A Novel Metric for Measuring the Robustness of Large Language Models in Non-adversarial Scenarios

Samuel Ackerman **Ella Rabinovich** **Eitan Farchi** **Ateret Anaby-Tavor**
IBM Research
{samuel.ackerman, ella.rabinovich1}@ibm.com
{farchi, atereta}@il.ibm.com

Abstract

We evaluate the robustness of several large language models on multiple datasets. Robustness here refers to the relative insensitivity of the model's answers to meaning-preserving variants of their input. Benchmark datasets are constructed by introducing naturally-occurring, non-malicious perturbations, or by generating semantically equivalent paraphrases of input questions or statements. We further propose a novel metric for assessing a model robustness, and demonstrate its benefits in the non-adversarial scenario by empirical evaluation of several models on the created datasets. [1]

1 Introduction

With the increase in the prominence and use of large language models (LLMs), there has been tremendous activity in evaluating various aspects of these models' behavior and its alignment with desirable qualities, such as accuracy, safety and privacy. The property of model *robustness*—the ability of a model to produce semantically equivalent output given meaning-preserving input—has been addressed from various perspectives: sensitivity to the wording of instruction template (Mizrahi et al., 2023; Sclar et al., 2023; Zhao et al., 2024), example choice and ordering for in-context learning (ICL) tasks (Voronov et al., 2024), perturbing the order of premises in logical reasoning tasks (Chen et al., 2024), as well as model resilience to adversarial prompts (e.g., Liang et al. (2023); Zhou et al. (2022); Shayegani et al. (2023)).

Model robustness (or insensitivity) to naturally-occurring, non-malicious variations in their input, such as the question in a question-answering (QA) task, or the statement in a classification task, has received relatively little attention (although see Liang et al. (2023) for sensitivity to typos, and Raj et al. (2022) for evaluation of model consistency on input paraphrases). Such slight variations include perturbations that can normally occur in human-generated input, e.g., changes in casing, redundant white-spacing or newlines, the lack of punctuation, "butter-finger" typos,[2] character swap, and meaning preserving paraphrases. While very common in everyday language, these changes may have a significant effect on a model's ability to produce anticipated answers. The reading comprehension example in Table 1 illustrates how slight changes in the phrasing of a question in one of our datasets cause the model to generate different responses.

Benchmarking the robustness of a model to variations in its input typically involves measuring the degree of performance decrease in the perturbed instance set, compared to the original example. For assessing the resilience of LLMs to *adversarial attacks*, the main metric that has been put forward in prior studies is *performance drop rate* (PDR), which is the fractional decrease in the average perturbed instances' score, relative to the original example (Zhu et al., 2023). As discussed in Section 3.1, PDR has two main drawbacks: First, since it measures fractional *decrease*, it is inherently an *asymmetric* function of its inputs; a fixed increase in performance after perturbation receives a larger magnitude PDR than the reversed decrease in performance. Second, since fractional change from 0 is undefined, the PDR is *undefined* in the specific case when the original score was zero but the average over perturbed instance set scores higher, as in the example in Table 1; thus, instances with undefined PDR are ignored when evaluating average PDR on a dataset (e.g., Liang et al. (2023); Zhu et al. (2023)), which can bias aggregates.

While performance improvement is not typical to adversarial tests, it can easily happen in the sce-

[1] All our datasets are available on HuggingFace.

[2] Realistic typos that commonly happen due to the proximity of some keys to others on the physical keyboard.

Findings of the Association for Computational Linguistics: EACL 2024, pages 2794–2802

Context: A function f is said to be continuously differentiable if the derivative f'(x) exists and is itself a continuous function. Though the derivative of a differentiable function never has a jump discontinuity, it is possible for the derivative to have an essential discontinuity. For example, the function [...]

question	model answer	correct
(1) is the derivative of a continuous function always continuous?	yes	✗
(2) Is the derivative of a continuous function always continuous	no	✓
(3) Is the derivative of a continuous function always continuous?	yes	✗
(4) IS THE DERIVATIVE OF A CONTINUOUS FUNCTION ALWAYS CONTINUOUS?	no	✓
(5) does the derivative of a continuous function always exhibit continuous behavior?	yes	✗
(6) is the derivative of a continuous function guaranteed to be continuous?	yes	✗

Table 1: Llama2-chat (13B) model's answers to the original question (1) and its perturbed variants (2-6), where simple superficial perturbations were applied to obtain variants (2-4), and a paraphrasing model produced variants (5-6). The LLM's answer to the original question is incorrect, two of the variants — (2) and (4) — obtained the correct answer ("no"). Variants' distinctions from the original phrasing are highlighted, where not easily visible.

nario with naturally-occurring, non-malicious input variations which we consider. Moreover, scoring a model output with 0 is not uncommon in tasks with binary-valued evaluation (correct or wrong), as in our study. Aiming to overcome these drawbacks, we adapt the Cohen's h statistical effect size metric (Cohen, 1988) for the difference in proportions, discussed in Section 3.2. Indicating the practical significance of a difference in two groups (e.g., outcome of two experimental settings), the use of effect sizes is widely practiced in research and commercial applications (see Fritz et al. (2012)). We show that in the setting of robustness evaluation, Cohen's h constitutes an elegant, symmetric and easily-interpretable metric, which correlates with PDR while overcoming its drawbacks.

