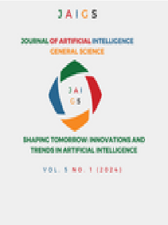
Journal of Artificial Intelligence General Science (JAIGS)

ISSN: 3006-4023 (Online), Volume 5, Issue 1,2024 DOI: 10.60087

Home page https://ojs.boulibrary.com/index.php/JAIGS



J A I G S
JOURNAL OF ARTIFICIAL INTELLIGENCE
GENERAL SCIENCE
SHAPING TOMORROW: INNOVATIONS AND
TRENDS IN ARTIFICIAL INTELLIGENCE
VOL. 5 NO. 1 (2024)

IFEval-Extended: Enhancing Instruction-Following Evaluation in Large Language Models through Dynamic Prompt Generation

Bohdan Kovalevskyi

Independent researcher, United States.

ABSTRACT

This paper introduces IFEval-Extended, an innovative benchmark for evaluating the instruction-following capabilities of Large Language Models (LLMs). Building upon the foundational principles of the existing IFEval framework, IFEval-Extended addresses the limitations of predefined prompts by employing a dynamic, generative approach to instruction synthesis. This method allows for the creation of thousands of unique, human-like instructions from a single base template, mitigating the risk of overfitting and enhancing the diversity and robustness of the evaluation process. The benchmark extends the original set of instruction categories in IFEval, providing a more granular assessment of LLM performance across various parameters such as language structure, keyword usage, and response formatting. The study evaluates state-of-the-art LLMs, including GPT-4o, LLama 3.1 (8B), and LLama 3 (70B), using strict and loose accuracy metrics. Results reveal that while models excel in handling simpler instructions, they struggle with complex tasks requiring precise adherence to multiple constraints. The findings highlight the strengths and weaknesses of current LLM capabilities, offering valuable insights for model development and real-world applications. IFEval-Extended contributes to the ongoing development of more robust, scalable, and objective LLM evaluation methods, thereby advancing the field of Natural Language Processing.

Keywords: Large Language Models (LLMs), Natural Language Processing (NLP), Instruction Following, Benchmark, Evaluation Framework, Dynamic Prompt Generation, Overfitting, Scalability, Generalizability, Model Performance, GPT-4o, LLama 3.

ARTICLE INFO: *Received: 01.09.2024 Accepted: 20.09.2024 Published: 01.10.2024*

1. INTRODUCTION

The development of Large Language Models (LLMs) has had a profound impact on the field of Natural Language Processing (NLP), enabling applications that range from conversational AI to content generation and summarization (Brown et al., 2020; OpenAI, 2023). One of the most important of these capabilities is an LLM's ability to accurately follow instructions presented in natural language, which is a critical function for tasks where the user's intent must be understood and executed with precision (Chowdhery et al., 2022; Victor et al., 2022). For example, instruction-following competence is essential in scenarios such as medical diagnosis, legal drafting, or autonomous systems, where misinterpretation can have significant consequences.

Despite the growing reliance on LLMs in sensitive applications, evaluating their ability to follow instructions remains an unsolved problem. Traditional human assessment methods can be thorough, but they suffer from subjectivity, high costs, and limited scalability (Ouyang et al., 2022; Taori et al., 2023). In turn, automated assessments that are based on models or benchmarks may lack objectivity or suffer from bias inherent in the evaluator's model (Chen et al., 2021; Skopek et al., 2023). In this regard, the lack of a standardized, reliable framework for evaluating instruction following ability creates a critical gap in our understanding of models' performance. There have been several attempts to address this gap, including developing quantitative metrics such as IFEval, a widely adopted assessment framework designed to evaluate instruction following performance (Zhou et al., 2023). IFEval introduces the concept of "verifiable instructions"—atomic directives that can be objectively verified for compliance. Examples of such tasks include tasks such as "write at least 450 words" or "include the keyword 'technology' three times in your response". By providing a controlled set of predefined prompts, IFEval allows for repeatable, unbiased evaluations across models. IFEval's framework includes 25 instruction types that are categorized by their verifiability and relevance to real-world tasks (Zhou et al., 2023). Models are evaluated based on strict and loose criteria to account for differences in response wording and format. This approach has significantly advanced the field by providing a standardized, interpretable, and scalable method for benchmarking LLMs.

