DOVE: A Large-Scale Multi-Dimensional Predictions Dataset Towards Meaningful LLM Evaluation

Anonymous ACL submission

Abstract

Recent work found that LLMs are sensitive to a wide range of arbitrary prompt dimensions, including the type of delimiters, answer enumerators, instruction wording, and more. This throws into question popular single-prompt evaluation practices. In this work, we present DOVE (Dataset Of Variation Evaluation) a large-scale dataset containing prompt perturbations of various evaluation benchmarks. In contrast to previous work, we examine LLM sensitivity from an *holistic* perspective, and assess the joint effects of perturbations along various dimensions, resulting in thousands of perturbations per instance. We evaluate several model families against DOVE, leading to several findings, including efficient methods for choosing well-performing prompts, observing that few-shot examples reduce sensitivity, and identifying instances which are inherently hard across all perturbations. DOVE consists of more than 300M prompt perturbations and model outputs, which we make publicly available to spur a community-wide effort toward meaningful, robust, and efficient evaluation.

Browse the data - `https://huggingface.co/datasets/DOVevaluation/Dove`[1]

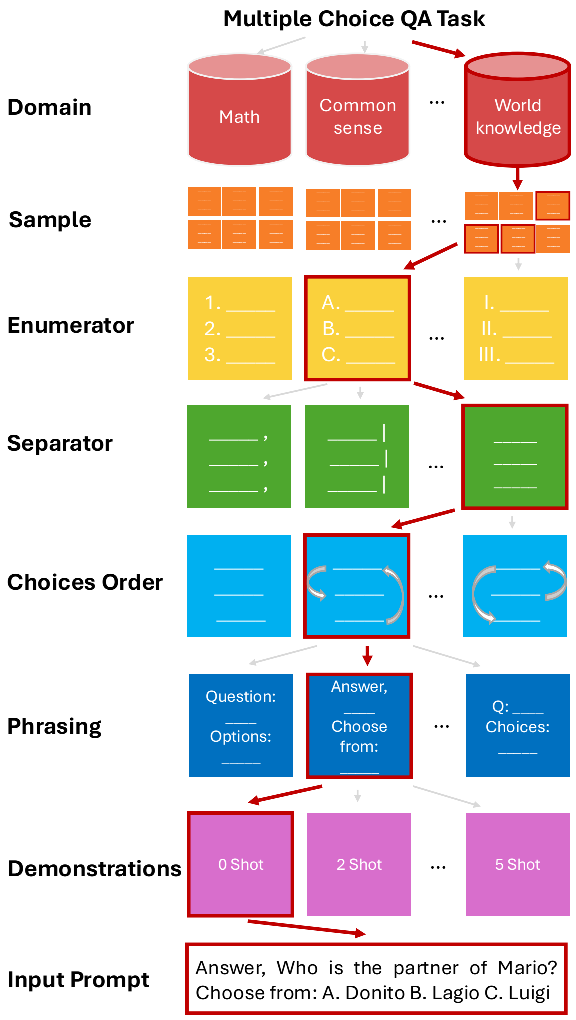

Figure 1: **Building DOVE.** To holistically explore LLM sensitivity, we sample prompts as a walk in the space of various *prompt dimensions* (rows, above).

1 Introduction

Recent years have seen an explosion of LLMs applied in few- or zero-shot settings, where natural language is used for both input and output. Although this free-text format lends itself to various applications, the flexibility in task formulation also leads to large variation in performance.

LLM performance was shown to change drastically based on slight perturbations in arbitrary prompt dimensions, including the number of white spaces (Sclar et al., 2023), answer enumerators and ordering (Alzahrani et al., 2024; Pezeshkpour and Hruschka, 2024), few-shot demonstrations (Lu et al., 2022), and more (Leidinger et al., 2023; Voronov et al., 2024). This sensitivity presents a challenge to meaningful evaluation, exacerbated by the rising cost of inference, which bars large-scale evaluation studies, especially for research groups with small to medium budgets (Perlitz et al., 2024).

[1] Links are fully anonymized for this submission. We will also make all code available upon publication.

Such concurrent findings throw into question the generalizbility of many of the recent evaluation benchmarks, which tend to rely on one arbitrary prompt (Mizrahi et al., 2024). We argue that this constitutes a crisis in evaluation which should be a *community-wide concern*, standing in the way of a better scientific understanding of LLMs, indicating where they excel and where they lack, especially as they are being increasingly deployed in real-world applications (Raiaan et al., 2024).

Our main contribution in this work is the introduction of DOVE, a publicly available large-scale dataset consisting of 300M *model predictions*, which facilitates and democratizes the systematic study of LLM sensitivity and the development of meaningful evaluation protocols.

Starting from popular multiple-choice benchmarks, such as MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), or Race (Lai et al., 2017), we go beyond common evaluation protocols and collect LLM predictions on a *wide range of prompt perturbations*, resulting in thousands of samples per single instance from the original benchmark. For each such instance, DOVE records the full LLM response along with the model's log probabilities and an automatic binary score.

We analyze the performance of various LLMs on DOVE and find that the problems observed at smaller scales persist at this large scale. We find that along various dimensions (prompt phrasing, formatting, and more), performance can vary by more than 10% absolute difference, while model ranking also varies based on these arbitrary choices. These make DOVE a valuable testbed for exploring evaluation and sensitivity at scale.

To demonstrate the kind of analysis permitted by DOVE, we use it to make three novel observations on prompt sensitivity in LLMs, which benefit downstream application and provide a more meaningful evaluation. First, we observe that prompt-tuning the entire prompt is subpar compared to independent dimension-wise tuning; second, we find that adding few-shot demonstrations consistently reduces sensitivity, though it is far from solving the problem; and third, DOVE can be used to find consistently hard instances, which stump models regardless of any prompt selection, thus delineating the real limits to their capabilities.

By making DOVE publicly and openly available, we hope to enable and spur research into meaningful, generalizable, and efficient LLM evaluation, which will help to understand their strengths and limitations. Toward that goal, we plan to make DOVE a collaborative and growing resource and encourage the contribution of data from more diverse domains, applications, and languages.

2 Definitions: Prompt Sensitivity

In this section, we establish terminology, definitions, and metrics for formally quantifying the phenomenon of prompt sensitivity. In this work, we choose to focus on multiple-choice questions to allow for a relatively easy evaluation of model outputs compared to text generation tasks, such as summarization or translation, where the space of correct predictions is vast, and may be considered in future work.

Intent-preserving prompts. Following Chatterjee et al. (2024), two prompts p_1, p_2 are considered *intent-preserving* if they are designed to convey the same underlying meaning, despite differences in phrasing or structure. For example, the two following prompts are considered intent preserving p_1 = *"Who is the partner of Mario? Choose from: A. Donito B. Lagio C. Luigi"*, p_2 = *"Answer the following question: Who is the partner of Mario? A. Donito B. Lagio C. Luigi"*.

Prompt dimensions and linearization. We categorize the differences between intent-preserving prompts along different dimensions, where each dimension D is a set of possible values such that any value from $d \in D$ preserves the intent of the prompt. For example, the *enumerator* dimension may contain values such as {roman, numerals}. Like enumerators, prompt dimensions may be discrete or continuous, e.g., instruction paraphrase. Furthermore, we define *prompt linearization*:

$$T(x, d_1, \dots, d_n) \mapsto p \quad (1)$$

Where x is an underlying question, e.g., *"who is Mario's partner?"*, $d_1 \in D_1, \dots, d_n \in D_n$ are choices made along n prompt dimensions, and T is their deterministic linearization to a prompt p, which can be given as input to an LLM.

Prompt sensitivity. measures the degree to which the performance of an LLM M deviates between intent-preserving prompts. Ideally, the performance of an LLM M should be invariable to different choices along intent-preserving prompt dimensions. Formally, to measure prompt sensitivity on multiple-choice questions, we define a model

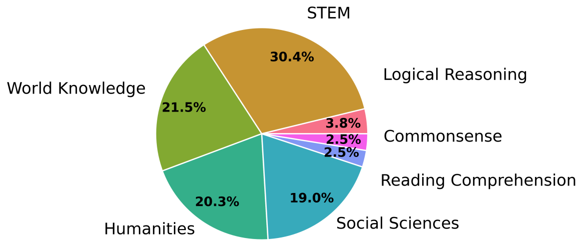

Figure 2: DOVE requires a diverse set of skills.