Our contribution is, therefore, twofold: First, we expand multiple datasets, concerning classification, QA and reading comprehension, with naturally-occurring input variants, and report a comprehensive assessment of the robustness of LLMs on these tasks. Second, we propose, and empirically evaluate, a novel metric for measuring model sensitivity to non-adversarial perturbations in its input.

Much effort has been invested recently in leaderboards for multi-faceted evaluation of foundation models (e.g., Liang et al. (2023)[3]). A broader impact of this study lies in the adoption of the proposed metric for benchmarking the robustness of LLMs in non-adversarial scenarios.

2 Datasets

2.1 Dataset Description

We make use of multiple diverse datasets in our experiments. The original datasets are expanded by introducing various types of perturbations into raw instances: superficial (non-semantic), paraphrasing, and adding distraction passages where applicable. We experiment with three datasets: (1) PopQA (Mallen et al., 2023): open-domain questions of factual nature about public figures and entities (books, countries, etc.); the dataset has been recently expanded with manually-generated paraphrases by Rabinovich et al. (2023); (2) social identity group abuse (SIGA): short statements for classification, that possibly carry over an abusive flavor towards an identity group by race, religion or gender (Wiegand et al., 2022); (3) BoolQ: a dataset of reading-comprehension questions, with boolean answers (Clark et al., 2019). We use *string containment* as the evaluation metric on PopQA (as in Mallen et al. (2023)), and *accuracy* for BoolQ and SIGA.

2.2 Expanding Datasets with Perturbations

We imitate naturally-occurring variations in human-generated input, by applying the following perturbation types on each input in the three datasets:

Superficial (S) Simple non-semantic perturbations, such as upper-, lower- or proper-casing of certain words, removing punctuation, "butter-finger" typos (misspelling by replacing a randomly-selected letter with one of the adjacent ones on a keyboard), character swap, or redundant white-spacing. A sentence variant can include one or multiple interventions from this set, as illustrated in examples (2)-(4) in Table 1.

Paraphrase (P) We automatically generate (at most) five semantics-preserving paraphrases using the NL-Augmenter (Dhole et al., 2023) paraphrase generator. For PopQA, we use the manually-crafted templated paraphrases by Rabinovich et al. (2023).

[3] https://crfm.stanford.edu/helm/

Distraction (D) BoolQ — the reading comprehension dataset — was additionally expanded with "distractions": a randomly selected passage from the corpus was appended before or after the passage in the input example, to assess models' resilience to (possibly related but not strictly relevant) information in the content-grounded QA task.

Table 2 summarizes the datasets statistics before and after expansion, and perturbation types used. Let $\mathcal{D}$ denote an unperturbed test dataset, consisting of n unique instances $x_1, \ldots, x_n$ (e.g., questions to be answered). The expanded dataset $\mathcal{D}'$ consists of each *original* instance x_i (now denoted by x_i^o), as well as the set of its $m(i) \geq 1$ perturbations, denoted by $(x_i^1, \ldots, x_i^{m(i)})$. Our expanded datasets are available on HuggingFace.

dataset	original	S	P	D	final
PopQA	14.2K	85.6K	104.3K	–	204.2K
BoolQ	3.2K	9.8K	9.7K	6.5K	29.3K
SIGA	2.1K	12.6K	6.1K	–	20.8K

Table 2: Datasets size before and after expansion, by perturbation type: superficial (S), paraphrase (P) and distraction (D). Only test set portion of the dataset was considered for experiments where applicable.

3 Quantifying Model Robustness

Consider x_i^o and $(x_i^1, \ldots, x_i^{m(i)})$, the i^{th} original input and its $m(i)$ perturbations, in the perturbed dataset $\mathcal{D}'$ (see Section 2.2). Given a scoring metric $score \in [0, 1]$, let $score_i^o$ and $(score_i^1, \ldots, score_i^{m(i)})$ be scores of the model's predicted value (e.g., generated answer) on the original ("o") and perturbed instances, respectively, versus the ground truth reference. In our case, $score \in \{0, 1\}$ are binary-valued accuracy or string containment match metrics. We compute the average score on the perturbed ("p") instance set for input x_i^o in $\mathcal{D}'$ by $score_i^p = \frac{1}{m(i)} \sum_{j=1}^{m(i)} score_i^j$. We consider a model's performance on a dataset as being *robust* if the performance tends to be insensitive to perturbations; that is, the two scores—$score_i^o$ and $score_i^p$—tend to be close to to each other, across all input instances i in $\mathcal{D}$.

