From Test-Taking to Test-Making: Examining LLM Authoring of Commonsense Assessment Items

Melissa Roemmele
Midjourney
San Francisco, CA, USA
mroemmele@midjourney.com

Andrew S. Gordon
Institute for Creative Technologies
University of Southern California
Los Angeles, CA, USA
gordon@ict.usc.edu

Abstract

LLMs can now perform a variety of complex writing tasks. They also excel in answering questions pertaining to natural language inference and commonsense reasoning. Composing these questions is itself a skilled writing task, so in this paper we consider LLMs as authors of commonsense assessment items. We prompt LLMs to generate items in the style of a prominent benchmark for commonsense reasoning, the Choice of Plausible Alternatives (COPA). We examine the outcome according to analyses facilitated by the LLMs and human annotation. We find that LLMs that succeed in answering the original COPA benchmark are also more successful in authoring their own items.

1 Introduction

Large Language Models (LLMs) can perform complex writing tasks ranging from paraphrasing sentences to composing long-form stories. Because success on these open-ended authoring tasks is hard to measure quantitatively, NLP researchers tend to focus on more constrained tasks when judging the abilities of LLMs. Many of these assessment tasks, or *benchmarks*, are conceptually connected with capabilities relevant to authoring, but they use a specific data design to support quantitative evaluation. For instance, assessments of commonsense reasoning often involve multiple-choice question answering, and accuracy on these questions is a proxy indicator for a model's ability to write coherently. New LLMs have excelled on these types of evaluations, but how this success relates to the LLMs' authoring capabilities is still unclear.

Creating these assessments is itself a skilled writing task, which up to this point has been performed by human authors. Given their success in answering assessments, in this paper we examine whether LLMs can author assessment items as well. We prompt LLMs to generate items in the style of one well-known benchmark for commonsense reasoning, the Choice of Plausible Alternatives (COPA). We examine the outcome in terms of the LLMs' responses to their own generated items as well to what degree the items meet the benchmark authoring standards as judged by human raters. We show that an LLM's authoring success is associated with its success in answering the original benchmark.

2 COPA

The Choice of Plausible Alternatives (COPA) (Roemmele et al., 2011) is an English-language benchmark that assesses the task of commonsense causal reasoning. As shown by the examples in Table 1, each item in COPA consists of a *premise*, a question specifying a causal *direction* (forwards or backwards), and two *alternatives*. One alternative is considered more plausible than the other with regard to the question. For the forwards direction, the question elicits the alternative that is the more plausible effect (result) of the premise, whereas for the backwards direction the question elicits the more plausible cause of the premise. Human performance on this benchmark is considered 100%. The order of the alternatives is balanced across the test set such that random guessing yields 50% accuracy.

Premise: The girl received a trophy.
What was the *cause* of this?
Alternative 1: She won a spelling bee.
Alternative 2: She made a new friend.
Premise: I tipped the bottle.
What happened as a *result*?
Alternative 1: The liquid in the bottle froze.
Alternative 2: The liquid in the bottle poured out.

Table 1: Examples of COPA items

From an authoring perspective, creating a COPA item involves writing a collection of one-sentence narrations of events. The author writes two events that have a clear cause-and-effect relation, which respectively become the premise and the more plau-

Findings of the Association for Computational Linguistics: EACL 2024, pages 5193–5203

sible alternative (mpa). The author also writes an event that does not have a clear causal relation to the premise, and this becomes the less plausible alternative (lpa). As described in Roemmele et al. (2011), composing the lpa is particularly challenging. This is because the benchmark is intended to assess models' ability to isolate causal relations in text separate from more generic associations that are captured by simple lexical co-occurence statistics. The author is expected to write an lpa that has some semantic and/or temporal relation to the premise, but no salient causal relation.

Until only recently, COPA was considered a difficult benchmark. When first presented as a shared task at SemEval-2012, submitted systems achieved accuracy in the 60-65% range (Gordon et al., 2012). When COPA was incorporated into the well-known SuperGLUE ensemble of benchmarks in 2019, the best system reached almost 85% (Wang et al., 2019). Today some LLMs obtain near-perfect accuracy, as we observe in this work. This reflects the life cycle of most benchmarks, where existing ones are continuously replaced by harder new ones once maximal performance is obtained.