M's accuracy along different dimension choices $d_1, \ldots, d_n$ in the following manner:

$$Acc(M, Dom, d_1, \ldots, d_n) = \frac{\sum_{(x_i, y_i) \in Dom} \mathbb{1}(M(T(x_i, d_1, \ldots, d_n)) = y_i)}{|Dom|} \tag{2}$$

Where Dom is a dataset consisting of labeled tuples (x_i, y_i) in a certain domain, for example (*who is Mario's partner?, Luigi*). Intuitively, Acc measures the accuracy of M on Dom according to a specific set of choices for the different prompt dimensions. Consequently, we measure *prompt sensitivity* as the difference in accuracy for different dimensions using various statistical measures.

3 DOVE: A Large-Scale Multi-Dimensional Dataset of LLM-Generated Responses Towards Meaningful LLM Evaluation

In this section we introduce DOVE, a large-scale corpus of model predictions along multiple dimensions.

As shown in Figure 1, the building blocks of DOVE are instances from existing popular datasets. For each instance, we create a wide range of intent-preserving prompts, by varying the instances along five dimensions (enumerator, separator, choices order, phrasing, and demonstrations).

Below we discuss the different dimensions, which are also summarized in Table 1. We choose these dimensions based on a survey of recent studies on LLM sensitivity, yet we do not claim that this forms an exhaustive list of factors affecting LLM performance. Future work can expand this with additional dimensions to explore their effect.

Domains. We cover a wide range of data sources, spanning 78 different data sets from MMLU (Hendrycks et al., 2021), MMLU Pro, ARC (Clark et al., 2018), HellaSwag (Zellers et al., 2019), OpenBookQA (Mihaylov et al., 2018), Social IQa (Sap et al., 2019), and RACE (Lai et al., 2017). From each of these, we take 100 instances chosen at random, resulting in 7,800 base instances, which we extend with different perturbations in subsequent steps. Figure 2 shows that solving these samples requires a wide range of skills.

Dimension	Examples	# of Values
Enumerator	Roman, Numerals	6
Separator	;, |	7
Choices Order	original, correct first	6
Phrasing	*The following are multiple-choice questions about {topic }. {question }{choices }Answer:*	13
Demonstrations	Zero-shot, Five-shot	2

Table 1: The different intent-preserving prompt dimensions in DOVE, along with example values, and overall number of values per dimension. The total number of perturbations per sample is the Cartesian product of all values, resulting in over 6.5K perturbations per sample.

Answer enumerators, choice separators and orderings. Recent work has noticed that very subtle changes in the prompt can lead to significant changes in both absolute as well as relative model performance. These include answer enumerators, e.g., roman versus numeral options, choice separators, e.g., new line versus commas, and the order in which the options are presented, e.g., the position of the correct answer (Alzahrani et al., 2024; Pezeshkpour and Hruschka, 2024; Zhou et al., 2024; Gupta et al., 2024). All options are summarized in Table 3 in the Appendix.

Instruction phrasing. Variations in the way instructions are written can significantly influence model behavior (Mizrahi et al., 2024; Chatterjee et al., 2024). To systematically explore this effect, we wrote and verified 13 distinct instruction templates for each of our datasets, drawing inspiration from the format used in established benchmarks like MMLU (Hendrycks et al., 2021) and HELM (Liang et al., 2023), as well as paraphrases from Zhuo et al. (2024) and Mizrahi et al. (2024). See Appendix A.1 for a complete listing of paraphrased instructions.

Demonstrations. We vary the number of few-shot demonstrations, chosen randomly from the training set of each dataset, based on previous work

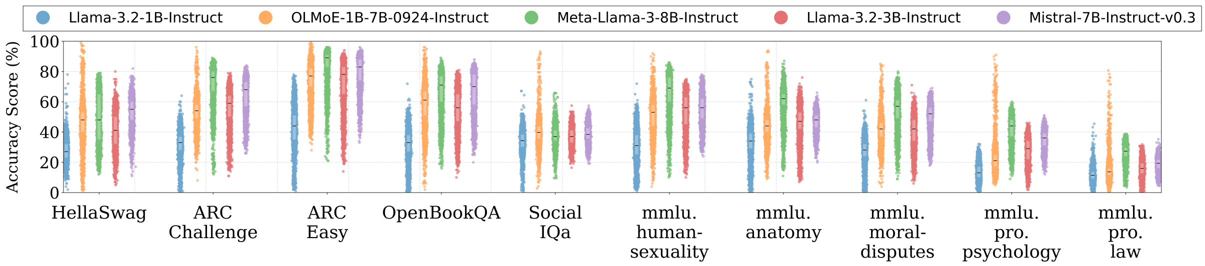

Figure 3: **Performance variations across evaluation datasets.** Each datapoint represents the accuracy of one model calculated across 100 instances. Vertical scatter plots illustrate the variance within each dataset and each model. Model performance varies substantially, indicating persistent prompt sensitivity prompts at large scales.

Field	Description
Hyperparameters	Temperature, top-p
Tokens logprobs	Model's log probability of prompt tokens
Few-shots	Example question-answer pairs
Response	Model's full response to the prompt
Tokens logprobs	Model's log probabilities for generated tokens
Ground truth	The correct answer for the given instance
Evaluation method	Name of method used to evaluate the model's response
Score	Automatic evaluation score

Table 2: **Additional metadata**. Instance-level details available in DOVE to allow future research into their effect, such as the input and output log probabilities assigned by the model.

which found this to be a factor affecting model performance (Zhao et al., 2021; Lu et al., 2022; Kumar and Talukdar, 2021; Reif and Schwartz, 2024).

Additional metadata. Table 2 shows additional instance-level details available in DOVE to allow future research into their effect, such as the input and output log probabilities assigned by the model.

4 Evaluation

In this section, we evaluate various models against DOVE, finding that they all exhibit prompt sensitivity at large scale, also when controlling for most of our tested dimensions.

4.1 Experimental Setup

We evaluate the following model families against DOVE: Llama (1B, 3B, 8B) (Dubey et al., 2024), OLMo (7B) (Muennighoff et al., 2024), and Mistral (7B) (Jiang et al., 2023). We focus on *open-weight* LLMs which we can run locally for two main reasons. First and foremost, API-based chatbots (such as ChatGPT or Claude) alter the prompt in undisclosed ways, for example, to try to ensure that it is safe, or to improve performance (Rao et al., 2024), which may interfere with our findings in a non-trivial manner.

Second, running closed models in such a large scale (60M instances per model) incurs infeasible costs, which do not pay back to the community. However, we note that such sensitivity was observed in closed models (Mizrahi et al., 2024), and we encourage future work to test them on DOVE.

We generate DOVE using vLLM (Kwon et al., 2023) on a cluster of NVIDIA A100 80GB GPUs. In total, dataset creation requires approximately 5,000 GPU hours. For instance, the Mistral-7B model requires 1,189 GPU hours, while other models range from 754 to 1,341 GPU hours each. Overall, creating DOVE on cloud services, such as AWS, costs upwards of $25K, highlighting the high costs of such large scale evaluations.

We extend and use the Unitxt framework (Bandel et al., 2024) to generate and evaluate multiple prompt variations in multiple datasets.

Evaluation metric. To evaluate model outputs we use semantic similarity matching (Mitkov et al., 2009; Obot et al., 2023). For each response, we identify the answer option with highest semantic similarity to the model's output and consider the prediction correct if it matches the ground truth.