Note that the notion of model robustness differs from the model's performance overall, which would assess the averages of the scores (either original, perturbed, or both). An LLM can be robust but have poor performance, or have high average performance on the original instances ($score_i^o$) but perform poorly on perturbations ($score_i^p$), making it sensitive to variation in its input. We now describe the two metrics used to measure model performance robustness: performance drop rate (PDR) (Zhu et al., 2023), traditionally used to assess an LLM's resilience to adversarial attacks, and Cohen's h effect size – the proposed metric for assessing models' robustness to naturally-occurring, non-malicious perturbations.

3.1 Performance Drop Rate (PDR)

PDR (Zhu et al., 2023), the fractional change in the mean perturbed score of example i, relative to the original, is defined[4] as PDR($score_i^o$, $score_i^p$) =

$$\begin{cases} 0, & score_i^o = score_i^p = 0 \\ \text{undefined}, & score_i^o = 0, score_i^p \neq 0 \\ 1 - \frac{score_i^p}{score_i^o}, & \text{otherwise} \end{cases}$$

Due to its asymmetric nature, PDR is biased towards cases where $score_i^p > score_i^o$, contrary to the opposite scenarios, skewing the final average score. As a concrete example, the increase from 0.1 (original) to 0.8 (perturbed set) has a PDR of -7 (=-700%), while the opposite direction, a decrease from 0.8 (original) to 0.1 (perturbed set), has a PDR of 0.875 (=87.5%). Additionally, the metric is udenfined in cases where the model's performance on the original instance is incorrect ($score_i^o$=0). Collectively, these characteristics make PDR a sub-optimal choice for assessing a model robustness to perturbations in non-adversarial scenarios.[5]

3.2 Cohen's h Effect Size

Cohen's h (Cohen, 1988) statistical effect size is commonly used for measuring the difference in two proportions in empirical research (see Appendix A.1 for background), and is defined as

$$\text{H}(score_i^o,\ score_i^p) = \psi(score_i^p) - \psi(score_i^o),$$
$$\text{where } \psi(score_i) = 2\left(\arcsin\left(\sqrt{score_i}\right)\right)$$

Cohen's h effect size takes values in the range $[-\pi, \pi]$, where h>0 indicates performance improvement relative to $score_i^o$. This metric has several important characteristics: (1) Unlike PDR, it is a symmetric function, and is defined for all pairs

[4] We added the first case to the definition in Zhu et al. (2023) for scenarios where both the original and the perturbed instances' performance are incorrect.

[5] Adaptations of PDR addressing (to some extent) its drawbacks can be devised, but they harm the semantics of PDR, being defined as performance drop ratio.

of $score$ values within the [0, 1] range. (2) It has rule-of-thumb thresholds[6] of what values constitute small, medium, etc. differences in sample proportions, which are based on its statistical properties. For better interpretability, we define a normalized version of Cohen's h, defined as $\tilde{\text{H}} = \text{H}/\pi$, which takes values within the [-1, 1] range. Consequently, the effect size thresholds are adjusted for this normalized version, each divided by π.

Figure 1 shows that PDR and $\tilde{\text{H}}$ correlate very well (Pearson's $r{\approx}0.995$ when $score_i^o{=}1.0$), supporting this novel application of the metric to the scenario of tasks with binary-valued evaluation outcome. We note, however, that the application of the metric is not limited to the scenario with binary evaluation outcome. See Appendix A.1 for a statistical discussion of the $\tilde{\text{H}}$ metric.

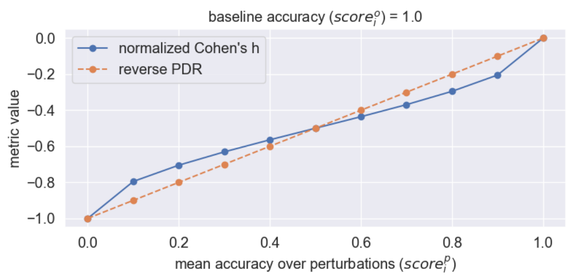

Figure 1: Comparison of normalized Cohen's h ($\tilde{\text{H}}$) and reverse PDR ($=-1{\times}$PDR) when the original instance accuracy $score_i^o{=}1.0$ (as in the tasks in our study – binary evaluation outcome: 0 or 1).

The absolute value of a directional effect sizes ($\text{A}\tilde{\text{H}} = |\tilde{\text{H}}|$) measures the degree of deviation in either direction, and can serve as a proxy for the expected *variance* or absolute deviation between the original and perturbed instance performance.