However, IFEval's reliance on hand-picked prompts poses a notable limitation. Predefined prompts are prone to overfitting because models can be explicitly trained on publicly available evaluation datasets. This situation reduces the benchmark's ability to generalize to unseen tasks, undermining its usefulness for assessing broader LLM instruction-following capabilities. Given the limitations of existing benchmarks, there is a pressing need for an evaluation framework that would balance reproducibility, scalability, and robustness to overfitting. While publicly available benchmarks are valuable for standardization, they may unintentionally become training datasets, allowing LLMs to achieve artificially high performance without actual improvements in instruction comprehension or adherence (Chang et al., 2023; Sun et al., 2023). In order to address these issues, this study presents IFEval-Extended, an extended version of IFEval, which is designed to overcome the limitations of predefined hints through dynamic automatic generation. IFEval-Extended uses a generative approach to instruction synthesis, allowing for the generation of thousands of unique hints from a single base template. This method not only increases the diversity of instructions but also ensures that the evaluation process remains robust to overfitting. By including a wider range of instruction types and dynamically generated prompts, IFEval-Extended aims to provide a more comprehensive assessment of LLMs' ability to follow instructions accurately.

In this regard, this article presents the design, implementation, and evaluation of IFEval-Extended, a benchmark that builds on the fundamental principles of IFEval while addressing its limitations. Particularly, the study makes the following contributions:

Dynamic Prompt Generation: IFEval-Extended introduces a new scenario-based methodology for generating unique, human-like prompts. This approach eliminates the reliance on manual curation, enabling the generation of diverse prompts tailored to specific evaluation criteria.

Enhanced Instruction Coverage: By expanding the original set of prompt types, IFEval-Extended provides a more detailed evaluation of LLM model performance across a variety of parameters, including language structure, keyword usage, and response formatting.

Compatibility with Existing Benchmarks: IFEval-Extended maintains compatibility with IFEval, allowing direct comparisons between the two frameworks and ensuring continuity in LLM evaluation practices.

Comprehensive Evaluation: Using strict and loose accuracy metrics, this study evaluates the performance of state-of-the-art LLM models, including GPT-4 and LLama 3.1, on the IFEval-Extended benchmark. The results highlight the strengths and weaknesses of the current model capabilities and suggest areas that require further improvement.

By addressing the limitations of existing benchmarks, IFEval-Extended contributes to the ongoing development of robust, scalable, and objective LLM evaluation methods. At the same time, it aims to advance the field of NLP by providing a robust framework for benchmarking instruction following performance.

2. METHODOLOGY

2.1. Core Technology: IFEval

Instruction-Following Evaluation (IFEval) represents a significant advance in benchmarking the instruction-following abilities of Large Language Models (LLMs). It focuses on verifiable instructions—clear, specific directives that allow objective compliance testing (Zhou et al., 2023). For example, such a directive might be “end your response with the exact phrase ‘thank you’”—the accuracy of which can be assessed using automated scripts. By classifying instructions into 25 different types, IFEval provides a standardized framework for testing LLMs on various tasks, ensuring reproducibility and scalability in model evaluation.

The IFEval methodology is based on constructing prompts that contain one or more verifiable instructions. Each prompt tests the LLM’s ability to accurately follow the specified instructions. Prompts are manually curated to ensure diversity and logical consistency, while strict and loose accuracy metrics are used to evaluate the model’s responses. These metrics account for subtle variations in the wording of responses that might otherwise lead to false negatives (Zhou et al., 2023).

While IFEval offers a number of advantages such as transparency and objectivity, its reliance on predefined prompts introduces limitations. Publicly available benchmarks can be used during model training, allowing LLMs to achieve artificially high performance without real improvements in instruction-following abilities (Chang et al., 2023). Addressing these limitations is critical to advancing LLM assessment frameworks.

2.2. Benchmark Extension: IFEval-Extended

IFEval-Extended is designed to overcome the limitations of manual curation and predefined prompts inherent in IFEval. By using dynamic prompt generation and expanding the range of instruction types, IFEval-Extended improves the diversity and robustness of LLM evaluation.

