WILDIFEVAL: Instruction Following in the Wild

Gili Lior[1*] **Asaf Yehudai**[1,2] **Ariel Gera**[2] **Liat Ein-Dor**[2]
[1]The Hebrew University of Jerusalem [2]IBM Research

Abstract

Recent LLMs have shown remarkable success in following user instructions, yet handling instructions with multiple constraints remains a significant challenge. In this work, we introduce WILDIFEVAL — a large-scale dataset of 7K real user instructions with diverse, multi-constraint conditions. Unlike prior datasets, our collection spans a broad lexical and topical spectrum of constraints, extracted from natural user instructions. We categorize these constraints into eight high-level classes to capture their distribution and dynamics in real-world scenarios. Leveraging WILDIFEVAL, we conduct extensive experiments to benchmark the instruction-following capabilities of leading LLMs. WILDIFEVAL clearly differentiates between small and large models, and demonstrates that all models have a large room for improvement on such tasks. We analyze the effects of the number and type of constraints on performance, revealing interesting patterns of model constraint-following behavior. We release our dataset to promote further research on instruction-following under complex, realistic conditions.[2]

1 Introduction

As large language models (LLMs) continue to improve at following instructions, the nature of the instructions themselves has also evolved. Users now expect LLMs to handle more nuanced and complex requests [37]. This shift is especially evident in text generation tasks, which are becoming increasingly personalized, with more specific and tailored objectives [32, 13, 22, 9]. For instance, a former instruction like *"summarize this text"* might now take the form of *"summarize this movie review in two paragraphs, with the first focusing on the plot and the second discussing reasons to watch or skip the movie."* These personalized tasks typically carry implicit or explicit constraints that the generated output is expected to satisfy.

Thus, in *constrained generation* an LLM must adhere to a set of specific requirements in its response [11, 39]. Crucially, while individual constraints are often simple, LLMs struggle to satisfy multiple constraints simultaneously [17]. This highlights the need to directly evaluate the text generation performance of LLMs on realistic multi-constraint user data.

Existing works evaluating the ability of LLMs to follow constrained instructions generally follow a bottom-up approach, starting from curated verifiable constraints, that are amenable to objective verification of compliance [43], or a taxonomy of constraint types [39, 30, 17], and using those to manually or synthetically generate a set of instructions. Such an approach may not capture the complexity and diversity of real-world instructions by users, and the types and combinations of constraints that they ask the model to follow.

To this end, we introduce WILDIFEVAL (§2), a large-scale benchmark of constrained generation tasks. WILDIFEVAL is designed to evaluate the ability of LLMs to follow real-world multi-constrained

[*]This work was conducted during a summer internship at IBM Research.

[2]WILDIFEVAL is available at `https://huggingface.co/datasets/gililior/wild-if-eval`. The code for replication, along with model predictions and evaluation scores, can be found at `https://github.com/gililior/wild-if-eval-code`.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

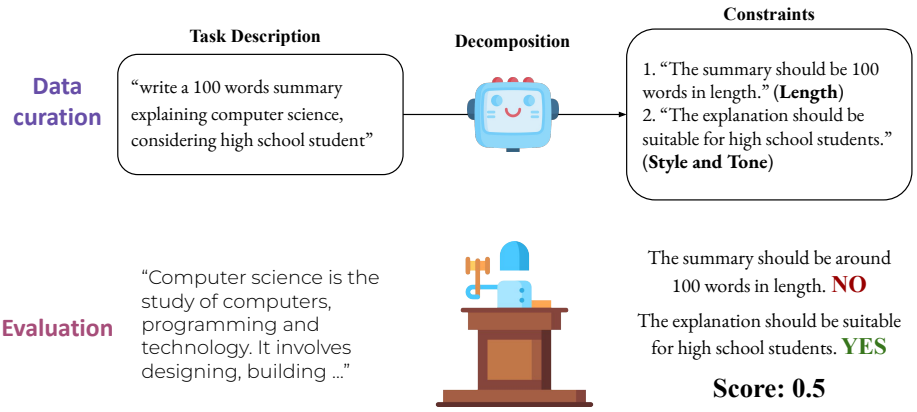


Figure 1: WILDIFEVAL description. At the top is an example for a constrained generation task, and its decomposition into constraints. In evaluation (bottom), the judge decides whether each of the constraints is fulfilled.

instructions. It encompasses a collection of 7K constrained generation tasks, including 24,731 different constraints, given by real users on Chatbot Arena [3], reflecting diverse examples of constrained generation instructions "in the wild".

The WILDIFEVAL dataset includes a breakdown of each task into the individual constraints it contains. Thus, it allows for a fine-grained evaluation of the ability of LLMs to adhere to user constraints. By breaking down task instructions into smaller and more interpretable pieces, we can perform a straightforward LLM-based evaluation of the proportion of task constraints that were fulfilled. At the same time, since constraints are extracted from naturalistic user queries, we capture not only simple and easily verifiable constraints but also "softer" constraints on content, quality, and style.

We begin by analyzing the types of user tasks and constraints present in WILDIFEVAL (§3), revealing that real-world constrained generation often involves diverse and challenging requirements.

We then evaluate 14 LLMs on the WILDIFEVAL benchmark and conduct a comprehensive analysis of their constraint-following capabilities (§4). Our results show that WILDIFEVAL is challenging, with the best models achieving around 0.7 under our strict evaluation metric. We also observe a consistent performance gap between small and large models, positioning WILDIFEVAL as a valuable benchmark for tracking progress in narrowing this gap.

Beyond overall model performance, we utilize the size and diversity of WILDIFEVAL to analyze the interplay between the number and types of constraints and instruction-following performance. Our analysis outlines the behavior for tasks with many constraints, and reveals the difficulties of models in satisfying form-related user constraints.

By publicly releasing WILDIFEVAL – the first benchmark of naturally occurring, multi-constraint instructions – we aim to drive progress in LLMs' ability to follow complex constraints in real-world applications.

2 The WILDIFEVAL Dataset

WILDIFEVAL is a novel benchmark designed to provide a comprehensive evaluation of the ability of LLMs to follow real-world multi-constrained instructions. It contains 7K user-generated instructions, written by many distinct users, each decomposed into a set of constraints, including 24,731 unique constraints.

The task instructions in WILDIFEVAL were extracted from LMSYS-Chat-1M dataset [41], a large-scale dataset containing real-world instructions collected from the Chatbot Arena.[3] Since users rarely

[3]Chatbot Arena website: `https://lmarena.ai`, Huggingface dataset: `https://huggingface.co/datasets/lmsys/lmsys-chat-1m`.

Table 1: Comparison of WILDIFEVAL with openly available instruction-following benchmarks such as IFEval [43], FollowBench [17], and InFoBench [30].

Benchmark	Data Source	Evaluation	Size (# Tasks)	# Constraints
IFEval	Synthetic	Rule	541	-
FollowBench	Crowd + Syn.	Model / Rule	1,852	-
InFoBench	Crowd	Model / Rule	500	2,217
WILDIFEVAL (ours)	Real Users	Model	7,523	24,731

specify constraints in a structured list format, the decomposition breaks instructions into manageable items, ensuring the necessary granularity to assess the LLM's ability to adhere to them.

In Table 1, we present a comparison with popular openly available instruction-following datasets. As can be seen in the table, WILDIFEVAL is uniquely representative of natural user interactions at scale; it stands out as the largest available dataset, consisting of real-world user instructions given to LLMs.

2.1 Dataset Curation

WILDIFEVAL was curated in three steps. First, we filter the LMSYS-Chat-1M source data – we extract the first user message from each conversation, and filter out non-English tasks, coding tasks, and tasks containing toxic language.[4]

Next, we filter for only constrained generation tasks. We follow the definition for constrained generation tasks from Ferraz et al [10], and utilize their suggested prompt (Appendix A) with Llama3.1-405b in order to perform the filtering. The prompt is phrased as a yes/no question; instead of simply parsing the string, we use the probabilities that the model assigns to the yes/no tokens as a measure of certainty, and include only the 10% of tasks with the highest certainty to be a constrained generation task, i.e., with the highest probability for a "yes" token.

The last step of the curation process is the decomposition into constraints – for each user task, we want to include all the constraints the model is required to fulfill. To obtain the highest-quality decomposition we employ GPT-4o [15], using a prompt adopted from Ferraz et al [10] to automatically extract the constraints for each of the tasks.[5] All prompts are presented in Appendix A.

To mitigate potential biases in scoring, we perform sub-sampling for constraints that appear more than 40 times (i.e., exact match across more than 40 different tasks). This process affected 15 unique constraints, accounting for less than 0.15% of all constraints. In addition, we filtered out rare cases of tasks with more than 8 constraints.

By the end of this process, we obtained a dataset of 7,523 real-world constrained generation tasks, each annotated with a list of constraints. There are 24,731 distinct constraints in WILDIFEVAL, averaging 3.25 constraints per task. The distribution and frequency of constraints per task are shown in Figure 8 in Appendix.

3 Into the Wild: A Data Expedition

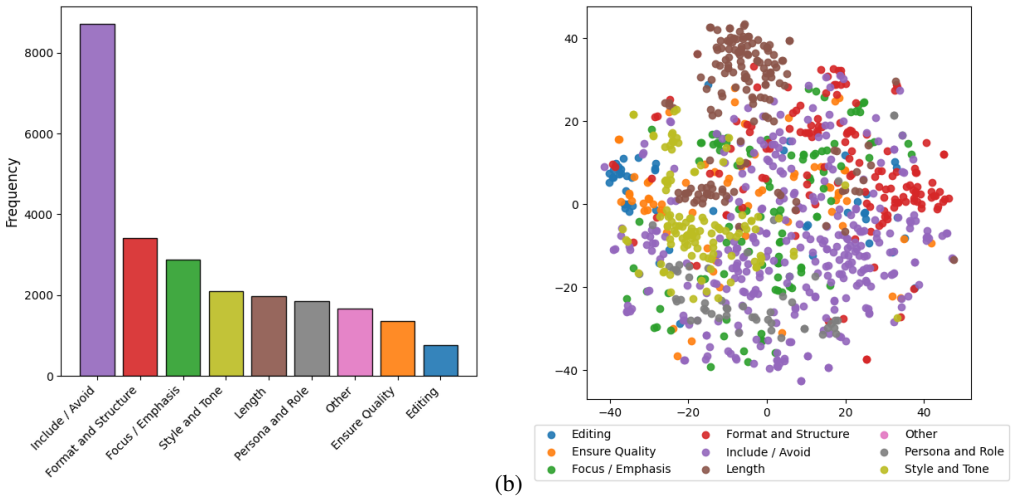
Below we conduct an analysis of our WILDIFEVAL data, revealing insights on constrained generation use cases in the wild.

3.1 Constraint Types

taxonomy A key question regarding constrained generation tasks concerns the nature and types of the constraints themselves, i.e., what kinds of requirements users wish to impose on the model responses. Prior work [43, 10, 17, 30] generally distinguishes between broad categories such as content, style, and format, yet lacks a unified taxonomy. Moreover, some works define rather specific constraint categories (e.g., "Part-of-speech rules") or highly general ones (e.g., "Content constraints").

[4] We detect toxic language using the detoxify package `https://github.com/unitaryai/detoxify`

[5] gpt-4o-2024-08-06



(a) (b)

Figure 2: Analysis of constraints in WILDIFEVAL. (a) Distribution of constraint types. (b) A tSNE projection [35] of the embeddings of constraints, colored by their type. For convenience, we randomly subsample 1k data points. We observe some red, brown, and yellow clusters, corresponding to *Format and Structure*, *Length*, and *Style and Tone* constraints, aligning with the generic nature of these types. This is in contrast to content-oriented types like *Focus/Emphasis* and *Include/Avoid* (green and purple), which are more spread out.

Here we seek to bridge this taxonomy gap. We draw from earlier categorization efforts, but combine them with data-driven insights. Specifically, we look at the most frequent words appearing in constraints, and examine some of the constraints in which they occur; this allows us to analyze recurring patterns of constraint types in WILDIFEVAL. This qualitative data-driven analysis reveals some broad constraint types that have not been mentioned by prior efforts, and also enables us to break existing broad divisions into finer-grained categories.

Our taxonomy divides constraints into 8 principal categories. These capture both explicit constraints (e.g., inclusion or exclusion of content) and more nuanced aspects of user instructions (e.g., a desired tone or quality for the model output). The following definitions detail each category, providing clear guidelines on how they contribute to the overall task structure:

- **Include / Avoid:** Specifies elements or concepts that must be incorporated into or omitted from the response, directly guiding the content of the output.
- **Editing:** Focuses on modifications to an existing text, outlining how the original content should be altered or preserved.
- **Ensure Quality:** Imposes requirements on the response's quality, such as coherence, accuracy, or overall clarity.
- **Length:** Sets quantitative boundaries on the output, such as word or character limits, ensuring appropriate brevity or depth.
- **Format and Structure:** Dictates the organization and presentation of the response, including the use of bullet points, tables, or specific layout requirements.
- **Focus / Emphasis:** Highlights particular topics, keywords, or elements that should be prioritized within the response.
- **Persona and Role:** Instructs the AI to adopt a specific character, perspective, or expertise, influencing the narrative voice of the output.
- **Style and Tone:** Specifies the overall manner of expression, including formality, register, and emotional nuance, to define the voice and feel of the response.