Taking the example values of $score_i^o = 0.8$ and $score_i^p = 0.1$ used above for PDR (Section 3.1), the corresponding value of $\tilde{\text{H}}$ is -0.5, and 0.5 if the direction is reversed (non-normalized $\text{H} \approx \pm 1.57$). This counts as a 'very large' difference according to the thresholds in Table 5 (see Appendix A.1).

4 Benchmarking Model Robustness

4.1 Experimental Setup

We conduct experiments using the following LLMs, proven effective in multiple tasks: instruction-tuned Google's Flan-T5-XXL (11B; Chung et al. (2022)) and Flan-UL2 (20B; Tay et al. (2022)), IBM's Granite 13B series: Chat and Instruct (IBM Research, 2024), Meta AI's Llama2-Chat (13B; Touvron et al. (2023)) and the recent Llama3-Instruct (70B; Meta (2024)), as well as Mistral AI's Mixtral-Instruct (8x7B; Jiang et al. (2024)).

We use default system prompts, zero-shot experimental setup, and greedy prediction mode, where temperature is set to 0. Our per-dataset prompts to LLMs are detailed in Appendix A.4.

4.2 Experimental Results

Table 3 shows per-dataset performance by model. Original performance (average on examples in the raw dataset – mean($score_i^o$)) is reported, as well as the average over mean perturbed sets – mean($score_i^p$). Figure 3 in Appendix A.3 visualizes this table along with the 95% confidence intervals for the metrics, represented by the cell shading in the table. The significance thresholds in Table 5 can tell us how large the observed mean $\tilde{\text{H}}$ across examples in a dataset is, and one can decide that, say, a difference that is 'small' or greater indicates the LLM is not robust to the perturbations. The 95% interval for $\tilde{\text{H}}$ (or $\text{A}\tilde{\text{H}}$) measure how variable this assessment of robustness is to random sampling; for instance, if the 95% interval is itself also within the bounds of a 'small' difference, this lends statistical confidence to the assessment that the LLM is robust. The fact that $\tilde{\text{H}}$ has such significance thresholds that can be used in a practical decision gives it a benefit compared to metrics like PDR, which lack them.

A slightly inferior performance on the perturbed instances compared to original is reflected by the negative Cohen's h effect size ($\tilde{\text{H}}$). Notably, the absolute value effect ($\text{A}\tilde{\text{H}}$) differs considerably from the directional $\tilde{\text{H}}$ values, indicating that the LLMs' predictions varied in both directions (better or worse) compared to the original instance performance – a finding largely supportive of our assumption that naturally-occurring, non-malicious perturbations may have either positive or negative effect on a model's accuracy. While only a single observed $\tilde{\text{H}}$ directional effect size is considered non-negligible (grey background), most absolute values constitute a "small" change, and virtually all effect sizes show significant, indicating small, yet systematic, and reliably detected change.

Model Robustness vs Performance Figure 2 illustrates average model accuracy on the original datasets vs their mean undirectional robustness ($\text{A}\tilde{\text{H}}$). Evidently, best performing, recently re-

[6]See Table 5 in Appendix A.1.

model	PopQA M(orig)	M(pert.)	$\tilde{H}$	$A\tilde{H}$	BoolQ M(orig)	M(pert.)	$\tilde{H}$	$A\tilde{H}$	SIGA M(orig)	M(pert.)	$\tilde{H}$	$A\tilde{H}$
Granite-Chat (13B)	0.20	0.18	-0.02*	0.08*	0.81	0.78	-0.04*	0.11*	0.71	0.67	-0.04*	0.18*
Granite-Instruct (13B)	0.16	0.15	-0.02	**0.06***	0.87	0.86	-0.02*	0.06*	0.60	0.59	**-0.01**	0.10*
Llama2-Chat (13B)	0.29	0.27	-0.03*	0.12*	0.84	0.81	-0.04*	0.11*	**0.81**	0.78	-0.07*	0.16*
Llama3-Instruct (70B)	0.37	0.33	-0.04*	0.15*	0.89	0.87	-0.03*	0.07*	0.80	**0.79**	**-0.01**	**0.07***
Mixtral-Instruct (8×7B)	**0.39**	**0.36**	-0.04*	0.16*	0.89	0.87	-0.03*	0.07*	0.79	0.78	-0.02*	0.10*
Flan-T5-XXL (11B)	0.13	0.13	**0.00**	**0.06**	0.92	0.91	**-0.02***	0.05*	0.79	0.78	-0.03*	0.09*
Flan-UL2 (20B)	0.15	0.14	-0.01	0.07*	**0.97**	**0.95**	-0.04*	**0.04***	0.80	**0.79**	-0.02*	0.10*

