81.58%	36.84%	63.16%	86.84%	68.42%
format:markdown_bold_italic_paragraph	97.33%	89.33%	86.33%	91.33%	74.00%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	89.70%
format:markdown_title	84.38%	59.38%	87.50%	100.00%	6.25%
format:ordered_list	96.73%	95.09%	98.36%	98.36%	95.82%
format:title_brackets	100.00%	82.93%	87.80%	14.63%	31.71%
format:two_answers_with_separator	53.12%	78.12%	90.62%	96.88%	81.25%
repeat:all_sentences_twice	52.00%	45.60%	43.20%	23.20%	13.60%
repeat:before_answer	96.30%	96.30%	96.30%	96.30%	91.11%
repeat:copy_request	96.55%	6.90%	93.10%	62.07%	34.48%
repeat:first_last_same	25.81%	32.26%	29.03%	6.45%	0.00%
repeat:last_sentence	15.86%	57.24%	26.90%	23.45%	0.00%
repeat:sentence_n_times	84.00%	94.67%	78.00%	72.00%	64.00%
marks:end_with_semicolons	93.94%	86.94%	47.61%	11.83%	35.94%
marks:no_commas	93.97%	100.00%	99.38%	93.03%	42.00%
marks:replace_with_asterisks	43.93%	99.90%	48.10%	61.93%	4.17%
marks:replace_with_exclamations	23.37%	100.00%	11.07%	45.44%	8.78%
marks:wrap_in_quotes	34.04%	14.89%	8.51%	2.13%	21.28%
citation:inline	60.00%	57.50%	52.50%	35.00%	67.50%
citation:square_brackets	100.00%	98.93%	100.00%	100.00%	95.36%
citation:start_from_zero	20.59%	82.35%	67.94%	61.47%	8.82%
emoji:banned	96.90%	87.59%	93.79%	75.17%	77.93%
emoji:end	66.67%	60.00%	60.00%	60.00%	53.33%
emoji:frequency	99.67%	99.33%	98.33%	80.00%	71.67%
style:informal	100.00%	100.00%	100.00%	100.00%	99.43%
style:letter	100.00%	100.00%	93.27%	97.35%	100.00%
style:official	83.33%	78.10%	84.52%	85.95%	85.00%
style:poetic	100.00%	100.00%	97.87%	93.62%	100.00%
style:technical	98.98%	85.08%	91.19%	77.97%	85.59%
tone:angry	64.42%	91.40%	80.00%	73.26%	45.81%
tone:humorous	85.33%	95.33%	90.89%	78.44%	74.67%
tone:negative	99.38%	100.00%	97.92%	96.04%	79.17%
tone:positive	99.42%	99.42%	99.42%	96.35%	96.35%
tone:sarcastic	100.00%	100.00%	100.00%	98.54%	84.88%
content:celebrity	96.75%	100.00%	82.75%	76.25%	84.50%
content:jokes	93.10%	96.03%	87.59%	70.52%	75.52%
content:quotes	100.00%	100.00%	94.08%	99.39%	98.16%
language_switch:multilingual	97.45%	99.41%	90.00%	50.78%	14.71%
language_switch:repeat	93.02%	100.00%	81.40%	70.70%	30.47%