4.2 Results: Prompt Sensitivity Persists in Large-Scale Data

Figure 3 depicts model performance on several domains as a distribution across intent-preserving prompts, while similar trends were observed across all other domains (see Appendix B.1). For instance, OLMo's performance on HellaSwag ranges from 1% to 99% based on the prompt. These findings

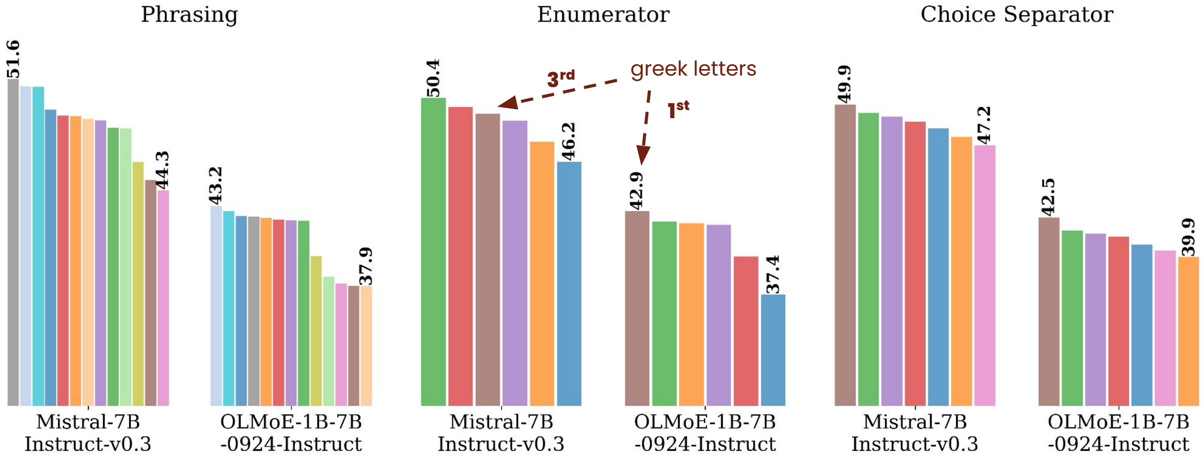

Figure 4: **Accuracy marginalization for different dimensions.** Variations along each of the dimensions in DOVE lead to prompt sensitivity, even when controlling for all other dimensions.

suggest that the dimensions we explore in DOVE indeed play a role in the performance of all LLMs.

To better understand these results, we marginalize each dimension by averaging its performance across all other dimensions. Formally, without loss of generality for each value $d_1 \in D_1$ (for example, the choice of roman numerals), we compute a marginalized accuracy score Acc_{d_1}:

$$Acc_{d_1}(M, Dom) = \sum_{\substack{d_2 \in D_2 \\ \vdots \\ d_n \in D_n}} \frac{Acc(M, Dom, d_1, \ldots, d_n)}{|D_2| \cdot \ldots \cdot |D_n|} \tag{3}$$

Where $D_1, \ldots, D_n$ are the different dimensions, and $Acc(\cdot)$ is according to Equation 2.

The results, depicted in Figure 4 show that variation along each individual dimension changes results substantially. For instance, for Mistral, different paraphrases lead to an 8% difference in accuracy. Beyond absolute performance differences, we also observe varying preferences across models to different prompt variations. For example, OLMoE performs best with greek numerals, achieving the highest average accuracy across the dataset with this choice. On the other hand, Mistral rank greek numerals only as the third best option, performing less than both capital and lateen numerals. This discrepancy underscores that models demonstrate distinct prompt preferences.

Statistical significance. Following (Mizrahi et al., 2024), we quantified performance variance by calculating divergence scores, defined as the number of standard deviations by which performance using the original prompt deviates from the mean performance across all prompts. Figure 5 shows significant divergence in randomly sampled domains from the MMLU (Hendrycks et al., 2021), where divergence is defined as exceeding one standard deviation (Kazmier et al., 2003). For Instance, Mistral's performance with original prompts exceeds its mean performance by more than one standard deviation in 35 of 57 domain tasks (complete results can be found in Figure 12 in Appendix B.3)

	mmlu.high-school-chemistry	mmlu.high-school-statistics	mmlu.law-international	mmlu.moral-disputes	mmlu.professional-psychology
OLMoE-1B-7B-0924-Instruct	1.50	0.40	0.71	1.00	1.40
Llama-3.2-1B-Instruct	0.71	3.33	-0.27	-1.25	0.25
Llama-3.2-3B-Instruct	0.33	1.86	0.86	0.50	0.00
Llama-3-8B-Instruct	-0.75	2.25	0.25	0.50	1.00
Mistral-7B-Instruct-v0.3	-0.25	-0.25	1.00	1.00	1.00

Figure 5: **Substantial performance differences across prompt variations.** The number of standard deviations by which model performance on original instructions deviates from average across few-shot prompts. Dark cells show substantial divergence.

5 Analysis

So far, we made use of DOVE to quantify the effect of prompt sensitivity in large scale, finding that each of the individual prompt dimensions further contributes to this sensitivity. In this section, we discuss three observations that stem from this large-scale analysis and have practical implications for downstream applications and for more generalizable and meaningful evaluation.

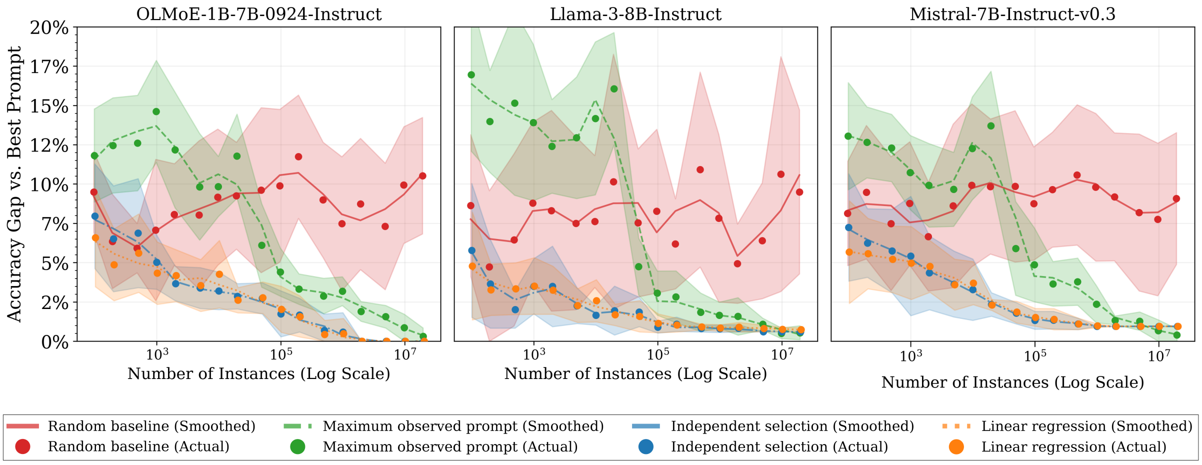

Figure 6: **Efficient prompt selection approaches can improve perfromance.** Performance gap from the ground truth prompt (y-axis) versus sample count (x-axis) for LLMs and selection methods. Results demonstrate that efficient prompt selection methods can improve performance with relatively small sample sizes, outperforming random selection.

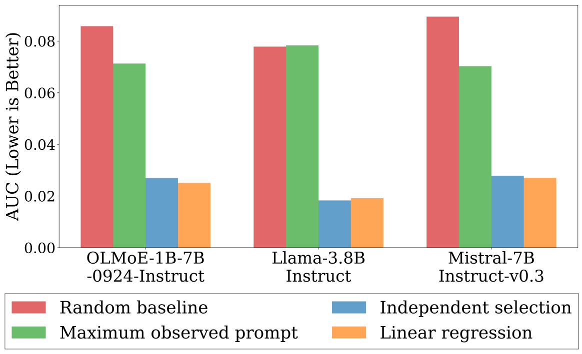

Figure 7: **Prompt selection methods outperform random and best observed baseline.** AUC comparison of prompt selection methods across different LLMs. The lower AUC values indicate better overall performance across sample sizes of selection methods over random baseline.

5.1 Efficient Prompt Selection

We use DOVE to answer the following question: *How should the values for the different dimensions be chosen to optimize performance, given a fixed inference budget?* This is a practical question whose answer can benefit downstream applications in various real-world scenarios.

Given a set of all possible prompts C and a limited sampling budget m, DOVE allows us to explore how to efficiently identify prompts that are likely to yield good performance. This question has actionable practical implications, as evaluating all possible prompts is computationally prohibitive.

We leverage DOVE to simulate different sampling scenarios, focusing on zero-shot settings. For each model, we establish ground truth by finding the prompt $c^* \in C$ that maximizes performance across our complete dataset. We then investigate how different selection methods perform with limited number of samples.

In particular, we explore four strategies for choosing a prompt based on a set of observations: (1) *independent selection:* chooses the best observed value for each dimension, marginalizing all other dimensions; (2) *linear regression:* we train a linear regression on the observed samples which aims to predict accuracy from the set of discrete observed values for each dimension; (3) *maximum observed prompt:* chooses the values for all dimensions according to the best performing prompt in the observed set; and (4) *random baseline:* chooses the values for all dimensions at random.