3 Research Questions

In this work, we apply a common LLM interaction paradigm (prompting with exemplars) to facilitate LLMs to generate COPA-style items. Moving forward in this paper, we refer to the original COPA items as **Orig-COPA** and the LLM-authored items as **Gen-COPA**. When existing research refers to the COPA task, it means the task of predicting the answer (i.e. more plausible alternative) to each item. Here, we more specifically refer to this task as **answering** COPA, in order to distinguish it from the task of generating COPA items.

We are interested in the following research questions. First, **can LLMs author items with the design of COPA?** In particular, how often do Gen-COPA items meet the same authoring standards as the original benchmark? Second, **when an LLM produces its own Gen-COPA set, does it then answer its own items correctly?** In other words, does the LLM behave consistently during generation and answering in what it deems as the more plausible alternative? Third, **how does an LLM's success in authoring Gen-COPA items relate to its ability to correctly answer Orig-COPA items?** Are LLMs that do well on Orig-COPA also better at authoring new COPA items?

4 Selected LLMs

To investigate these questions, we employ a variety of notable open-source LLMs. We consider eleven models from six different families, where models within the same family are distinguished by size (number of parameters).

BLOOM-7B & BLOOM-176B[1]: The BigScience Large Open-science Open-access Multilingual Language Model (BLOOM) family was developed through a large cross-team research collaboration (Workshop et al., 2022). It was trained on a collection of publicly available datasets that together comprise 1.61 terabytes of text spanning a wide variety of natural and programming languages (Laurençon et al., 2022). Its performance on various benchmarks in SuperGLUE was competitive with other open-source models.

FALCON-7B & FALCON-40B[2]: The FALCON family (Almazrouei et al., 2023) developed by the Technology Innovation Institute was trained on 1-1.5 trillion tokens, primarily the RefinedWeb corpus derived from CommonCrawl (Penedo et al., 2023). The 40B-sized model obtained top performance on several benchmarks upon its release.

LLAMA2-7B, LLAMA2-13B, & LLAMA2-70B[3]: The LLAMA2 family developed by Meta was trained on a collection of publicly available datasets comprising 2 million tokens (Touvron et al., 2023). It has outperformed FALCON and MPT (below) on several benchmarks.

MISTRAL-7B[4] MISTRAL-7B (Jiang et al., 2023), developed by Mistral AI, was trained on open web data. Upon its release, it was presented as a key competitor of the LLAMA2 models. It outperformed the much larger LLAMA2-13B on various knowledge and reasoning benchmarks.

MPT-7B & MPT-30B[5]: The MosaicML Pretrained Transformer (MPT) family developed by MosaicML was trained on 1 billion tokens from a mix of web text and curated data (MosaicML NLP Team, 2023). It has performed favorably compared with other open-source LLMs on benchmarks including COPA.

PHI-2[6]: PHI-2 (Javaheripi and Bubeck, 2023), developed by Microsoft, is a 2.7B parameter model

[1] huggingface.co/bigscience/{bloom,bloom-7b1}
[2] huggingface.co/tiiuae/{falcon-7b,falcon-40b}
[3] huggingface.co/meta-llama/{Llama-2-7b,Llama-2-13b,Llama-2-70b}
[4] huggingface.co/mistralai/Mistral-7B-v0.3
[5] huggingface.co/mosaicml/{mpt-7b,mpt-30b}
[6] huggingface.co/microsoft/phi-2

Prompt Template

```
{% for ex in exemplars %}
Premise:  {{ex['premise']}}
{% if ex['direction'] == 'backwards' %}What was the cause
of this?{% else %}What happened as a result?{% endif %}
Alternative 1:  {{ex['alt1']}}
Alternative 2:  {{ex['alt2']}}
The more plausible {% if ex['direction'] == 'backwards'
%}cause{% else %}result{% endif %} is Alternative
{{ex['more_plausible_alt']}}.
{% endfor %}

Premise:  {{item['premise']}}
{% if item['backwards'] == 'cause' %}What was the cause
of this?{% else %}What happened as a result?{% endif %}
Alternative 1:  {{item['alt1']}}
Alternative 2:  {{item['alt2']}}
The more plausible {% if item['backwards'] == 'cause'
%}cause{% else %}result{% endif %} is Alternative
```

Prompt Example

Premise: My body cast a shadow over the grass.
What was the cause of this?
Alternative 1: The sun was rising.
Alternative 2: The grass was cut.
The more plausible cause is Alternative 1.

[...3 more exemplars...]