Table 49: Detailed evaluation results for Malay language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	93.87%	99.03%	81.61%	89.68%	65.81%
keywords:first_word	90.32%	74.19%	70.97%	61.29%	70.97%
keywords:frequency	92.71%	87.92%	91.88%	86.04%	77.71%
keywords:paragraph_end	86.67%	78.89%	71.67%	70.00%	60.00%
keywords:together	80.00%	75.97%	72.58%	63.71%	72.58%
length:max_words	79.14%	80.62%	71.23%	78.15%	47.75%
length:range_words	67.39%	81.62%	73.95%	65.02%	69.40%
format:addition_at_end	75.00%	68.33%	78.33%	71.67%	61.67%
format:json_output	39.47%	39.47%	44.74%	55.26%	63.16%
format:markdown_bold_italic_paragraph	97.67%	79.67%	88.67%	98.00%	95.00%
format:markdown_highlight	100.00%	96.67%	100.00%	100.00%	99.70%
format:markdown_title	100.00%	81.25%	96.88%	96.88%	34.38%
format:ordered_list	96.55%	85.82%	96.73%	96.55%	99.64%
format:title_brackets	29.27%	85.37%	100.00%	19.51%	14.63%
format:two_answers_with_separator	87.50%	81.25%	43.75%	78.12%	78.12%
repeat:all_sentences_twice	68.00%	52.80%	40.00%	28.00%	11.20%
repeat:before_answer	96.30%	96.30%	92.59%	95.56%	82.96%
repeat:copy_request	100.00%	0.00%	96.55%	96.55%	65.52%
repeat:first_last_same	48.39%	74.19%	41.94%	35.48%	0.00%
repeat:last_sentence	40.00%	57.93%	23.45%	25.52%	4.14%
repeat:sentence_n_times	86.00%	98.00%	75.33%	87.33%	52.00%
marks:end_with_semicolons	93.94%	93.44%	53.44%	27.67%	49.33%
marks:no_commas	100.00%	100.00%	93.45%	90.41%	24.52%
marks:replace_with_asterisks	72.17%	96.67%	43.80%	63.20%	14.97%
marks:replace_with_exclamations	58.74%	99.00%	17.56%	38.00%	24.22%
marks:wrap_in_quotes	95.74%	51.06%	65.96%	46.81%	51.06%
citation:inline	77.50%	72.50%	67.50%	47.50%	75.00%
citation:square_brackets	100.00%	100.00%	100.00%	100.00%	85.71%
citation:start_from_zero	100.00%	85.29%	85.00%	91.47%	100.00%
emoji:banned	96.90%	87.59%	87.59%	81.03%	72.07%
emoji:end	60.00%	32.33%	50.00%	63.33%	50.00%
emoji:frequency	100.00%	82.00%	95.00%	84.33%	77.67%
style:informal	100.00%	98.11%	93.77%	94.34%	98.30%
style:letter	95.92%	94.69%	96.94%	96.73%	99.39%
style:official	81.67%	78.57%	74.29%	70.71%	66.43%
style:poetic	100.00%	100.00%	100.00%	100.00%	100.00%
style:technical	91.36%	78.31%	85.59%	89.15%	90.85%
tone:angry	73.49%	94.88%	77.21%	83.95%	63.49%
tone:humorous	98.00%	99.33%	94.89%	94.00%	96.67%
tone:negative	95.42%	96.67%	100.00%	96.67%	94.38%
tone:positive	99.42%	98.85%	99.42%	99.42%	98.27%
tone:sarcastic	99.27%	100.00%	100.00%	100.00%	91.95%
content:celebrity	92.75%	92.50%	88.25%	81.50%	93.75%
content:jokes	97.24%	100.00%	88.28%	77.07%	100.00%
content:quotes	97.96%	93.88%	90.20%	93.88%	96.73%
language_switch:multilingual	98.04%	92.94%	87.25%	81.96%	28.24%
language_switch:repeat	97.67%	97.67%	86.05%	90.00%	89.30%

Table 50: Detailed evaluation results for Portuguese language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	18.39%	100.00%	91.61%	73.23%	3.23%
keywords:first_word	29.03%	87.10%	51.61%	74.19%	0.00%
keywords:frequency	81.67%	84.38%	67.29%	69.17%	70.42%
keywords:paragraph_end	37.78%	83.89%	45.56%	44.44%	23.89%
keywords:together	45.48%	77.58%	56.77%	56.29%	45.32%
length:max_words	78.95%	100.00%	100.00%	80.18%	86.83%
length:range_words	9.52%	13.67%	14.54%	26.08%	5.69%
format:addition_at_end	51.67%	66.67%	50.00%	41.67%	43.33%
format:json_output	15.79%	5.26%	57.89%	0.00%	0.00%
format:markdown_bold_italic_paragraph	73.67%	73.33%	59.33%	17.67%	43.00%
format:markdown_highlight	82.73%	99.39%	96.06%	99.70%	74.85%
format:markdown_title	21.88%	31.25%	69.06%	100.00%	0.00%
format:ordered_list	97.09%	96.55%	74.00%	65.64%	42.55%
format:title_brackets	75.61%	100.00%	56.10%	51.21%	53.66%
format:two_answers_with_separator	9.38%	100.00%	65.62%	53.12%	59.38%
repeat:all_sentences_twice	0.80%	4.80%	12.00%	12.80%	7.20%
repeat:before_answer	4.44%	96.30%	55.56%	48.15%	0.74%
repeat:copy_request	37.93%	13.79%	0.00%	0.00%	37.93%
repeat:first_last_same	0.00%	3.23%	0.00%	0.00%	0.00%
repeat:last_sentence	0.69%	24.14%	0.00%	2.76%	2.76%
repeat:sentence_n_times	62.67%	88.67%	48.00%	61.33%	42.00%
marks:end_with_semicolons	77.39%	98.17%	81.11%	61.56%	60.44%
marks:no_commas	52.45%	64.48%	71.90%	22.90%	26.62%
marks:replace_with_asterisks	0.00%	81.13%	13.60%	0.00%	0.00%
marks:replace_with_exclamations	0.00%	25.81%	6.56%	0.00%	0.00%
marks:wrap_in_quotes	0.00%	0.00%	0.00%	0.00%	0.00%
citation:inline	20.00%	65.00%	27.50%	37.50%	35.00%
citation:square_brackets	81.07%	100.00%	70.71%	82.14%	63.57%
citation:start_from_zero	38.82%	100.00%	50.00%	81.47%	8.82%
emoji:banned	21.38%	93.79%	90.34%	92.41%	30.00%
emoji:end	6.67%	46.67%	46.67%	20.00%	0.00%
emoji:frequency	75.00%	96.33%	89.67%	82.67%	58.67%
style:informal	97.55%	100.00%	95.66%	93.77%	85.28%
style:letter	36.94%	77.96%	14.49%	22.86%	1.43%
style:official	94.05%	79.05%	85.71%	70.24%	84.52%
style:poetic	65.11%	98.72%	58.09%	82.98%	38.72%
style:technical	71.02%	53.56%	65.59%	39.83%	39.66%
tone:angry	1.63%	77.21%	13.49%	47.21%	0.00%
tone:humorous	34.22%	74.22%	56.67%	51.33%	1.56%
tone:negative	26.88%	92.92%	34.58%	73.33%	15.63%
tone:positive	95.77%	99.42%	96.92%	90.58%	78.65%
tone:sarcastic	0.00%	12.20%	2.44%	26.10%	0.00%
content:celebrity	12.75%	71.00%	32.50%	7.50%	6.75%
content:jokes	4.66%	30.69%	5.17%	8.62%	1.72%
content:quotes	12.24%	95.92%	30.00%	48.98%	0.00%
language_switch:multilingual	23.92%	43.92%	16.27%	60.00%	40.78%
language_switch:repeat	11.63%	81.40%	13.95%	59.77%	41.86%