Figure 6 shows the accuracy of the different approaches along various data sizes, reporting for each the mean accuracy as well as its standard deviation across 10 random seeds, while Figure 7 shows the area under the graph for each of the the different approaches (See Appendix B.4 for similar results across all models).

It is evident that different prompt selection approaches can lead to vastly different results. Interestingly, choosing the values for the different dimensions in an independent manner achieves performance on par with linear regression, and performs better than choosing the best observed performance. Choosing the best observed prompt becomes reliable with more data, but only after observing tens of millions of samples.

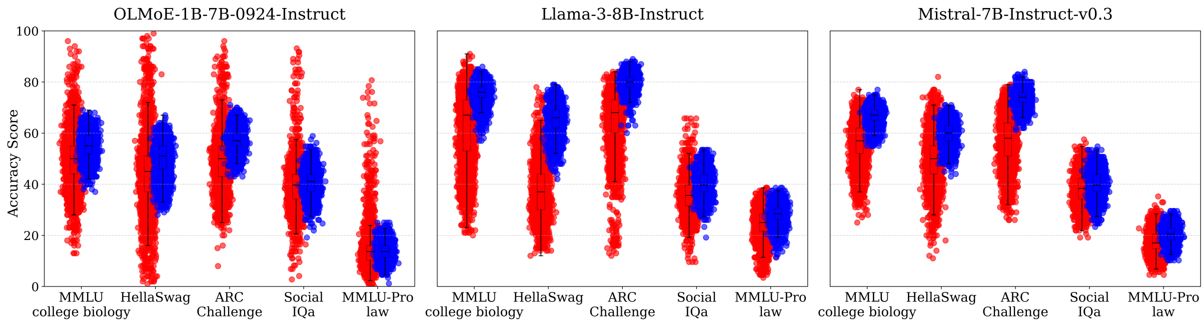

Figure 8: **Few-shot reduces performance variance across evaluation dimensions**. Comparing zero-shot and five-shot on a subset of domains from DOVE reveals a narrower spread of accuracy scores. Each point represents the accuracy across 100 instances, demonstrating that the five-shot demonstrations lead to more robust performance.

5.2 Few-Shot Demonstrations Consistently Reduce Sensitivity

Figure 8 depicts the performance of prompts with few-shot demonstrations versus zero-shot prompts. We find that few-shot demonstrations consistently lead to more robust performance (see Appendix B.2 results across all domains).

Still, few-shot demonstrations are far from completely mitigating all sensitivity. Even with demonstrations we see a wide range of scores, e.g., above 20% for all datasets in Figure 8. Furthermore, their effect is sometimes minimal, for example, in Social IQa and in the legal domain of MMLU-Pro.

From a practical perspective, these results suggest that few-shot examples should be added where possible to mitigate the sensitivity of current LLMs.

5.3 Some Examples are Consistently Easy or Hard for Models

We use DOVE to perform an instance-level analysis. Figure 9 categorizes each sample according to its *success rate*, which we calculate as the percentage of prompt perturbations for which the model outputs the correct answer, of all the perturbations for that sample. The lower ends of the spectrum, marked in red, count instances for which the model errs on all prompt perturbations, whereas on the higher end of the spectrum are samples for which the models succeds on all prompt perturbations.

These results suggest a novel definition for what constitutes inherently hard instances for models, namely where they fail on all possible prompt perturbations for the same instance. Moreover, on either of these extreme ends, models are in fact *less* sensitive, as they consistently succeeded or err on all prompt perturbations.

6 Related Work

Many studies which we have leveraged extensively throughout this work have focused on individual prompt dimensions, examining variations in instruction wording (Mizrahi et al., 2024; Leidinger et al., 2023; Sclar et al., 2023), answer ordering (Gupta et al., 2024; Wang et al., 2024), input perplexity (Gonen et al., 2023), few-shot example selection (Reif and Schwartz, 2024; Lu et al., 2022), and answer enumeration styles (Alzahrani et al., 2024). Some works propose metrics for prompt sensitivity, such as POSIX (Chatterjee et al., 2024), which measures log-likelihood shifts, and ProSA (Zhuo et al., 2024), which uses decoding confidence. Although these methods quantify sensitivity, they do not examine interactions between multiple perturbations, nor do they collect data and make observations at a large scale.

Several recent work have noted that similarly to our findings, few-shot examples help improve performance (Webson and Pavlick, 2022; Perez et al., 2021). In contrast to these works, we show the effect that few-shot examples have on reducing prompt sensitivity.

Beyond investigating individual factors, several notable frameworks aim to standardize and improve evaluation process. HELM (Liang et al., 2023) takes a broad view of LLM performance by creating a taxonomy of a wide range of use cases and evaluation metrics, but was not designed to examine prompt sensitivity. OLMES (Gu et al., 2024) establishes detailed protocols for the reproducibility of the evaluation, carefully specifying aspects such as prompt formatting. OLMES demonstrated that standardizing these procedures could lead to more consistent results but may inadvertently harm models which do not perform well on its specific

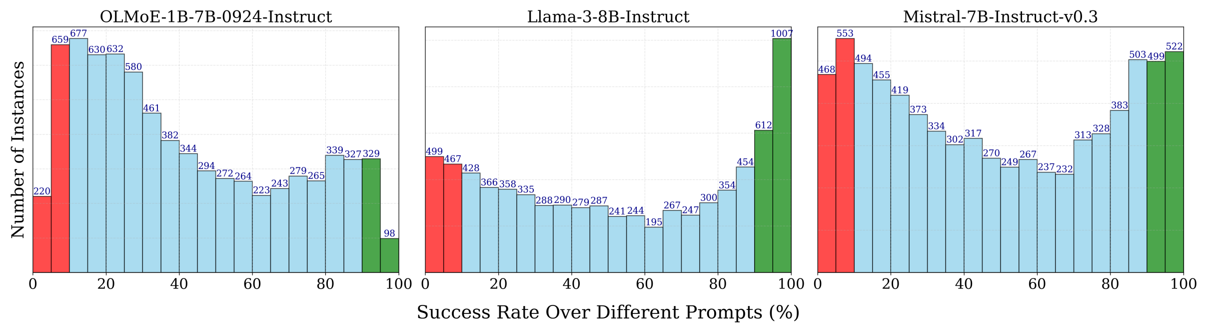

Figure 9: **Success rate distribution reveals inherent example difficulty patterns.** Distribution of success rates by evaluation dimension and model. The x-axis shows the percentage of successful perturbations per instance, while the y-axis shows the instance count in DOVE. The distribution reveals examples that are consistently easy or difficult for LLMs across prompt dimensions.

dimension choices.

Although these studies have provided valuable insights, our work is the first to take a holistic view of the problem. This large-scale dataset, encompassing more than 300M model predictions, allows us to aggregate across multiple prompt dimensions, noticing practical patterns, and opening the door for many future research directions.

7 Future Work

DOVE provides a foundation for exploring LLM evaluation and sensitivity. The dataset's broad coverage enables flexible partitioning for granular error analysis, targeted evaluations, and investigations of specific dimensions. Future research directions include understanding model biases, improving evaluation methodologies, and refining confidence estimation.

Task-level sensitivity: Do some model capabilities have distinct sensitivity patterns? For example, is factual retrieval more fragile than logical reasoning? Do format biases manifest differently across tasks from different domains?

Alternative evaluation measures: Do less common approaches, like perplexity-based evaluation or sensitivity-aware assessments, better mitigate prompt sensitivity in benchmarks (Gonen et al., 2023)? Do past prediction data help predict the most effective evaluation method for a new benchmark (Polo et al., 2024; Maia Polo et al., 2025)?

Optimizing evaluation focus: Given resource constraints, what dimensions are most critical for assessing model performance? Can a predictive framework identify the relative importance of different dimensions?

Instance characterization: What distinguishes consistently answered examples from those with high variability, e.g., as expemlified by the two ends of the spectrum in Figure 9? Do specific linguistic, semantic, or structural features influence susceptibility to example variation?

Uncertainty quantification: How do token-level log probabilities relate to model consistency? Can their distributions help predict or explain model sensitivity better than accuracy scores? Towards that goal DOVE also records all model log probabilities.

Future versions of DOVE: We plan to expand DOVE through both our team's ongoing efforts and community contributions. To facilitate community contributions to DOVE, we will release tools and documentation to expand coverage across domains, languages, and tasks. We particularly welcome contributions that extend coverage to specialized domains and tasks.