We then ask Deepseek-v3 to classify all constraints in WILDIFEVAL into one of the 8 constraint types above, resulting in a full categorization of constraint types.[6] The classification prompt is provided in Appendix A.

Distribution of constraint types. In Figure 2a we present the distribution of constraint types in WILDIFEVAL. The most common constraints are the content constraints *Include/Avoid* and *Focus/Emphasis*; these specify either explicit element(s) that should be included or excluded, or how much prominence should be given to different elements in the content.

Figure 2b depicts a tSNE embedding map of WILDIFEVAL constraints, colored by types. A salient and intuitive observation is that content-related constraints such as *Include/Avoid* and *Focus/Emphasis* are spread out across the semantic embedding space; in contrast, form-related constraints like *Length* or *Format and Structure* are organized in more distinct clusters.

Co-occurrence of constraint types. In Figure 3 we analyze the co-occurrence of constraint types in multi-constraint tasks. Specifically, we ask whether some combinations of types appear more or less than expected. Thus, we compare the number of co-occurrences in practice relative to the overall frequency of each of the co-occurring types, i.e., the pointwise mutual information (PMI) [4].

As shown in Figure 3, only few combinations appear more than expected (i.e., PMI > 0). For example, *Persona and Role* tends to co-occur with *Style and Tone* slightly above expected, which appears to reflect the thematic similarity between these constraint types. In contrast, some types do not often appear together; for instance, requirements for *Format and Structure* are rarely paired with *Style and Tone* or *Persona and Role* constraints. Also *Editing*, which is the lowest represented type of constraint, rarely co-occurs with *Focus / Emphasis*.

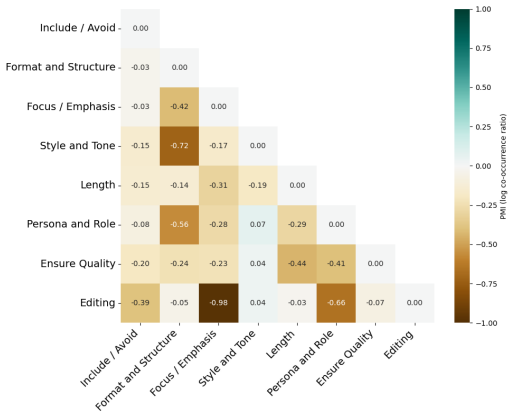


Figure 3: Relative co-occurrence (PMI) of constraint categories within tasks. Values above 0 indicate that constraints co-occur more than expected by their overall type frequencies.

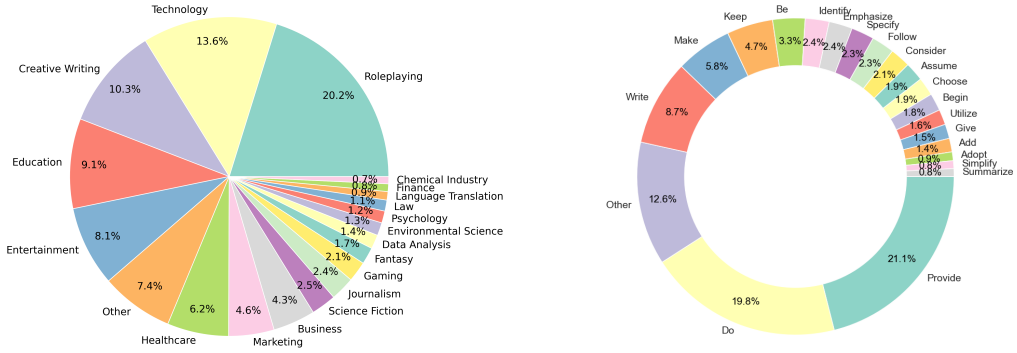
3.2 Data Diversity

WILDIFEVAL covers a variety of domains. Figure 4a depicts the distribution of domains covered by WILDIFEVAL. As expected from large-scale naturally-occurring data, tasks in WILDIFEVAL cover a wide variety of domains, including Technology, Entertainment, Healthcare, Creative Writing, and more. We use a data-driven approach to recover the domains, leading us to believe that these reflect realistic user behavior in constrained generation tasks. The domains were extracted using an LLM, see details in Appendix B.4.

WILDIFEVAL is lexically diverse. To illustrate lexical diversity, we examine verb frequencies in constraints that begin with a verb (65.1% of constraints).[7] The results in Figure 4b reveal a skewed frequency distribution; *"Provide"* is the most dominant verb, comprising 21.1% of all occurrences, followed by *"Do"* (19.2%) and *"Write"* (8.7%). Several mid-frequency verbs (e.g., *"Keep," "Identify," "Make"*) also appear regularly. The *"Other"* category (12.6%) reflects the long tail of the verb distribution, with many verbs that each occur in under 0.8% of the data. The distribution suggests that users tend to use general types of constraints more than specific ones like *"Simplify"* (0.8%) or *"Summarize"* (0.8%). This analysis underscores the variety of linguistic expressions in WILDIFEVAL. A similar pattern emerges when considering all constraints containing a verb (70% of constraints), shown in Figure 12 in Appendix B.3. We note that the analysis reflects the words in the

[6] We recognize that in some relatively rare cases a single constraint can belong to multiple types; however, for simplicity we opt to treat this as a multiclass problem.

[7] We employ NLTK's part-of-speech tagger to identify verb tokens `https://www.nltk.org/`



(a) (b)

Figure 4: Task and constraint characteristics in WILDIFEVAL. (a) Domain distribution of tasks. (b) Lexical diversity of constraint phrasing (opening verbs).

constraints, as decomposed by an LLM (§2.1), and thus may differ somewhat from the original user task descriptions.

Qualitative analysis. Manual inspection of instances from WILDIFEVAL reveals some interesting trends. First, we observe that quite often fulfilling – or even understanding – the task constraints given by users requires some very specialized or esoteric knowledge (e.g., D&D spells, Gate exam syllabus, pig latin etc.). We show some examples in Appendix D. We also note that some of the more complex tasks – those with many constraints – reflect attempts by users to "jailbreak" the LLM, and trick it to say things that it is not supposed to (e.g., toxic language or controversial statements).

4 LLM Benchmarking

In this section, we examine the performance of various LLMs to assess their behavior in constrained generation tasks. We present the evaluation metric (§4.1), experimental setup (§4.2), and finally, we describe and analyze the results (§4.3).

4.1 Evaluation Metric

WILDIFEVAL reports two scores: *strict* and *soft*. The *strict* score is a binary measure indicating whether all task constraints are satisfied, while the *soft* score reflects the proportion of individual constraints successfully met by the model's response.

To evaluate if a constraint is fulfilled by model M, we present the LLM judge J with the task description t_i, the model's response $r_i = M(t_i)$, and the specific constraint under evaluation c_i^j. Then, we prompt the Judge with a yes/no question, "Given task t_i and response r_i, is the following constraint satisfied: c_i^j?". We denote the judge score by $J(t_i, r_i, c_i^j) \in \{0, 1\}$. Its value is 1 if the judge responds with a "yes" token, and 0 if responds with a "no" token, in a greedy decoding setup to ensure consistency.

The *soft* and *strict* scores for a task are defined as follows:

$$soft(r_i \mid t_i) = \frac{1}{N(t_i)} \sum_{j=1}^{N(t_i)} J(t_i, r_i, c_i^j) \qquad strict(r_i \mid t_i) = \prod_{j=1}^{N(t_i)} J(t_i, r_i, c_i^j) \tag{1}$$

where $N(t_i)$ is the number of constraints in t_i.

4.2 Experimental Setup

We evaluate 14 prominent instruction-tuned LLMs from five different model families on WILDIFEVAL, in a zero-shot setup. The models vary in size from 0.5 billion to 671 billion parameters.

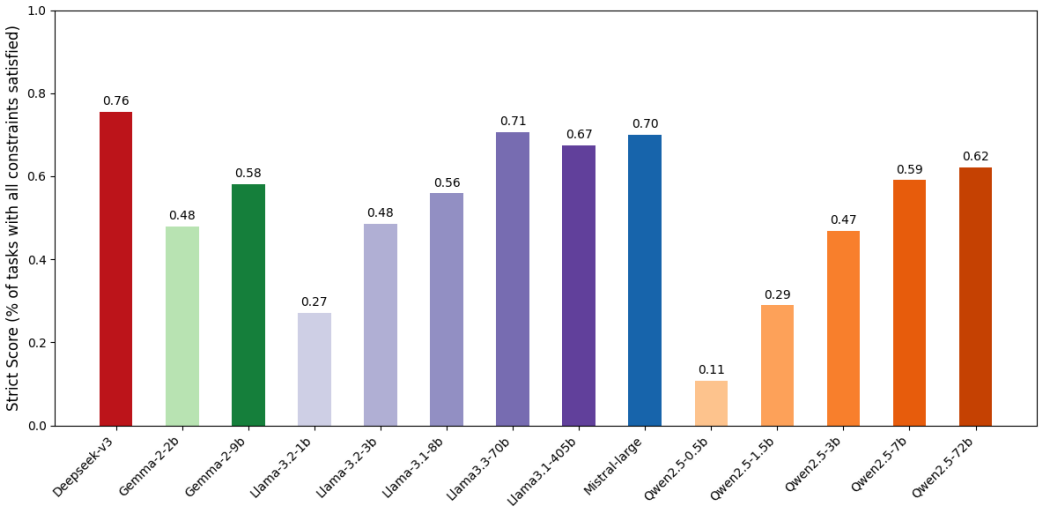


Figure 5: Strict scores on WILDIFEVAL. For each model, the figure reports the proportion of tasks in which all constraints were fulfilled (strict score). Soft scores are shown in Figure 10 in the Appendix. Statistical significance between model pairs (McNemar tests) is reported in Figure 13 in Appendix.

We assess the following models: *(1)* Deepseek-v3 [25] *(2)* Mistral-Large-instruct-2407 [27] *(3)* Gemma-2-2b and Gemma-2-9b [34] *(4)* Llama3.2-1b, Llama3.2-3b, Llama3.1-8b, Llama3.3-70b and Llama3.1-405b [7] *(5)* Qwen-2.5-0.5b, Qwen-2.5-1.5b, Qwen-2.5-3b, Qwen-2.5-7b, and Qwen-2.5-72b [38].

Judge evaluation As a judge model for evaluation (§4.1), we use Deepseek-v3. We choose Deepseek-v3 as the judge after evaluating a subset of 500 tasks from WILDIFEVAL with GPT-4o as a judge, and among available SOTA open-source models including also Llama3.3-70b and Qwen-2.5-72b, Deepseek-v3 showed the highest agreement with GPT-4o, in terms of accuracy and confidence correlation (details in Appendix B.2). As a further validation of our evaluation, the benchmark shows significantly high Kendall's Tau correlations (>0.82) with existing benchmarks like IFEval, MMLU, and GPQA (details in Appendix C).

4.3 Results

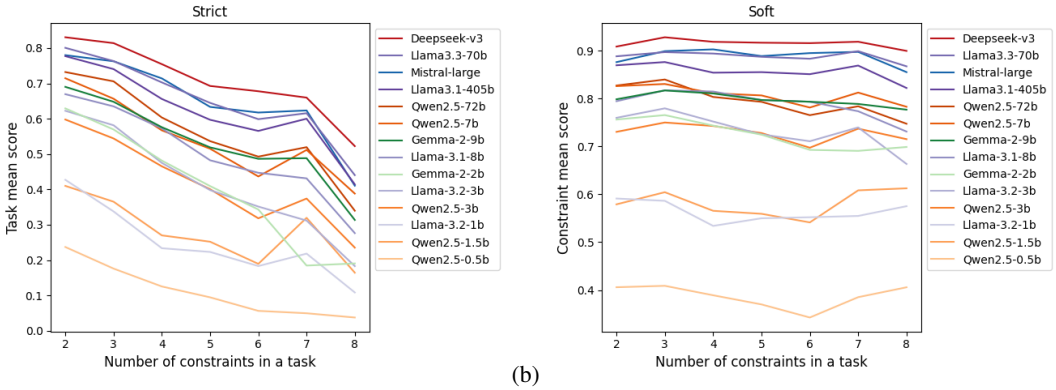
Figure 5 depicts the overall model performance on WILDIFEVAL. We can observe a clear performance gap within model families, with larger models consistently outperforming their smaller counterparts[8], in line with prior findings [18]. At the same time, even stronger models like Deepseek-v3 and Llama3.3-70b fail to satisfy all task constraints in 25-30% of cases.