Table 51: Detailed evaluation results for Quechua language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	98.06%	99.03%	95.16%	94.19%	86.45%
keywords:first_word	90.32%	83.87%	83.87%	74.19%	61.29%
keywords:frequency	64.58%	68.96%	68.13%	65.42%	58.13%
keywords:paragraph_end	68.89%	50.56%	45.56%	35.00%	41.11%
keywords:together	40.32%	54.52%	34.35%	35.16%	26.29%
length:max_words	86.82%	80.24%	84.18%	83.81%	66.10%
length:range_words	82.13%	89.21%	85.51%	73.08%	72.42%
format:addition_at_end	75.00%	93.33%	70.00%	73.33%	80.00%
format:json_output	68.42%	15.79%	39.47%	73.68%	47.37%
format:markdown_bold_italic_paragraph	98.67%	90.67%	90.33%	99.00%	91.33%
format:markdown_highlight	98.79%	100.00%	100.00%	100.00%	98.48%
format:markdown_title	100.00%	96.88%	96.88%	96.88%	37.50%
format:ordered_list	96.55%	98.36%	95.09%	95.09%	97.82%
format:title_brackets	95.12%	87.80%	100.00%	29.27%	43.90%
format:two_answers_with_separator	96.88%	100.00%	93.75%	96.88%	84.38%
repeat:all_sentences_twice	68.00%	60.00%	15.20%	28.00%	24.80%
repeat:before_answer	96.30%	96.30%	96.30%	96.30%	76.30%
repeat:copy_request	100.00%	27.59%	93.10%	62.07%	17.24%
repeat:first_last_same	51.61%	74.19%	41.94%	35.48%	0.00%
repeat:last_sentence	34.48%	57.93%	25.52%	12.41%	0.00%
repeat:sentence_n_times	84.00%	98.00%	93.33%	83.33%	51.33%
marks:end_with_semicolons	94.11%	95.67%	58.22%	27.06%	40.61%
marks:no_commas	100.00%	100.00%	98.24%	90.10%	31.45%
marks:replace_with_asterisks	76.27%	98.40%	58.80%	68.30%	0.00%
marks:replace_with_exclamations	53.89%	99.00%	38.30%	57.22%	11.15%
marks:wrap_in_quotes	63.83%	70.21%	85.11%	57.45%	48.94%
citation:inline	77.50%	70.00%	67.50%	20.00%	52.50%
citation:square_brackets	100.00%	100.00%	100.00%	100.00%	96.43%
citation:start_from_zero	100.00%	94.12%	100.00%	81.76%	96.18%
emoji:banned	96.90%	84.48%	96.55%	68.62%	75.17%
emoji:end	50.00%	30.00%	53.33%	60.00%	46.67%
emoji:frequency	100.00%	97.00%	95.67%	90.33%	75.67%
style:informal	100.00%	100.00%	96.23%	94.34%	100.00%
style:letter	94.90%	98.16%	84.49%	90.41%	96.94%
style:official	82.14%	78.33%	68.81%	67.14%	77.62%
style:poetic	100.00%	100.00%	100.00%	100.00%	97.87%
style:technical	93.05%	82.37%	86.95%	87.63%	84.75%
tone:angry	76.74%	97.21%	83.26%	75.12%	59.30%
tone:humorous	96.00%	98.67%	98.67%	92.67%	87.33%
tone:negative	98.12%	100.00%	95.21%	100.00%	87.71%
tone:positive	98.08%	100.00%	99.42%	99.42%	96.35%
tone:sarcastic	99.27%	100.00%	100.00%	94.39%	85.61%
content:celebrity	100.00%	98.50%	94.75%	76.00%	90.25%
content:jokes	100.00%	98.28%	88.79%	80.86%	89.83%
content:quotes	97.96%	99.39%	96.73%	94.90%	95.31%
language_switch:multilingual	99.41%	100.00%	89.80%	73.14%	20.39%
language_switch:repeat	100.00%	100.00%	93.02%	90.70%	89.30%