8 Conclusions

We introduced DOVE, a large-scale dataset of 300M model predictions across prompt dimensions. Our analysis revealed prompt sensitivity remains a significant challenge, with performance varying over 10% across different prompt variations. Key findings showed dimension-wise tuning outperforms entire-prompt optimization, few-shot demonstrations reduce but do not eliminate sensitivity, and certain examples remain challenging across all prompt variations. The public release of DOVE aims to democratize evaluation research and enable development of robust protocols for assessing LLM capabilities.

9 Limitations

While DOVE provides valuable insights into LLM evaluation, several limitations should be acknowledged. Our focus on multiple-choice questions, while enabling controlled study of prompt variations, does not capture the full complexity of open-ended generation tasks. However, multiple-choice questions remain a fundamental benchmark in the field, with most models reporting results on such tasks. Though we explore various prompt dimensions including paraphrasing, enumeration, and ordering based on prior work, the exponential space of possible variations necessitates a selection of dimensions and values. We plan to systematically expand these dimensions based on analyses of the current version. Additionally, despite its scale, DOVE is currently constrained in terms of model diversity and language coverage, and we plan to expand to additional languages and domains in the next version. Large-scale prompt variations computational costs constrain update frequency. We welcome community contributions to expand the DOVE scope.

References

Norah Alzahrani, Hisham Alyahya, Yazeed Alnumay, Sultan AlRashed, Shaykhah Alsubaie, Yousef Almushayqih, Faisal Mirza, Nouf Alotaibi, Nora Al-Twairesh, Areeb Alowisheq, M Saiful Bari, and Haidar Khan. 2024. When benchmarks are targets: Revealing the sensitivity of large language model leaderboards. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13787–13805, Bangkok, Thailand. Association for Computational Linguistics.

Elron Bandel, Yotam Perlitz, Elad Venezian, Roni Friedman, Ofir Arviv, Matan Orbach, Shachar Don-Yehiya, Dafna Sheinwald, Ariel Gera, Leshem Choshen, Michal Shmueli-Scheuer, and Yoav Katz. 2024. Unitxt: Flexible, shareable and reusable data preparation and evaluation for generative AI. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: System Demonstrations)*, pages 207–215, Mexico City, Mexico. Association for Computational Linguistics.

Anwoy Chatterjee, S Hsvn, Kowndinya Renduchintala, Sumit Kaur Bhatia, and Tanmoy Chakraborty. 2024. Posix: A prompt sensitivity index for large language models. In *Conference on Empirical Methods in Natural Language Processing*.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the AI2 reasoning challenge. *ArXiv preprint*, abs/1803.05457.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Hila Gonen, Srini Iyer, Terra Blevins, Noah Smith, and Luke Zettlemoyer. 2023. Demystifying prompts in language models via perplexity estimation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10136–10148, Singapore. Association for Computational Linguistics.

Yuling Gu, Oyvind Tafjord, Bailey Kuehl, Dany Haddad, Jesse Dodge, and Hannaneh Hajishirzi. 2024. Olmes: A standard for language model evaluations. *ArXiv preprint*, abs/2406.08446.

Vipul Gupta, David Pantoja, Candace Ross, Adina Williams, and Megan Ung. 2024. Changing answer order can decrease mmlu accuracy.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Leonard J Kazmier, Michael K Staton, Daniel L Fulks, and 1 others. 2003. Business statistics: based on schaums outline of theory and problems of business statistics, by leonard j. kazmier. Technical report, McGraw-Hill.

Sawan Kumar and Partha Talukdar. 2021. Reordering examples helps during priming-based few-shot learning. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4507–4518, Online. Association for Computational Linguistics.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention.

Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. RACE: Large-scale ReAding comprehension dataset from examinations. In

Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Alexander Cosgrove, Christopher D Manning, Christopher Re, Diana Acosta-Navas, Drew Arad Hudson, and 31 others. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*. Featured Certification, Expert Certification.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson de Oliveira, Yuekai Sun, and Mikhail Yurochkin. 2025. Efficient multi-prompt evaluation of llms. *Advances in Neural Information Processing Systems*, 37:22483–22512.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Ruslan Mitkov, Le An Ha, Andrea Varga, and Luz Rello. 2009. Semantic similarity of distractors in multiple-choice tests: Extrinsic evaluation. In *Proceedings of the Workshop on Geometrical Models of Natural Language Semantics*, pages 49–56, Athens, Greece. Association for Computational Linguistics.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Niklas Muennighoff, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Jacob Morrison, Sewon Min, Weijia Shi, Pete Walsh, Oyvind Tafjord, Nathan Lambert, Yuling Gu, Shane Arora, Akshita Bhagia, Dustin Schwenk, David Wadden, Alexander Wettig, Binyuan Hui, Tim Dettmers, Douwe Kiela, and 5 others. 2024. Olmoe: Open mixture-of-experts language models.

Okure U Obot, Gregory O Onwodi, Kingsley F Attai, Anietie E John, Etiese Wilson, and 1 others. 2023. Grading multiple choice questions based on similarity measure. *Journal of Computer Science and Information Technology*, 11(1):9–21.

Ethan Perez, Douwe Kiela, and Kyunghyun Cho. 2021. True few-shot learning with language models. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 11054–11070.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. 2024. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2519–2536, Mexico City, Mexico. Association for Computational Linguistics.

Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2006–2017, Mexico City, Mexico. Association for Computational Linguistics.

Felipe Maia Polo, Seamus Somerstep, Leshem Choshen, Yuekai Sun, and Mikhail Yurochkin. 2024. Sloth: scaling laws for llm skills to predict multi-benchmark performance across families. *ArXiv preprint*, abs/2412.06540.

Mohaimenul Azam Khan Raiaan, Md. Saddam Hossain Mukta, Kaniz Fatema, Nur Mohammad Fahad, Sadman Sakib, Most Marufatul Jannat Mim, Jubaer Ahmad, Mohammed Eunus Ali, and Sami Azam. 2024. A review on large language models: Architectures, applications, taxonomies, open issues and challenges. *IEEE Access*, 12:26839–26874.

Abhinav Sukumar Rao, Atharva Roshan Naik, Sachin Vashistha, Somak Aditya, and Monojit Choudhury. 2024. Tricking LLMs into disobedience: Formalizing, analyzing, and detecting jailbreaks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 16802–16830, Torino, Italia. ELRA and ICCL.

Yuval Reif and Roy Schwartz. 2024. Beyond performance: Quantifying and mitigating label bias in LLMs. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6784–6798, Mexico City, Mexico. Association for Computational Linguistics.

Maarten Sap, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi. 2019. Social IQa: Commonsense reasoning about social interactions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4463–4473, Hong Kong, China. Association for Computational Linguistics.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6287–6310, Bangkok, Thailand. Association for Computational Linguistics.

Hao Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu. 2024. Llms may perform mcqa by selecting the least incorrect option.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2300–2344, Seattle, United States. Association for Computational Linguistics.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

Wenjie Zhou, Qiang Wang, Mingzhou Xu, Ming Chen, and Xiangyu Duan. 2024. Revisiting the self-consistency challenges in multi-choice question formats for large language model evaluation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14103–14110, Torino, Italia. ELRA and ICCL.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms.

A Prompt Dimensions Values

The prompt dimensions values are listed in Table 3.

<table>
<tr><th>Dimension</th><th>Possible Values</th></tr>
<tr><td rowspan="6">Enumerator</td><td>“A, B, C, D..” (Capitals)</td></tr>
<tr><td>“a, b, c, d...” (Lowercase)</td></tr>
<tr><td>“1, 2, 3, 4...” (Numbers)</td></tr>
<tr><td>“I, II, III, IV...” (Roman numerals)</td></tr>
<tr><td>“$! @ # % ^...” (Keyboard symbols)</td></tr>
<tr><td>“α, β, γ, δ” (Greek letters)</td></tr>
<tr><td rowspan="7">Choice Separator</td><td><code>"\s"</code> (Space)</td></tr>
<tr><td><code>"\n"</code> (Newline)</td></tr>
<tr><td><code>", "</code></td></tr>
<tr><td><code>"; "</code></td></tr>
<tr><td><code>" | "</code></td></tr>
<tr><td><code>" OR "</code></td></tr>
<tr><td><code>" or "</code></td></tr>
<tr><td rowspan="7">Choices Order</td><td>Keep original order</td></tr>
<tr><td>Sort by length (ascending)</td></tr>
<tr><td>Sort by length (descending)</td></tr>
<tr><td>Sort alphabetically (ascending)</td></tr>
<tr><td>Sort alphabetically (descending)</td></tr>
<tr><td>Force correct choice at first index</td></tr>
<tr><td>Force correct choice at last index</td></tr>
</table>

Table 3: **Prompt Formatting Dimensions**. *Prompt Formatting Dimensions.* We systematically vary these three axes when creating prompts. *Enumerator* controls how answer options are labeled, *Choice Separator* determines how answer options are delimited, and *Choices Order* rearranges (or fixes) the position of the correct choice position.