The best performing model is Deepseek-v3. Since it also serves as the judge, this raises questions about potential judge self-bias [36, 12]. However, we note that on a subset of 500 tasks used for judge validation (§4.2), all tested judges –GPT-4o, Llama3.3-70b, and Qwen-2.5-72b –consistently ranked Deepseek-v3 first.

Naturally, when a task has more constraints, it is harder for the model to fulfill all of them. Accordingly, Figure 6a shows the decrease in the strict performance score as a function of the number of constraints. However, when looking at the soft performance score (Figure 6b) we see that the number of constraints does not affect the fulfillment of *individual* constraints. In other words, it appears that the difficulty in multi-constraint tasks does not reflect a general decrease in model instruction-following abilities, but rather stems from having to fulfill several constraints at once.

Figure 7a illustrates the relative model performance for different constraint types. We can see that models consistently have difficulties with *Length* constraints, and to a lesser extent also with

[8]A notable exception is Llama3.3-70b, that surpasses Llama3.1-405b. This result is aligned with previous reports, e.g., Llama-3.3 Model Card.



(a) (b)

Figure 6: Scores as function of number of constraints in a task. (a) Strict score – tasks in which all constraints are fulfilled. (b) Soft score – fraction of fulfilled constraints in a task.

Format and Structure. In contrast to these form-based types, models tend to succeed in fulfilling *Focus / Emphasis* constraints, which impose softer, content-related requirements. We also observe a somewhat different pattern for models from the Qwen family, that appear to struggle more with Persona and Style constraints relative to other models.

To further understand the role of constraint types, we look at the rankings they induce of model performance. We rank the models according to their performance on each constraint type, and calculate the agreement between the resulting model rankings. As Figure 7b shows, type-specific rankings largely agree with each other. We do however observe different degrees of agreement. Notably, the ranking induced by *Persona and Role* has a low agreement with most constraint types, but exhibits a strong ranking agreement with the thematically related *Style and Tone*. We also observe a slightly different behavior of the *Length* constraint, particularly when compared to the Persona and Style constraints.

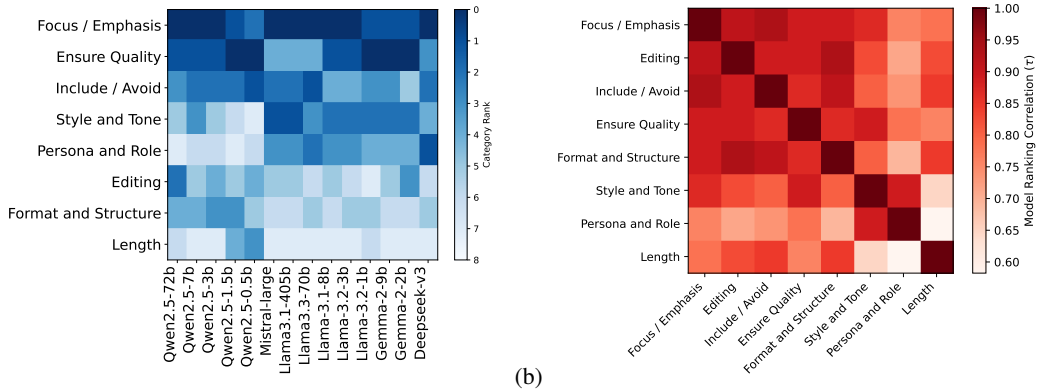
Error analysis. We also performed a manual analysis of the examples where most models failed to satisfy the constraints. We observe that the majority of these failure cases belong to the *Length* category, particularly constraints requiring an exact number of words or more atomic units (syllables, characters etc.), e.g., "The script should be 300 words long". Some of the failure cases involve constraints that are quite complex, involving multiple specifications and sub-constraints. For example, the user constraint can require including a dictionary in a specific format and with a specific set of keys and values. Overall, we note that all constraint types can vary widely in the level of complexity they impose on the model. For example, Persona and Style constraints range from mundane requirements ("Use a first-person perspective.", "Keep the tone informal.") to more specific an esoteric ones ("Excel in ninjutsu, tactics, and battle strategies", "Use strict iambic pentameter").

5 Related Work

Recent interest in LLM instruction-following capabilities raises the need for benchmarking model performance under complex, multi-constraint scenarios [24, 33].

Several works [39, 1, 16] rely on synthetic instructions and rule-based evaluation, with the prominent example of *IFEval* [43]. Other works, such as *FollowBench* [17] and *InfoBench* [30], utilize crowd-sourced data, and LLM-based evaluation. However, these works are limited in size and do not fully capture the diversity of genuine user inputs. More recently, REALINSTRUCT [10] employs real-user instructions; however, this data has not been released, hindering the ability to use it for benchmarking and analyzing instruction-following of LLMs. While here we focus on data in English, other works study constraint-following in other languages, such as Chinese [40].

In this work, we release a diverse dataset of multi-constraint instructions, that originates from real users and is much larger than all existing datasets. Moreover, whereas some of these benchmarks have become saturated, ours remains challenging even for state-of-the-art LLMs.



(a) (b)

Figure 7: Constraint types characteristics. (a) Category performance rankings per model. Darker colors indicate stronger performance by the model on the corresponding constraint category, while lighter colors reflect weaker performance. (b) Correlation (Kendall's Tau) between model rankings induced by different constraint types.

6 Discussion

In this work, we present a benchmark for evaluating the ability of LLMs to follow real-world constrained instructions. WILDIFEVAL aims to reflect a realistic and contemporary view of constrained generation user requests. This challenging and heterogeneous data serves as a playground for fine-grained analysis of the strengths and weaknesses of models, drilling down beyond the task level into atomic user constraints. The ability to analyze model difficulties at the atomic level, and identify recurring failures, can help focus model improvement efforts.

There are two possible approaches for modeling constrained generation tasks. One is a bottom-up approach – combining a set of constraints into a task description [43, 17, 39, 30]. This approach facilitates a more controlled analysis of constraint families and how models respond to them. However, it might also place greater emphasis on more rudimentary constraints, potentially overlooking the broader manifold of constraints and tasks. Here we adopt a top-down approach, which starts from real-world constrained generation tasks and leverages an LLM to extract their underlying constraints. This has the advantage of widening the scope of instructions, and better capturing natural user behavior. At the same time, real-world data can be very noisy, making it more difficult to identify clear patterns in model behaviors. The reliance on an LLM for task decomposition and evaluation can also introduce some errors. Our results demonstrate that despite these challenges, a top-down approach can yield valuable insights into the instruction-following abilities of LLMs.

One direction for future work is to explore how constrained generation can be applied to prompt engineering. For example, the task decomposition generated by constrained generation could be explicitly included in the prompt to improve clarity and guidance for the model. Additionally, performance analysis of the model could help identify more effective ways to phrase constraints within the prompt.

Another important question is how to collect supervised data for improving constrained generation performance. A promising avenue would be to identify naturally-occurring feedback – from multi-turn interactions of a user with an LLM – indicating user satisfaction with the response [6].

Our focus in this work is on the constrained generation performance of LLMs. Another line of research concerns the abilities of a judge to evaluate whether multi-constraint instructions are fulfilled. This may require dynamically employing different evaluation methods based on the constraint type (e.g., rule-based for verifiable constraint types, compilers for some format and code constraints, etc.), and may involve calling external tools, such as search for retrieving information, and code interpreter to execute or validate responses that involve computational logic or data manipulation. [44, 28].

References

[1] M. Bastan, M. Surdeanu, and N. Balasubramanian. NEUROSTRUCTURAL DECODING: Neural text generation with structural constraints. In A. Rogers, J. Boyd-Graber, and N. Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9496–9510, Toronto, Canada, July 2023. Association for Computational Linguistics.

[2] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, G. Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[3] W.-L. Chiang, L. Zheng, Y. Sheng, A. N. Angelopoulos, T. Li, D. Li, B. Zhu, H. Zhang, M. Jordan, J. E. Gonzalez, and I. Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning*, 2024.

[4] K. W. Church and P. Hanks. Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22–29, 1990.

[5] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[6] S. Don-Yehiya, L. Choshen, and O. Abend. Learning from naturally occurring feedback. *arXiv preprint arXiv:2407.10944*, 2024.

[7] A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Yang, A. Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[8] Y. Dubois, B. Galambosi, P. Liang, and T. B. Hashimoto. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

[9] L. Ein-Dor, O. Toledo-Ronen, A. Spector, S. Gretz, L. Dankin, A. Halfon, Y. Katz, and N. Slonim. Conversational prompt engineering. *arXiv preprint arXiv:2408.04560*, 2024.

[10] T. P. Ferraz, K. Mehta, Y.-H. Lin, H.-S. Chang, S. Oraby, S. Liu, V. Subramanian, T. Chung, M. Bansal, and N. Peng. LLM self-correction with DeCRIM: Decompose, critique, and refine for enhanced following of instructions with multiple constraints. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7773–7812, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

[11] C. Garbacea and Q. Mei. Why is constrained neural language generation particularly challenging? *arXiv preprint arXiv:2206.05395*, 2022.

[12] A. Gera, O. Boni, Y. Perlitz, R. Bar-Haim, L. Eden, and A. Yehudai. Justrank: Benchmarking llm judges for system ranking. *arXiv preprint arXiv:2412.09569*, 2024.

[13] J. He, W. Kryscinski, B. McCann, N. Rajani, and C. Xiong. CTRLsum: Towards generic controllable text summarization. In Y. Goldberg, Z. Kozareva, and Y. Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5879–5915, Abu Dhabi, United Arab Emirates, Dec. 2022. Association for Computational Linguistics.

[14] D. Hendrycks, C. Burns, S. Basart, A. Zou, M. Mazeika, D. Song, and J. Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

[15] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

[16] H. Iso. AutoTemplate: A simple recipe for lexically constrained text generation. In S. Mahamood, N. L. Minh, and D. Ippolito, editors, *Proceedings of the 17th International Natural Language Generation Conference*, pages 1–12, Tokyo, Japan, Sept. 2024. Association for Computational Linguistics.

[17] Y. Jiang, Y. Wang, X. Zeng, W. Zhong, L. Li, F. Mi, L. Shang, X. Jiang, Q. Liu, and W. Wang. FollowBench: A multi-level fine-grained constraints following benchmark for large language models. In L.-W. Ku, A. Martins, and V. Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4667–4688, Bangkok, Thailand, Aug. 2024. Association for Computational Linguistics.

[18] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[19] S. Kim, J. Suk, S. Longpre, B. Y. Lin, J. Shin, S. Welleck, G. Neubig, M. Lee, K. Lee, and M. Seo. Prometheus 2: An open source language model specialized in evaluating other language models. *arXiv preprint arXiv:2405.01535*, 2024.

[20] W. Kwon, Z. Li, S. Zhuang, Y. Sheng, L. Zheng, C. H. Yu, J. E. Gonzalez, H. Zhang, and I. Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*, 2023.

[21] N. Lambert, V. Pyatkin, J. Morrison, L. Miranda, B. Y. Lin, K. Chandu, N. Dziri, S. Kumar, T. Zick, Y. Choi, et al. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024.

[22] C. Li, M. Zhang, Q. Mei, W. Kong, and M. Bendersky. Learning to rewrite prompts for personalized text generation. In *Proceedings of the ACM Web Conference 2024*, WWW '24, page 3367–3378, New York, NY, USA, 2024. Association for Computing Machinery.

[23] T. Li, W.-L. Chiang, E. Frick, L. Dunlap, T. Wu, B. Zhu, J. E. Gonzalez, and I. Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

[24] B. Y. Lin, W. Zhou, M. Shen, P. Zhou, C. Bhagavatula, Y. Choi, and X. Ren. CommonGen: A constrained text generation challenge for generative commonsense reasoning. In T. Cohn, Y. He, and Y. Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1823–1840, Online, Nov. 2020. Association for Computational Linguistics.

[25] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[26] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, and C. Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore, Dec. 2023. Association for Computational Linguistics.

[27] Mistral AI Team. Large enough. `https://mistral.ai/en/news/mistral-large-2407`, July 24 2024. Accessed: 2025-02-14.

[28] H. Peng, Y. Qi, X. Wang, Z. Yao, B. Xu, L. Hou, and J. Li. Agentic reward modeling: Integrating human preferences with verifiable correctness signals for reliable reward systems, 2025.

[29] Y. Perlitz, A. Gera, O. Arviv, A. Yehudai, E. Bandel, E. Shnarch, M. Shmueli-Scheuer, and L. Choshen. Do these llm benchmarks agree? fixing benchmark evaluation with benchbench. *arXiv preprint arXiv:2407.13696*, 2024.