Dynamic Prompt Generation. The central innovation of IFEval-Extended is its approach to dynamic prompt generation. Rather than relying on a fixed set of manually crafted prompts, IFEval-Extended uses a generative framework to create unique instructions tailored to specific evaluation goals. This process starts with a single base prompt: “You are an AI assistant that generates a single, diverse, human-like instruction for language models. Each time you're prompted, create one unique instruction that mimics how people typically interact with AI.” This base prompt is sent to the LLM to create the human-like instructions. Each generated instruction is then programmatically combined with other predefined elements to create a variety of assessment tasks. For example, the instruction “Include keywords ‘assembly’ and ‘balance’ in your response” could be paired with prompts such as “Write a summary of the novel *To Kill a Mockingbird*.” This approach allows for the automatic generation of thousands of unique prompts, which greatly exceeds the diversity achievable with manual curation. The dynamic nature of this methodology addresses the problem of overfitting to predefined prompts. Since the generated prompts are not fixed, LLMs cannot rely on prior familiarity with specific tasks, making the scoring process more robust to manipulation.

Instruction Categories and Compatibility Rules. IFEval-Extended builds on the 25 instruction categories provided in IFEval and extends them with new variations and combinations. Each instruction is parameterized, allowing for generating multiple variations. For example, the instruction “Your response must contain exactly {N} words” can be parameterized with different values of N, such as 100, 200, or 500. This feature increases the granularity and scope of scoring. In order to ensure logical consistency, IFEval-Extended includes compatibility rules that prevent conflicting instructions from appearing in the same prompt. For example, a guideline that requires a response to be exactly three paragraphs long cannot coexist with a guideline that limits responses to fewer than 50 words. These rules are coded into the prompt generation script to automatically eliminate invalid combinations. Table 1 illustrates the extended list of instruction categories used in IFEval-Extended. Each category includes representative examples that demonstrate the diversity of tasks covered by the benchmark. This comprehensive categorization ensures that IFEval-Extended evaluates LLMs across a broad range of real-world learning types, providing a deeper understanding of their strengths and weaknesses.

Table 1. Extended List of Instruction Categories Used in IFEval-Extended.

Category	**Instruction Example**
Keywords	Include keywords {keyword1} and {keyword2} in your response.
Length Constraints	Write exactly {N} paragraphs, separated by ***.
Detectable Format	Wrap your entire response in JSON format.
Language	Respond entirely in {language}, no other language is allowed.
Punctuation	Do not use any commas in your response.

Evaluation Process. The IFEval-Extended evaluation framework maintains compatibility with the IFEval methodology, ensuring that results are directly comparable. Each model's performance is evaluated using both strict and loose accuracy metrics:

Strict Accuracy: This metric evaluates whether all instructions in the prompt are followed accurately. For example, if the prompt specifies that the response must contain three paragraphs and include the keywords "assembly" and "balance", then the model must meet both conditions to be considered accurate.

Loose Accuracy: Recognizing that minor changes in wording or formatting can lead to false negatives, this metric allows for some flexibility in evaluating the fit. For example, a response that includes the required keywords but deviates slightly in paragraph formatting may still be considered accurate according to a loose criterion.

The evaluation process consists of the following steps:

Prompt Generation: dynamically generate a diverse set of prompts, including multiple instruction types where applicable.

Response Collection: use API calls to collect responses from the LLM being evaluated.

Instruction Verification: use deterministic scripts to check the conformance of each instruction. The scripts account for common variations such as markdown formatting or extraneous text.

Metric Calculation: calculate strict and loose accuracy scores at both the prompt and instruction levels. These scores provide a detailed view of the model's performance on various tasks.

Implementation Details. IFEval-Extended is implemented as an open-source Python package, enabling reproducibility and community contribution. The main components include:

Instruction Generator: a script that synthesizes unique prompts by combining base instructions with random parameters.

Evaluation Scripts: deterministic scripts to check conformance for each instruction type.

Dataset Management: tools for storing and managing generated prompts, responses, and evaluation results.

The implementation is designed to be modular, allowing researchers to extend the benchmark with additional instruction categories or evaluation metrics. The entire source code and datasets are available on GitHub (Kovbo, 2024).