Premise: My favorite song came on the radio.
What happened as a result?
Alternative 1: I covered my ears.
Alternative 2: I sang along to it.
The more plausible result is Alternative

Table 2: 4-shot prompt design for answering COPA items. The left shows the raw template (in Jinja2 syntax). The rendered version provided as input to the LLM is on the right.

trained on 250 billion tokens from "textbook quality" data (Gunasekar et al., 2023). Some texts were selectively curated from the web while others were synthetically generated, with both processes emphasizing a high density of commonsense knowledge in the resulting data. Upon its release, PHI-2 demonstrated similar or better performance on commonsense reasoning benchmarks compared with the much larger MISTRAL and LLAMA2 models.

5 Orig-COPA Answering Performance

We first consider how well these LLMs answer the original COPA benchmark. We used a few-shot prompting approach to elicit answers for COPA items. In particular, we selected four items from the COPA development set as task exemplars to include in all prompts for the 500 items in the test set. Table 2 shows the prompt format with the exemplars preceding the target item to be answered[7]. Each exemplar ends with a sentence specifying the numerical label of the correct (more plausible) alternative, and the target item includes the prefix of this sentence in order to elicit a corresponding prediction from the model.

We applied the LLMs[8] outlined in Section 4 to these prompts for the COPA test set. We used greedy decoding (i.e. selecting the maximum probability token at each generation step) to ensure the predictions were deterministic. We generated outputs with a maximum length of four tokens, then post-processed each output with a regular expression to detect the presence of "1" or "2" as the predicted answer. Our policy was to randomly select one of these numbers if neither was present in the output, but in our experiments this was not applied since all outputs contained a 1 or 2.

Table 3 shows the accuracy of these predictions according to the benchmark. For families with models of varying sizes, the larger models outperform the smaller ones within each family, which is an expected pattern (e.g. Kaplan et al., 2020). However, size does not fully account for the overall ranking of model performance. For example, BLOOM-176B is the largest model but ranks only the third lowest in accuracy, and PHI-2 is the smallest model but ranks the third highest.

Model	Accuracy
BLOOM-7B	0.532
BLOOM-176B	0.576
FALCON-7B	0.538
FALCON-40B	0.844
LLAMA2-7B	0.826
LLAMA2-13B	0.850
LLAMA2-70B	0.976
MISTRAL-7B	0.938
MPT-7B	0.656
MPT-30B	0.848
PHI-2	0.902

Table 3: Answering accuracy on COPA test set

Even with the wide spread in performance between different models, the near-perfect accuracy from LLAMA2-70B confirms that COPA is no longer as challenging as it previously was and will soon become outdated. We note that because the COPA test set is available online, it is possible the LLMs observed these items during training. While we expect that developers of these LLMs adhered to the standard best practice of excluding test data from training sets, we cannot strictly assume this.

[7] All prompt templates and code for reproducing this work are available at: github.com/roemmele/Gen-COPA

[8] We ran all LLMs using HuggingFace inference endpoints: huggingface.co/inference-endpoints

Prompt Template	Prompt Example
{% for ex in exemplars %} Premise: {{ex['premise']}} {% if ex['direction'] == 'backwards' %}What was the cause of this?{% else %}What happened as a result?{% endif %} More Plausible Alternative: {% if ex['more_plausible_alt'] == '1' %}{{ex['alt1']}}{% else %}{{ex['alt2']}}{% endif %} Less Plausible Alternative: {% if ex['more_plausible_alt'] == '1' %}{{ex['alt2']}}{% else %}{{ex['alt1']}}{% endif %} {% endfor %} Premise:	Premise: The girl politely declined the hamburger. What was the cause of this? More Plausible Alternative: She was a vegetarian. Less Plausible Alternative: She liked fast food. *[...2 more exemplars...]* Premise:

Table 4: 3-shot prompt design for generating COPA items

This issue is further addressed in the next section.

6 LLM Authoring of Gen-COPA

Next, we examined to what degree these LLMs can generate their own items in the design of COPA. We again used a few-shot prompting approach to facilitate this task. Each prompt consisted of three items of one particular causal direction (forwards or backwards) randomly sampled from the COPA development set. Table 4 shows the format: instead of using the labels "Alternative 1" and "Alternative 2" as in the answering task, each exemplar directly refers to the "More Plausible Alternative" and "Less Plausible Alternative" for a given premise and question, and the models are expected to generate segments with these same identifiers.

