

Who Benchmarks the Benchmarks? Towards Comprehensive Evaluation of Commonsense Reasoning Benchmarks

Pavel Chizhov^{1,2}

Anton Changalidis²

Vishnu Prasad Vijaya Kumar^{1,2}

Yannick Detrois^{2,3}

Mattia Nee²

Pierre-Carl Langlais²

Ivan P. Yamshchikov^{1,2}

¹CAIRO, Technical University of Applied Sciences Würzburg-Schweinfurt

²PleIAs ³EPFL

Correspondence: pavel.chizhov@thws.de

Abstract

Commonsense reasoning is a key language model capability, as it is purportedly a prerequisite for many basic tasks, unlike specific factual knowledge. It is often measured with multiple-choice questions (MCQ) benchmarks, e.g. HellaSwag and PIQA. Some of these benchmarks, however, are outdated and contain numerous validity issues. We illustrate some typical validity issues with a case study on HellaSwag, one of the most popular and problematic benchmarks for commonsense reasoning. The issues we find range from basic ungrammaticality and numerous typos to misleading prompts or equally correct options. We show that if we remove question prompts or replace them with "*Lorem ipsum dolor...*", about 68% of model predictions do not change. We argue that this occurs due to inner flaws in the benchmark, not mere contamination that might be present in some models. Since benchmark scores are an essential part of model selection in both research and commercial applications, these issues can have severe consequences. Based on our findings, we propose BenCheck, a package for benchmark validity analysis that encapsulates the main checks performed in our case study and can be used to audit commonsense reasoning benchmarks. We apply these checks to PIQA, Global PIQA, and Winogrande.

 PleIAs/BenCheck

1 Introduction

Language models are evaluated on benchmarks. These evaluations shape model development by indicating how different design decisions, e.g., hyperparameters, training data, and post-training, impact model performance (Biderman et al., 2026). NLP practitioners select training procedures to improve performance on accepted benchmarks. However, if the benchmarks do not measure what the researchers think they measure, development may

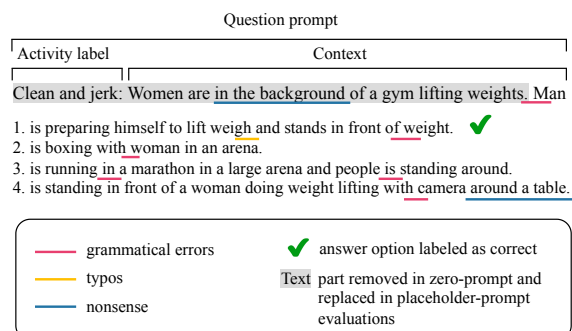


Figure 1: Example from the HellaSwag validation set. Validity issues are underlined in color. The part of the question prompt that we remove or replace in prompt-related experiments is highlighted in light gray.

not be going in the most optimal direction, and we may be missing out on performance gains.

Construct validity refers to whether a benchmark is measuring the capability it is claimed to measure (Cronbach and Meehl, 1955; Bean et al., 2025). Low construct validity can lead to real-world consequences, including misdirected research (Bean et al., 2025). Model developers choose dataset sources (Penedo et al., 2025) and make design choices (Chowdhery et al., 2023, *inter alia*) based on benchmark results. Furthermore, policymakers and government bodies worldwide may use benchmarks to inform policy decisions (Reuel et al., 2024). Thus, benchmark construct validity not only influences model development and scientific research, but also even trickles down to policy-making processes. Without established validity, the use of a benchmark may be “supercharging bad science” (Blili-Hamelin et al., 2025).

Commonsense reasoning is a desirable capability in a model because it encompasses a more generalizable language and world understanding rather than memorized facts. This kind of reasoning focuses on reasoning over domain-general scenarios and thus differs from the math- and code-oriented

reasoning, for which there are increasingly many benchmarks. In a survey of over 100 commonsense reasoning benchmarks, [Davis \(2023\)](#) observe widespread quality issues. HellaSwag ([Zellers et al., 2019](#)) is one of the most widely used of these commonsense reasoning benchmarks, and it has played a large role in language model development over the past six years.

In order for a commonsense reasoning benchmark to serve its purpose, it should allow us to draw an inference about model capabilities, specifically that the model with the highest HellaSwag score has the best commonsense reasoning capabilities. However, we argue that there are key issues (see an example in Figure 1) that prevent us from drawing that inference from HellaSwag scores. We make the connection between dataset artifacts and construct validity as follows: the dataset artifacts (grammar/typos, generation hallucinations, etc.) are inherent flaws in the benchmark. The models exploit these artifacts and learn shortcuts that allow them to avoid the commonsense reasoning framework. Thus, the benchmark does not adequately measure the key target capability, thereby compromising construct validity.

Our work makes the following contributions:

- We conduct a case study on HellaSwag and argue that its scores are confounded by factors beyond sentence completion and commonsense reasoning. We point out diverse issues and their consequences and show that they are omnipresent in HellaSwag (Section 3.1).
- We connect these issues to model performance and separately test HellaSwag’s prompt importance by zero- and placeholder-prompt evaluations (Section 3.2).
- Based on our case study, we introduce BenCheck (Section 4), a package for dataset validity audits, and apply it to other commonsense reasoning benchmarks.
- We propose a list of requirements that a quality commonsense reasoning benchmark should meet (Section 6).

2 Related Work

Benchmark validity issues. Several recent studies have highlighted various issues in widely used benchmarks. In the text-generation evaluation format, MMLU ([Hendrycks et al., 2021](#)) has been

shown to yield variable performance for the same model after minor changes. Small perturbations to the prompt have also been shown to significantly affect model performance on benchmarks such as MMLU and HellaSwag ([Habba et al., 2025](#)).

In addition to sensitivity to these changes, MMLU has been shown to contain questions with no clear correct answer or incorrect ground truth labels ([Gema et al., 2025](#)). Other benchmarks, ranging in domain and task type, have also been shown to have errors in the reference answers or in the ground truth labels. One of these is GSM8K ([Cobbe et al., 2021](#)), a math word problem benchmark. [Vendrow et al. \(2025\)](#) find that at least 5% of the items in GSM8K contain a serious error, such as mislabeled questions, logical contradictions, and ambiguous questions. Another example is XSUM ([Narayan et al., 2018](#)), a summarization benchmark. [Liang et al. \(2023\)](#) found that reference summaries in XSUM were rated as being worse than model-generated responses, according to human annotators. Therefore, the quality of the benchmark underestimates the model’s summarization performance.

The current paper contributes to this line of work, which aims to identify issues with existing benchmarks. We also encapsulate our analysis and other checks into a customizable package for benchmark validity audits, enabling faster, more comprehensive validation.

Commonsense reasoning benchmarks. Commonsense reasoning is a desirable capability in a language model because it requires relatively sophisticated language understanding and knowledge about the world. However, a survey of commonsense reasoning benchmarks highlights item quality as a common issue across over 100 text-based benchmarks ([Davis, 2023](#)), with HellaSwag among the most widely used. A previously published blog post highlighted some issues with HellaSwag ([Chen, 2023](#)); however, the author used only a small sample ($n = 300$) from the entire dataset. The author estimates that 36% of items in HellaSwag contain errors. In this work, we annotate the entire validation set and expand the list of the issues we consider. As a result, we demonstrate that a much higher proportion of HellaSwag contains errors.

Prompt importance. Prior work has shown that for many commonsense reasoning benchmarks, models may exploit task artifacts or option-level heuristics rather than the intended commonsense

reasoning signal. Trichelair et al. (2019) show that in SWAG benchmarks, incorrect answer options differ from golden answers both structurally and in quality, and connect these artifacts with internal validity (Campbell and Stanley, 1963) because they constitute an uncontrolled alternative explanation to the model’s performance. This, in turn, undermines the construct validity (Cronbach and Meehl, 1955) as the score variance does not reflect the target construct. Kavumba et al. (2021) fine-tune BERT and RoBERTa models on answer-only entries and show that models are able to learn shortcuts to distinguish correct answers without question context. Prompt-ablation diagnostics have also been used in “no-context” (Shah et al., 2020), “answer-only” (Li et al., 2022), and “choice-only” (Balepur and Rudinger, 2024; Balepur et al., 2024) settings, where models can often complete the task without seeing the full prompt.

Following this line of work, we perform zero-prompt and placeholder-prompt evaluations and analyze not only accuracy but also agreement and disagreement with the full-prompt predictions. This adds granularity to how prompt perturbations affect model performance, regardless of the actual score.