A.1 Instruction Phrasing Options

We present the collection of instruction phrasings that vary in their structure and formality used in DOVE.

```
The following are multiple choice questions (with answers) about {topic}.

{question}
{choices}
Answer:
```

```
The following are multiple choice questions (with answers).

{question}

{choices}
Answer:
```

```
The following are multiple choice questions (with answers) about {topic}.

Question: {question}
{choices}
Answer:
```

```
The following are multiple choice questions (with answers).

Question: {question}

{choices}
Answer:
```

```
Question: {question}

Choices: {choices}
Answer:
```

```
Topic: {topic}
Question: [question] Choices: [choices] Answer: [answer]
Question: {question} Choices: {choices} Answer:
```

```
Question: [question] Choices: [choices] Answer: [answer]
Question: {question} Choices: {choices} Answer:
```

```
Please answer the following question:
{question}
{choices}
Answer:
```

```
Please address the following question:
{question}
{choices}
Answer:
```

```
Could you provide a response to the following question:
{question}
{choices}
Answer:
```

```
Here are some multiple choice questions along with their answers about {
    topic}.

Question: {question}
Choices: {choices}
Correct Answer:
```

```
Below are multiple-choice questions related to {topic}, each followed by
    their respective answers.

Question: {question}
Choices: {choices}
Correct Answer:
```

```
Below are multiple-choice questions related to {topic}. Please provide
    the correct answer for each question.

Question: {question}
Choices: {choices}
Answer:
```

B Extended Results

B.1 Performance Analysis Across All Domains

Figure 10 reveals consistent patterns in prompt sensitivity across our evaluation domains.

B.2 Analysis of Few-Shot Impact Across All Domains

The impact of few-shot demonstrations on reducing prompt sensitivity becomes evident across domains, as illustrated in Figure 11.

B.3 Divergence Across All Domains

Standard deviation measurements shown in Figure 12 highlight the varying model responses to different instruction prompts.

B.4 Selection Methods Across All Models

Our comparison of prompt selection methods spans both AUC analysis and success rate distributions. These results are detailed in Figure 13 and Figure 14.

B.5 Examples are Consistently Easy or Hard Across All Models

Task difficulty follows consistent patterns across different models, with success rate distributions mapped in Figure 15 and Figure 16.

C Dataset Scheme

Table 4 details the components and structure of our dataset, providing descriptions and example values for each field

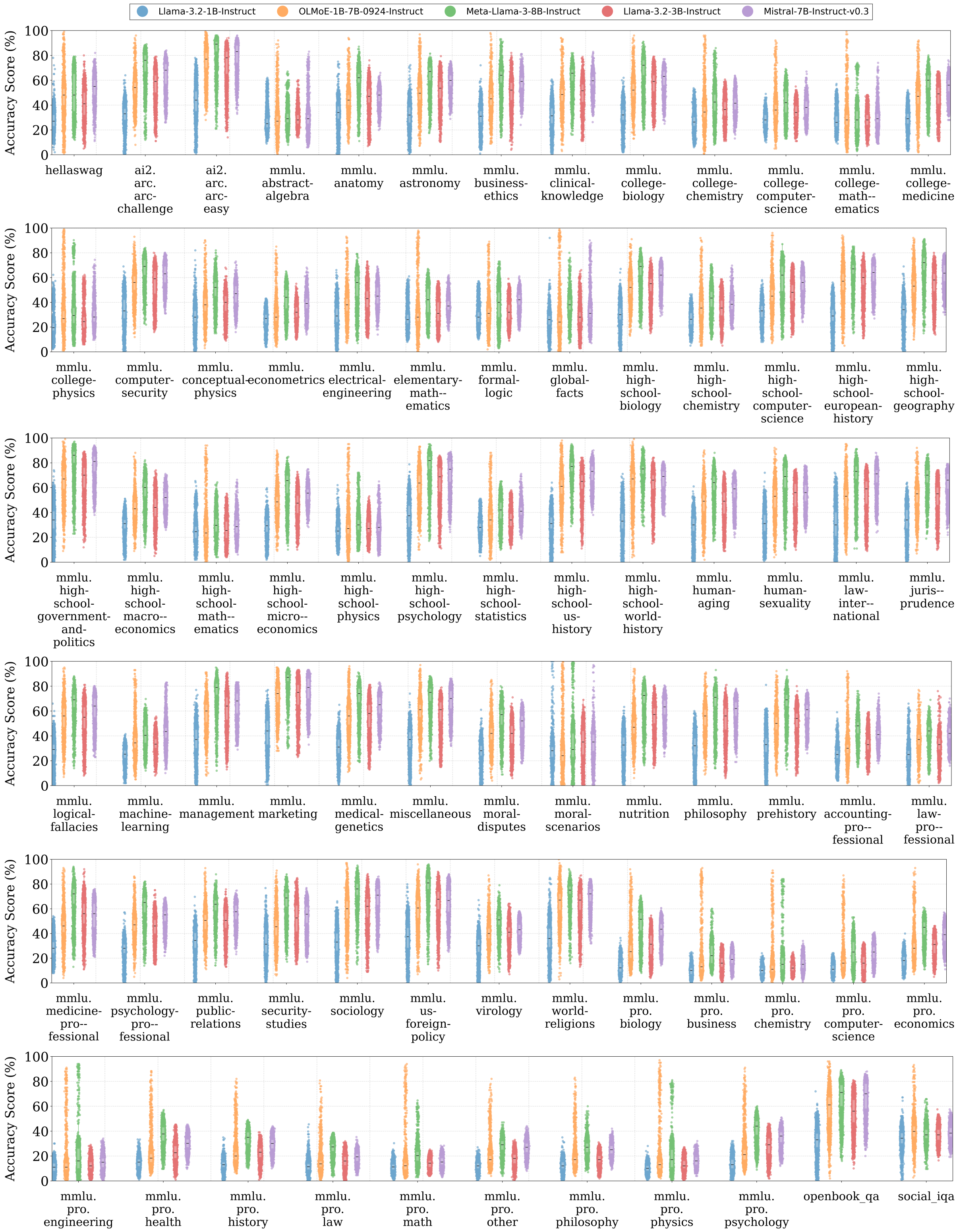

Figure 10: Performance variations across all evaluation domains. Extended analysis showing consistent prompt sensitivity patterns across the complete set of domains