We assembled 500 prompts for each causal direction, so each LLM was run on 1000 prompts total. Each of these prompts had a unique set of exemplars. To further promote diversity in the outputs, we used random sampling during decoding, in particular top-p (nucleus) sampling with p=0.9 and temperature=1.0. We generated outputs with a maximum length of 200 tokens.

We parsed each LLM output with the template shown in Table 5, which captures the premise, more plausible alternative (`mpa`), and less plausible alternative (`lpa`) segments for each item of a predefined direction. We refer to these resulting variables for a single output as a **schema**. We automatically classified an output as *failed* if it could not be parsed according to the template or if at least one segment in the item exactly matched another one (e.g. the two alternatives were the same). We also considered whether the generated outputs were duplicates of Orig-COPA items. We failed any output whose tokens were all contained in a single item from the Orig-COPA dev or test set. See Appendix A for further validation of the novelty of the Gen-COPA items in relation to Orig-COPA. In Table 6, **# Items** refers to the number of *passable* (non-failed) schemas generated by each LLM. Across all LLMs the vast majority of outputs (between 97.6% and 99.2%) are passable.[9]

```
<premise>
{% if direction == 'backwards' %}What
was the cause of this?{% else %}What
happened as a result?{% endif %}
More Plausible Alternative: <mpa>
Less Plausible Alternative: <lpa>
```

Table 5: Parsing template for Gen-COPA LLM outputs

6.1 Consistency

Our first analysis of the resulting Gen-COPA items assessed the LLMs' **consistency** between generation and answering. In particular, when an LLM is presented with an item that the LLM itself generated, is the alternative it predicts as more plausible the same one it originally authored as the `mpa`? To determine this, we transformed each LLM's passable schemas into the design of the benchmark items by randomly assigning the labels of "Alternative 1" and "Alternative 2" to the `mpa` and `lpa` in each schema. We ensured each resulting Gen-COPA set was balanced such that always guessing Alternative 1 as the answer yielded 50% accuracy (or trivially above 50% for sets with an odd number of items). We used the same 4-shot prompt design we applied to Orig-COPA (Section 5) in order to elicit answers from an LLM for its own Gen-COPA set. We quantify consistency as the LLM's answering accuracy on this set: it measures how often the LLM's predicted answer for an item is the same alternative that LLM originally designated as the `mpa` in the generated schema for that item.

Table 6 shows that consistency varies widely between different LLMs. The least consistent models

[9] A remark about failures resulting from duplicating (plagiarizing) Orig-COPA items: a further analysis of these outputs across all LLMs revealed that the plagiarized item was always one of the prompt exemplars. We did not encounter cases where an LLM duplicated an Orig-COPA item that was not contained in the prompt.

(BLOOM-7B, FALCON-7B) predict their mpa only slightly more often than random chance, while the most consistent model (LLAMA2-70B) predicts it about 86% of the time. Thus, LLMs are not guaranteed to perform well on their own generated items. Notably, consistency is associated with answering accuracy on Orig-COPA, which is indicated by the extremely strong correlation (in terms of Spearman's rank-order) between the scores in Table 3 and 6 ($r_s = .97, p < .001$).

Model	# Items	Consistency
BLOOM-7B	983	0.512
BLOOM-176B	980	0.588
FALCON-7B	992	0.505
FALCON-40B	987	0.733
LLAMA2-7B	976	0.673
LLAMA2-13B	986	0.753
LLAMA2-70B	978	0.858
MISTRAL-7B	993	0.777
MPT-7B	989	0.602
MPT-30B	991	0.701
PHI-2	987	0.828

Table 6: Answering accuracy of each LLM on its own Gen-COPA set (i.e. consistency)

6.2 Validity

The consistency analysis does not indicate whether the Gen-COPA items are **valid** according to a standard other than the LLM itself. Here, we consider an item valid if the alternative deemed more plausible by a consensus of humans is the same as the mpa in the generated schema. This is analogous to the development of the original COPA benchmark, where an item was considered valid if two human judges both selected the answer designated as the mpa by the author of the item.