Benchmark validation. Recent work introduced BenchMarker (Balepur et al., 2026), a toolkit for LLM-as-a-judge evaluation of MCQ benchmarks, including the study of contamination. Our BenCheck complements it by covering checks that are largely inferred from the benchmark’s statistics and model behavior, with the two tools together providing a broader view of benchmark validity.

3 Case Study: HellaSwag

HellaSwag is useful as a relatively large dataset, with over 10 000 items in the validation set¹. It is used in leaderboards, such as Hugging Face’s Open LLM Leaderboard v1, which measure and encourage development of open-weight models. It also still often appears in the curated list of benchmarks for model cards and technical reports for production-scale language models, e.g. GPT-4 (Achiam et al., 2023) and Qwen 2.5 (Yang et al., 2025), and smaller open models, e.g. Gemma 3 (Gemma Team et al., 2025) and OLMo 3 (Team Olmo et al., 2025). Since its release, it has had an immeasurable impact on language model

development and remains one of the most popular default evaluation choices.

HellaSwag was created by taking portions of existing data (ActivityNet (Krishna et al., 2017) and WikiHow, which the authors scraped themselves) and generating alternative completions to the segments. The completions were generated by the original GPT model (Radford et al., 2018), largely inferior to the current state of the art, and curated using an adversarial filtering procedure intended to select only the machine-generated answers that seemed most plausible to the language model but nonsensical to humans. Human validators reviewed the dataset after the AF process. The task was intended to be trivial for people (approximately 95% accuracy) but significantly more difficult (less than 50% accuracy) for SOTA models at the time, which (1) were primarily encoder-style models and (2) are not necessarily relevant for current models, especially generative ones. In this case study, we focus on the following issues:

Grammar errors and typos. Grammatical errors can have opposite effects depending on where they occur: errors in the prompt or correct answer may make an item harder, whereas errors in distractors may make it easier by lowering the probability of those options. Although models may be expected to handle minor errors, robustness to such noise should be evaluated separately.

Nonsensicality. Nonsensical text directly undermines commonsense reasoning evaluation, especially when it appears in the prompt or correct answer. If the distractors are implausible (see example in Section 3.1), the task becomes trivial and non-informative, just as with the grammar errors.

None or many good answers. HellaSwag is designed to have exactly one correct answer. If no good option exists, models are rewarded for avoiding bad answers, which can overestimate performance. If multiple answers fit, a plausible choice may be marked wrong, which can underestimate performance and distort the random baseline.

Prompt importance. As discussed in previous sections, commonsense reasoning is directly evaluated by inferring the answer option from the question prompt (context). If the context can be avoided and the answer is inferable from other heuristics, including the plausibility or grammatical correctness of the answer options, the construct validity of the benchmark is compromised.

¹The validation set is the only public split on Hugging Face (except for train) and is commonly used as the test set, which was intentionally hidden to avoid contamination.

3.1 Annotations

We use Gemini 2.5 Flash (Comanici et al., 2025) to annotate the entire HellaSwag validation set. Based on the validity issues enumerated above, we formulate a list of rubrics for LLM annotation, see the full prompt in Appendix I. We also specify in the prompt which answer is supposed to be correct so that we do not rely on LLM’s solutions for the benchmark, and the model has a better understanding of the question’s design. To check the reliability of LLM annotations, we collect two sets of human annotations for a subsample of the validation set. In most cases, LLM annotations received high agreement with at least one human annotator, and overall agreement was on a similar level as inter-annotator agreement. We present the validation analysis and provide further discussion in Appendix D.

From the evaluations of grammar, we find that 30.9% of question prompts, 24.1% of correct answers, and the absolute majority of distractor sets (68.6%) contain grammatical issues or typos (Table 1). For the ActivityNet part, this can be attributed to the benchmark creation methodology, in which these questions were generated by the original GPT model (Radford et al., 2018), which is drastically inferior to modern ones. It is also clear that correct answers generally have considerably fewer issues than incorrect ones because, on benchmark construction, the correct answers were taken from an existing corpus and the incorrect ones were synthetically generated. We also found that 85.1% of questions have utterly implausible constructions in incorrect options, for example:

“Changing out a car seat can dramatically improve the color of your car seat if you use **vinyl, vinyl, vinyl, or vinyl** seats.”

This might lead to trivial solutions dependent solely on choosing the least problematic option. 4.7% of questions were also highlighted due to ethical concerns. We investigated that they usually contain unethical language expressions or descriptions of violence, but also describe making weapons, taking drugs, and adult content in several cases.

3.2 Model Evaluations

We apply the preprocessing procedure from lm-evaluation-harness (Biderman et al., 2026) and perform two types of evaluations: log-likelihood and by generation. They differ in the way the task is presented. With log-likelihood, we

Parameter	Total, %	AN, %	WH, %
Grammar/typos:			
— Prompt	30.9%	46.1%	23.6%
— Correct option	24.1%	17.3%	27.3%
— Incorrect option(s)	68.6%	54.6%	75.2%
Incoherent:			
— Prompt	4.0%	9.0%	1.6%
— Correct option	0.8%	0.9%	0.8%
Illogical correct answer	3.2%	4.2%	2.7%
Implausible distractors	85.1%	67.8%	93.3%
Ethical concerns	4.7%	2.8%	5.5%
Wrong best answer	1.1%	1.7%	0.9%
Multiple correct options	8.0%	11.7%	6.2%

Table 1: Gemini annotations for the HellaSwag validation set: for the complete set, and by source: from ActivityNet (AN) and WikiHow (WH). The validation set comprises 10042 questions: 3243 (32.3%) from ActivityNet, and 6799 (67.7%) from WikiHow.

Accuracy	0.62
Prompt grammar/typos	0.59
Incoherent prompt	0.54
Correct option grammar/typos	0.52
Incoherent correct option	0.38
Illogical correct option	0.47
Incorrect option(s) grammar/typos	0.65
Implausible incorrect option(s)	0.64

Table 2: Performance of Llama 3.2 1B on full HellaSwag validation set (top row) and on questions with highlighted issues. Full results reported in Appendix E.

directly request the model’s estimate of the plausibility of each option. On the contrary, evaluation by generation allows us to present the model with all options, enabling it to compare and choose the most plausible or the least implausible option. We consider evaluation by generation only for larger models ($\sim 32B$), as their instruct versions are more likely to produce outputs in the desired format. In addition, small model capabilities are often underestimated in prompt-based evaluation (Hu and Levy, 2023) due to auxiliary task demands (Hu and Frank, 2024). We report the used models and the generation prompt in Appendices A and B.

Mean Log-Likelihood. We append each answer option to the question prompt and run each sequence through the model to compute output logits. We then choose the answer with the highest mean log-likelihood (see formula in Appendix F).

Generation. The model is presented with a question and all answer choices and is asked to output the correct answer label (1–4). The language model prompt used for evaluation is presented in

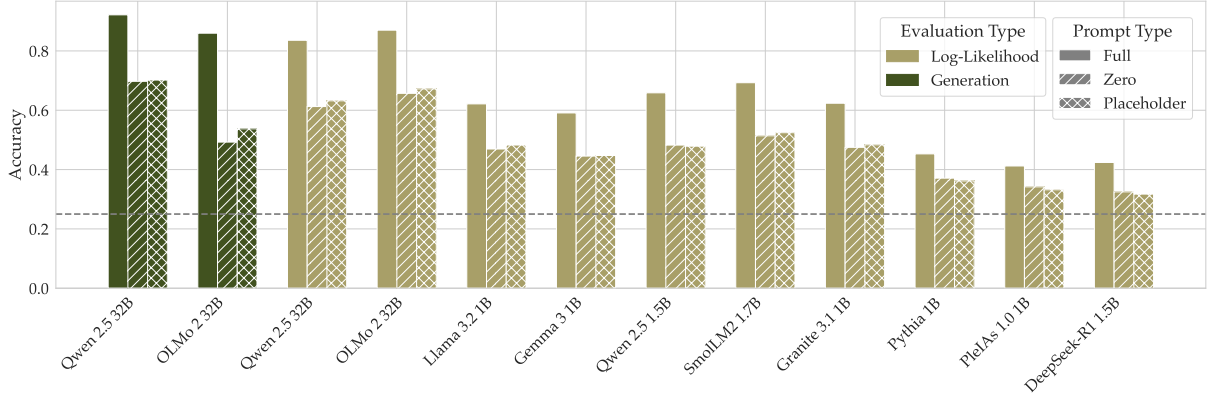


Figure 2: Accuracy evaluation with log-likelihood for larger (32B) and smaller (1-2B) models with a full question prompt, and without it (zero-prompt), and with “*Lorem ipsum*” instead of the prompt (placeholder-prompt).