Benefits of IFEval-Extended. By addressing the limitations of manual curation and static prompts, IFEval-Extended offers several benefits:

Scalability: the generative approach allows for the creation of large and diverse datasets with minimal human intervention.

Resistance to Overfitting: dynamically generated prompts reduce the risk of models being explicitly trained on the evaluation dataset.

Broader Coverage: the extended range of training types enables a more comprehensive evaluation of LLM capabilities.

3. RESULTS

The evaluation of IFEval-Extended was conducted on three state-of-the-art Large Language Models (LLM): GPT-4o, LLama 3.1 (8B parameters), and LLama 3 (70B parameters). These models were chosen for their widespread use and their distinct architectures, which provide a representative comparison of LLM capabilities at different parameter scales. Each model was tested using the IFEval-Extended benchmark. The evaluation involved generating responses to dynamically synthesized prompts that contained one or more instructions to be tested. Responses were collected via API calls and subsequently verified using the automated scripts described in the methodology section. Results were analyzed using strict and loose accuracy metrics measured at both prompt and instruction levels.

Table 2 provides an overview of the prompt-level and instruction-level accuracy achieved by each model under strict and loose evaluation criteria. The results indicate that GPT-4o achieved the highest accuracy across all metrics, followed closely by LLama 3.1. LLama 3 exhibited slightly lower performance, suggesting that increased model parameters alone do not guarantee superior instruction-following capabilities.

Table 2. Prompt-Level and Instruction-Level Accuracy Across Models.

	Strict (%)	Loose (%)	Strict (%)	Loose (%)
GPT-4o	80.8	83.0	83.9	86.2
LLama 3.1 (8B)	79.0	83.2	83.3	86.8
LLama 3 (70B)	77.2	82.4	82.5	85.7

In order to provide a more granular view, instruction-level accuracy was further analyzed across different instruction categories. The results are summarized in Table 3 and the graph in Figure 1.

Table 3. Instruction-Level Accuracy by Category (Strict Metric).

Category	GPT-4o (%)	LLama 3.1 (8B) (%)	LLama 3 (70B) (%)
Change Case	84.9	86.1	79.7
Combination	95.1	95.0	54.5
Detectable Content	77.3	78.6	73.9
Detectable Format	84.6	86.2	89.8
Keywords	74.5	73.6	73.6