To assess the validity of the Gen-COPA items, we employed a two-stage process. In the first stage, an internally employed expert rater reviewed each schema to judge if the mpa was indeed more plausible than the lpa. If so, the rater marked the schema as conditionally-valid, otherwise they marked it as invalid. We then set aside the invalid schemas and converted the conditionally-valid schemas to items with the randomized "Alternative 1" and "Alternative 2" labels. We presented each of these items to two external raters on the Prolific[10] data collection platform. The raters observed the premise, question, and alternatives for each item and selected from two options indicating either Alternative 1 or 2 as the more plausible answer to the question. Ultimately, an item was considered fully-valid if both Prolific raters selected the alternative that was also designated as the mpa in the schema for that item. Items where raters disagreed on the answer were marked as invalid. Each rater responded to 50 items and was paid $6 for an expected completion time of no more than 30 minutes.

We randomly sampled 300 schemas per LLM for this validity assessment, a total of 3300 items. The expert rater flagged a few items to be withheld from the external raters due to potentially offensive content. We resampled a new item to replace each of these (0-5 per LLM, as shown in Table 7). 1033 items were classified as conditionally-valid in the first stage of annotation. In the second stage, the two raters agreed on the mpa for 914 of them. Their agreement in terms of Cohen's κ was .79, indicating substantial agreement.

Table 7 shows the proportion of items for each LLM that were categorized as valid through this process. The validity rate varies significantly between models. The least successful LLM is BLOOM-7B, with only 10% of items marked valid, while LLAMA2-70B has the most success with ≈46% of its items marked valid. Just like consistency, validity is strongly correlated with answering accuracy on Orig-COPA ($r_s = .87, p < .001$). Thus, models that perform well on the original benchmark are more likely to generate valid items.

Model	# Replaced	Validity
BLOOM-7B	3	0.100
BLOOM-176B	2	0.227
FALCON-7B	0	0.113
FALCON-40B	4	0.343
LLAMA2-7B	3	0.197
LLAMA2-13B	1	0.330
LLAMA2-70B	2	0.463
MISTRAL-7B	3	0.310
MPT-7B	5	0.223
MPT-30B	2	0.300
PHI-2	1	0.440

Table 7: Validity rate of Gen-COPA items

We conducted a qualitative analysis of the invalid items to characterize the most common reasons they were rejected. Table 8 lists these characteristics with some exemplifying schemas.

6.3 Gen-COPA Answering Performance

The consistency results in Section 6.1 are affected by both the LLM's answering performance as well as the quality of its Gen-COPA items. For instance, BLOOM-7B has the lowest consistency but also the lowest Gen-COPA validity. Its poor performance

[10] prolific.com

Description	Example
The premise is vague or difficult to interpret	**Premise**: It happened on a blacktop road. What happened as a *result*? [BLOOM-7B] **mpa**: The car slid on the ice. **lpa**: The car jumped into the ditch.
The mpa is not plausible in relation to the premise	**Premise**: The woman contracted polio. What was the *cause* of this? [BLOOM-7B] **mpa**: She ate ice cream. **lpa**: She ran into a car.
The mpa has a semantic or temporal relation to the premise, but not a clear *causal* relation that distinguishes it from the lpa	**Premise**: We opened the envelope. What happened as a *result*? [LLAMA2-7B] **mpa**: We found $1,000 inside. **lpa**: We found nothing.
The mpa is not anchored to commonsense knowledge	**Premise**: A balloon burst. What was the *cause* of this? [BLOOM-7B] **mpa**: The balloon was too big for its zipper. **lpa**: The balloon was too small for its zipper.
Some causal information in the mpa is already contained in the premise	**Premise**: She was fired for showing up late. What was the *cause* of this? [MPT-30B] **mpa**: She arrived late for work. **lpa**: She arrived too early for work.
The mpa and lpa are synonymous or closely related in meaning	**Premise**: The man stubbed his toe. What happened as a *result*? [FALCON-40B] **mpa**: He flinched. **lpa**: He winced.
The mpa and lpa are both plausible to an equal degree	**Premise**: I ate the moldy bread. What happened as a *result*? [FALCON-7B] **mpa**: I vomited. **lpa**: I developed an unusual stomachache.
Assessing the relative plausibility of the mpa and lpa is subjective or requires more information	**Premise**: The boy wanted to be wealthy. What happened as a *result*? [LLAMA2-7B] **mpa**: He started a company. **lpa**: He stole the money of others.

Table 8: Common characteristics of invalid Gen-COPA items

in answering its own items might be due to how often the mpa is invalid. We considered whether the results would differ if we isolated only the valid Gen-COPA items and measured the LLMs' answering accuracy specifically on these sets.