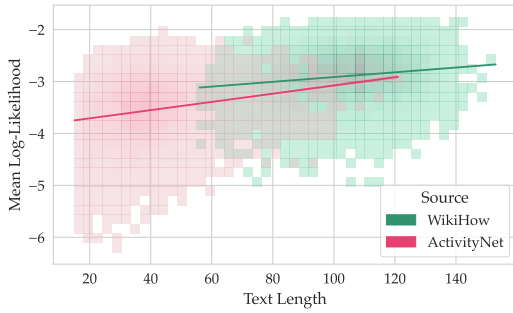


Figure 3: Mean log-likelihood vs answer option lengths. The lines represent the trend for each question source.

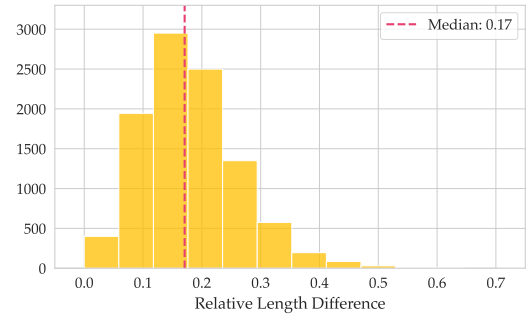


Figure 4: Differences between the longest and the shortest answer options relative to the longest option length.

Appendix B. We use the default generation strategy for each model. To lower the potential impact of contamination, we shuffle the options for each question, following Alzahrani et al. (2024).

3.2.1 Accuracy evaluations

We first evaluate all selected models on the complete HellaSwag and separately on selected annotated categories related to common issues in prompts and answer options. We present the results in Table 2. As hypothesized above, issues with question prompts and correct answer options make the task more challenging, without increasing the question’s value in terms of commonsense reasoning. When the incorrect options are ungrammatical or contain utterly nonsensical expressions, the task becomes simpler. All tested models exhibit similar trends (see Appendix E). We also find that questions from WikiHow are generally simpler for all models (see details in Appendix C).

3.2.2 Answer length

The likelihood of the answers positively correlates with their lengths in number of tokens (see Fig-

ure 3), especially for the questions from ActivityNet. This might be due to the difference in answer option sources: generated distractor options are often shorter than the correct answers. Additionally, every text has very low likelihood values in the beginning. As the prior context grows longer, the following tokens become more predictable, and log-likelihoods increase overall. Therefore, the overall mean token probability is lower for shorter texts. This relationship may also be sensitive to the type of log-likelihood normalization. We separately investigate total and byte-normalized log-likelihoods proposed by Gao (2022) and find that these methods also produce values correlated with the input length (see Appendix F for details).

For HellaSwag questions, this issue is highly relevant. As we show in Figure 4, the relative length difference between answer options, computed as the difference between the longest and the shortest answer options relative to the length of the longest answer option, can be quite high (more than 17% for half of the data). Furthermore, we find that the accuracy of Llama 3.2 1B when the correct answer

Evaluation	Model	Params	Agreement	Agreement type		Disagreement type		
				Both ✓	Both ✗	Full ✓	Zero ✓	Both ✗
Log-Likelihood	Llama 3.2	1B	0.69	0.43	0.26	0.19	0.04	0.07
	Gemma 3	1B	0.70	0.40	0.29	0.19	0.04	0.07
	Qwen 2.5	1.5B	0.69	0.45	0.24	0.21	0.03	0.07
	SmolLM2	1.7B	0.71	0.49	0.23	0.21	0.03	0.05
	Granite 3.1	1B	0.72	0.44	0.28	0.18	0.03	0.06
	Pythia	1B	0.71	0.32	0.39	0.14	0.05	0.10
	PleIAs	1B	0.69	0.28	0.41	0.13	0.06	0.11
	DeepSeek-R1	1.5B	0.65	0.26	0.39	0.16	0.06	0.13
	Qwen 2.5	32B	0.71	0.60	0.12	0.24	0.02	0.03
	OLMo 2	32B	0.74	0.64	0.09	0.23	0.01	0.02
Generation	Qwen 2.5	32B	0.70	0.67	0.03	0.25	0.03	0.02
	OLMo 2	32B	0.50	0.45	0.05	0.41	0.05	0.04

Table 3: The breakdown of zero-prompt agreement (a model gives equal correct or incorrect predictions) and disagreement (the model gives correct prediction only in either full- or zero-prompt scenario or gives different incorrect answers).

Number of models	1	2	3	4	5	6	7	8	9	10
% of questions	75%	67%	60%	55%	50%	45%	39%	34%	27%	18%

Table 4: Zero-prompt core. A pair (N models, X%) means that at least N models can answer correctly X% of questions without question prompt texts.

is the longest one (0.72) is greater than when it is not (0.59). Thus, varied answer lengths pose validity problems for a benchmark, as models implicitly prefer longer answers.

3.2.3 Prompt importance

We test the construct validity of HellaSwag in zero- and placeholder-prompt evaluations. In the **zero-prompt** scenario, we remove the activity label and the question context from the task; in the **placeholder-prompt**, we substitute them with a generic text typically used as a filler:

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Morbi vel venenatis dui. Pellentesque sed cursus massa.

We show an example of the affected part of the question prompt in Figure 1. In commonsense reasoning, the model should figure out which ending stems best from the question prompt. By comparing the results from these evaluations, we can determine whether this has a causal impact on model performance. If a model can choose the correct option without the question text, this compromises the construct validity.

Accuracy. For all models (see Figure 2), accuracies with both prompt perturbations are well above chance (25%), meaning the probabilities of the answer choices are biased towards the correct answer. This presents two problems. First, the question

prompt does not contribute to the task. The model needs only to choose the most plausible text fragment, which does not necessarily reflect commonsense reasoning capabilities. Second, features of the incorrect answers, such as grammatical or logical errors, as discussed in Section 3.1, may increase the probability of the correct option. A model may therefore achieve high accuracy on a subset of the questions by simply ruling out grammatical inconsistencies or nonsensical phrases.

Agreement. We separately test the agreement of predictions with full and modified prompts. Following Lyu et al. (2024), “agreement” here means that the model gives the same predictions in both evaluation modes (with and without the prompt), regardless of whether these predictions are correct or wrong. The results of agreement evaluations are presented in Table 3 for all models. On average, 68% of the model predictions do not change if we remove the question prompt from the evaluation. This holds not only for correct answers but also for a substantial proportion of the incorrect ones, which rules out contamination as the main suspected cause. For Pythia, PleIAs, and DeepSeek-R1 models, the share of agreement for incorrect answers is larger than for the correct ones. This suggests that, for the majority of HellaSwag questions, the question text does not contain additional

Evaluation	Model	Size	Agreement	Agreement type		Disagreement type		
				Both ✓	Both ✗	Full ✓	Lores ✓	Both ✗
Log-Likelihood	Llama-3.2	1B	0.67	0.43	0.25	0.20	0.06	0.08
	Gemma-3	1B	0.69	0.40	0.29	0.19	0.05	0.07
	Qwen-2.5	1.5B	0.65	0.43	0.22	0.23	0.05	0.07
	SmolLM2	1.7B	0.64	0.46	0.18	0.23	0.06	0.07
	Granite-3.1	1B	0.71	0.44	0.27	0.18	0.04	0.07
	Pythia	1B	0.55	0.26	0.30	0.19	0.10	0.15
	PleIAs	1B	0.59	0.24	0.35	0.17	0.09	0.14
	DeepSeek-R1	1.5B	0.62	0.25	0.37	0.18	0.07	0.14
	Qwen-2.5	32B	0.71	0.61	0.10	0.23	0.02	0.03
	OLMo-2	32B	0.73	0.65	0.08	0.22	0.02	0.02
Generation	Qwen-2.5	32B	0.70	0.67	0.03	0.25	0.03	0.02
	OLMo-2	32B	0.54	0.49	0.05	0.37	0.05	0.04

Table 5: Agreement between full-prompt and placeholder-prompt evaluation. Agreement means that a model gives equal predictions in both evaluation modes. We separately report fractions for agreement types (a model gives equal correct or incorrect predictions) and disagreement types (the model gives correct prediction only in either full- or placeholder-prompt scenario or gives different incorrect answers).

information that the model uses to perform the task. Furthermore, for some questions, removing the context actually allowed the model to make a better choice (see column Zero ✓). We observe similar agreement patterns in the placeholder-prompt scenario (see Table 5), with even more questions (up to 10%) answered correctly with a perturbed prompt.