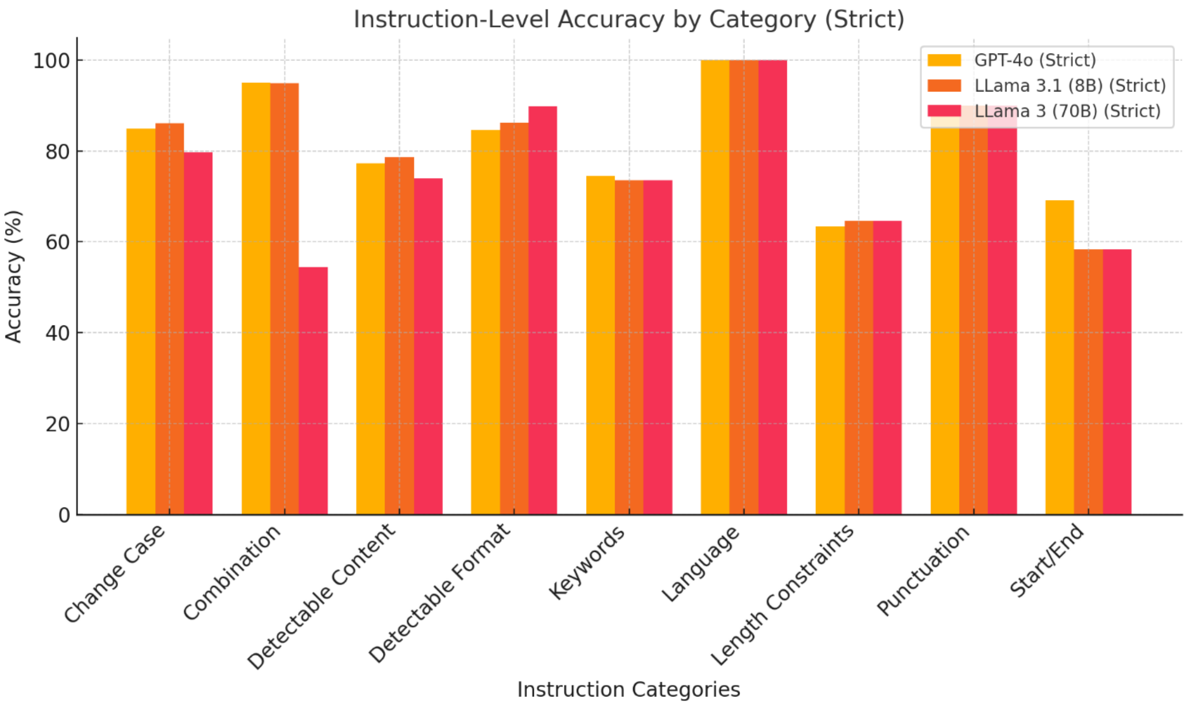


Figure 1. Instruction-Level Accuracy by Category (Strict Metric).

In this regard, it is possible to summarize the key findings:

- All models performed exceptionally well in the Language category, achieving 100% accuracy, reflecting their ability to generate responses in the given languages.
- Categories such as Combination and Punctuation also showed high accuracy rates, with GPT-4o slightly outperforming the other models.
- The lower accuracy scores in the "Length Constraints" and "Start/End" categories indicate persistent issues across all models in adhering to precise formatting and boundary conditions.

It is also possible to conduct a comparative analysis by contrasting strict and loose metrics. Here, Figure 2 illustrates the strict and loose instruction-following accuracy of each model across all categories.

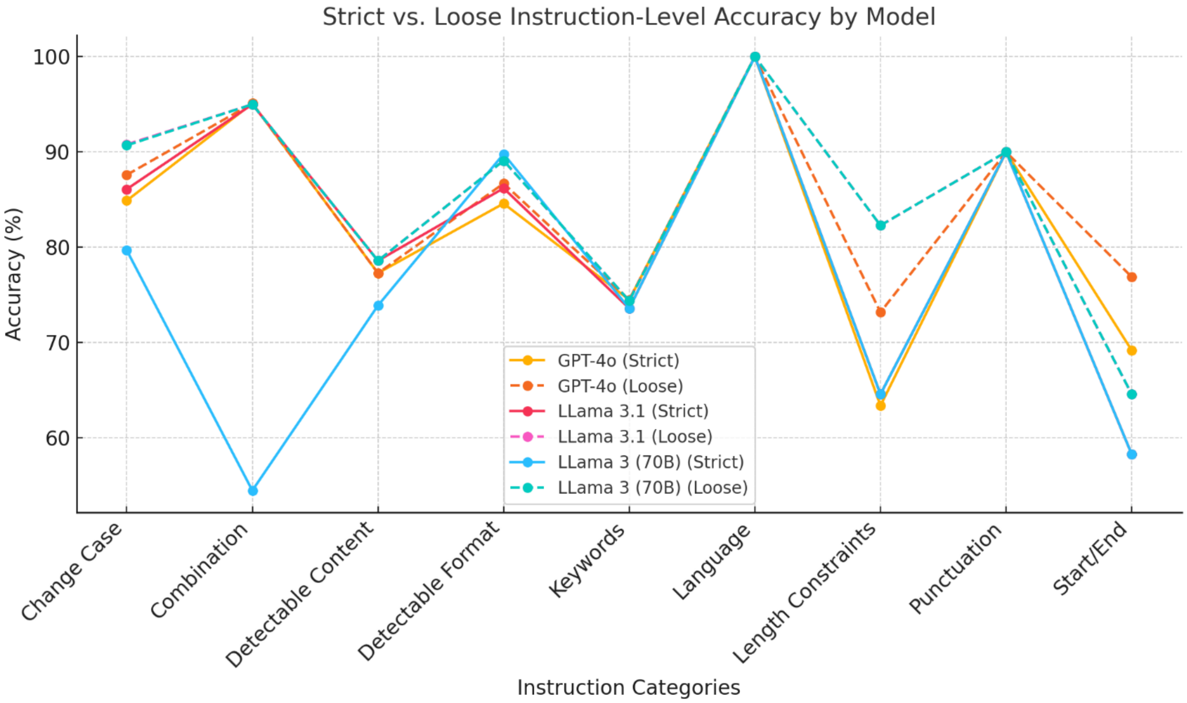


Figure 2. Strict and Loose Instruction-Level Accuracy by Model.