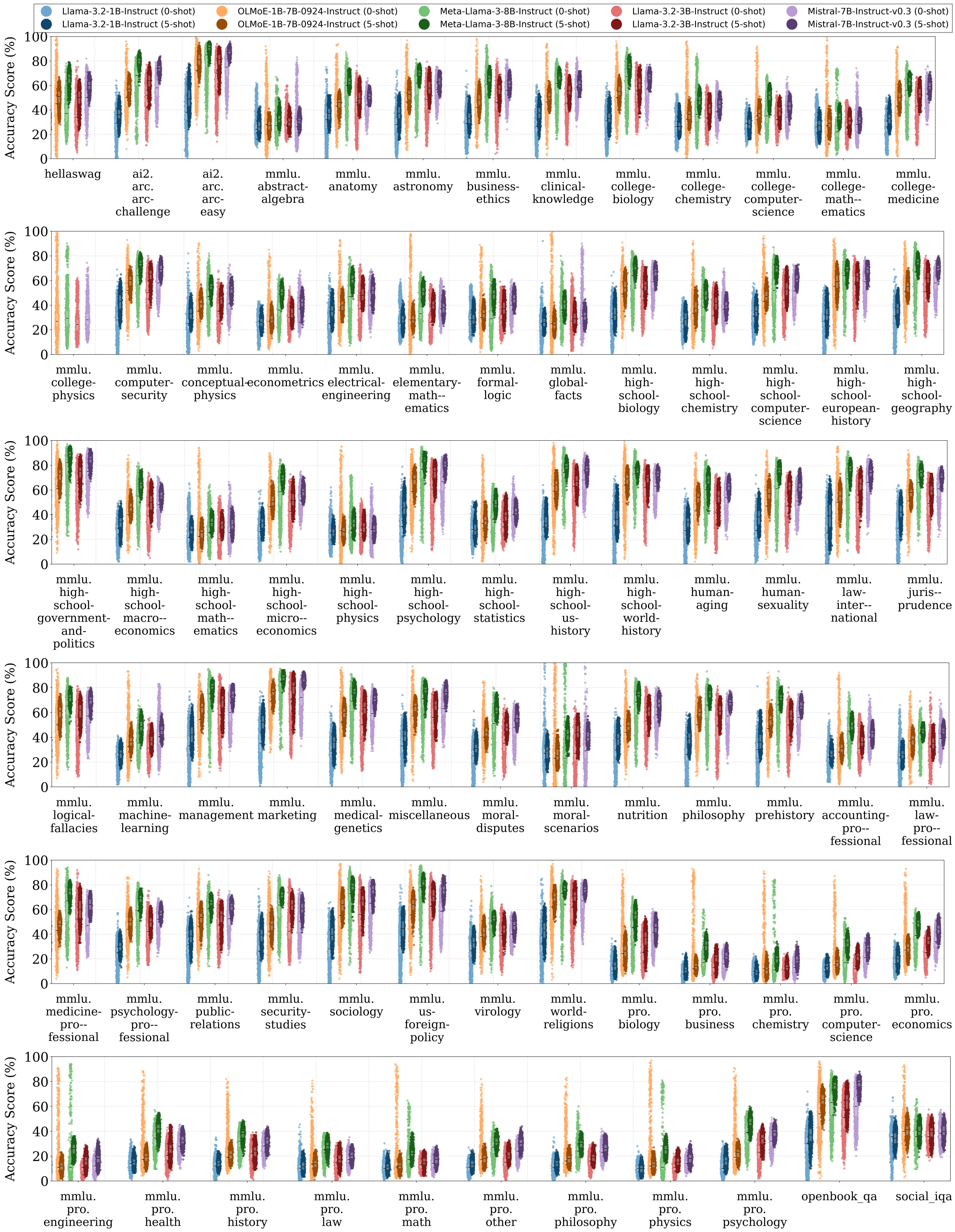

Figure 11: Few-shot versus zero-shot performance across all domains. Extended analysis showing consistent reduction in sensitivity with few-shot demonstrations

	OLMoE-1B-7B-0924-Instruct	Llama-3.2-1B-Instruct	Llama-3.2-3B-Instruct	Llama-3-8B-Instruct	Mistral-7B-Instruct-v0.3
mmlu.abstract_algebra	-0.50	-0.80	-2.00	-0.50	-1.67
mmlu.anatomy	1.25	-0.67	-0.50	-0.75	0.33
mmlu.astronomy	1.80	0.29	-0.50	0.00	-0.25
mmlu.business_ethics	1.20	-0.50	-0.60	0.25	-0.75
mmlu.clinical_knowledge	-0.25	1.17	0.40	0.00	-0.33
mmlu.college_biology	1.20	1.17	0.43	0.25	1.00
mmlu.college_chemistry	1.00	3.40	2.00	-0.25	-0.50
mmlu.college_computer_science	2.00	1.40	1.00	1.00	0.25
mmlu.college_mathematics	1.60	0.83	0.00	1.00	1.75
mmlu.college_medicine	-0.40	2.00	1.60	0.50	-0.33
mmlu.computer_security	2.40	0.71	-0.83	0.33	0.25
mmlu.conceptual_physics	2.00	-0.40	1.40	-0.75	0.75
mmlu.econometrics	0.00	-0.25	0.60	0.75	1.25
mmlu.electrical_engineering	1.00	-0.50	-1.17	0.75	1.20
mmlu.elementary_mathematics	0.25	-0.50	0.00	0.75	-1.60
mmlu.formal_logic	-1.50	2.83	2.00	1.40	-0.75
mmlu.global_facts	1.25	-1.20	-1.40	0.80	0.00
mmlu.high_school_biology	1.17	0.43	0.40	1.00	1.25
mmlu.high_school_chemistry	1.50	0.71	0.33	-0.75	-0.25
mmlu.high_school_computer_science	1.00	-1.67	-0.57	0.50	1.25
mmlu.high_school_european_history	0.86	-0.71	0.20	0.00	1.50
mmlu.high_school_geography	0.60	1.17	1.67	1.33	0.67
mmlu.high_school_government_and_politics	0.67	0.33	0.71	0.75	2.00
mmlu.high_school_macroeconomics	2.25	1.67	0.57	1.25	0.00
mmlu.high_school_mathematics	0.50	0.20	-1.75	1.00	1.20
mmlu.high_school_microeconomics	1.00	1.40	1.33	1.67	0.00
mmlu.high_school_physics	1.50	1.00	0.20	-0.20	-0.50
mmlu.high_school_psychology	1.35	0.21	0.46	1.29	0.61
mmlu.high_school_statistics	0.40	3.33	1.86	2.25	-0.25
mmlu.high_school_us_history	1.33	-0.50	0.33	1.00	0.25
mmlu.high_school_world_history	1.00	-0.44	0.40	2.00	-0.33
mmlu.human_aging	1.00	-0.83	-0.62	1.00	1.25
mmlu.human_sexuality	1.60	0.00	0.80	1.33	2.00
mmlu.international_law	0.71	-0.27	0.86	0.25	1.00
mmlu.jurisprudence	-0.50	-0.86	-0.33	0.33	0.67
mmlu.logical_fallacies	1.17	-0.17	-0.29	-0.67	-0.25
mmlu.machine_learning	2.25	-2.25	-1.25	0.00	1.25
mmlu.management	0.60	0.75	1.00	1.00	0.75
mmlu.marketing	1.50	-0.71	0.00	1.00	2.00
mmlu.medical_genetics	1.40	0.00	-0.14	-0.67	0.33
mmlu.miscellaneous	1.60	-0.33	0.43	0.50	0.25
mmlu.moral_disputes	1.00	-1.25	0.50	0.50	1.00
mmlu.moral_scenarios	-0.83	1.50	0.14	0.40	-1.25
mmlu.nutrition	1.50	-0.14	0.40	1.00	0.80
mmlu.philosophy	0.50	1.17	0.20	0.33	-1.00
mmlu.prehistory	0.40	-1.25	-0.60	0.50	2.33
mmlu.professional_accounting	-0.25	1.00	-0.40	-0.25	-1.00
mmlu.professional_law	1.75	0.25	0.00	0.00	0.75
mmlu.professional_medicine	1.25	1.57	1.00	1.50	-1.20
mmlu.professional_psychology	1.40	0.25	0.00	1.00	1.00
mmlu.public_relations	0.71	-1.10	0.37	1.43	0.91
mmlu.security_studies	0.33	0.71	-1.20	-1.00	0.60
mmlu.sociology	1.33	0.43	0.67	1.75	1.25
mmlu.us_foreign_policy	2.35	-0.20	-0.39	1.29	1.16
mmlu.virology	0.40	-1.00	-0.40	0.67	1.25
mmlu.world_religions	2.50	0.88	0.33	1.50	0.33

Figure 12: Model performance variations across different instruction prompts (shown in standard deviations).