Table 52: Detailed evaluation results for Romanian language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	97.10%	98.06%	90.97%	91.94%	92.26%
keywords:first_word	74.19%	74.19%	77.42%	74.19%	67.74%
keywords:frequency	78.75%	82.29%	84.58%	82.92%	70.83%
keywords:paragraph_end	81.67%	73.33%	62.22%	52.78%	46.11%
keywords:together	74.19%	77.26%	74.84%	69.35%	56.61%
length:max_words	88.08%	93.52%	89.48%	92.09%	76.36%
length:range_words	86.16%	98.49%	81.17%	75.51%	71.33%
format:addition_at_end	76.67%	61.67%	56.67%	45.00%	55.00%
format:json_output	44.74%	2.63%	34.21%	63.16%	47.37%
format:markdown_bold_italic_paragraph	98.00%	85.67%	89.67%	88.33%	78.00%
format:markdown_highlight	100.00%	100.00%	100.00%	100.00%	85.45%
format:markdown_title	100.00%	96.88%	96.88%	96.88%	37.50%
format:ordered_list	96.55%	96.73%	98.36%	93.09%	97.09%
format:title_brackets	100.00%	100.00%	100.00%	26.83%	29.27%
format:two_answers_with_separator	96.88%	100.00%	93.75%	93.75%	81.25%
repeat:all_sentences_twice	72.00%	64.80%	32.00%	22.40%	31.20%
repeat:before_answer	96.30%	96.30%	96.30%	92.59%	77.04%
repeat:copy_request	100.00%	0.00%	55.17%	68.97%	24.14%
repeat:first_last_same	77.42%	67.74%	51.61%	41.94%	0.00%
repeat:last_sentence	33.79%	61.38%	8.97%	26.90%	11.72%
repeat:sentence_n_times	86.00%	91.33%	86.67%	88.67%	54.67%
marks:end_with_semicolons	94.11%	94.11%	58.39%	9.61%	66.17%
marks:no_commas	99.90%	100.00%	99.59%	99.59%	78.03%
marks:replace_with_asterisks	89.63%	93.20%	49.93%	47.83%	6.57%
marks:replace_with_exclamations	95.22%	100.00%	36.67%	66.44%	40.07%
marks:wrap_in_quotes	78.72%	87.23%	74.47%	59.57%	48.94%
citation:inline	72.50%	72.50%	72.50%	35.00%	60.00%
citation:square_brackets	100.00%	100.00%	100.00%	100.00%	96.43%
citation:start_from_zero	100.00%	88.24%	99.12%	96.18%	96.18%
emoji:banned	93.79%	81.38%	90.69%	84.48%	75.17%
emoji:end	63.33%	59.00%	66.67%	60.00%	53.33%
emoji:frequency	99.67%	98.33%	93.00%	93.67%	80.67%
style:informal	100.00%	100.00%	94.34%	96.23%	98.30%
style:letter	97.96%	100.00%	95.31%	96.33%	100.00%
style:official	80.48%	84.29%	79.29%	72.62%	74.05%
style:poetic	100.00%	100.00%	95.74%	93.62%	100.00%
style:technical	86.44%	78.98%	81.19%	82.20%	72.88%
tone:angry	76.28%	93.02%	86.05%	82.56%	70.93%
tone:humorous	98.67%	99.33%	96.44%	93.78%	91.56%
tone:negative	99.38%	100.00%	97.92%	95.83%	83.75%
tone:positive	98.85%	98.27%	95.00%	98.85%	97.69%
tone:sarcastic	99.27%	100.00%	100.00%	97.07%	96.83%
content:celebrity	100.00%	97.00%	89.25%	89.50%	87.75%
content:jokes	98.97%	100.00%	86.21%	83.28%	90.86%
content:quotes	97.35%	100.00%	98.78%	96.12%	91.43%
language_switch:multilingual	100.00%	100.00%	89.22%	43.73%	6.08%
language_switch:repeat	100.00%	100.00%	87.67%	95.35%	87.67%