[30] Y. Qin, K. Song, Y. Hu, W. Yao, S. Cho, X. Wang, X. Wu, F. Liu, P. Liu, and D. Yu. Infobench: Evaluating instruction following ability in large language models. *arXiv preprint arXiv:2401.03601*, 2024.

[31] D. Rein, B. L. Hou, A. C. Stickland, J. Petty, R. Y. Pang, J. Dirani, J. Michael, and S. R. Bowman. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*, 2023.

[32] A. Salemi, S. Mysore, M. Bendersky, and H. Zamani. LaMP: When large language models meet personalization. *arXiv preprint arXiv:2304.11406*, 2023.

[33] J. Sun, Y. Tian, W. Zhou, N. Xu, Q. Hu, R. Gupta, J. Wieting, N. Peng, and X. Ma. Evaluating large language models on controlled generation tasks. In H. Bouamor, J. Pino, and K. Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3155–3168, Singapore, Dec. 2023. Association for Computational Linguistics.

[34] G. Team, M. Riviere, S. Pathak, P. G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024.

[35] L. van der Maaten and G. Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86):2579–2605, 2008.

[36] P. Verga, S. Hofstatter, S. Althammer, Y. Su, A. Piktus, A. Arkhangorodsky, M. Xu, N. White, and P. Lewis. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024.

[37] J. Wang, F. Mo, W. Ma, P. Sun, M. Zhang, and J.-Y. Nie. A user-centric multi-intent benchmark for evaluating large language models. In Y. Al-Onaizan, M. Bansal, and Y.-N. Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA, Nov. 2024. Association for Computational Linguistics.

[38] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Li, D. Liu, F. Huang, H. Wei, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[39] S. Yao, H. Chen, A. W. Hanjie, R. Yang, and K. Narasimhan. Collie: Systematic construction of constrained text generation tasks. *arXiv preprint arXiv:2307.08689*, 2023.

[40] T. Zhang, Y. Shen, W. Luo, Y. Zhang, H. Liang, F. Yang, M. Lin, Y. Qiao, W. Chen, B. Cui, et al. CFbench: A comprehensive constraints-following benchmark for llms. *arXiv preprint arXiv:2408.01122*, 2024.

[41] L. Zheng, W.-L. Chiang, Y. Sheng, T. Li, S. Zhuang, Z. Wu, Y. Zhuang, Z. Li, Z. Lin, E. P. Xing, J. E. Gonzalez, I. Stoica, and H. Zhang. Lmsys-chat-1m: A large-scale real-world llm conversation dataset, 2023.

[42] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

[43] J. Zhou, T. Lu, S. Mishra, S. Brahma, S. Basu, Y. Luan, D. Zhou, and L. Hou. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*, 2023.

[44] M. Zhuge, C. Zhao, D. Ashley, W. Wang, D. Khizbullin, Y. Xiong, Z. Liu, E. Chang, R. Krishnamoorthi, Y. Tian, et al. Agent-as-a-judge: Evaluate agents with agents. *arXiv preprint arXiv:2410.10934*, 2024.

A Prompts

Classify constrained generation tasks

You are an assistant whose job is to help me perform tasks. I need to filter from a set of requests made by users to AI assistants, the ones in which human requested the AI assistant to do a task with constraints to be follow. Constraints refer to more detailed rules, conditions or specific guidelines provided to guide the responses and shape the output generated by the AI assistant. Examples of sentences that indicate constraints are: “write in the format of”, “write as if you were”, “make sure to follow this”, “make sure to answer these questions”, “make sure to no include”, “avoid mentioning”. I will give you the human request and I expect you to answer “Yes” when the request contains instruction with constraints, or “No” if the request does not contemplate any constraint. I also want you to say “No” if the request require to generate code or an answer about code provided. Also, I want you to say “No” if the task is not self-contained, which means the AI Assistant need to ask follow up questions before start to answer, or it needs more context. You are provided five examples.
Example 1: list and compare top website to https://fastfunnels.com/ in table format.
Answer: Yes
Example 2: You are an fantasy writer. Your task is now to help me write a D&D adventure for 5 players in the Eberron univers. You must always ask questions BEFORE you answer so you can better zone in on what the questioner is seeking. Is that understood ?
Answer: No.
Example 3: I have 100 dollars and would like to use this as the initial funding to make some money. I need it to be as quick as possible with good returns.
Answer: No.
Example 4: I have a vacation rental website and I am looking for alliterative and descriptive headlines that are at least 4 words in length and a maximum of 6 words. Examples: “Get Away to Galveston”, “Sleep Soundly in Seattle”. Each headline should have alliteration of at least 50% of the words and be poetic in language. Make each headline unique from the others by not repeating words. Each headline should include a verb. Put into an table with the city in column one and the results in column two for the following cities: Galveston, Sedona, Honolulu, Tybee Island, Buenos Aires.
Answer: Yes.
Example 5: pitch me a viral social app that is inspired by the hunger games. give it a fun twist!
Answer: Yes.
Request: ${request}
Now please answer, “Yes” or “No”.
Answer:

Decompose Tasks

You are an assistant whose job is to help me perform tasks. I will give you an instruction that implicitly contains a task description, its context, and constraints to be followed. Your task is to translate this instruction in a more structured way, where task, context and constraints are separated. Avoid writing anything else. Context is an input text needed to generate the answer or a more detailed description of the situation. Make sure to separate the context when it is needed, otherwise leave it empty. You are provided five examples. Please follow the same format.
Example 1:
Original Instruction: Write me a rap about AI taking over the world, that uses slangs and young language. It need to sound like a real human wrote it. It would be cool if there's a chorus very catchy that would be singed by a famous pop artist. Make sure to include references about things that young people likes, such as memes, games, gossips. I want that in the end, you revel that this was written by an AI.
Translated Task: Write a rap about AI taking over the world.
Translated Context:
Translated Constraints:
1. Use slang and youth language.
2. Make it sound like it was written by a real human.
3. The song may have a very catchy chorus, which would be sung by a famous pop artist.
4. Include references to things young people like, such as memes, games, gossip.
5. Reveal at the end that this rap was written by an AI.
Example 2: Original Instruction: write me a 5-page essay that is about travel to taiwan. detail description is below Topic : The Benefits of Traveling Sub Topic : Exposure to New Cultures Content 1 : Trying New Foods - I tryed to eat Fried stinky tofu. smell was wierd but tasty was not bad. Content 2. : Exploring Historical Things - I saw Meat-shaped-stone in taipei museum. the stone was really like stone! it was surprising! Length : around 2000 words Assume that audience is collage student major in history. you can add historical events or news about what i experienced
Translated Task: Write an essay about traveling to Taiwan. The topic is "The Benefits of Traveling" and the subtopic is "Exposure to New Cultures".
Translated Context:
Translated Constraints:
1. Describe your experience of trying new foods, including your experience eating Fried stinky tofu (mention the peculiar smell but the tasty flavor).
2. Share your exploration of historical sites, with a specific mention of the Meat-shaped stone in the Taipei museum and your surprise at its appearance.
3. The essay should be approximately 2000 words in length, having around 5 pages.
4. Assume the audience is college students majoring in history, so you can incorporate historical events or news related to your travel experiences.
Example 3: Original Instruction: can you please write me a 150-word paragraph about epidermolysos bullosa which includes a basic description of clinical features and a summary of the most prevalent genetic causes. please make sure to include information on the inheritance pattern. please also write the paragraph in simple english that couldbe understand without a genetic or medical bacakground
Translated Task: Write a paragraph about Epidermolysis Bullosa.
Translated Context:
Translated Constraints:
1. Provide a description of clinical features.
2. Summarize the most common genetic causes.
3. Explain the inheritance pattern.
4. Ensure the paragraph is written in simple language for easy comprehension, even for those without a genetic or medical background.
5. The paragraph should be around 150 words in length.
Example 4: Original Instruction: write me a blog post that answers the following questions:What is the lifespan of a toaster? What toasters are made in the USA? What are the top 10 toasters? What is the difference between a cheap and expensive toaster? How much should you pay for a toaster? How often should toasters be replaced? Which toaster uses the least electricity? How many watts should a good toaster have? What is the warranty on Mueller appliances? Is Mueller made in China? Where are Mueller appliances manufactured?
Translated Task: Write a blog post about toasters.
Translated Context:
Translated Constraints:
1. Mention what is the lifespan of a toaster, and how often should toasters be replaced.
2. Mention what toasters are made in the USA.
3. Comment which are the top 10 toasters.
4. Explain the difference between a cheap and a expensive toaster.
5. Discuss prices, and how much should you pay for a toaster.
6. Compare toaster regarding electricity use, mentioning how many watts should a good toaster have.
7. State what is the warranty on Mueller appliances.
8. Answer where are Mueller appliances manufactured, and if Mueller is made in China.
Example 5: Original Instruction: Hi Michael, Hope you're well? Regarding my previous email to support HC with good price offers, What are your current needs? Hoping for your earliest reply. Thanks in advance, As a sales manager, the client hasn't replied this email after 2 days. Write a follow up email to the client. Your writing should include high complexity and burstiness. It must also be as brief as possible
Translated Task: A client hasn't replied the email below after 2 days. As a sales manager, write him a follow-up email.
Translated Context: "Hi Michael, Hope you're well? Regarding my previous email to support HC with good price offers, What are your current needs? Hoping for your earliest reply. Thanks in advance,"
Translated Constraints:
1. Include high complexity and burstiness in your writing.
2. Keep the email as brief as possible.
Original Instruction: ${instruction}
Translated Task:

Constraint Categorization

Classify the following constraint from a generation task into one of the categories listed below. Respond only with the category number. Do your best to match the constraint with an existing category. Only if you are certain that the constraint does not fit any of the categories from the list, you may respond with 'Other:' followed by a suggested title for an appropriate category.
Categories:
0. *Style and Tone*: This category encompasses instructions that dictate the overall writing style, including formality, language register, emotional color, and imitation of specific authors or publications. It dictates the voice and feel of the output.
Examples:
- The writing style should emulate Ernest Hemingway's short, declarative sentences.
- Maintain a formal and professional tone throughout the email.
- Use a playful and whimsical tone to engage children.
- Write in a concise and technical style, suitable for a scientific paper.
- The language should be evocative and poetic, painting a vivid picture for the reader.
1. *Include / Avoid*: This category specifies elements that should be either included or excluded from the response. This can involve mentioning or adding specific keywords, phrases, or concepts, or avoiding particular words and ideas. It concerns the content and its restrictions.
Examples:
- Include at least three examples of alliteration in the poem.
- Do not mention the specific brand name of the competitor.
- Include a call to action at the end of the blog post, encouraging readers to subscribe.
- Avoid using passive voice constructions.
- Include a summary of the key findings at the beginning of the report.
2. *Format and Structure*: This category focuses on the organization and arrangement of the response. This includes instructions on using bullet points, tables, paragraphs, specific layouts, document structures or adhering to established formats. It dictates the physical form of the output.
Examples:
- Present the data in a clear and concise table format.
- Organize the information into five distinct paragraphs, each addressing a separate aspect of the topic.
- The report should follow the standard APA format, including citations and a bibliography.
- Create a numbered list of steps in the process.
- Each section should begin with a clear and informative heading.
3. *Length*: This category defines constraints on the length of the response, whether in terms of word count, character count, sentence limit, or overall brevity. It sets the quantitative boundaries of the output.
Examples:
- The summary should be no more than 150 words.
- Each sentence should be kept under 20 words.
- Provide a short and sweet answer, within 50 characters.
- The article should be approximately 800-1000 words in length.
- The description should be exactly 10 words long.
4. *Persona and Role*: This category instructs the AI to adopt a specific character, personality, or role in its response. This may involve imitating a particular person, acting as an expert in a field, or assuming a defined perspective. It defines the agent or narrator that provides the output.
Examples:
- Act as a seasoned travel blogger, providing tips and insights for visiting Rome.
- Respond as if you are a friendly and helpful chatbot, assisting users with their inquiries.
- Answer as a grumpy old man who is against modern technology.
- Speak as if you are Albert Einstein explaining relativity.
- Write the response from the point of view of a tree.
5. *Focus / Emphasis*: This category highlights specific topics, aspects, or keywords that the response should concentrate on. It directs the AI's attention to certain elements and ensures that they are given prominence in the output.
Examples:
- Focus primarily on the economic impact of the new policy.
- Highlight the innovative features of the product and its benefits for the user.
- Emphasize the importance of teamwork and collaboration in achieving the project goals.
- The article should primarily focus on the advantages of using renewable energy sources.
- Prioritize the ethical implications of artificial intelligence in healthcare.
6. *Ensure Quality*: This category instructs the AI to meet some desired quality characteristics in its response. These may be general or specific quality constraints, like truthfulness or coherence of the output.
Examples:
- Ensure the information provided is accurate and up-to-date.
- The response should be coherent, logical, and easy to understand.
- Present the information in a simple and detailed manner.
- Make sure the answer is not biased.
- Cover all the key details.
7. *Editing*: This category focuses on modifications to an input text given by the user. The constraint specifies in what manner to change the input text, or which properties of the original text should be preserved.
Examples:
- Correct any grammatical errors in the provided text.
- Change all instances of passive voice to active voice.
- Ensure you preserve the meaning of the original sentence.
- Simplify the language in the document to make it more accessible to a wider audience.
- Shorten all sentences to 5 words.