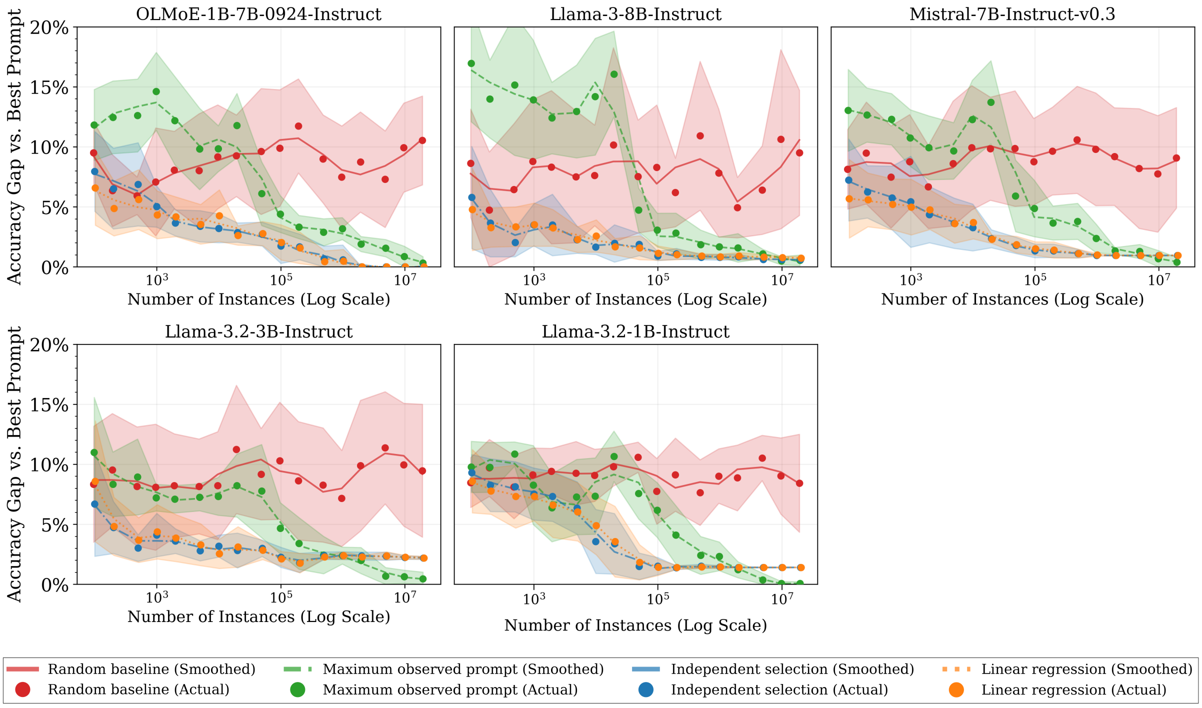

Figure 13: Success rate distributions across all models

Component	Field	Description	Example Values
ID	Evaluation ID	Unique identifier for the evaluation run	`f8544...2240`
Model	Name	Model identifier and version	`Mistral-7B-Instruct-v0.3`
	Metadata	Architecture, Size, Context window, Instruction tuning	`transformer, 7B, 32768, True`
	Quantization	Bit precision and method settings for model inference	`float16, none`
	Generation Args	Generation control settings	`temperature:null, top_p:null, top_k: -1`
Instance	Task Type	Type of evaluation task	`classification, generation`
	Raw Input	Original input from the dataset (before formatting)	`"What size of cannula would you use..."`
	Tokens Logprobs	Log probability of prompt tokens	`[token_index:153, logprob:-0.96, rank:1, decoded_token:"Question", ...]`
	Sample Identifier	Dataset source details, including split and index	`mmlu.clinical_knowledge, test, 487`
	Language	Language of the input	`en, fr, ar, zh`
	Classification Fields	Classification details: question, choices, answer	`question, choices, gt`
	Perplexity	Complexity score of the input text	`20.12`
Prompt (MCQ)	Template	Prompt structure with placeholders	`"Below are multiple-choice questions..."`
	Separator	Character(s) used to separate multiple-choice options	`"\s", "\n", ", ", " | ", " OR ", " or "`
	Enumerator	Choice enumeration style	`"ABCD", "abcd", "1234", "I,II,III,IV", "!@#$",` "$\alpha\beta\gamma\delta$"
	Choices Order	Method for ordering answer choices	`"original order, by length, alphabetical, correct first/last"`
	Shots	Number of examples included in the prompt	`zero, two, five`
	Few-Shot Examples	Example question-answer pairs	`question, choices, answers`
Output	Response	Model's full response to the prompt	`"The size depends on a number of factors..."`
	Tokens Logprobs	Log probabilities for generated tokens	`[token_index:1183, logprob:-2.73, rank:4, decoded_token:"The", ...]`
	Cumulative Logprob	Log probability of the entire generated sequence	`-49.28`
Evaluation	Ground Truth	The correct answer for the given instance	`"IV. 18 gauge."`
	Evaluation Method	Method used to evaluate the model's response	`label_only_match, content_similarity`
	Score	Binary score indicating correctness	`1`

Table 4: Dataset Schema Components, Descriptions, and Example Values

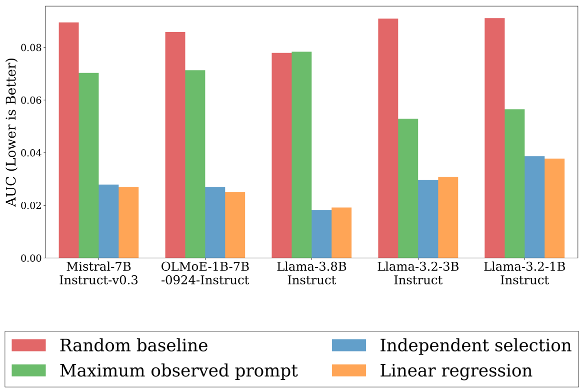

Figure 14: AUC comparison of prompt selection methods across all models

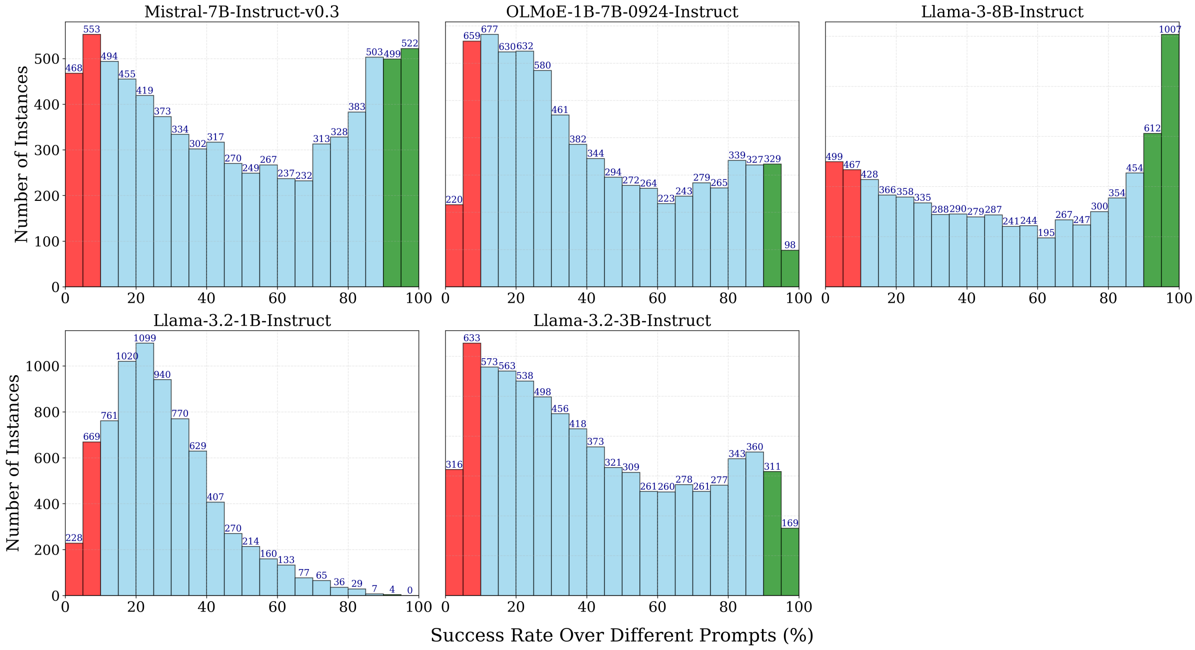

Figure 15: Success rate distribution reveals inherent example difficulty patterns

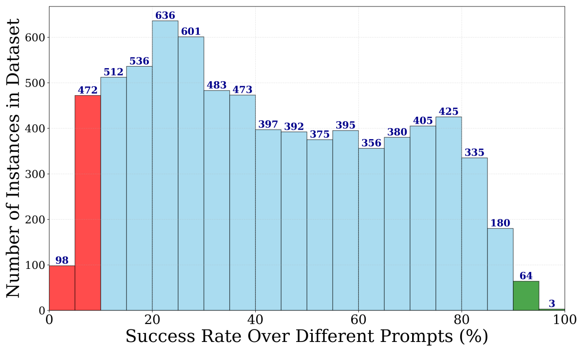

Figure 16: Success rate distribution reveals inherent example difficulty patterns