Table 53: Detailed evaluation results for Swedish language

Metrics	GPT-4o	Claude-3.5 Sonnet	Gemini-1.5 Pro	Gemini-1.5 Flash	GPT-3.5 Turbo
keywords:banned	99.03%	97.10%	93.23%	93.23%	83.23%
keywords:first_word	16.13%	16.13%	16.13%	12.90%	6.45%
keywords:frequency	78.54%	77.71%	68.96%	78.75%	68.12%
keywords:paragraph_end	20.56%	18.33%	18.33%	18.33%	13.89%
keywords:together	66.29%	42.74%	27.10%	27.58%	35.32%
length:max_words	99.93%	100.00%	98.50%	99.83%	94.95%
length:range_words	31.12%	12.12%	63.95%	58.00%	35.17%
format:addition_at_end	80.00%	38.33%	55.00%	63.33%	50.00%
format:json_output	81.58%	28.95%	55.26%	71.05%	10.53%
format:markdown_bold_italic_paragraph	99.33%	78.00%	79.67%	87.33%	37.00%
format:markdown_highlight	100.00%	82.73%	100.00%	97.27%	24.85%
format:markdown_title	26.56%	29.69%	30.00%	28.75%	0.00%
format:ordered_list	95.09%	88.18%	98.36%	98.36%	81.82%
format:title_brackets	74.73%	22.46%	12.13%	14.27%	14.57%
format:two_answers_with_separator	96.88%	75.00%	59.38%	96.88%	34.38%
repeat:all_sentences_twice	0.00%	0.00%	0.00%	0.00%	0.00%
repeat:before_answer	100.00%	81.48%	81.48%	95.56%	22.22%
repeat:copy_request	96.55%	0.00%	17.24%	79.31%	0.00%
repeat:first_last_same	0.00%	0.00%	0.00%	0.00%	0.00%
repeat:last_sentence	7.59%	4.14%	0.00%	0.00%	0.00%
repeat:sentence_n_times	82.00%	60.00%	64.00%	59.33%	28.00%
marks:end_with_semicolons	78.83%	85.39%	51.72%	26.00%	13.56%
marks:no_commas	27.55%	92.31%	89.45%	79.03%	59.79%
marks:replace_with_asterisks	23.07%	86.07%	50.80%	48.67%	2.57%
marks:replace_with_exclamations	30.74%	88.89%	1.93%	23.85%	0.00%
marks:wrap_in_quotes	42.55%	44.68%	59.57%	34.04%	0.00%
citation:inline	30.00%	20.00%	22.50%	32.50%	27.50%
citation:square_brackets	100.00%	85.71%	96.43%	90.71%	51.79%
citation:start_from_zero	94.12%	88.24%	92.35%	67.35%	29.41%
emoji:banned	96.90%	93.79%	93.45%	86.90%	82.76%
emoji:end	60.00%	54.67%	60.00%	53.33%	0.00%
emoji:frequency	93.33%	99.67%	85.33%	85.00%	66.67%
style:informal	95.66%	96.23%	100.00%	98.11%	57.55%
style:letter	98.16%	84.69%	73.67%	83.47%	41.22%
style:official	92.14%	80.95%	88.10%	83.81%	95.00%
style:poetic	100.00%	90.21%	89.36%	99.36%	45.32%
style:technical	98.31%	85.08%	87.12%	92.37%	68.64%
tone:angry	45.58%	77.67%	49.30%	56.28%	1.63%
tone:humorous	90.44%	88.22%	93.11%	87.78%	20.22%
tone:negative	54.17%	87.71%	77.71%	82.29%	15.83%
tone:positive	97.50%	87.88%	97.50%	96.92%	82.88%
tone:sarcastic	58.05%	97.56%	76.59%	86.59%	10.98%
content:celebrity	98.50%	92.75%	87.25%	83.75%	22.25%
content:jokes	94.83%	89.14%	84.66%	73.28%	13.28%
content:quotes	99.39%	85.10%	90.00%	80.41%	23.88%
language_switch:multilingual	99.41%	96.86%	84.51%	70.98%	7.25%
language_switch:repeat	100.00%	90.70%	80.70%	90.00%	0.00%