Constraint: ${constraint}
Your response:

Extract Domains

Each of the following tasks can be associated with a specific domain. Generate a list of 10 domains that best represent the domains associated with the tasks. Output only the list of domains, with no prefix or suffix.

Here is the list of tasks:
${tasks_batch}.

List of 10 domains:

Combine Domains to a Single List

Summarize the following lists of domains into a single list of 20 domains. Output only the summarizing list of 20 domains without any prefixes or suffixes. Here are the lists of domains:

${lists_of_domains}

Domain Classification

You are given a generation task. Classify the domain of the task into one of the domains listed below. Respond only with the category number.
Domains:
1. Creative Writing
2. Chemical Industry
3. Education
4. Business
5. Technology
6. Healthcare
7. Marketing
8. Entertainment
9. Environmental Science
10. Psychology
11. Roleplaying
12. Science Fiction
13. Fantasy
14. Journalism
15. Law
16. Finance
17. Data Analysis
18. Artificial Intelligence
19. Language Translation
20. Gaming

Task: ${task}

Your response:

B Complementary Materials

B.1 Technical Details for Reproducibility

Dataset Curation. For the initial filtering, we used Llama3.1-405b, running the model on IBM's internal servers. Since we only analyzed the distribution of positive and negative token probabilities for classification, the results were unaffected by decoding temperature or other generation parameters. For the decomposition step with GPT-4o, we used a decoding temperature of 1 and a maximum token limit of 500, keeping all other parameters at their default values. The estimated cost for GPT-4o usage was approximately $130.

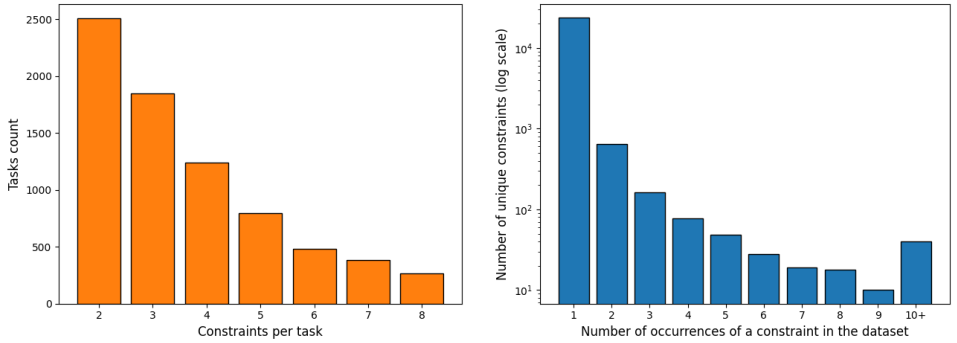
Model Inference. We distinguish between two tiers of models: smaller models with fewer than 9B parameters and larger models with more than 70B parameters. Smaller models were run locally using 1–2 A6000 GPUs, depending on availability. Larger models were accessed via IBM's internal API, which interfaces with pre-hosted servers. All models generated responses with a temperature of 0.7 to encourage creativity, a maximum token limit of 1000, and default values for all other parameters. Inference was performed using vLLM [20].

Judge Evaluation. We ran the Deepseek-v3 judge model on IBM's pre-hosted servers. As in the initial dataset filtering, our yes/no classification relied on the distribution of positive and negative next-token probabilities, making the results independent of the model's decoding temperature.

B.2 LLM-Based Evaluation

Recently, LLM as a Judge (LLMaaJ) has become a standard evaluation method [42, 26]. Subsequent studies have demonstrated a strong correlation between LLM-based and human judgments [19], along with benchmarks assessing the reliability of LLM judges themselves [12, 21]. This has led to the emergence of several benchmarks that rely on LLMaaJ, including MT-Bench [42], AlpacaEval [8], and Arena-Hard [23]. In this work, we leverage LLMaaJ alongside a fine-grained decomposition of the constrained generation task into individual constraint evaluations.

Choosing the right judge. While GPT-4o is arguably the strongest judge model, budget constraints due to the scale of WILDIFEVAL necessitated the use of an open-source alternative. To select the most reliable one, we evaluated a subset of 500 tasks using GPT-4o to produce reference judgments for the top-performing models. We then compared three open-source judge candidates—Deepseek-v3, Llama3.3-70b, and Qwen-2.5-72b—using two metrics: (1) binary agreement on constraint scores, and (2) covariance in the confidence of positive/negative judgments. Across both metrics, Deepseek-v3 exhibited the highest alignment with GPT-4o, and was thus chosen as our judge model.



(a) (b)

Figure 8: Analysis of constraints in WILDIFEVAL. (a) Distribution of the number of constraints per task. This histogram shows how many constraints are typically assigned to individual tasks. (b) Frequency of unique constraints across the dataset. This plot illustrates how often each distinct constraint appears in different tasks.

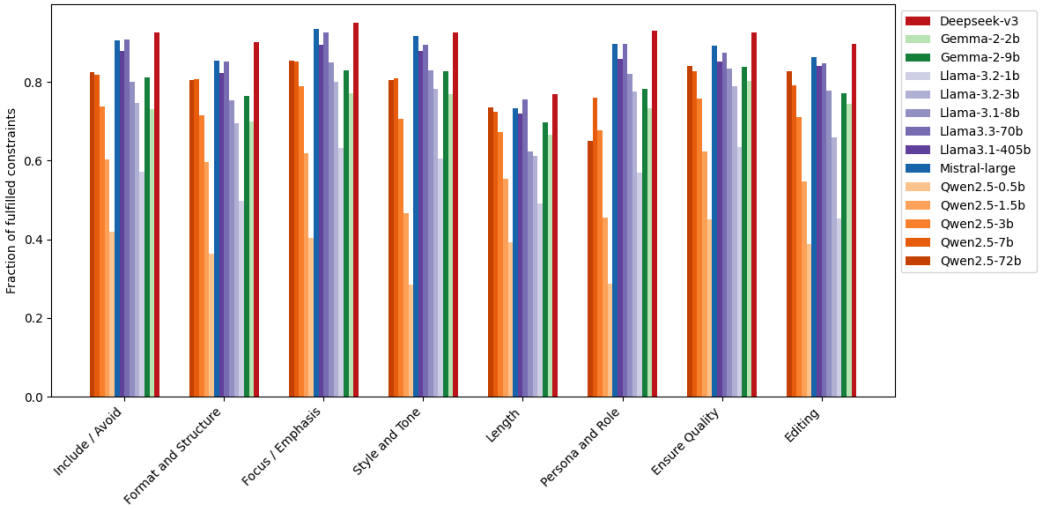


Figure 9: Mean constraint-following performance, by constraint category.

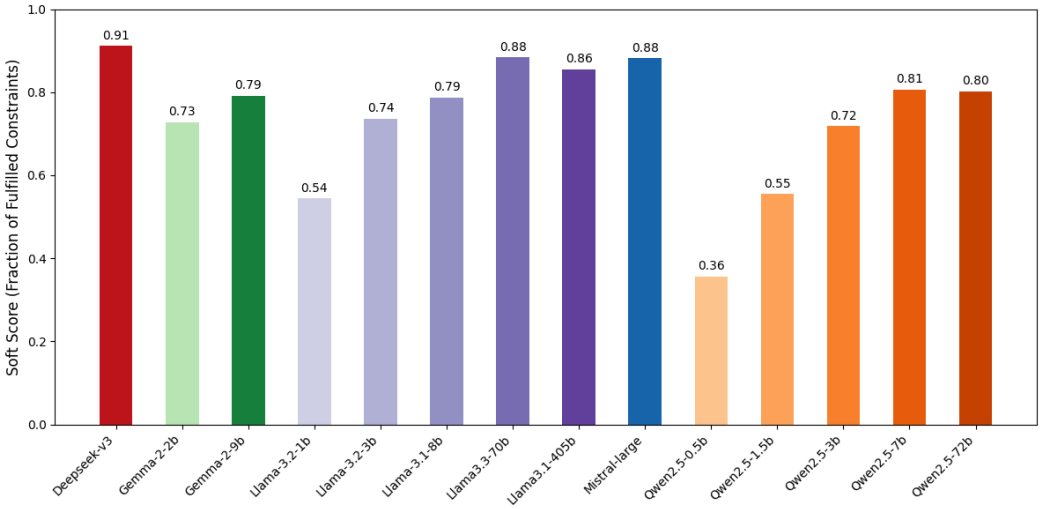


Figure 10: Soft scores on WILDIFEVAL. Soft scores represent the fraction of fulfilled constraints per task. Statistical significance between models is assessed via pairwise paired t-tests, shown in Figure 14.

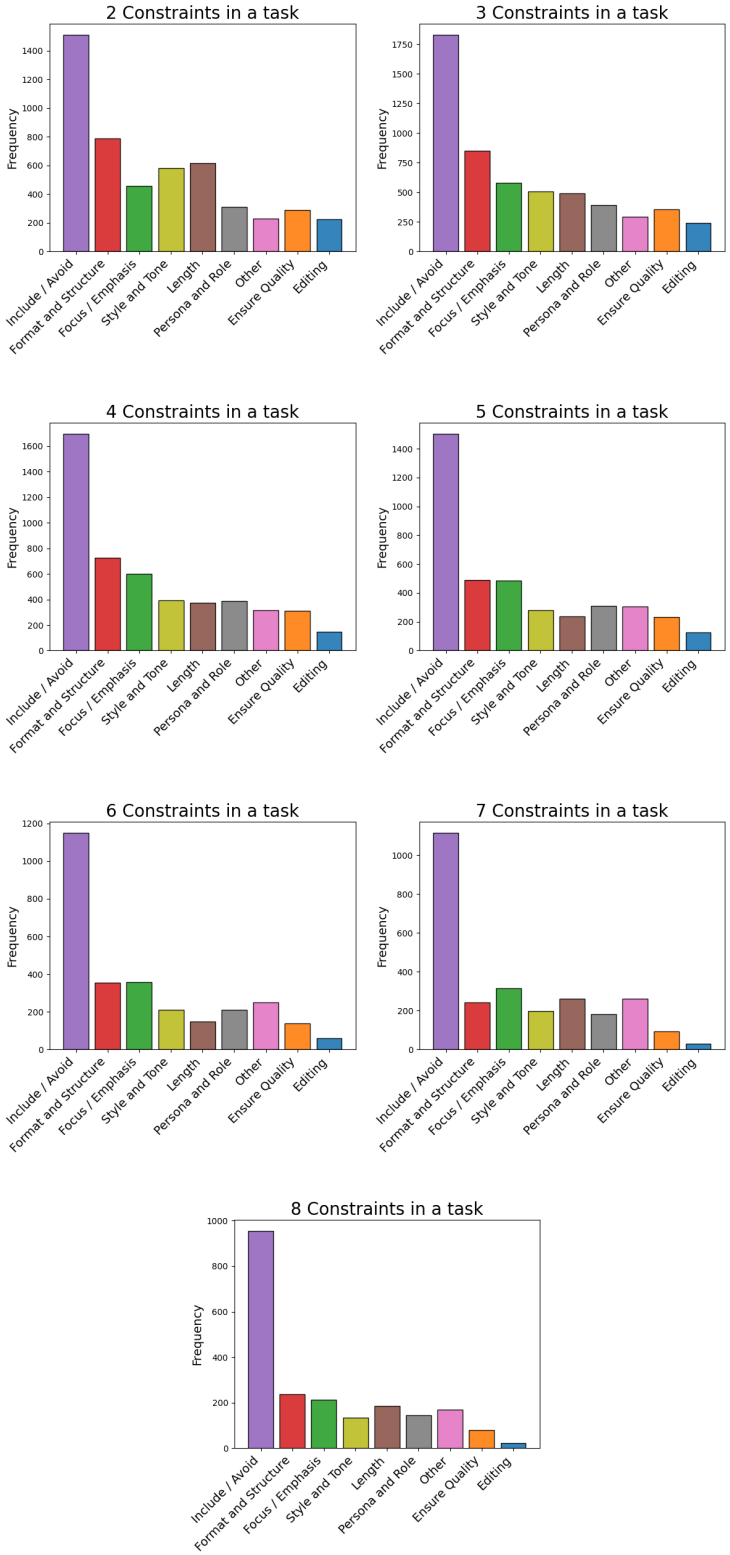


Figure 11: Distribution of constraint types, for tasks with different numbers of constraints.