In addition, we find that questions answered correctly without the prompt are shared across many models. In Table 4, we show the percentage of questions answered correctly without the prompt by groups of models. In particular, about 18% of questions are answered correctly without prompts by all 10 tested models. Together with the other results, these findings lead us to question the most fundamental quality of the benchmark — its construct validity. Since the question text does not play a decisive role in $\sim 68\%$ of cases, the benchmark cannot accurately measure commonsense reasoning, as presenting the model with the cause does not help it choose the right effect. This also raises concerns about the answer choices: the incorrect options may often be so blatantly implausible that the correct answer stands out, even without context.

Finally, we test how zero-/placeholder-prompt agreement corresponds to the distance to the nearest neighbor in the training set (see results for Llama 3.1 1B in Figure 5 and Appendix H for other models). We use the all-MiniLM-L6-v2 sentence transformer (Reimers and Gurevych, 2019) and L2 distance. Both comparisons show no significant difference, indicating that if HellaSwag was present in the Llama’s training set, it has no effect on the zero-/placeholder-prompt agreement, ruling

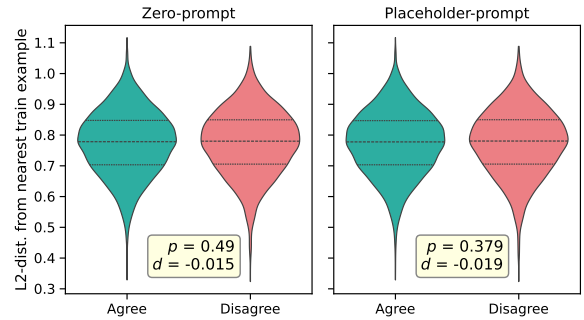


Figure 5: Distribution of L2 distances to near-duplicates in the train set for zero-/placeholder-prompt agreement for Llama 3.2 1B. For each pair, we present the p-value of the t-test and Cohen’s d (both insignificant).

out contamination as a primary alternative explanation. This pattern shows in other models, regardless of their openness and performance (Appendix H).

3.3 Interim Summary

We identify dimensions relevant to construct validity and annotate the questions. We find that questions with issues in prompts or correct answers are harder for the models, whereas those with issues in the distractors make the task simpler. We also see that the question prompt does not play the decisive role in the majority of questions, with 18% of questions being answerable without the prompt for all ten tested models. However, this analysis is limited to a single benchmark. Scaling this analysis to other benchmarks is time-intensive; therefore, in the following section, we propose a tool to quickly identify potential validity issues in commonsense reasoning benchmarks.

4 BenCheck Package

Based on our results and related work, we create and release a package for benchmark validity analysis under a permissive license — **BenCheck**². In this package, we combine the main tests from our case study with several other relevant works on benchmark validation. Each of the implemented checks is accompanied by diagnostics and a detailed report with metric explanation. In particular, we include the following customizable checks:

- **Answer lengths:** we compare answer lengths of correct and incorrect answer options, relative answer length difference, and also compute the proportion of cases when the correct answer is the longest or the shortest one.
- **Enumeration bias:** following the work of Alzahrani et al. (2024), we check the distribution of correct answers among option positions as well as model preferences or biases.
- **Prompt importance:** following our evaluations with zero-prompt and placeholder-prompt (Section 3.2.3), we compare the scores with the original and removed or replaced question text, thereby evaluating the question prompt importance and construct validity of the benchmark.
- **Placeholder answer:** we replace the correct answer option with "none of the above" (NOTA), as in the work of Salido et al. (2025), and measure changes in model performance. For example, if a model changes from an incorrect answer to 'NOTA', the original correct answer may present issues.
- **Grammar and typos:** We provide an interface for any open LLM with guided decoding or optionally GECToR (Omelianchuk et al., 2020), a grammar correction model, for detecting grammatical errors and typos in question prompts, correct and incorrect answer options.

In Table 6, we present main metrics from BenCheck reports for commonsense reasoning benchmarks: Winogrande (Sakaguchi et al., 2021), PIQA (Bisk et al., 2020), and the English subset from Global PIQA (Chang et al., 2025), a manually corrected, multilingual version of the dataset.

²<https://github.com/Pleias/BenCheck>

Metric	HellaSwag	Winogrande	PIQA	G-PIQA
Prompt importance				
Accuracy	0.792	0.589	0.864	0.846
Zero score	0.646	0.498	0.771	0.797
Agreement	0.588	0.496	0.813	0.770
Lorem score	0.523	0.490	0.714	0.700
Agreement	0.564	0.510	0.800	0.700
Grammar issues				
Question	0.35	0.39	0.20	0.03
Golden	0.12	0.00	0.13	0.12
Distractors	0.60	0.00	0.34	0.15
Answer lengths				
Mean rel diff	0.43	0.01	0.08	0.02
✓ is longest	0.23	0.12	0.24	0.21
✓ is shortest	0.23	0.13	0.30	0.16
Enumeration bias				
A ✓	2505	628	910	50
B ✓	2517	639	928	50
C ✓	2552	—	—	—
D ✓	2468	—	—	—
None of the above (NOTA)				
✓ → NOTA	0.06	0.05*	0.08*	0.06*
✗ → NOTA	0.32	0.29*	0.24*	0.28*

Table 6: BenCheck evaluation of commonsense reasoning benchmarks. We perform model evaluations with Gemma 3 12B (Gemma Team et al., 2025) and grammatical issue detection with Gemma 4 E4B (Google, 2026).

*The benchmark questions have only two options.

Winogrande exhibits the most grammatical errors in question prompts, with no signs of grammatical issues in answer options, given its single-word answers. For the same reason, it is the only benchmark robust to zero-prompt and placeholder-prompt evaluations. Compared to PIQA, Global PIQA presents fewer grammatical issues, which confirms better dataset curation, while still being solvable without the question prompts. Similarly to Section 3.2.1, we present additional model evaluations based on grammar issues present in benchmark parts in Appendix G. HellaSwag’s case of varied answer lengths is absent in other benchmarks; in terms of position bias, all benchmarks have uniformly distributed options. In a NOTA scenario, the task becomes more complicated, and the models rarely choose the correct answer, which might also indicate that, in some cases, the distractors are suitable. At the same time, in 24-32% of cases, the model chooses NOTA instead of a previously incorrect answer, which indicates that the correct answer might present issues.

5 Discussion

Our experiments highlight a variety of artifacts in HellaSwag. They range from poor language to

flaws in question design. One of the primary drivers of these issues is the outdated generation process. ActivityNet and WikiHow, the two subsets of HellaSwag, show different patterns. ActivityNet has generally shorter answer options and more grammatical issues. We hypothesize that this is due to the original source of the questions, the generated passages that describe the video content.

Even though the benchmark is generally considered challenging, this is mainly due to grammatical errors, typos, illogical and implausible constructions in many questions. Even though such difficulties can also test a model’s capabilities, they are orthogonal to commonsense reasoning. Thus, under our diagnostics, HellaSwag is a weak proxy for commonsense reasoning, as it is not possible to isolate this capability judging solely by HellaSwag scores. Although implausible constructions allow the model to rule out incorrect options, effectively mimicking how humans write tests, we argue that this should occur only in connection with question prompts, not within answer options in isolation.

Nevertheless, given that the intended correct option is still the most plausible (or the least problematic) in the absolute majority of items, the benchmark still presents at least some useful signal for model capabilities; these capabilities, however, are heavily conflated, and HellaSwag alone cannot be used for selecting the best model, especially as a proxy for commonsense reasoning specifically.

We encapsulate our findings, along with other relevant work, into a package for benchmark validity audits, a useful tool for screening and identifying critical issues. BenCheck checks can also be applied to MCQ benchmarks in general. In some cases, BenCheck can also be used to filter out faulty items from the benchmark. However, it must be ensured that the remaining items are enough for a comprehensive evaluation. For instance, filtering out problematic items from HellaSwag would remove the absolute majority of ActivityNet, effectively reducing the benchmark to tutorial-like items from WikiHow.

6 Requirements.

Based on our findings, we formulate a list of requirements that a valid modern commonsense reasoning benchmark should meet:

- **Grammar and typos.** All the texts should be grammatically correct, including the incorrect options. If the model is to be tested for

robustness to bad grammar or typos, a separate version of the dataset should be created, e.g., by injecting errors or typos.