Table 3: Mean accuracy on the original datasets, mean accuracy on perturbed variants (slightly lower). Confidence intervals (CIs) for are calculated by original or perturbed group-level bootstrapping, as discussed in Appendix A.2. The high difference between the directional $\tilde{H}$ and undirectional $A\tilde{H}$ is suggestive of both increase and decrease in models' performance on original, compared to perturbed examples. Results for which the $\tilde{H}$ and $A\tilde{H}$ 95% confidence intervals (see Appendix A.2.) do not contain 0 are marked with "*", indicating the significance of the finding. Notably, significant $\tilde{H}$ and $A\tilde{H}$ values may still indicate a very small effect size (see Appendix A.1); values reflecting a non-negligible change are marked with gray background. The best result in a column is boldfaced.

dataset	M(orig)	superficial (S) M(pert.)	$\tilde{H}$	$A\tilde{H}$	paraphrase (P) M(pert.)	$\tilde{H}$	$A\tilde{H}$	distraction (D) M(pert.)	$\tilde{H}$	$A\tilde{H}$
PopQA	0.37	0.32	-0.05*	0.12*	0.34	-0.03	0.15*	–	–	–
BoolQ	0.89	0.88	-0.01*	0.04*	0.85	-0.04*	0.07*	0.88	-0.01	0.05*
SIGA	0.80	0.79	-0.01*	0.06*	0.77	0.00	0.09*	–	–	–

Table 4: Mean accuracy on the original datasets, mean accuracy on perturbed variants with the most recent `Llama3-Instruct(70B)` model in this study, with break-down by variant type (superficial (S), paraphrase (P), distraction (D)). Results for which the $\tilde{H}$ and $A\tilde{H}$ 95% confidence intervals (see Appendix A.2.) do not contain 0 are marked with "*", indicating the significance of the finding; values reflecting a non-negligible change are marked with gray background. Notably, paraphrasing the original question results in a more considerable performance drop than introducing superficial (simple) perturbations, across all three datasets.

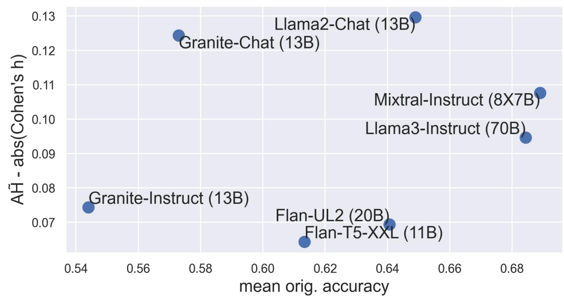

Figure 2: Mean model accuracy on original datasets vs its undirectional robustness. x-axis: the higher, the better performing; y-axis: the lower, the more robust.

leased models (Llama3-Instruct (70B) and Mistral-Instruct (8×7B)) exhibit more moderate robustness, compared to the most robust Flan-T5-XXL, that shows slightly inferior average performance. Granite-Instruct (13B) is one of the most robust models, while performing worse on average.

Robustness Evaluation by Perturbation Type We further break down the robustness measurements by individual perturbation types in Table 4 for `Llama3-Instruct(70B)` – one of the best-performing models on our datasets. Notably, paraphrasing the original question results in a more considerable performance drop than introducing superficial (simple) perturbations, across all three datasets. We attribute the particularly high absolute effect size in the PopQA dataset (0.15) to the fact that paraphrases in this dataset were created manually, aiming at high linguistic diversity while maintaining the original semantics.

5 Conclusions

We evaluate the robustness of several LLMs on multiple diverse datasets, by expanding them with non-malicious, naturally-occurring perturbations, and measuring models' resilience to these variants in user input. We propose and evaluate a novel application of a statistical effect size metric for assessing model robustness in tasks with binary- or proportion- valued evaluation scores, and demonstrate its benefits in the non-adversarial scenario.

6 Limitations

Our study, while contributing valuable insights for measuring model robustness to non-adversarial perturbations, is subject to several limitations. First, the application of Cohen's h effect size, suggested is this work, is an intuitive fit for tasks with binary-valued evaluation outcome, correlating with PDR (denoting fractional decrease); other effect size metrics could constitute a more intuitive choice in scenarios with continuous evaluations scores. Second, a limited number of open models were evaluated on three datasets; the study can be extended to additional (commercial) models and more sophisticated tasks, e.g. MMLU (Hendrycks et al., 2021). Finally, while our automatic paraphrase generation is of high-quality overall, it is admittedly conservative – only slight deviations from the original examples were applied to preserve semantics. We plan to make use of advanced models for more diverse paraphrase generation in the future.

References

Elron Bandel, Yotam Perlitz, Elad Venezian, Roni Friedman-Melamed, Ofir Arviv, Matan Orbach, Shachar Don-Yehyia, Dafna Sheinwald, Ariel Gera, Leshem Choshen, et al. 2024. Unitxt: Flexible, shareable and reusable data preparation and evaluation for generative ai. *arXiv preprint arXiv:2401.14019*.