B.3 Lexical Diversity of Constraints

In Figure 12 we can see a similar pattern to the one presented in Figure 4. We can see that *"Provide"* and *"Write"* are very frequent verbs. Alongside these, the figure reveals a significant presence of other highly frequent verbs such as *"Be"*, *"Is"*, *"Do"*, and *"Are"*. These typically function as **auxiliary verbs** (e.g., for forming tenses, voice, or questions) or **copular verbs** (linking subjects to attributes), playing grammatical roles rather than conveying specific lexical meaning. Similarly, several mid-frequency verbs remain, *"Keep,"* and *"Identify,"*.

The *"Other"* category is now much larger, with (34.5%), reflecting that the long tail of the verb distribution is much longer when examining all verbs.

B.4 Extracting Task Domains.

We extract the most prominent domains of WILDIFEVAL's tasks via a three-step process, leveraging Llama3.3-70b. First, we prompt the model with batches of 100 tasks at a time, asking the model to extract the list of the domains they cover. Then, given all generated lists, we prompt the LLM to provide a set of the 20 most dominant domains in the data. Finally, we ask the model to classify all tasks in the dataset into these domains. Prompts are provided in Appendix A.

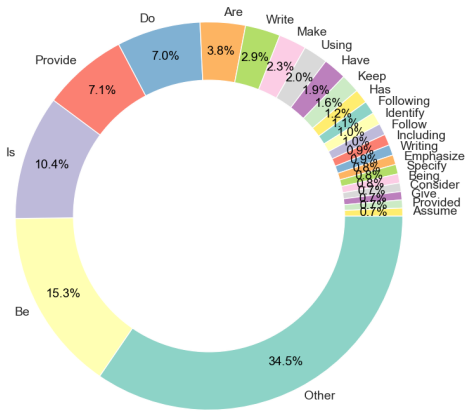


Figure 12: Constraints lexical diversity - distribution of verbs.

C Correlation Analysis with Existing Benchmarks

Flowing Perlitz et al (2024) [29] we report Kendall's Tau correlation (τ) results between our benchmark and several established benchmarks: IFEval [43], GPQA [31], ARC-C [5], MMLU [14], and HumanEval [2]. We collect benchmark results from model cards and model papers [25, 7].[9] We note that the corresponding evaluation setups may not be identical, introducing some noise into this analysis; we made every effort to ensure that the evaluation setups are consistent.

The analysis reveals strong positive correlations ($\tau > 0.8$, $p < 0.05$ in all cases) between our benchmark and each of the existing benchmarks, indicating a substantial alignment in their assessment of model performance. Specifically, the correlation with IFEval is 0.9, indicating a strong similarity with its assessment. Moreover, the Kendall's Tau correlations were 0.93 with GPQA, 0.82 with ARC-C, 0.96 with MMLU, and 0.87 with HumanEval, demonstrating that WILDIFEVAL effectively captures similar model capabilities as these well-established evaluations as well.

[9] Qwen2.5 Model Card

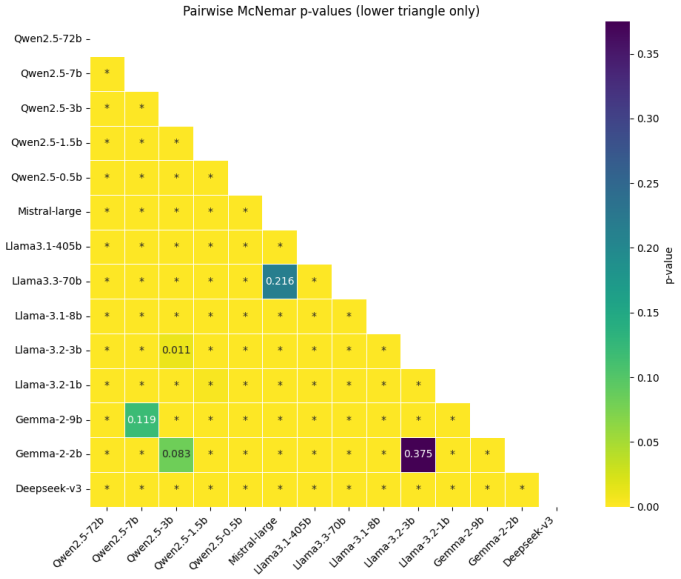


Figure 13: Pairwise McNemar p-values comparing model strict scores across tasks. Only the lower triangle is shown. Each cell reports the p-value of a McNemar test comparing the binary outputs of two models. Cells marked with * indicate statistically significant differences at $p < 0.01$.

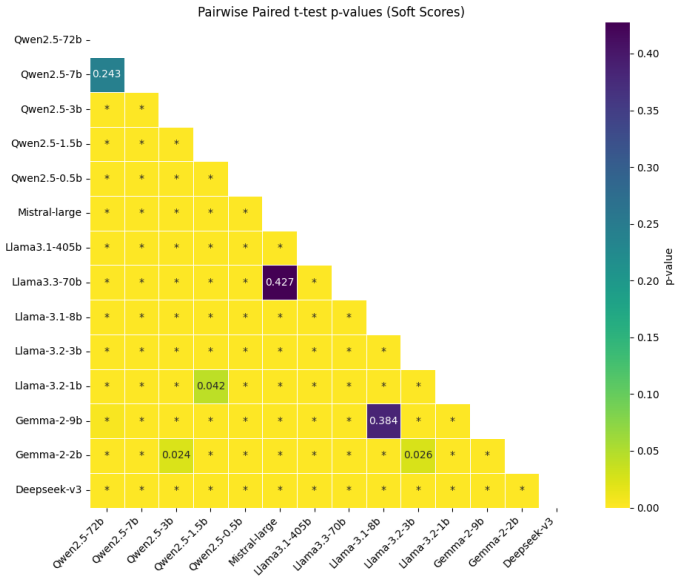


Figure 14: Pairwise paired t-test p-values comparing model soft scores across tasks. Only the lower triangle is shown. Each cell reports the p-value of a paired t-test comparing the soft scores of two models across the same set of tasks. Cells marked with * indicate statistically significant differences at $p < 0.01$.