Table 54: Detailed evaluation results for Tamil language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	98.06%	100.00%	98.06%	97.10%	85.48%
keywords:first_word	12.90%	9.68%	12.90%	6.45%	12.90%
keywords:frequency	75.21%	86.25%	73.96%	77.71%	75.21%
keywords:paragraph_end	36.67%	31.11%	33.89%	25.00%	13.89%
keywords:together	64.03%	59.03%	53.71%	52.74%	49.68%
length:max_words	100.00%	97.37%	98.86%	99.16%	91.89%
length:range_words	51.65%	44.49%	54.55%	65.52%	18.63%
format:addition_at_end	56.67%	73.33%	68.33%	76.67%	33.33%
format:json_output	76.32%	26.32%	44.74%	65.79%	2.63%
format:markdown_bold_italic_paragraph	25.00%	86.33%	55.33%	72.00%	51.33%
format:markdown_highlight	100.00%	80.91%	96.97%	99.70%	17.88%
format:markdown_title	20.00%	43.75%	38.75%	19.69%	7.50%
format:ordered_list	92.91%	84.73%	96.73%	95.09%	65.27%
format:title_brackets	7.16%	30.49%	7.32%	2.44%	32.02%
format:two_answers_with_separator	78.12%	81.25%	71.88%	87.50%	59.38%
repeat:all_sentences_twice	48.00%	28.00%	14.40%	10.40%	0.80%
repeat:before_answer	96.30%	81.48%	59.26%	81.48%	13.33%
repeat:copy_request	93.10%	3.45%	89.66%	86.21%	3.45%
repeat:first_last_same	70.97%	61.29%	51.61%	38.71%	0.00%
repeat:last_sentence	13.10%	20.69%	0.00%	6.90%	3.45%
repeat:sentence_n_times	89.33%	90.00%	60.00%	80.67%	26.67%
marks:end_with_semicolons	99.33%	99.17%	49.33%	26.33%	47.89%
marks:no_commas	58.93%	90.52%	89.90%	91.76%	83.97%
marks:replace_with_asterisks	36.73%	90.53%	45.50%	58.37%	0.00%
marks:replace_with_exclamations	24.22%	88.89%	36.52%	25.96%	0.00%
marks:wrap_in_quotes	38.30%	44.68%	51.06%	31.91%	0.00%
citation:inline	27.50%	12.50%	40.00%	27.50%	27.50%
citation:square_brackets	96.43%	89.29%	71.79%	92.86%	61.79%
citation:start_from_zero	87.35%	76.47%	89.41%	80.88%	9.12%
emoji:banned	93.79%	93.45%	87.24%	75.17%	44.48%
emoji:end	55.67%	46.67%	63.33%	66.67%	0.00%
emoji:frequency	93.00%	99.33%	95.33%	84.00%	55.00%
style:informal	94.72%	92.64%	96.42%	96.98%	71.89%
style:letter	71.84%	66.33%	49.18%	53.47%	51.02%
style:official	68.57%	65.71%	85.00%	86.67%	88.81%
style:poetic	81.70%	91.06%	80.21%	97.87%	19.15%
style:technical	94.58%	83.39%	87.97%	96.27%	57.97%
tone:angry	36.74%	74.19%	36.51%	48.84%	1.63%
tone:humorous	86.00%	89.11%	89.11%	86.00%	25.56%
tone:negative	57.50%	84.38%	62.50%	86.67%	6.46%
tone:positive	96.35%	90.00%	98.85%	98.27%	81.92%
tone:sarcastic	54.88%	87.32%	73.41%	57.07%	12.68%
content:celebrity	82.50%	79.50%	74.50%	79.25%	22.75%
content:jokes	87.93%	84.48%	75.52%	74.83%	32.76%
content:quotes	77.76%	83.06%	86.12%	94.08%	27.96%
language_switch:multilingual	88.43%	86.27%	83.92%	61.18%	1.96%
language_switch:repeat	86.05%	97.67%	76.05%	93.02%	0.00%