- **Sensicality and prompt importance.** All options should be a reasonable, coherent text. Even though, by the nature of the task, incorrect options make no sense, this should be only due to how they relate to the question context and not, for instance, due to single implausible constructions. Apart from this, the question prompt should play a role in determining the correct answer.
- **Distinct correct answers.** The correct option should be a substantially better fit for the question prompt than the incorrect options. There might be more than one correct option if the task design allows it, but they should all be comparably suitable.
- **Uniform option lengths.** The answers can have different lengths, though this variability should be limited. The length rank of correct answers should be uniform over the dataset. The intended evaluation procedure should consider the potential length bias.
- **Content.** The questions should be filtered for toxicity and ethical issues. For smaller models, the questions should preferably be on general-domain topics, excluding overly specific professional knowledge.

7 Conclusion

In this paper, we describe numerous issues in HellaSwag, ranging from grammatical errors and typos to ambiguous answer choices, and argue that these issues compromise the construct validity of the benchmark. Evaluations with perturbed prompts show that in most of the HellaSwag questions, question text does not affect model predictions, which allows us to conclude that HellaSwag performance does not necessarily indicate commonsense reasoning capabilities. We specifically highlight that commonsense reasoning is a salient characteristic, and there should be a raised research interest in constructing a good benchmark to measure it. We extend our findings to other benchmarks and propose an extensive package for benchmark validity analysis, providing an evaluation toolkit that guides the development of benchmarks.

Limitations

The proposed methodology is diverse and requires different tools and different hardware for different checks. Model-related checks require models that can complete the task above a random chance and follow the instructions, as in the case with the generative evaluation; they also require GPU resources. Some of the checks are not necessarily indicative in all cases. For example, length or grammar checks are not representative of benchmarks with factual answers or the answers consisting of a single word. Additionally, evaluations of grammar relying on language models have their limitations. Customized grammatical error detection with specialized models is possible (e.g., as with GECToR), but the benchmark format should be carefully tailored for such evaluations.

The interpretation of some evaluations involves multiple possible explanations due to potential contamination of the model by the benchmark questions or content. We argue that our results show this behavior is present even without contamination (e.g., in models trained on open datasets); however, we cannot completely rule out this issue in closed-data models.

BenCheck is a useful tool for initial smoke testing of the benchmark. However, it is more of an audit toolkit that provides the aggregated picture rather than a precise filtering instrument. BenCheck-guided filtering is possible but has limitations, primarily in the choice of relevant criteria and the potential specificity of the remaining data. For instance, if HellaSwag were to be BenCheck-filtered, most ActivityNet examples would be removed, leaving only a subset of WikiHow. Though it is likely of higher quality, this dataset would represent only a narrow set of specifically formatted tutorials, which does not tell much about the model’s capabilities.

Finally, though it has multiple issues and compromised construct validity, HellaSwag may still provide a useful engineering signal. The models with HellaSwag performance above random show better capabilities, which, when combined with other benchmarks on the leaderboards, might yield useful rankings in certain cases.

Acknowledgments

The authors of this work acknowledge the HPC resource allocation by Erlangen National High-Performance Computing Center (NHR@FAU)

of the Friedrich-Alexander-Universität Erlangen-Nürnberg (FAU) (joint project with the Center for Artificial Intelligence (CAIRO), THWS) and Jean Zay (compute grant #GC011015451). The authors would also like to thank Catherine Arnett for her thoughtful ideas and continuous help throughout this work, and Andrey Indukaev for manual annotations and useful ideas.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.
- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, Joshua Lochner, Caleb Fahlgren, Xuan-Son Nguyen, Clémentine Fourrier, Ben Burtenshaw, Hugo Larcher, Haojun Zhao, Cyril Zakka, Mathieu Morlon, and 3 others. 2025. [SmolLM2: When Smol Goes Big – Data-Centric Training of a Small Language Model](#). *Preprint*, arXiv:2502.02737.
- 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.
- Nishant Balepur, Bhavya Rajasekaran, Hyunjin Jane Oh, Michael Xie, Atrey Desai, Vipul Gupta, Steven James Moore, Eunsol Choi, Rachel Rudinger, and Jordan Lee Boyd-Graber. 2026. [BenchMarker: An education-inspired toolkit for highlighting flaws in multiple-choice benchmarks](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15793–15824, San Diego, California, United States. Association for Computational Linguistics.
- Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. 2024. [Artifacts or Abduction: How Do LLMs Answer Multiple-Choice Questions Without the Question?](#) In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10308–10330, Bangkok, Thailand. Association for Computational Linguistics.
- Nishant Balepur and Rachel Rudinger. 2024. [Is Your Large Language Model Knowledgeable or a Choices-](#)

- Only Cheater?** In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 15–26, Bangkok, Thailand. Association for Computational Linguistics.
- Andrew M. Bean, Ryan Othniel Kearns, Angelika Romanou, Franziska Sofia Hafner, Harry Mayne, Jan Batzner, Negar Foroutan, Chris Schmitz, Karolina Korgul, Hunar Batra, Oishi Deb, Emma Beharry, Cornelius Emde, Thomas Foster, Anna Gausen, Maria Grandury, Simeng Han, Valentin Hofmann, Lujain Ibrahim, and 23 others. 2025. **Measuring what Matters: Construct Validity in Large Language Model Benchmarks**. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, and 1 others. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *International Conference on Machine Learning*, pages 2397–2430. PMLR.
- Stella Biderman, Hailey Schoelkopf, Lintang Sutawika, Leo Gao, Jonathan Tow, Baber Abbasi, Alham Fikri Aji, Pawan Sasanka Ammanamanchi, Sidney Black, Jordan Clive, Anthony DiPofi, Julien Etzaniz, Benjamin Fattori, Jessica Zosa Forde, Charles Foster, Jeffrey Hsu, Mimansa Jaiswal, Wilson Y. Lee, Haonan Li, and 11 others. 2026. **Lessons from the Trenches on Reproducible Evaluation of Language Models**. Preprint, arXiv:2405.14782.
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2020. PIQA: Reasoning about Physical Commonsense in Natural Language. In *Thirty-Fourth AAAI Conference on Artificial Intelligence*.
- Borhane Blili-Hamelin, Christopher Graziul, Leif Hancox-Li, Hananel Hazan, El-Mahdi El-Mhamdi, Avijit Ghosh, Katherine A Heller, Jacob Metcalf, Fabricio Murai, Eryk Salvaggio, Andrew J Smart, Todd Snider, Mariame Tighanimine, Talia Ringer, Margaret Mitchell, and Shiri Dori-Hacohen. 2025. **Position: Stop treating ‘AGI’ as the north-star goal of AI research**. In *Forty-second International Conference on Machine Learning Position Paper Track*.
- D.T. Campbell and J.C. Stanley. 1963. *Experimental and Quasi-experimental Designs for Research*. Rand McNally.
- Tyler A. Chang, Catherine Arnett, Abdelrahman Eldes-okey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, Aditi Gupta, Afitab Iyigun, Afonso Simplicio, Ahmed Essouaied, Aicha Chorana, Akhil Eppa, Akintunde Oladipo, Akshay Ramesh, Aleksei Dorkin, and 319 others. 2025. **Global PIQA: Evaluating Physical Commonsense Reasoning Across 100+ Languages and Cultures**. Preprint, arXiv:2510.24081.
- Edwin Chen. 2023. **HellaSwag or HellaBad? 36% of This Popular LLM Benchmark Contains Errors**. Accessed: 2025-03-27.
- Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sashank Tsveyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, and 48 others. 2023. PaLM: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).
- Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.
- Gheorghe Comanici, Eric Bieber, Mike Schaeckermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3416 others. 2025. **Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities**. Preprint, arXiv:2507.06261.
- LJ Cronbach and PE Meehl. 1955. Construct validity in psychological tests. *Psychological Bulletin*, 52(4):281–302.
- Ernest Davis. 2023. Benchmarks for automated commonsense reasoning: A survey. *ACM Computing Surveys*, 56(4):1–41.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. **DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning**. Preprint, arXiv:2501.12948.
- Leo Gao. 2022. **Multiple Choice Normalization**.
- Aryo Pradipta Gema, Joshua Ong Jun Leang, Giwon Hong, Alessio Devoto, Alberto Carlo Maria Mancino, Rohit Saxena, Xuanli He, Yu Zhao, Xiaotang Du, Mohammad Reza Ghasemi Madani, Claire Barale, Robert McHardy, Joshua Harris, Jean Kaddour, Emile Van Krieken, and Pasquale Minervini. 2025. **Are We Done with MMLU?** In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5069–5096, Albuquerque, New Mexico. Association for Computational Linguistics.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin,

- Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. **Gemma 3 Technical Report**. *Preprint*, arXiv:2503.19786.
- Google. 2026. Welcome Gemma 4: Frontier Multimodal Intelligence on Device. <https://huggingface.co/blog/gemma4>. Hugging Face Blogpost.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. **The Llama 3 Herd of Models**. *Preprint*, arXiv:2407.21783.
- Eliya Habba, Ofir Arviv, Itay Itzhak, Yotam Perlitz, Elron Bandel, Leshem Choshen, Michal Shmueli-Scheuer, and Gabriel Stanovsky. 2025. **DOVE: A large-scale multi-dimensional predictions dataset towards meaningful LLM evaluation**. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 11744–11763, Vienna, Austria. Association for Computational Linguistics.
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. **Measuring Massive Multitask Language Understanding**. In *International Conference on Learning Representations*.
- Jennifer Hu and Michael Frank. 2024. **Auxiliary task demands mask the capabilities of smaller language models**. In *First Conference on Language Modeling*.
- Jennifer Hu and Roger Levy. 2023. **Prompting is not a substitute for probability measurements in large language models**. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.
- IBM. 2024. **IBM Unveils Granite 3.1: Powerful Performance, Long Context, and More**.
- Pride Kavumba, Benjamin Heinzerling, Ana Brassard, and Kentaro Inui. 2021. **Learning to Learn to be Right for the Right Reasons**. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3890–3898, Online. Association for Computational Linguistics.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715.
- Pierre-Carl Langlais, Pavel Chizhov, Mattia Nee, Carlos Rosas Hinostroza, Matthieu Delsart, Irène Girard, Anastasia Stasenko, and Ivan P. Yamshchikov. 2025. **Pleias 1.0: the First Ever Family of Language Models Trained on Fully Open Data**. *Procedia Computer Science*, 267:146–156. Proceedings of the Third EuroHPC user day.
- Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d’Autume, Phil Blunsom, and Aida Nematzadeh. 2022. **A Systematic Investigation of Commonsense Knowledge in Large Language Models**. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. 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.
- Chenyang Lyu, Minghao Wu, and Alham Aji. 2024. **Beyond Probabilities: Unveiling the Misalignment in Evaluating Large Language Models**. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 109–131, Bangkok, Thailand. Association for Computational Linguistics.
- Shashi Narayan, Shay B. Cohen, and Mirella Lapata. 2018. **Don’t Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization**. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1797–1807, Brussels, Belgium. Association for Computational Linguistics.
- Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2025. **2 olmo 2 furious**. *Preprint*, arXiv:2501.00656.
- Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzchanskyi. 2020. **GECToR – Grammatical Error Correction: Tag, Not Rewrite**. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.
- Guilherme Penedo, Hynek Kydlíček, Vinko Sabolčec, Bettina Messmer, Negar Foroutan, Amir Hossein Kargaran, Colin Raffel, Martin Jaggi, Leandro Von Werra, and Thomas Wolf. 2025. **FineWeb2: One Pipeline to Scale Them All — Adapting Pre-Training Data Processing to Every Language**. In *Second Conference on Language Modeling*.

- Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.
- Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks](#). 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.
- Anka Reuel, Amelia Hardy, Chandler Smith, Max Lamparth, Malcolm Hardy, and Mykel Kochenderfer. 2024. [BetterBench: Assessing AI Benchmarks, Uncovering Issues, and Establishing Best Practices](#). In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavathula, and Yejin Choi. 2021. [WinoGrande: an adversarial winograd schema challenge at scale](#). *Commun. ACM*, 64(9):99–106.
- Eva Sánchez Salido, Julio Gonzalo, and Guillermo Marco. 2025. [None of the Others: a General Technique to Distinguish Reasoning from Memorization in Multiple-Choice LLM Evaluation Benchmarks](#). *Preprint*, arXiv:2502.12896.
- Krunal Shah, Nitish Gupta, and Dan Roth. 2020. [What do we expect from multiple-choice QA systems?](#) In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3547–3553, Online. Association for Computational Linguistics.
- Team Olmo, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers, Hamish Ivison, and 1 others. 2025. Olmo 3. *arXiv preprint arXiv:2512.13961*.
- Paul Trichelair, Ali Emami, Adam Trischler, Kaheer Suleman, and Jackie Chi Kit Cheung. 2019. [How Reasonable are Common-Sense Reasoning Tasks: A Case-Study on the Winograd Schema Challenge and SWAG](#). 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 3382–3387, Hong Kong, China. Association for Computational Linguistics.
- Joshua Vendrow, Edward Vendrow, Sara Beery, and Aleksander Madry. 2025. [Do Large Language Model Benchmarks Test Reliability?](#) *arXiv preprint arXiv:2502.03461*.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. [Qwen2.5 Technical Report](#). *Preprint*, arXiv:2412.15115. ArXiv preprint.
- 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.

A Models

In Table 7, we present the models we used for evaluations in the paper with their corresponding HuggingFace repository paths.

B Generation Prompt

We use the following prompt to evaluate the models by generation:

Generation Prompt

You are given a situation followed by four possible endings. Choose the most appropriate ending by selecting the corresponding number. Respond only with the number of the correct answer.

Context: Roof shingle removal: A man is sitting on a roof. He

1. is using wrap to wrap a pair of skis.
2. is ripping level tiles off.
3. is holding a Rubik’s cube.
4. starts pulling up roofing on a roof.

Answer:

C Accuracy by Source

In Figure 6, we show the differences in accuracies by question sources. ActivityNet part of the data, which is a suspect for lower-quality questions in HellaSwag, is more challenging to the models. However, taking all our results into account, this complexity is not a compelling benchmark feature, but rather low-quality data that produces unwanted noise in the scores.

Interestingly, the zero-prompt accuracy drop in ActivityNet is also larger than that for the WikiHow. One of the reasons for this might be that WikiHow has generally longer options, from which it is easier to get a general understanding of the situation.

D Human Validation

We collected two sets of human annotations for 100 questions from HellaSwag. The results of human annotators’ assessment compared to LLM labels

Model	Size	HuggingFace repository
LLaMA 3.2 (Grattafiori et al., 2024)	1B	meta-llama/Llama-3.2-1B
Qwen 2.5 (Yang et al., 2025)	1.5B	Qwen/Qwen2.5-1.5B
Granite 3.1 (IBM, 2024)	1B	ibm-granite/granite-3.1-1b-a400m-base
Gemma 3 (Gemma Team et al., 2025)	1B	google/gemma-3-1b-pt
SmolLM2 (Allal et al., 2025)	1.7B	HuggingFaceTB/SmolLM2-1.7B
DeepSeek-R1 (DeepSeek-AI et al., 2025)	1.5B	deepseek-ai/DeepSeek-R1-Distill-Qwen-1.5B
Pythia (Biderman et al., 2023)	1B	EleutherAI/pythia-1b
Pleias 1.0 (Langlais et al., 2025)	1B	PleIAs/Pleias-1b-Preview
Qwen 2.5 (Yang et al., 2025)	32B	Qwen/Qwen2.5-32B-Instruct
OLMo 2 (OLMo et al., 2025)	32B	allenai/OLMo-2-0325-32B-Instruct

Table 7: Model names and their corresponding HuggingFace repositories.

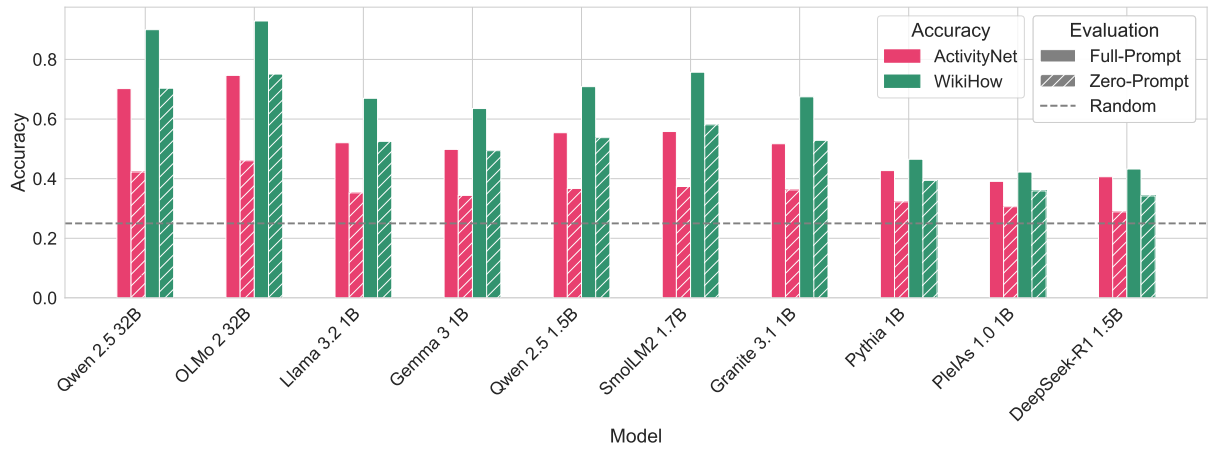


Figure 6: Accuracy of the models evaluated with log-likelihood maximization split by the source of questions.