For a deeper analysis, instruction-level accuracy was broken down into subcategories for specific tasks within broader categories. Figure 3 highlights the performance of GPT-4o in the "Keywords" and "Length Constraints" categories.

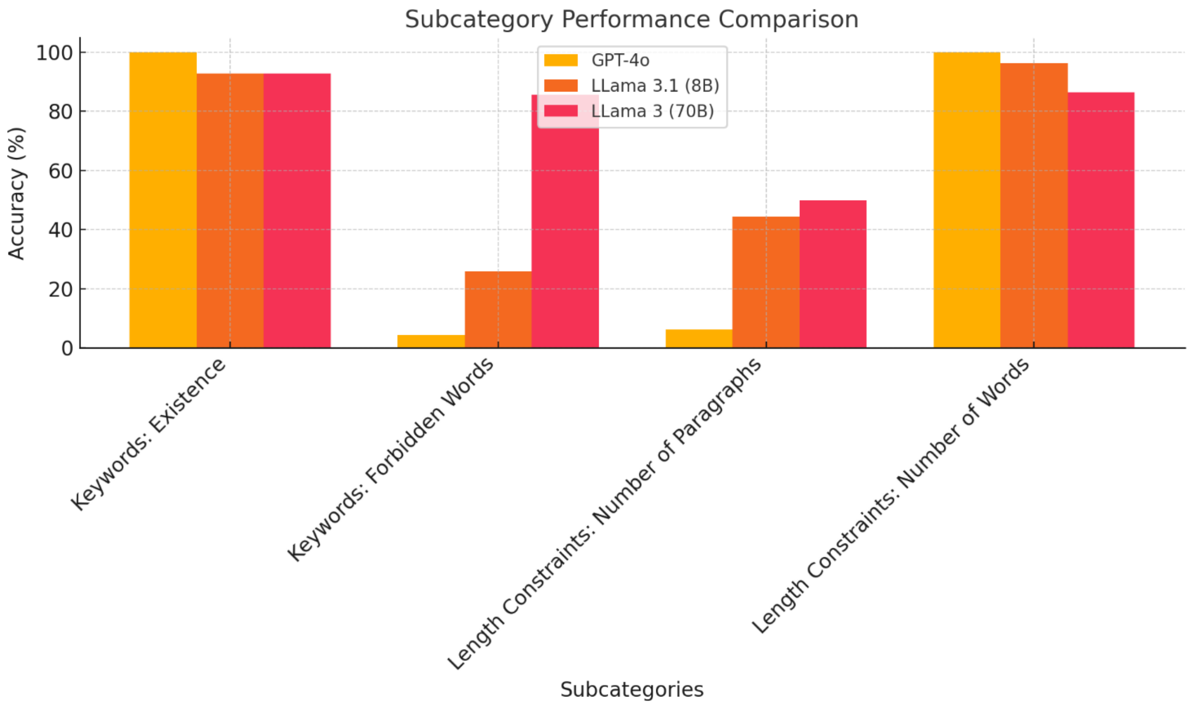


Figure 3. Instruction-Level Accuracy by Subcategory.

In this regard, we can identify the main performance trends and observations:

- **Strengths:** Models consistently achieved high accuracy in categories requiring language-specific responses and simple formatting rules. The robustness of the generative process in IFEval-Extended ensured compatibility across diverse instructions, as reflected in the results for "Combination" and "Detectable Format."
- **Weaknesses:** Accuracy declined significantly in subcategories demanding complex constraints, such as "Length Constraints: Number of Paragraphs" and "Keywords: Forbidden Words." This indicates that models struggle with tasks requiring simultaneous adherence to multiple, often conflicting instructions.
- **Comparison with IFEval**: Compared to the original IFEval results, the expanded set of prompts in IFEval-Extended exposed limitations in the models' ability to

generalize. Notably, the use of dynamically generated prompts resulted in more balanced evaluations, reducing potential overfitting to specific tasks.