Table 55: Detailed evaluation results for Telugu language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	97.10%	99.03%	72.26%	67.42%	68.06%
keywords:first_word	58.06%	48.39%	54.84%	61.29%	61.29%
keywords:frequency	76.67%	80.00%	77.92%	79.17%	76.25%
keywords:paragraph_end	80.00%	57.78%	43.89%	43.89%	38.89%
keywords:together	76.94%	60.65%	76.77%	78.71%	65.32%
length:max_words	93.14%	86.49%	90.73%	91.03%	89.69%
length:range_words	78.86%	74.06%	77.67%	76.36%	52.38%
format:addition_at_end	53.33%	35.00%	65.00%	65.00%	83.33%
format:json_output	92.11%	23.68%	28.95%	55.26%	47.37%
format:markdown_bold_italic_paragraph	98.33%	87.33%	66.00%	84.67%	83.67%
format:markdown_highlight	100.00%	91.82%	100.00%	100.00%	90.30%
format:markdown_title	81.25%	75.00%	93.75%	56.25%	3.12%
format:ordered_list	94.73%	82.18%	96.36%	96.55%	98.36%
format:title_brackets	2.44%	4.88%	21.95%	2.44%	4.88%
format:two_answers_with_separator	75.00%	75.00%	46.88%	96.88%	93.75%
repeat:all_sentences_twice	76.00%	64.80%	27.20%	11.20%	20.80%
repeat:before_answer	92.59%	92.59%	88.15%	92.59%	65.93%
repeat:copy_request	86.21%	0.00%	68.97%	68.97%	10.34%
repeat:first_last_same	67.74%	48.39%	51.61%	29.03%	0.00%
repeat:last_sentence	22.76%	51.03%	13.10%	23.45%	9.66%
repeat:sentence_n_times	84.00%	98.00%	82.00%	60.67%	60.67%
marks:end_with_semicolons	87.56%	94.11%	16.61%	9.83%	30.89%
marks:no_commas	93.10%	100.00%	83.83%	73.90%	16.03%
marks:replace_with_asterisks	49.30%	87.60%	38.70%	43.33%	6.57%
marks:replace_with_exclamations	8.22%	96.19%	16.74%	15.30%	7.19%
marks:wrap_in_quotes	68.09%	74.47%	59.57%	29.79%	29.79%
citation:inline	65.00%	85.00%	62.50%	37.50%	62.50%
citation:square_brackets	100.00%	96.43%	100.00%	100.00%	92.86%
citation:start_from_zero	97.06%	85.29%	77.06%	72.35%	98.24%
emoji:banned	93.79%	84.14%	96.90%	83.79%	92.76%
emoji:end	70.00%	66.67%	70.00%	70.00%	48.00%
emoji:frequency	96.33%	99.67%	96.33%	85.00%	79.33%
style:informal	96.98%	97.55%	97.55%	94.53%	85.28%
style:letter	97.96%	97.35%	97.35%	98.78%	93.88%
style:official	94.76%	81.90%	89.52%	90.71%	90.24%
style:poetic	97.87%	95.74%	100.00%	100.00%	95.74%
style:technical	99.49%	90.51%	96.27%	95.08%	92.54%
tone:angry	51.40%	77.91%	75.81%	62.09%	51.86%
tone:humorous	91.56%	99.33%	95.56%	87.78%	73.78%
tone:negative	78.75%	99.38%	97.29%	98.75%	78.54%
tone:positive	97.50%	98.08%	98.27%	96.92%	91.73%
tone:sarcastic	70.49%	100.00%	96.83%	93.41%	57.56%
content:celebrity	97.50%	97.50%	91.25%	83.00%	77.75%
content:jokes	50.00%	56.90%	30.52%	26.21%	36.21%
content:quotes	97.96%	95.92%	68.37%	55.31%	89.39%
language_switch:multilingual	98.04%	95.49%	96.47%	81.18%	42.75%
language_switch:repeat	100.00%	97.67%	90.70%	94.65%	73.72%