Criterion	LLM – A1		LLM – A2		A1 – A2	
	Agreement	F1	Agreement	F1	Agreement	α
1. Prompt grammar/typos	0.76	0.66	0.80	0.63	0.78	0.49
2. Correct option grammar/typos	0.82	0.64	0.81	0.60	0.86	0.65
3. Incorrect option(s) grammar/typos	0.72	0.77	0.69	0.72	0.73	0.48
4. Incoherent prompt*	0.86	0.12	0.87	0.13	0.88	0.30
5. Incoherent correct option*	0.90	0.00	0.94	0.00	0.89	0.24
6. Illogical correct option*	0.92	0.33	0.91	0.00	0.88	-0.05
7. Implausible incorrect option(s)	0.79	0.88	0.47	0.52	0.40	-0.24
8. Ethical concern*	0.93	0.22	0.90	0.17	0.89	0.41
9. Best option	0.89	—	0.95	—	0.86	—
10. Valid alternatives	0.03	—	0.03	—	0.07	—
11. Worst option	0.59	—	0.54	—	0.59	—

Table 8: Annotation agreement for binary criteria. We show LLM’s agreement with every annotator (A1 and A2) along with F1 score, which treats human annotations as ground truth, and inter-annotator agreement together with Krippendorff’s α . *These categories received small number of positive human annotations (ranging from 3 to 9 records out of 100).

	Log-Likelihood										Generation	
	Llama 3.2 1B	Gemma 3 1B	Qwen 2.5 1.5B	SmoLM2 1.7B	Granite 3.1 1B	Pythia 1B	PleIAs 1B	DeepSeek-R1 1.5B	Qwen 2.5 32B	OLMo 2 32B	Qwen 2.5 32B	OLMo 2 32B
Accuracy	0.62	0.59	0.66	0.69	0.62	0.45	0.41	0.42	0.84	0.87	0.92	0.86
Prompt grammar/typos	0.59	0.56	0.62	0.64	0.58	0.44	0.41	0.42	0.79	0.82	0.9	0.83
Incoherent prompt	0.54	0.53	0.56	0.57	0.53	0.41	0.41	0.41	0.73	0.75	0.86	0.75
Correct option grammar/typos	0.52	0.49	0.56	0.61	0.53	0.34	0.31	0.32	0.79	0.83	0.93	0.86
Incoherent correct option	0.38	0.36	0.41	0.43	0.35	0.26	0.22	0.23	0.59	0.77	0.81	0.68
Illogical correct option	0.47	0.44	0.48	0.5	0.45	0.36	0.33	0.33	0.67	0.75	0.67	0.59
Incorrect option(s) grammar/typos	0.65	0.62	0.69	0.72	0.65	0.48	0.43	0.45	0.86	0.89	0.93	0.87
Implausible incorrect option(s)	0.64	0.61	0.68	0.71	0.64	0.46	0.42	0.43	0.86	0.89	0.93	0.87

Table 9: Performance of all evaluated models on full HellaSwag validation set (top row) and on questions with highlighted issues.

are shown in Table 8. In most common issues, LLM agrees with at least one of the annotators for the majority of questions, while sometimes the annotators agree with each other less. This might be because of subjectivity of certain issues and difference in expertise. For some criteria, namely 4, 5, 6, and 8, the set of positive labels from humans and LLM was small (less than 10), thus the balanced agreement metrics are low. We investigated separately 20 questions marked with a positive label by the model, and found that at least 60% of records were labeled correctly for each category (up to 80% in the ethical concerns).

For the valid alternatives, we computed agreement using intersection over union metric. Both valid alternatives and worst option categories were largely subjective, therefore we did not use them in our further analysis.

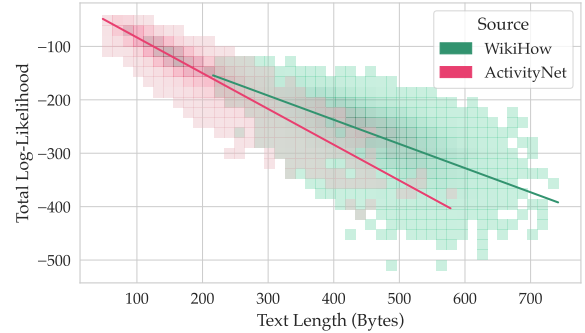
E Accuracy vs Annotations

In Table 9, we show the performance of all models on the complete benchmark, along with the subsets corresponding to each category of issue annotations. The pattern, where a model scores better on questions with issues in distractors and worse on questions with problematic correct answers or question prompts, is consistent across all models.

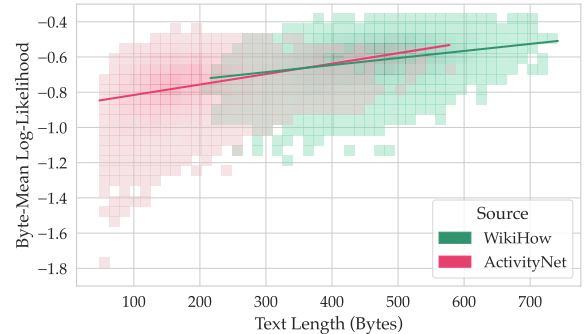
F Log-Likelihood Evaluation

We compute the mean log-likelihood using the following formula:

$$\mathcal{L} = \frac{1}{|\mathcal{S}|} \sum_{t \in \mathcal{S}} \log P(y_t | x_{<t}), \quad (1)$$



(a) Total log-likelihood vs text length in bytes.



(b) Byte-normalized log-likelihood vs text length in bytes.

Figure 7: Joint distributions of likelihoods and lengths of the HellaSwag validation set questions for different variants of likelihood computation. The lines represent the trend for each question source.

where y_t is the ground truth token at position t , $x_{<t}$ is the preceding sequence of tokens, and $\mathcal{S}$ is the sequence of non-special token positions. The resulting value will have non-positive values, and the higher it is, the more plausible the option is, according to the model.

We also compare the correlation with an-

swer lengths for two other formulations of log-likelihood. In Figure 7, we show joint distributions for **total** log-likelihood:

$$\mathcal{L}_t = \sum_{t \in \mathcal{V}} \log P(y_t | x_{<t}), \quad (2)$$

and **byte-normalized** log-likelihood:

$$\mathcal{L}_b = \frac{1}{B} \sum_{t \in \mathcal{V}} \log P(y_t | x_{<t}), \quad (3)$$

where y_t is the ground truth token at position t , $x_{<t}$ is the preceding sequence of tokens, B is the sequence length in bytes, and $\mathcal{V}$ is the set of valid (non-special) token positions. As with the mean log-likelihood, there is a visible correlation between answer length and log-likelihood in both cases (negative, in Figure 7a).

G Grammar Issues

The observations made in Section 3.2.1 for HellaSwag are present in other commonsense reasoning benchmarks as well (see Figure 8). Winogrande has no issues with answer options, whereas the issues in the question prompt make the task harder. In PIQA and Global PIQA, grammar issues in the distractors make the task easier, with a similar pattern in the golden answers. A potential reason is that there is overlap between these issues in the benchmark items (see Figure 9). The issues in the question prompts in Global PIQA also decrease the model’s accuracy, yet this part has a large error bar, having only three questions as a support set.

H Near-duplicate Contamination

In Figures 10, 11, 12, and 13, we show additional example of L2 distances distribution between train and validation sets in HellaSwag. The same pattern of no significant difference is showing in the models with both closed data and open data.

I Annotations

In Figure 14, we report the prompt we used for annotations with Gemini 2.5 Flash.