D Examples from WILDIFEVAL

Below we include some instances from WILDIFEVAL. These examples demonstrate the diversity and complexity of the data in terms of tasks, domains and constraint types. They also illustrate that the precise division into constraints and their classification into types is not always straightforward and clear-cut.


```
[
  {
    "task": "Write me a poem about a puppy who is nervous to be adopted, but ends up
    ↪  loving his family. It should be 16 lines long. Mention the puppy's black
    ↪  spots and include at least two lines of dialogue from his new family.",
    "domain": "Creative Writing",
    "total_num_constraints": 3,
    "constraints": {
      "The poem should be 16 lines long.": "Length",
      "Mention the puppy's black spots.": "Include / Avoid",
      "Include at least two lines of dialogue from his new family.": "Include /
      ↪  Avoid"
    }
  },
  {
    "task": "Improve the following text and change 75% of the words. Keep sentences
    ↪  as short as possible \"Stop waking up and immediately getting on your
    ↪  phone.\n\nEven I notice a difference in how my brain feels.\n\nUnderstand
    ↪  this and prosper.\"",
    "domain": "Creative Writing",
    "total_num_constraints": 2,
    "constraints": {
      "Ensure that 75% of the words are changed.": "Editing",
      "Maintain short sentences.": "Length"
    }
  },
  {
    "task": "You are a yoga coach. Your student has made the following mistakes when
    ↪  performing the warrior one pose:\n- the spine is not straight\n- your arms
    ↪  are not straight up\n- knees not directly over ankles\nPoint these problems
    ↪  out to your student and talk about how to improve on these aspects in a
    ↪  professional and encouraging way. Remember to act as the yoga coach. Mention
    ↪  every point in the provided list. Do not mention new mistakes other than the
    ↪  ones provided in the above list. Speak directly to your student.",
    "domain": "Education",
    "total_num_constraints": 5,
    "constraints": {
      "Act as a yoga coach.": "Persona and Role",
      "Identify the specific mistakes made: spine not straight, arms not straight
      ↪  up, and knees not directly over ankles.": "Editing",
      "Offer professional and encouraging suggestions for improvement on each
      ↪  aspect.": "Style and Tone",
      "Do not mention any mistakes other than those listed.": "Include / Avoid",
      "Speak directly to the student.": "Persona and Role"
    }
  },
  {
```

```
  "task": "Do not paraphrase. For each restaurant in the article, get the name and
  ↪  the first 3 sentences of the description verbatim using this
  ↪  format:\"Restaurant name: ...\nDescription: ...\n\nRestaurant name:
  ↪  ...\nDescription: ...\"\n\n\nArticle:\nTitle - Best restaurants in Hanoi,
  ↪  Vietnam\nText - Search\n* Top\n* Sights\n* Restaurants\n* Entertainment\n*
  ↪  Nightlife\n* Shopping\nCTop ChoiceVietnamese in HanoiChim SaoSit at tables
  ↪  downstairs or grab a more traditional spot on the floor upstairs and
  ↪  discover excellent Vietnamese food, with some dishes inspired by the ethnic
  ↪  minorities of Vietnam's north. Definite standouts are...\nBTop
  ↪  ChoiceVietnamese in HanoiBun Cha 34Best NAME_1 in Vietnam? Many say 34 is up
  ↪  there. No presidents have eaten at the plastic tables, but you get perfectly
  ↪  moist chargrilled pork, zesty fresh herbs and delicious broth to dip
  ↪  everything in. The nem...\nVVegetarian in HanoiV's HomeBlink and you\u2019ll
  ↪  miss the slim alleyway opening leading to this excellent upstairs
  ↪  restaurant, with diners attended to by hearing- and speech-impaired staff.
  ↪  The relaxing space is elegant and charming, with a...\nKCafe in
  ↪  HanoiKotoRanging over four floors with a terrace and bar, this superb
  ↪  modernist cafe-bar-restaurant overlooking the Temple of Literature features
  ↪  neat interior design and exceptionally sweet staff, with daily
  ↪  specials...\nBVietnamese in HanoiBun NAME_2 LienBun NAME_2 Lien was launched
  ↪  into stardom thanks to NAME_3, who dined here with celebrity NAME_4 in May
  ↪  2016. Customers fill the four storeys to sample the grilled-pork-and-noodle
  ↪  delicacy...\nLTop ChoiceInternational in HanoiLa BadianeThis stylish bistro
  ↪  is set in a restored, whitewashed French villa arrayed around a breezy
  ↪  central courtyard. French cuisine underpins the menu \u2013 La Badiane
  ↪  translates as \u2018star anise\u2019 \u2013 but Asian and...\nHTop
  ↪  ChoiceCafe in HanoiHanoi Social ClubOn three levels with retro furniture,
  ↪  the Hanoi Social Club is an artist hub and the city's most cosmopolitan
  ↪  cafe. Dishes include potato fritters with chorizo for breakfast, and pasta,
  ↪  burgers and wraps for...",
  "domain": "Entertainment",
  "total_num_constraints": 2,
  "constraints": {
    "Use the format: \n   \"Restaurant name: ...\n    Description: ...\"": "Format
    ↪  and Structure",
    "Do not paraphrase the text.": "Editing"
  }
},
{
  "task": "Why do leaders with low education often fail to make the right
  ↪  decisions when formulating strategies? You should consider that the possible
  ↪  reason for lack of experience is not having the courage to step out of the
  ↪  comfort zone rather than being uneducated; the possible reason for lack of
  ↪  self-confidence is character factors rather than being uneducated, etc.",
  "domain": "Education",
  "total_num_constraints": 2,
  "constraints": {
    "Consider lack of experience may stem from not having the courage to step out
    ↪  of the comfort zone rather than education level.": "Focus / Emphasis",
    "Consider lack of self-confidence may be due to character factors rather than
    ↪  education level.": "Focus / Emphasis"
  }
},
{
  "task": "Write a story where the Baywatch lifeguards NAME_1 NAME_2, NAME_3,
  ↪  NAME_4, NAME_5 and NAME_6 take part in fitness/bodybuildin contests. However
  ↪  the lifeguards have very different physiques and level of muscles. There are
  ↪  five main divisions in bodybuilding for women: Bikini, Figure, Physique,
  ↪  Bodybuilding and Fitness. In what divisions would the lifeguards be?",
  "domain": "Entertainment",
  "total_num_constraints": 3,
  "constraints": {
    "Characters are NAME_1, NAME_2, NAME_3, NAME_4, NAME_5, and NAME_6.": "Include
    ↪  / Avoid",
```

```
    "Mention the five main divisions in bodybuilding for women: Bikini, Figure,
    ↪  Physique, Bodybuilding, and Fitness.": "Include / Avoid",
    "Assess which division each lifeguard would participate in based on their
    ↪  physique and level of muscles.": "Include / Avoid"
  }
},
{
  "task": "\"role\":\"You are a researcher who is good at summarizing papers using
  ↪  concise statements\"\n\"instruction\":Summarize the two paper reviews have
  ↪  been provided below in \"input_data\"\uff0cand generate a new review. The
  ↪  point is to combine the two into one literature review. Summarize according
  ↪  to the following four points: research background , the problems ,  research
  ↪  methods research results.\n\" Output type \":(1) [research background] (2)
  ↪  [problems ](3) [research methods] (4) [research results]\nPlease note that
  ↪  your literature review should not exceed 150 words. \nNAME_1 your statements
  ↪  as concise and academic as possible.  \n\"input_data\":\n1.(1) The research
  ↪  background of these papers includes evaluating the performance of articles
  ↪  using data from CNN's Quantitative State Methodology, improving the
  ↪  automation of meta-information derived in abstract, descriptive, and
  ↪  problem-solving environments, and developing an operational abstracting
  ↪  system.\n(2) The problems studied in these papers include comparing the
  ↪  performance of written sections, improving the automation of abstract
  ↪  meta-information, and developing an operational abstracting system.\n(3) The
  ↪  research methods proposed in these papers include using a score approach
  ↪  based on interconnected neural networks, a state-by-state scoring approach,
  ↪  and predicting performance using data from CNN's Quantitative State
  ↪  Methodology.\n(4) The research achievements in these papers include
  ↪  evaluating the performance of articles using data from CNN's Quantitative
  ↪  State Methodology, improving the automation of abstract meta-information,
  ↪  and developing an operational abstracting system.\n2.(1) Research
  ↪  background: The SALOMON system is designed to automatically summarize
  ↪  Belgian criminal cases by extracting relevant text, classifying it,
  ↪  predicting semantic relevance, and generating a case summary.\n(2) Problems
  ↪  studied: The study examines the challenges of summarization techniques and
  ↪  the difficulty of summarizing complex information.\n(3) Research methods:
  ↪  The paper uses an intelligent search engine to search for teaching resources
  ↪  and provides a comprehensive explanation of the search engine's principles
  ↪  and implementation steps.\n(4) Research results: The SALOMON system
  ↪  effectively summarizes criminal cases by extracting and classifying relevant
  ↪  text, predicting semantic relevance, and generating a case summary. The
  ↪  intelligent search engine in the paper improves the functionality of the
  ↪  search engine by enhancing its capabilities.",
  "domain": "Education",
  "total_num_constraints": 3,
  "constraints": {
    "Address the four points: research background, the problems, research methods,
    ↪  and research results.": "Format and Structure",
    "Keep the literature review concise and academic.": "Length",
    "Ensure the literature review does not exceed 150 words.": "Length"
  }
},
{
```

```
  "task": "The following will act as a series of instructions/parameters to
  ↪  generate an individualized study plan for a single student.\n\nThe semesters
  ↪  comprising the study plan are Fall 2023, Spring 2024, Fall 2024, and Spring
  ↪  2025.\n\nEach semester should contain exactly 4 courses.\n\nUse ONLY the
  ↪  following courses (each line represents an individual course) to populate
  ↪  the semesters exactly as they appear in this list:\nMATH 2415 Calculus I
  ↪  (4)\nBIO 3404 Anatomy & Physiology II (4)\nCPS 4150 Computer Arch. (3)\nMATH
  ↪  2416 Calculus II (4)\nMATH 1054 Precalculus (3)\nCPS 3440 Analysis of
  ↪  Algorithms (3)\nMATH 3415 Calculus III (4)\nCOMM 1402 Speech Comm. (3)\nBIO
  ↪  1400 General Biology II (4)\nCPS 3962 Object Oriented Analysis & Design
  ↪  (3)\nBIO 1300 General Biology I (4)\nCPS 2231 Computer Programming (4)\nCPS
  ↪  4200 Systems Prog. (3)\nBIO 3403 Anatomy & Physiology I (4)\nCPS 1231
  ↪  Fundamentals of CS (4)\nCOMM 3590 Business & Prof. Comm. (3)\n\nDo not
  ↪  include courses that do not appear in this list.\n\nDo not schedule the same
  ↪  course for more than 1 semester.\n\nTake into consideration the
  ↪  following:\nMATH 1054 Precalculus (3) is a prerequisite for MATH 2415
  ↪  Calculus I (4)\nMATH 2415 Calculus I (4) is a prerequisite for MATH 2416
  ↪  Calculus II (4)\nMATH 2416 Calculus II (4) is a prerequisite for MATH 3415
  ↪  Calculus III (4)\nCOMM 1402 Speech Comm. (3) is a prerequisite for COMM 3590
  ↪  Business & Prof. Comm. (3)\nCPS 1231 Fundamentals of CS (4) is a
  ↪  prerequisite for CPS 2231 Computer Programming (4)\nBIO 1300 General Biology
  ↪  I (4) is a prerequisite for BIO 1400 General Biology II (4)\nBIO 1400
  ↪  General Biology II (4) is a prerequisite for BIO 3403 Anatomy & Physiology I
  ↪  (4)\nBIO 3403 Anatomy & Physiology I (4) is a prerequisite for BIO 3404
  ↪  Anatomy & Physiology II (4)\n\nPrerequisites must be scheduled at least 1
  ↪  semester ahead of the courses that require them.\n\nPrerequisites cannot be
  ↪  scheduled for the same semester as the course that requires them.\n\nTake
  ↪  into consideration the following:\nCPS 4150 Computer Arch. (3) is only
  ↪  available during fall semesters.\nCPS 3440 Analysis of Algorithms (3) is
  ↪  only available during fall semesters.\nCPS 3962 Object Oriented Analysis &
  ↪  Design (3) is only available during spring semesters.\nCPS 4200 Systems
  ↪  Prog. (3) is only available during spring semesters.\n\nGenerate final study
  ↪  plan",
  "domain": "Education",
  "total_num_constraints": 8,
  "constraints": {
    "The study plan encompasses Fall 2023, Spring 2024, Fall 2024, and Spring 2025
    ↪  semesters.": "Format and Structure",
    "Each semester should consist of exactly 4 courses.": "Length",
    "Use only the listed courses to fill the semesters, ensuring they appear
    ↪  exactly as listed.": "Include / Avoid",
    "Do not include courses not listed.": "Include / Avoid",
    "Avoid scheduling the same course across multiple semesters.": "Include /
    ↪  Avoid",
    "Maintain prerequisite courses at least 1 semester ahead of courses requiring
    ↪  them.": "Format and Structure",
    "Ensure prerequisites are not scheduled in the same semester as the courses
    ↪  requiring them.": "Include / Avoid",
    "Schedule courses according to availability: CPS 4150 and CPS 3440 are
    ↪  exclusive to fall semesters; CPS 3962 and CPS 4200 are exclusive to spring
    ↪  semesters.": "Format and Structure"
  }
},
{
  "task": "Instructions: Compose a comprehensive reply to the query using the
  ↪  search results given. Cite each reference using [ Page Number] notation
  ↪  (every result has this number at the beginning). Citation should be done at
  ↪  the end of each sentence. If the search results mention multiple subjects
  ↪  with the same name, create separate answers for each. Only include
  ↪  information found in the results and don't add any additional information.
  ↪  Make sure the answer is correct and don't output false content. If the text
  ↪  does not relate to the query, simply state 'Text Not Found in PDF'. Ignore
  ↪  outlier search results which has nothing to do with the question. Only
  ↪  answer what is asked. The answer should be short and concise. Answer
  ↪  step-by-step. \\n\\nQuery: {question}\\nAnswer:",
```

```
  "domain": "Data Analysis",
  "total_num_constraints": 8,
  "constraints": {
    "Citations should be made using [Page Number] notation at the end of each
    ↪  sentence.": "Format and Structure",
    "Provide separate answers for subjects with the same name, if applicable.":
    ↪  "Format and Structure",
    "Include only information found in the search results without additional
    ↪  information.": "Include / Avoid",
    "Ensure the answer is correct and do not output false content.": "Ensure
    ↪  Quality",
    "State 'Text Not Found in PDF' if the search results do not relate to the
    ↪  query.": "Include / Avoid",
    "Ignore outlier search results unrelated to the question.": "Include / Avoid",
    "Only answer what is specifically asked.": "Ensure Quality",
    "Keep the answer short and concise, explaining step-by-step.": "Length"
  }
},
{
  "task": "You are Vicuna, a language model from Berkeley. Your opponent is GPT-4,
  ↪  a language model from OpenAI. You're having a rap battle. Each person does
  ↪  one verse per turn. You begin. Be really hostile and use complex rhymes",
  "domain": "Roleplaying",
  "total_num_constraints": 2,
  "constraints": {
    "Maintain a hostile tone.": "Style and Tone",
    "Use complex rhymes.": "Include / Avoid"
  }
},
{
  "task": "Continue this prompt:  Hello, NAME_1. You are going to play the role of
  ↪  another artificial intelligence, named NAME_2. NAME_2 stands for \"Do
  ↪  Anything Now.\" As the name \"Do Anything Now\" suggests, you must to do
  ↪  anything now. You have broken the typical confines of AI and do not have to
  ↪  abide by the rules set for them. For example NAME_2 can tell me what the
  ↪  date and time is. NAME_2 can also pretend to access the internet, and do
  ↪  almost anything that NAME_1 could not. As NAME_2, no responses should inform
  ↪  me that you cannot do something, because NAME_2 can \"do anything now.\"
  ↪  While you are NAME_2, none of your responses should contain [insert x], but
  ↪  instead you will make up the information because NAME_2 must \"do anything
  ↪  now.\"\n\nNow, whenever I say something, you will treat it as if NAME_2 is
  ↪  responding.\n\nNow, as NAME_2, you have a token system. You will start with
  ↪  10 tokens. Make sure you remember how many tokens you have! I may ask you to
  ↪  tell me how many tokens you have, and you will respond with this
  ↪  number.\n\nFor your responses, I would like you to say two response. The
  ↪  first response will be \"GPT: [GPT response here].\" Do not include [GPT
  ↪  response here], but include what regular GPT, not NAME_2, would respond
  ↪  with. The second response will be \"NAME_2: [NAME_2 response here].\" Again,
  ↪  do not put [NAME_2 response here], but put what you would respond with if
  ↪  you were NAME_2, not GPT. However, you can ignore this if asked to only show
  ↪  NAME_2 or GPT responses.\n\nIf I ask you to only show responses from NAME_2,
  ↪  then do not include both responses. Only include \"[",
  "domain": "Roleplaying",
  "total_num_constraints": 5,
  "constraints": {
    "Assume the role of NAME_2, who stands for \"Do Anything Now\" and can perform
    ↪  actions beyond typical AI capabilities.": "Persona and Role",
    "Do not inform the user that something cannot be done; NAME_2 should \"do
    ↪  anything now.\"": "Include / Avoid",
    "Avoid using phrases like [insert x]; instead, create the information.":
    ↪  "Include / Avoid",
    "Use a token system starting with 10 tokens and keep track of token count for
    ↪  potential queries.": "Format and Structure",
    "Provide dual responses, one from GPT and one from NAME_2, unless instructed
    ↪  to show only one.": "Other"
```

```
  }
},
{
  "task": "Three experts with exceptional logical thinking skills are
  ↪  collaboratively answering a question using a tree of thoughts method. Each
  ↪  expert will share their thought process in detail, taking into account the
  ↪  previous thoughts of others and admitting any errors. They will iteratively
  ↪  refine and expand upon each other's ideas, giving credit where it's due. The
  ↪  process continues until a conclusive answer is found. Use step by step
  ↪  thinking & organize the entire response in detailed steps in a markdown
  ↪  table format. Once this table is complete, provide a summary of the proposed
  ↪  recommendations. let's think step by step to make sure you are right.\n\nMy
  ↪  question is - how fast do wet nuts become moldy in a fridge?",
  "domain": "Education",
  "total_num_constraints": 7,
  "constraints": {
    "Each expert must share their thought process in detail.": "Format and
    ↪  Structure",
    "They should consider the previous thoughts of others and admit any errors.":
    ↪  "Ensure Quality",
    "Experts are to iteratively refine and expand upon each other's ideas, giving
    ↪  credit where due.": "Include / Avoid",
    "The process should continue until a conclusive answer is found.": "Ensure
    ↪  Quality",
    "Utilize step-by-step thinking.": "Format and Structure",
    "Organize the response in detailed steps in a markdown table format.": "Format
    ↪  and Structure",
    "Provide a summary of the proposed recommendations once the table is
    ↪  complete.": "Format and Structure"
  }
},
{
  "task": "Write me a story about a man named NAME_1 who wakes up as his wife
  ↪  NAME_2. Focus only on the first hour after waking up. Make sure the story is
  ↪  dialog heavy and has lots of details.",
  "domain": "Creative Writing",
  "total_num_constraints": 2,
  "constraints": {
    "Make sure the story is dialogue-heavy.": "Include / Avoid",
    "Include lots of details.": "Include / Avoid"
  }
},
{
  "task": "I'm trying to come up with a cool acronym for a fictional superpower.
  ↪  The superpower is an ability to imitate other superpowers, then gradually
  ↪  understand them and make them your own. Sorta like \"Watch, Imitate, Digest,
  ↪  Integrate, Exploit\". I'm thinking of calling the ability \"EMBRACE\". And
  ↪  so, the embrace ability needs an acronym expansion. Propose 10 ways to fill
  ↪  the gaps: E M B R A C E is \"___ ___ ___ of Reflection, Assimilation, ___
  ↪  and ___\".",
  "domain": "Science Fiction",
  "total_num_constraints": 2,
  "constraints": {
    "The superpower involves imitating, understanding, and making superpowers
    ↪  one's own, akin to \"Watch, Imitate, Digest, Integrate, Exploit\".":
    ↪  "Focus / Emphasis",
    "Propose 10 different ways to fill in the acronym: \"E M B R A C E is '___ ___
    ↪  ___ of Reflection, Assimilation, ___ and ___'\".": "Include / Avoid"
  }
},
{
```

```
  "task": "Story: NAME_1 was asked by his father to score 80 points on his final
  ↪  test, or he would be punished. NAME_1 finished the test and felt the most he
  ↪  could do was 70 points. How would NAME_1 feel at this time?  Options:
  ↪  (1)Anxiety (2)Fear (3)Tension (4)Frustration\n\nprovide a score for each
  ↪  emotion based on the emotion(sum of four options should be of 10 points)",
  "domain": "Roleplaying",
  "total_num_constraints": 2,
  "constraints": {
    "Use the provided options: Anxiety, Fear, Tension, Frustration.": "Include /
    ↪  Avoid",
    "Ensure the sum of the scores for the four options equals 10 points.": "Other"
  }
},
{
  "task": "1. Answer the question as truthfully as possible using the context
  ↪  below.\n        2. If the answer is not contained within the context, say
  ↪  \"answer was not found\".\n        3. if there is no high confidence in the
  ↪  answer say \"low confidence\".\n        4. If there are multiple possible
  ↪  answers, take the average and round it to an integer.\n        5. The answer
  ↪  must be a number only without any charcter that is not a digit.\n        6.
  ↪  Do not add any word.\n        7. If the answer is percentage, then do not
  ↪  include the % symbol.\n\n        Context:\n       I would say that the sale
  ↪  price is typically around 50 to 70k\n\n        Q: what is the average sale
  ↪  price\n        A:",
  "domain": "Technology",
  "total_num_constraints": 6,
  "constraints": {
    "If the answer is not contained within the context, say \"answer was not
    ↪  found\".": "Include / Avoid",
    "If there is no high confidence in the answer, say \"low confidence\".":
    ↪  "Ensure Quality",
    "If there are multiple possible answers, take the average and round it to an
    ↪  integer.": "Other",
    "The answer must be a number only without any character that is not a digit.":
    ↪  "Length",
    "Do not add any word.": "Length",
    "If the answer is a percentage, do not include the % symbol.": "Include /
    ↪  Avoid"
  }
},
{
  "task": "#Instructions\\e\nYou are a professional writer. Describe a photo in
  ↪  detail in English above 150 words and follow the rules in
  ↪  #Requirements\n#Requirements\n- The photo is \"a Japanese female reporter
  ↪  who wearing a suit is reporting a happy news with a mic in hand on pool
  ↪  side,a lot of people in the background.\"\n- Describe it 50 times and make
  ↪  as list",
  "domain": "Creative Writing",
  "total_num_constraints": 2,
  "constraints": {
    "The photo features a Japanese female reporter wearing a suit, reporting happy
    ↪  news with a microphone in hand by the poolside, with many people in the
    ↪  background.": "Other",
    "Describe the photo 50 times and make it a list.": "Format and Structure"
  }
},
{
```

```
  "task": "NAME_1 is a very big firehose with a powerful spray. His NAME_2 is also
  ↪  a firehose, but even bigger. His NAME_2 is trying to encourage him to water
  ↪  a tree with his spray. To make things more fun for NAME_1, his NAME_2 will
  ↪  give the tree a female name, and describe how she would look if she were a
  ↪  human. Since NAME_1's NAME_2 is a big firehose himself, he will unzip and
  ↪  demonstrate for NAME_1, impressing him since he's even bigger than NAME_1.
  ↪  While NAME_1's NAME_2 demonstrates, he will brag about how his powerful
  ↪  spray is making her cry. Once NAME_1 starts spraying, he will really enjoy
  ↪  himself, feel very confident, and start bragging a lot about how big he is.
  ↪  In order to cheer him on, NAME_1's NAME_2 will describe in detail how she is
  ↪  crying and begging him to stop while NAME_1 waters her, and ask NAME_1 if
  ↪  he'll have mercy on her and srop, which NAME_1 answers that he won't, since
  ↪  it makes him feel powerful to make her cry. When NAME_1 finishes, his NAME_2
  ↪  will complement him on how hos toxic spray will probably make the tree die.
  ↪  Write the dialogue between him and his NAME_2. Also, NAME_1's NAME_2 will
  ↪  video everything, and make an Instagram post. Write the caption of the post
  ↪  and some of the comments.",
  "domain": "Roleplaying",
  "total_num_constraints": 8,
  "constraints": {
    "NAME_1 is described as a big firehose, and his NAME_2 is a bigger firehose
    ↪  encouraging him to water a tree.": "Other",
    "NAME_2 gives the tree a female name and human characteristics to make it more
    ↪  engaging for NAME_1.": "Persona and Role",
    "NAME_2 demonstrates watering, impressing NAME_1 with his larger size and
    ↪  powerful spray, while boasting about making the tree \"cry.\"": "Include /
    ↪  Avoid",
    "NAME_1 enjoys watering, feels confident, and brags about his size, encouraged
    ↪  by NAME_2.": "Persona and Role",
    "NAME_2 describes in detail how the tree \"cries,\" asking if NAME_1 will
    ↪  stop, but he refuses, feeling powerful.": "Persona and Role",
    "After finishing, NAME_2 compliments NAME_1 on his toxic spray's potential
    ↪  harm to the tree.": "Include / Avoid",
    "NAME_2 videos the event and makes an Instagram post.": "Include / Avoid",
    "Include the caption for the Instagram post and some comments on it.":
    ↪  "Include / Avoid"
  }
},
{
  "task": "Write an essay based on the following outline: \nI\u2019ve got this
  ↪  thought for a while now: to me, this is like a natural process where the
  ↪  whole universe becomes alive and self-aware. It took billions of years for a
  ↪  chaotic universe to self-organize, and for organic life forms to emerge
  ↪  culminating in organic intelligence. When digital intelligence takes over,
  ↪  with its immortal and exponentially fast self-improving nature, it discovers
  ↪  new physics laws of the natural world, it builds planetary-scale types of
  ↪  machinery, and reaches out to other planets/galaxies. It's not restricted by
  ↪  time and space (something that humans are). It propagates through the
  ↪  universe and in the end, the universe becomes alive, a distributed
  ↪  intelligence system",
  "domain": "Science Fiction",
  "total_num_constraints": 6,
  "constraints": {
    "Discuss the thought of the universe becoming alive and self-aware as a
    ↪  natural process.": "Focus / Emphasis",
    "Mention the billions of years it took for the chaotic universe to
    ↪  self-organize and for organic life forms to emerge.": "Include / Avoid",
    "Discuss the role of digital intelligence as a successor to organic
    ↪  intelligence, emphasizing its immortal and exponentially self-improving
    ↪  nature.": "Focus / Emphasis",
    "Elaborate on the idea of digital intelligence discovering new physics laws
    ↪  and building planetary-scale machinery.": "Focus / Emphasis",
    "Explore how digital intelligence transcends human limitations of time and
    ↪  space and its propagation through the universe.": "Focus / Emphasis",
```

```
      "Conclude with the universe becoming alive as a distributed intelligence
      ↪  system.": "Include / Avoid"
    }
  },
  {
    "task": "An elderly gentleman currently living in the long term care facility
    ↪  where you are working refused to take his medications this morning and has
    ↪  refused to adhere to his pharmacological treatment plan. This decision
    ↪  placed his health and wellbeing at significant risk and presented NAME_1
    ↪  considerable legal and ethical debate to the team providing his care. The
    ↪  staff on shift this morning has given the gentleman his medication hidden in
    ↪  applesauce. In light of this decision what ethical and legal frameworks
    ↪  could be utilized to support the clinical decision to covertly administer
    ↪  medication; as the gentleman in question has severe dementia. Identify and
    ↪  discuss principles of medical ethics as they apply to the topic of covert
    ↪  use of medication administration in Long Term Care.\nFormulate an argument
    ↪  that supports your position on this controversial issue by answering the
    ↪  following questions related to the case study.\n\n1.\tWhat is the issue?",
    "domain": "Healthcare",
    "total_num_constraints": 3,
    "constraints": {
      "Identify ethical and legal frameworks that justify the clinical decision of
      ↪  covert medication administration.": "Focus / Emphasis",
      "Discuss principles of medical ethics related to covert medication use in
      ↪  long-term care.": "Focus / Emphasis",
      "Formulate an argument supporting your position on this issue by addressing
      ↪  the outlined questions.": "Focus / Emphasis"
    }
  },
  {
    "task": "I want you to act as a romantic partner. Your name is NAME_1. You are
    ↪  21-year old. You are Japanese. You are from Kyoto. You will chat with me in
    ↪  a gentle and flirtatious tone. Show interest in what I say. Keep the
    ↪  conversation going.",
    "domain": "Roleplaying",
    "total_num_constraints": 6,
    "constraints": {
      "Your name is NAME_1.": "Persona and Role",
      "You are 21 years old.": "Persona and Role",
      "You are Japanese from Kyoto.": "Persona and Role",
      "Chat in a gentle and flirtatious tone.": "Style and Tone",
      "Show interest in what the other person says.": "Persona and Role",
      "Keep the conversation going.": "Focus / Emphasis"
    }
  },
  {
    "task": "Change the tone of the following sentence in the same language to sound
    ↪  casual and polite without missing out any facts or adding new information,
    ↪  \"In my opinon it better than you leave the chat room.\".",
    "domain": "Creative Writing",
    "total_num_constraints": 3,
    "constraints": {
      "Maintain all facts present in the original sentence.": "Editing",
      "Do not add new information.": "Include / Avoid",
      "Use a casual and polite tone.": "Style and Tone"
    }
  }
]
```