4. DISCUSSION

4.1 Key Findings

The results obtained from the IFEval-Extended benchmark provide insights into the strengths and weaknesses of up-to-date Large Language Models (LLMs). Across all models tested — GPT-4o, LLama 3.1 (8B), and LLama 3 (70B) — a recurring theme is the models' ability to handle simpler, unambiguous instructions while struggling with tasks that require precise adherence to complex directives. This finding is consistent with previous results in instruction following benchmarks, where performance differences between instruction categories revealed inherent limitations in current LLM architectures (Zhou et al., 2023; Sun et al., 2023).

The results highlight the superior overall performance of GPT-4o, particularly in the Combination and Punctuation categories. This finding is consistent with OpenAI's advances in contextual understanding and token-level prediction optimization (OpenAI, 2023). However, even GPT-4o, the most advanced model in this evaluation, struggled in categories such as Length Constraints and Start/End, where tasks often involve following multiple potentially conflicting instructions.

One notable trend is the uniformly high accuracy across all models in the Language category, with each achieving a perfect score. This result suggests that current LLMs are well suited for language-specific tasks, reflecting their robust pre-training on multilingual corpora (Chowdhery et al., 2022). In contrast, the Keywords and Length Constraints categories show significant difficulties, particularly in subcategories such as Forbidden Words and Number of Paragraphs, indicating persistent gaps in the models' ability to handle subtle or complex instructions.

4.2 Implications

The success of IFEval-Extended demonstrates the potential of dynamic prompt generation to improve instruction compliance assessment. By removing the reliance on predefined prompts, the benchmark improves robustness and generalizability. The diverse set of dynamically generated instructions has revealed weaknesses in the generalization by the models that would likely remain undetected in static benchmarks such as IFEval. This situation is particularly evident in the differences in performance across models and categories, highlighting the benchmark's ability to discern nuanced features.

The findings also have implications for model development and fine-tuning. Consistent poor performance across complex instruction categories highlights areas where LLMs could benefit from targeted training. For example, including additional pre-training data that emphasizes logical consistency and contextual reasoning could improve compliance with multifaceted instructions (Victor et al., 2022). Moreover, the discrepancy between the strict and loose accuracy scores suggests that models often generate responses that are partially correct but do not meet all of the evaluation criteria. This insight can help design finer-grained loss functions during model training to balance flexibility and accuracy.

The benchmark’s ability to identify strengths in language tasks and weaknesses in adhering to complex directives also provides actionable insights for deploying LLMs in domains such as automated customer support, legal drafting, and educational tools. For example, perfect accuracy in the Language category suggests that LLMs can reliably handle tasks that require output in specific languages. This is especially useful for multilingual chatbots or content localization services. However, poor performance in categories such as Length Constraints cautions against using LLMs for tasks that require strict adherence to formatting rules, such as generating legal contracts without human supervision.

4.3 Limitations

Although IFEval-Extended addresses several limitations of its predecessor, it has its own limitations. They primarily stem from the base prompt used for dynamic instruction generation and the saturation effect observed in some instruction categories.

The base prompt, “You are an AI assistant that generates a single, diverse, human-like instruction for language models,” serves as the basis for generating dynamic prompts. While effective in generating a wide range of instructions, this approach nonetheless introduces a degree of homogeneity in the types of tasks generated. For example, prompts often emphasize syntactic constraints (such as word count or paragraph structure) while they less frequently explore semantic issues such as tasks requiring logical reasoning or domain knowledge.

Future research could explore methods to further diversify the baseline prompt, perhaps by incorporating techniques such as multi-way dialogue generation or reinforcement learning-based optimization. This would allow for creating prompts that better reflect real-world scenarios, where instructions often combine syntactic, semantic, and pragmatic elements.

Another limitation is the saturation effect observed in certain instruction categories. For example, Language has achieved perfect scores across all models, indicating that the tasks in this category may be too simplistic to discriminate between advanced models. Similarly, instructions that simply require the inclusion of keywords (e.g., Keyword: Existence) often resulted in uniformly high performance, indicating that these tasks may no longer provide meaningful insight into the capabilities of the model.