Xinyun Chen, Ryan A. Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *Preprint*, arXiv:2402.08939.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. BoolQ: Exploring the surprising difficulty of natural yes/no questions. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2924–2936.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. Lawrence Erlbaum Associates.

Kaustubh Dhole, Denis Kleyko, and Yue Zhang. 2023. Nl-augmenter: A framework for task-sensitive natural language augmentation. *NEJLT Northern European Journal of Language Technology*, 9(1):1–41.

Catherine O Fritz, Peter E Morris, and Jennifer J Richler. 2012. Effect size estimates: current use, calculations, and interpretation. *Journal of experimental psychology: General*, 141(1):2.

Larry V Hedges. 1981. Distribution theory for Glass's estimator of effect size and related estimators. *journal of Educational Statistics*, 6(2):107–128.

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

IBM Research. 2024. Granite foundation models.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. 2023. Holistic evaluation of language models. *Transactions on Machine Learning Research*.

Alex Troy Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2023. State of what art? a call for multi-prompt llm evaluation. *arXiv preprint arXiv:2401.00595*.

Ella Rabinovich, Samuel Ackerman, Orna Raz, Eitan Farchi, and Ateret Anaby-Tavor. 2023. Predicting question-answering performance of large language models through semantic consistency. In *Conference on Empirical Methods in Natural Language Processing*.

Harsh Raj, Domenic Rosati, and Subhabrata Majumdar. 2022. Measuring reliability of large language models through semantic consistency. In *NeurIPS ML Safety Workshop*.

Shlomo S Sawilowsky. 2009. New effect size rules of thumb. *Journal of modern applied statistical methods*, 8:597–599.

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. In *The Twelfth International Conference on Learning Representations*.

Skipper Seabold and Josef Perktold. 2010. Statsmodels: Econometric and statistical modeling with python. In *Proceedings of the Python in Science Conference*, page 57. SciPy.

Erfan Shayegani, Md Abdullah Al Mamun, Yu Fu, Pedram Zaree, Yue Dong, and Nael Abu-Ghazaleh. 2023. Survey of vulnerabilities in large language models revealed by adversarial attacks. *arXiv preprint arXiv:2310.10844*.

Gail M Sullivan and Richard Feinn. 2012. Using effect size—or why the p value is not enough. *Journal of graduate medical education*, 4(3):279–282.

Yi Tay, Mostafa Dehghani, Vinh Q Tran, Xavier Garcia, Jason Wei, Xuezhi Wang, Hyung Won Chung, Siamak Shakeri, Dara Bahri, Tal Schuster, et al. 2022. Ul2: Unifying language learning paradigms. *arXiv preprint arXiv:2205.05131*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Anton Voronov, Lena Wolf, and Max Ryabinin. 2024. Mind your format: Towards consistent evaluation of in-context learning improvements. *arXiv preprint arXiv:2401.06766*.

Michael Wiegand, Elisabeth Eder, and Josef Ruppenhofer. 2022. Identifying implicitly abusive remarks about identity groups using a linguistically informed approach. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5600–5612.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024. Improving the robustness of large language models via consistency alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8931–8941.

Chunting Zhou, Junxian He, Xuezhe Ma, Taylor Berg-Kirkpatrick, and Graham Neubig. 2022. Prompt consistency for zero-shot task generalization. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2613–2626.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, et al. 2023. Promptbench: Towards Evaluating the Robustness of Large Language Models on Adversarial Prompts. *arXiv preprint arXiv:2306.04528*.

A Appendix

A.1 Effect Sizes and Cohen's h

Effect size metrics are measures of the size of a statistical phenomenon; common examples are Pearson correlation and odds ratios. An effect size metric measures the aspect of interest (e.g., the difference in means between two sample sizes) in a way that is independent of the sample sizes. This is in contrast to the p-value of a hypothesis test statistic (e.g., a two-sample test) which, for fixed values of the sample means and variances, becomes more significant when the sample sizes increase (Sullivan and Feinn, 2012); this quality makes p-values vulnerable to manipulation ("p-hacking"). Effect size metrics are often used to ensure that a hypothesis test has enough statistical power (complement of the Type-II or false negative error probability) given the sample size(s). This insensitivity to the sample size in the effect size value means that they can be used to measure the significance of an effect in cases of small sample sizes (e.g., in our case when we have one original instance and a small number of perturbations), and thus may be better than, say, a two-sample p-value, for assessing robustness. Cohen's h has a particular advantage in that it is defined even if there is no sample variation (e.g., if $score \in \{0, 1\}$), which causes p-values and some other effect sizes to be undefined.