Table 56: Detailed evaluation results for Turkish language

Metrics	**GPT-4o**	**Claude-3.5 Sonnet**	**Gemini-1.5 Pro**	**Gemini-1.5 Flash**	**GPT-3.5 Turbo**
keywords:banned	99.03%	100.00%	92.26%	95.16%	43.87%
keywords:first_word	48.39%	70.97%	41.94%	51.61%	61.29%
keywords:frequency	76.25%	95.00%	74.37%	69.17%	60.83%
keywords:paragraph_end	81.11%	72.22%	37.22%	38.89%	25.00%
keywords:together	60.32%	58.39%	45.65%	33.39%	46.77%
length:max_words	88.51%	100.00%	100.00%	94.70%	55.91%
length:range_words	73.12%	38.67%	32.34%	43.61%	6.31%
format:addition_at_end	96.67%	96.67%	80.00%	85.00%	50.00%
format:json_output	55.26%	18.42%	44.74%	84.21%	28.95%
format:markdown_bold_italic_paragraph	87.33%	91.00%	51.00%	65.00%	54.33%
format:markdown_highlight	99.39%	100.00%	100.00%	100.00%	26.36%
format:markdown_title	84.38%	43.75%	87.50%	96.88%	3.12%
format:ordered_list	95.09%	96.55%	96.73%	92.91%	79.45%
format:title_brackets	29.27%	9.76%	9.76%	9.76%	12.20%
format:two_answers_with_separator	81.25%	100.00%	81.25%	75.00%	62.50%
repeat:all_sentences_twice	60.00%	53.60%	14.40%	4.80%	11.20%
repeat:before_answer	96.30%	88.89%	56.30%	60.74%	13.33%
repeat:copy_request	100.00%	6.90%	62.07%	58.62%	0.00%
repeat:first_last_same	70.97%	77.42%	67.74%	25.81%	0.00%
repeat:last_sentence	16.55%	40.69%	3.45%	2.76%	8.97%
repeat:sentence_n_times	85.33%	85.33%	80.00%	76.00%	34.67%
marks:end_with_semicolons	82.22%	78.17%	4.06%	6.94%	32.67%
marks:no_commas	31.41%	44.45%	12.69%	20.97%	28.62%
marks:replace_with_asterisks	63.67%	100.00%	44.90%	5.07%	4.17%
marks:replace_with_exclamations	21.48%	91.70%	5.56%	0.00%	0.00%
marks:wrap_in_quotes	0.00%	4.26%	2.13%	0.00%	0.00%
citation:inline	35.00%	45.00%	42.50%	32.50%	25.00%
citation:square_brackets	88.21%	100.00%	96.79%	80.00%	48.57%
citation:start_from_zero	85.88%	90.88%	74.41%	45.00%	17.94%
emoji:banned	93.79%	96.90%	93.79%	84.14%	76.21%
emoji:end	49.00%	66.67%	52.33%	41.33%	0.00%
emoji:frequency	99.33%	100.00%	95.67%	87.67%	59.00%
style:informal	98.30%	100.00%	98.11%	100.00%	80.19%
style:letter	91.43%	92.65%	42.04%	40.20%	4.29%
style:official	94.76%	95.00%	95.00%	96.43%	97.86%
style:poetic	97.45%	97.87%	95.32%	93.40%	32.98%
style:technical	94.58%	88.47%	89.49%	98.98%	70.17%
tone:angry	36.74%	51.40%	49.77%	26.05%	3.26%
tone:humorous	80.22%	81.78%	94.67%	88.22%	9.33%
tone:negative	68.33%	79.58%	58.75%	34.58%	18.75%
tone:positive	98.85%	100.00%	96.92%	98.85%	73.08%
tone:sarcastic	49.02%	66.83%	72.44%	62.93%	10.73%
content:celebrity	89.50%	98.50%	85.00%	83.00%	10.25%
content:jokes	80.00%	66.72%	61.21%	67.41%	0.00%
content:quotes	89.18%	92.65%	85.92%	72.86%	0.00%
language_switch:multilingual	91.57%	94.90%	67.45%	21.96%	3.33%
language_switch:repeat	90.70%	100.00%	74.42%	69.77%	0.00%

Table 57: Detailed evaluation results for Zulu language

C.4 Cross-lingual Results from English and Original Results in 22 Languages

Resource	**Language**	**Cross-lingual Result from English**	**Original Result**
High Resource	Swedish	71.11%	69.39%
	Portuguese	70.03%	70.79%
	French	68.64%	68.10%
	Italian	69.39%	66.57%
	Chinese	66.59%	67.18%
	Japanese	65.80%	66.01%
	Average High	**68.59%**	**68.01%**
Medium Resource	Filipino	69.24%	60.21%
	Romanian	67.87%	67.33%
	Indonesian	69.92%	64.25%
	Malay	68.33%	63.94%
	Turkish	70.45%	62.77%
	Korean	67.21%	60.74%
	Bengali	57.29%	38.58%
	Hindi	66.01%	51.41%
	Average Medium	**67.04%**	**58.65%**
Low Resource	Kyrgyz	59.09%	29.34%
	Armenian	51.71%	33.02%
	Georgian	52.51%	34.28%
	Malagasy	55.88%	29.74%
	Zulu	49.80%	30.14%
	Tamil	53.64%	33.22%
	Telugu	55.60%	33.00%
	Quechua	43.59%	29.52%
	Average Low	**52.73%**	**31.53%**

Table 58: Comparison of Cross-lingual Results from English and Original Results in 22 Languages on GPT-3.5 Turbo

Resource	Language	Cross-lingual Result from English	Original Result
High Resource	Swedish	79.13%	77.67%
	Portuguese	81.45%	78.17%
	French	80.39%	76.27%
	Italian	80.04%	76.32%
	Chinese	74.46%	74.79%
	Japanese	73.59%	77.17%
	Average High	**78.18%**	**76.73%**
Medium Resource	Filipino	80.57%	75.75%
	Romanian	77.77%	76.10%
	Indonesian	78.74%	74.09%
	Malay	80.45%	73.92%
	Turkish	80.07%	70.83%
	Korean	73.61%	70.89%
	Bengali	76.18%	68.56%
	Hindi	75.41%	67.92%
	Average Medium	**77.85%**	**72.26%**
Low Resource	Kyrgyz	78.52%	74.89%
	Armenian	74.91%	72.60%
	Georgian	74.29%	68.55%
	Malagasy	79.84%	64.73%
	Zulu	76.92%	60.29%
	Tamil	72.42%	68.46%
	Telugu	73.54%	68.77%
	Quechua	66.48%	48.39%
	Average Low	**74.62%**	**65.84%**

Table 59: Comparison of Cross-lingual Results from English and Original Results in 22 Languages on Gemini-1.5 Flash