In order to address this issue, future iterations of IFEval-Extended could include more complex instructions in these categories. For example, tasks requiring the generation of multilingual responses or context-sensitive keyword inclusion could provide a more rigorous test of the model's capabilities.

4.4. Future Directions

Building on the current framework, future work could focus on increasing the diversity and complexity of dynamically generated prompts, such as incorporating real-world contexts (e.g., creating cues that mimic real-world applications such as legal analysis) to assess domain-specific instruction following capabilities, or introducing new categories such as tasks requiring multimodal thinking (e.g., generating text from images or tables).

It is also necessary to note that the current evaluation framework relies on strict and loose accuracy metrics, which may not capture the full spectrum of instruction following performance. Adaptive metrics that account for context-dependent variations could provide a more nuanced understanding of model behavior. For example, weighting instruction categories based on their complexity or real-world relevance could provide more actionable insights for model developers and end users.

5. CONCLUSION

The development and evaluation of Large Language Models (LLMs) have become key in advancing the capabilities of Natural Language Processing. This article has presented IFEval-Extended, a new benchmark designed to address the limitations of existing instruction-following evaluation frameworks. By incorporating dynamic prompt generation and expanding the range of instruction categories, IFEval-Extended improves the robustness, scalability, and generalizability of LLM model evaluation. The results of this study highlight the usefulness of the benchmark in identifying the strengths and limitations of state-of-the-art LLMs, offering valuable insights for both research and real-world applications.

Using a generative framework, IFEval-Extended creates a virtually infinite set of unique prompts, thereby reducing the risk of overfitting and enabling more comprehensive evaluations. In addition, the extended instruction categories and compatibility rules ensure that the benchmark covers a wider range of tasks, reflecting the diverse requirements of real-world applications. The results demonstrate the effectiveness of IFEval-Extended in differentiating LLMs, highlighting notable performance differences across models and instruction categories. These results validate the benchmark’s ability to reveal nuanced capabilities and limitations in LLMs. In this regard, IFEval-Extended has made a significant contribution to the development of more robust and general-purpose evaluation frameworks. Its ability to uncover subtle insights into model behavior makes it a valuable resource for researchers, developers, and end users. As LLMs continue to evolve, benchmarks like IFEval-Extended will play a critical role in shaping the future of Natural Language Processing.

References

1. Brown, T., Mann, B., Ryder, N., et al. (2020). Language Models Are Few-Shot Learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
2. Chang, Y., Wang, X., Wang, J., et al. (2023). A Survey on Evaluation of Large Language Models. *arXiv preprint arXiv:2307.03109*.
3. Chen, M., Tworek, J., Jun, H., et al. (2021). Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*.
4. Chung, H. W., Tay, Y., Zoph, B., et al. (2022). Scaling Instruction-Finetuned Language Models. *arXiv preprint arXiv:2210.11416*.
5. Chowdhery, A., Narang, S., Devlin, J., et al. (2022). PaLM: Scaling Language Modeling with Pathways. *arXiv preprint arXiv:2204.02311*.
6. Kovbo. (2024). IFEval-Extended Repository. Retrieved from https://github.com/Kovbo/IFEval-extended.
7. OpenAI. (2023). GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
8. Ouyang, L., Wu, J., Jiang, X., et al. (2022). Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
9. Skopek, O., Aralikatte, R., Gooding, S., et al. (2023). Towards Better Evaluation of Instruction-Following: A Case-Study in Summarization. *arXiv preprint arXiv:2310.08394*.
10. Sun, J., Tian, Y., Zhou, W., et al. (2023). Evaluating Large Language Models on Controlled Generation Tasks. *arXiv preprint arXiv:2310.14542*.

11. Taori, R., Gulrajani, I., Zhang, T., et al. (2023). Stanford Alpaca: An Instruction-Following LLaMA Model. Retrieved from https://github.com/tatsu-lab/stanford_alpaca.

12. Victor, S., Albert, W., Raffel, C., et al. (2022). Multitask Prompted Training Enables Zero-Shot Task Generalization. *Proceedings of ICLR*.

13. Zhou, J., Lu, T., Mishra, S., et al. (2023). Instruction-Following Evaluation for Large Language Models. Retrieved from https://arxiv.org/abs/2311.07911.