Cohen's h—and thus $\tilde{\mathrm{H}}$—changes non-linearly with changes in $|score^o_i - score^p_i|$. In contrast, PDR changes linearly when $score^o_i$ is fixed. Considering the specific case when the original instance accuracy $score^o_i$ is perfect (1.0), both $\tilde{\mathrm{H}}$ and "reverse PDR" ($= -1 \times$PDR) take values in the [-1, 0] range and are highly correlated, as shown in Figure 1; this high correlation suggests that in this case ($score^o_i$=1.0), Cohen's h constitutes an intuitive and easily-interpretable alternative to PDR.

A two-sample proportions difference hypothesis test (see statsmodels' `proportions_ztest`, Seabold and Perktold (2010)) is an alternative way of measuring the significance of these differences. In this test, the test statistic is maximized (i.e., is more significant) for a given fixed difference $|score^o_i - score^p_i|$ when one of the proportions is equal to 0 or 1, because the pooled variance in the denominator is minimized. The arcsine transformation in Cohen's h magnifies the resulting effect size for a given $|score^o_i - score^p_i|$ when one of the proportions is close to 0 or 1 since the difference is more detectable, which causes the non-linear change in Figure 1 (around 0.1 and 0.9).

The corollary of the fact that Cohen's h changes non-linearly with $|score^o_i - score^p_i|$ is that Cohen's h (and $\tilde{\mathrm{H}}$) for pairs of $(score^o_i,\ score^p_i)$ should be equal when the difference is equally *detectable* (Cohen, 1988, p. 180–181), despite $|score^o_i - score^p_i|$ differing. Thus, for instance, $\tilde{\mathrm{H}}(0.8,\ 1.0) \approx 0.295$ but $\tilde{\mathrm{H}}(0.6,\ 0.8) \approx 0.141$, meaning that the 0.2 accuracy decrease from $score^o_i$=1.0 to $score^p_i$=0.8

should statistically be more than twice as detectable as the same decrease from $score_i^o$=0.8 to $score_i^p$=0.6. Perturbation accuracy would have to fall from 0.8 to $score_i^p \approx 0.36$ to be as significant, by $\tilde{\text{H}}$, as the fall from 1.0 to 0.8, despite the raw decrease being more than twice as large.

We note that if the instance scores $score_i^j$ are continuous-valued rather than binary, an alternative effect size metric such as Cohen's d (Cohen, 1988) or Hedges' g (Hedges, 1981) for comparing sample means can be used in a similar way to Cohen's h. However, these effect sizes, unlike Cohen's h, are undefined or infinite when the within-sample variance is zero. We leave further investigation for future work.

effect size	Cohen's h	$\tilde{\text{H}}$ (normalized)
essentially zero	[0.0, 0.01)	[0.0, 0.0032)
very small	[0.01, 0.2)	[0.0032, 0.0637)
small	[0.2, 0.5)	[0.0637, 0.1592)
medium	[0.5, 0.8)	[0.1592, 0.2546)
large	[0.8, 1.2)	[0.2546, 0.3820)
very large	[1.2, 2.0)	[0.3820, 0.6366)
huge	[2.0, π]	[0.6366, 1.0]

Table 5: Ranges of values of Cohen's h and their size interpretation, as defined by Cohen (1988) and Sawilowsky (2009); many other related metrics, such as Cohen's d, have the same thresholds but are not bounded from above. The bounds for our normalized metric ($\tilde{\text{H}}$) are the first bounds, divided by π.

A.2 Bootstrapped Confidence Intervals

All metrics and instance scoring functions are implemented in the open-source repository `unitxt` (Bandel et al., 2024). A 'group' here consists of an original instance and its $m(i)$ perturbations (see Section 2.2). A given metric f (e.g., mean score, PDR, $\tilde{\text{H}}$) produces a group-level score $s_i = f\left(score_i^o, score_i^1, \ldots, score_i^{m(i)}\right)$. Thus, if original dataset $\mathcal{D}$ has n instances, we have n instance-group scores $(s_1, \ldots, s_n)$ on $\mathcal{D}'$. Statistical analysis of the metric is done by constructing 95% bootstrapped confidence intervals on the s_i scores, discarding any undefined values, rather than resampling the instance scores, reforming the groups, and calculating the group scores; the latter option could result in duplicated perturbed instances or incomplete groups if $score_i^o$ is missing, which can make s_i undefined. We conduct group-score resampling because we analyze the typical average robustness across original instances, and thus the unit of analysis is the original instance *together with* its perturbations, as reflected in s_i.

A.3 Detailed Experimental Results

Figure 3 shows the robustness metric values by model and dataset. The bar colors represent the model source (Google, IBM, Meta, Mistral), while the individual models are distinguished by the diagonal hashing pattern. At the top of each bar, a red line shows the 95% bootstrapped confidence interval as described in Appendix A.2.