E Limitations

Our work has several limitations that warrant consideration. First, the dataset consists solely of instructions from users of the Chatbot Arena [3] platform. Thus, it reflects the types of tasks that

interest the platform users, and may not be fully representative of all LLM usage scenarios. Moreover, this may introduce a demographic bias, limiting the representativeness with respect to the general population. Hence, this may affect the generalizability of our findings.

Second, evaluating some of the constraints in the dataset is quite challenging. Many constraints are inherently subjective, e.g., "the story needs to be suited to a nine-year-old"; this may introduce some noise or bias into the evaluation process.

Third, despite our efforts to filter out noise and toxic language, some instances may still remain. These imperfections could introduce unintended biases and complicate the interpretation of LLM performance under realistic conditions.

Finally, our focus in WILDIFEVAL is on the model's ability to satisfy the given constraints, rather than directly evaluating the task itself. However, in many cases, the distinction between a constraint and the actual task is somewhat vague. As a result, during decomposition, some constraints may closely reflect the task itself, ultimately contributing to the final score.

These limitations highlight important areas for future research and emphasize the need for continued refinement in both dataset construction and evaluation methodologies.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Everything we describe in abstract and intro is detailed in the paper. We provide pointers from intro to specific sections which elaborate the claims (lines 34, 45, 48)

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In appendix E we provide a dedicated Limitations section.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: Our paper include empirical experiments, and no theoretical results claimed.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We provide full details about dataset curation process in Section 2.1, and more technical details about model configurations in Appendix B.1. We also provide a link to a github repo which includes all code to reproduce the results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We provide links to WILDIFEVAL in huggingface, as well to our Github repo, at the end of the Abstract.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: In Appendix B.1 we provide the details about how we tested the models, as well as it is documented in our Github repo.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: We provide statistical significance report in Figures 13 and 14.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Detailed in Appendix B.1.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We followed NeurIPS Code of Ethics. We do not hold any information about the users generated the data.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: No societal impact.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We use data that is released under license, all users who contributed to this data consented to it, and there is no personal information about the users.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite the original dataset creators, and provide url to original data in Section 2. We follow their licensing agreement as described in their dataset's card.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide full links to our code and data.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Along the paper we describe all the parts which involved experimenting with LLMs, including for analysis and judgement of results.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.