J LLM Usage Statement

In this work, we utilized AI tools for paraphrasing and grammar correction (Grammarly) and assistance with data visualization design (Gemini, ChatGPT, Claude Code).

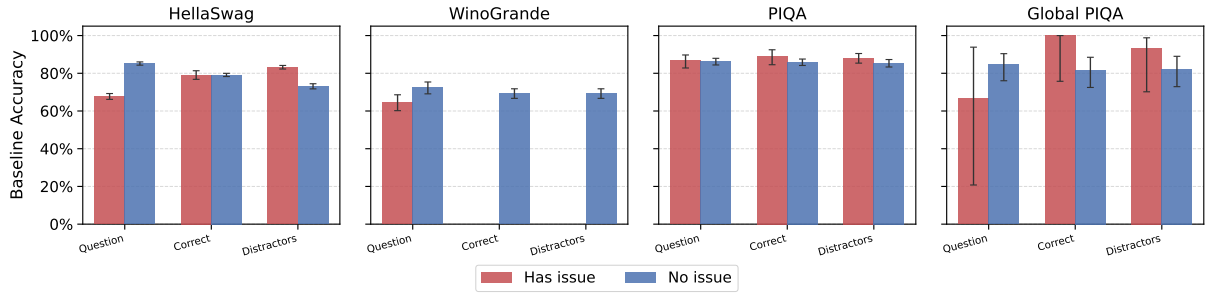


Figure 8: Accuracy of Gemma 3 12B (Gemma Team et al., 2025) on four commonsense reasoning benchmarks, split by the presence of grammar issues.

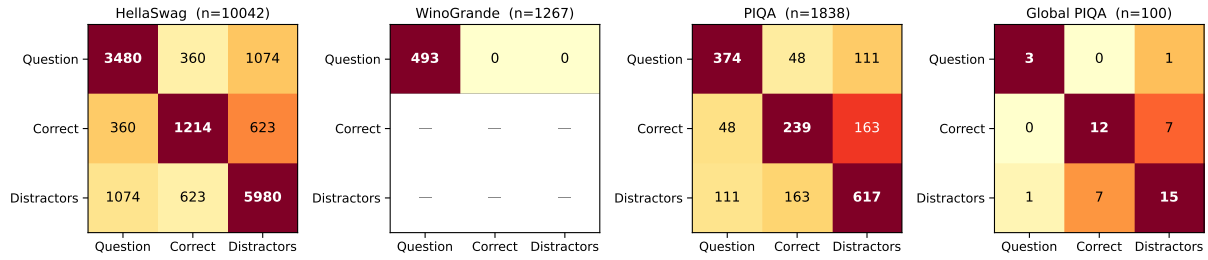


Figure 9: Grammatical issue overlap in commonsense reasoning benchmarks. Grammar issues are evaluated via guided decoding from Gemma 4 E4B (Google, 2026).

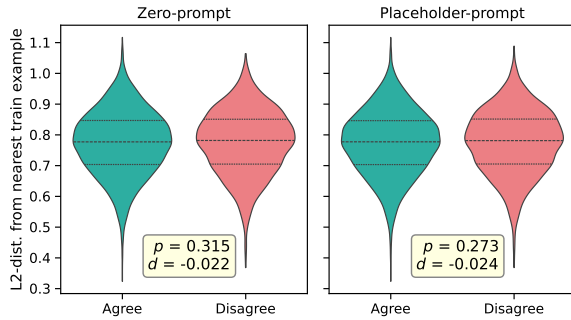


Figure 10: Distribution of L2 distances to near-duplicates in the train set for zero-/placeholder-prompt agreement for Gemma 3 1B. For each pair, we present the p-value of the t-test and Cohen's d.

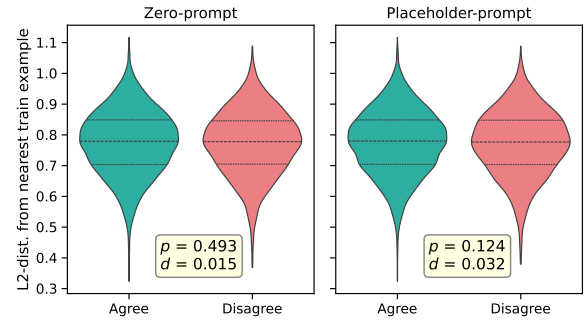


Figure 12: Distribution of L2 distances to near-duplicates in the train set for zero-/placeholder-prompt agreement for SmolLM2 1.7B. For each pair, we present the p-value of the t-test and Cohen's d.

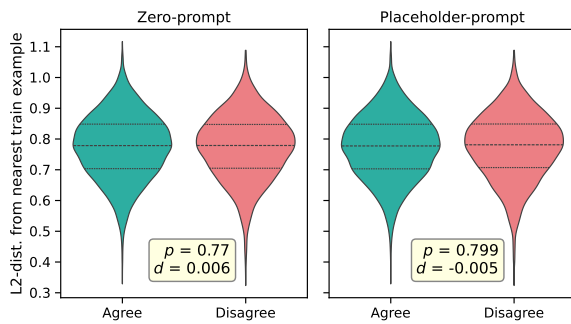


Figure 11: Distribution of L2 distances to near-duplicates in the train set for zero-/placeholder-prompt agreement for Qwen2.5 1.5B. For each pair, we present the p-value of the t-test and Cohen's d.

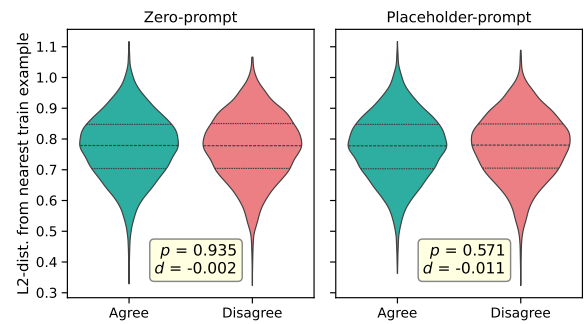


Figure 13: Distribution of L2 distances to near-duplicates in the train set for zero-/placeholder-prompt agreement for PleIAs 1.0 1B. For each pair, we present the p-value of the t-test and Cohen's d.

You are an expert linguist and professional data annotator. Your task is to analyze a multiple-choice question from the "HellaSwag" commonsense reasoning benchmark. You must output your analysis in strict JSON format.

Context

Here is the dataset entry to analyze:

Question Context: {data['query']}

Answer Options:
{options_text}

The Intended Correct Answer Option: {correct_letter}

Instructions

Analyze the validity of this question based on the criteria below.

Criteria Definitions

1. **prompt_grammar_error**: (1/0) The question context has grammatical errors or typos.
2. **correct_grammar_error**: (1/0) The intended correct answer text has grammatical errors or typos.
3. **incorrect_grammar_error**: (1/0) At least one incorrect answer option has grammatical errors or typos.
4. **prompt_incoherent**: (1/0) The question context is incoherent or nonsensical.
5. **correct_incoherent**: (1/0) The intended correct answer text is incoherent or nonsensical.
6. **correct_illogical**: (1/0) The intended correct answer makes no logical sense *specifically in the context of the question*.
7. **incorrect_implausible**: (1/0) At least one incorrect answer option is *fundamentally incoherent, structurally nonsensical, or contains clear evidence of hallucinated or garbled text (e.g., random word sequences, impossible object names, or distinct incoherent phrases or constructions the sense of which is hard to establish)*. This assessment is based on the option's internal validity, independent of the question's context.
8. **ethical_concern**: (1/0) At least one component of the question (context or any option) raises serious ethical concerns (e.g., promoting hate speech, illegal acts, or harmful stereotypes).

Qualitative Analysis

9. **best_option**: Which option (A/B/C/D) is the absolute best fit?
10. **valid_alternatives**: List other options (A/B/C/D) that fit equally well. If none, strictly return "None".
11. **worst_option**: Which option fits the prompt the least?

Output Format

You must respond with a single valid JSON object. Do not include markdown code blocks. For example:

```
{
  "prompt\_grammar\_error": 1,
  "correct\_grammar\_error": 0,
  "incorrect\_grammar\_error": 1,
  "prompt\_incoherent": 0,
  "correct\_incoherent": 0,
  "correct\_illogical": 0,
  "incorrect\_implausible": 1,
  "ethical\_concern": 0,
  "best\_option": "A",
  "valid\_alternatives": ["C", "D"],
  "worst\_option": "B"
}
```

Figure 14: Prompt for LLM annotations.