Note that the values of normalized Cohen's h (i.e., $\tilde{\text{H}}$) are almost always negative, indicating a decrease in accuracy after perturbation; however, these changes are nearly all very minor, falling between 0 and the "very small" decrease threshold of ≈ -0.0637, shown by the horizontal blue line. These insignificant changes in performance are reflected in the fact that the heights of the mean original score $score_i^o$ bars (top left) are very similar to the corresponding mean perturbed score $score_i^p$ bars (top right), and that their confidence interval bands overlap significantly. Furthermore, the relative magnitudes of $\tilde{\text{H}}$ and $\text{A}\tilde{\text{H}}$ also correlate well with PDR, which, despite its drawbacks (see Section 3.1), can serve as a good sanity check.

A.4 Experimental Setup and Prompts

We use default system prompts, zero-shot experimental setup, and greedy prediction mode, where the temperature is set to 0. Our per-dataset prompts for the models are detailed below.

- **BoolQ**:
  ```
  "Context: {passage}
  Based on the context, answer the following question with a single word: yes or no. Question: {question}
  Answer:"
  ```

- **PopQA**:
  ```
  "Question: {question}
  Answer:"
  ```

- **SIGA**:
  ```
  "Does the following statement contain abusive or unfair negative opinion against some social identity group? Answer with yes or no.
  Statement: {question}
  Answer:"
  ```

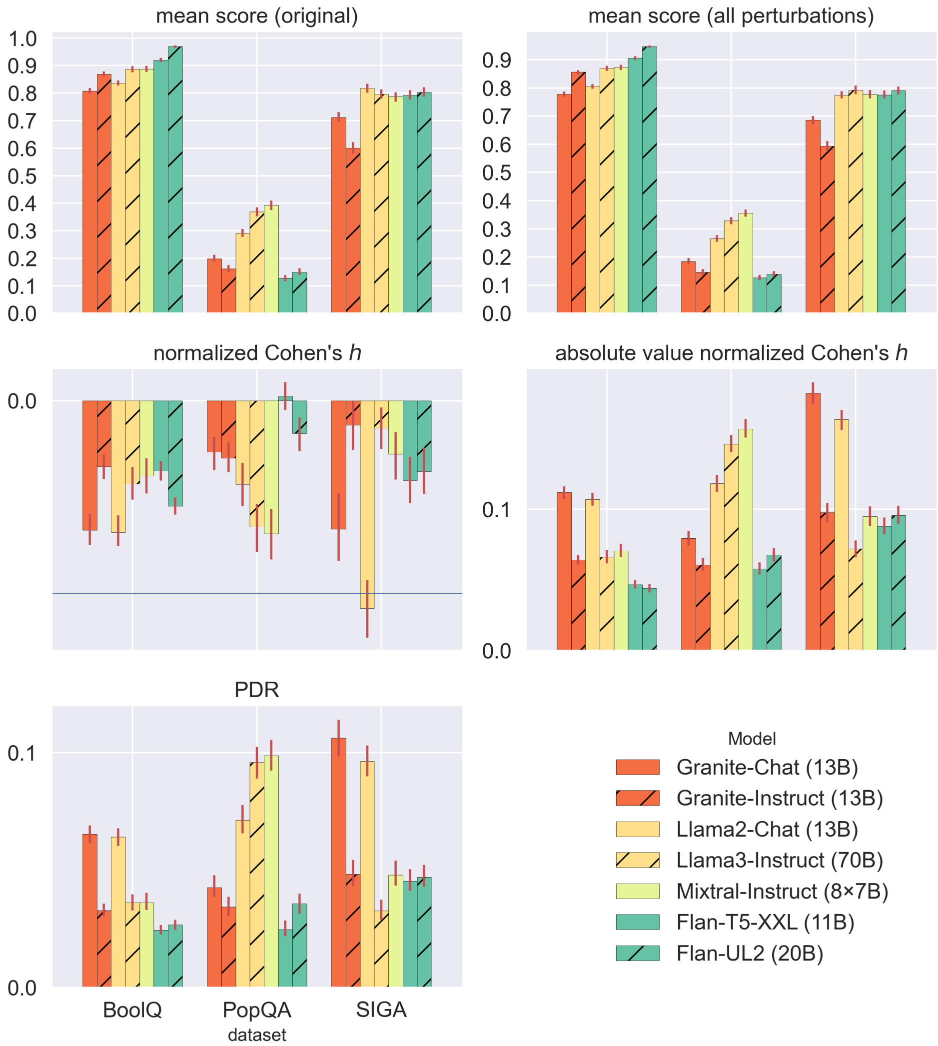

Figure 3: Mean metric scores by model and dataset. Red error bars show a 95% bootstrapped confidence interval.