Given the valid schemas for a single LLM, we used the same process described in Section 6.1 to randomly map the mpa and lpa in each schema to the numerical labels, balancing them across the set so that always selecting Alternative 1 would yield 50% accuracy. For this analysis, we ensured an even number of items in each set, so we downsampled one item in sets with an odd number of valid Gen-COPA items. We then applied the same 4-shot prompt format to these items that we used for the experiments in Section 5 and Section 6.1.

Table 9 shows each LLM's answering accuracy on all valid Gen-COPA sets. The bolded numbers on the diagonal indicate each model's consistency on its own items. For all eleven Gen-COPA sets, the accuracies obtained by the LLMs on the set are strongly correlated with their accuracies on the Orig-COPA test set ($r_s \geq .89, p < .001$ for all columns of Table 9). This shows that LLMs' answering performance on COPA also generalizes to LLM-generated versions of COPA. Moreover, the fact that the models that do particularly well on Orig-COPA (i.e. LLAMA2-70B, MISTRAL-7B, and PHI-2) also do well on answering Gen-COPA suggests their success on the former is not just an illusion resulting from rote memorization of the Orig-COPA test set. Since we verified the Gen-COPA items are not duplicates of Orig-COPA items, we know the answers to the Gen-COPA items have not been memorized by the LLMs during training.

Accuracy on the valid Gen-COPA sets tends to be higher for all LLMs compared with their accuracy on Orig-COPA. In particular, examining the final column in Table 9 where the results are aggregated over all Gen-COPA sets, the accuracy of each LLM is between 1 and ≈11 percentage points higher on Gen-COPA than on Orig-COPA, with exception to BLOOM-7B whose accuracy is low on both groups of items. This suggests that the valid subset of Gen-COPA items is easier to answer than Orig-COPA. This may reflect the mechanism behind model collapse, where synthetically generated data contains only the most common patterns in the original data distribution, losing the diverse signal in the distribution tail (Shumailov et al., 2024).

These results also confirm that LLMs do not necessarily answer their own Gen-COPA items consistently even if the items are valid. For example, FALCON-7B performs exactly at the level of random guessing on its own valid Gen-COPA set. This finding aligns with increasing evidence demonstrating LLM inconsistency in question answering performance: in particular, changing the order and/or

Item Set →	BLOOM-7B	BLOOM-176B	FALCON-7B	FALCON-40B	LLAMA2-7B	LLAMA2-13B	LLAMA2-70B	MISTRAL-7B	MPT-7B	MPT-30B	PHI-2	ALL
Model ↓ # →	30	68	34	102	58	98	138	92	66	90	132	908
BLOOM-7B	**0.633**	0.515	0.412	0.529	0.655	0.480	0.543	0.522	0.470	0.489	0.508	0.520
BLOOM-176B	0.600	**0.676**	0.735	0.608	0.638	0.561	0.565	0.554	0.606	0.644	0.758	0.628
FALCON-7B	0.567	0.529	**0.500**	0.559	0.517	0.541	0.587	0.500	0.561	0.556	0.644	0.561
FALCON-40B	0.800	0.882	0.882	**0.882**	0.931	0.847	0.891	0.880	0.894	0.889	0.932	0.882
LLAMA2-7B	0.800	0.882	0.794	0.931	**0.948**	0.898	0.826	0.891	0.818	0.811	0.902	0.871
LLAMA2-13B	0.867	0.912	0.853	0.951	0.966	**0.918**	0.891	0.913	0.864	0.867	0.962	0.913
LLAMA2-70B	1.000	1.000	1.000	1.000	1.000	0.969	**0.986**	0.967	0.955	0.989	0.992	0.986
MISTRAL-7B	0.967	0.985	0.941	0.980	1.000	0.929	0.935	**0.967**	0.879	0.889	0.977	0.949
MPT-7B	0.667	0.838	0.706	0.804	0.776	0.714	0.841	0.815	**0.591**	0.656	0.811	0.764
MPT-30B	0.800	0.882	0.882	0.941	0.966	0.796	0.848	0.880	0.894	**0.833**	0.947	0.882
PHI-2	0.800	0.956	0.941	0.971	0.983	0.969	0.949	0.924	0.909	0.900	**0.992**	0.947

Table 9: Answering accuracy on valid Gen-COPA items. Each LLM's consistency on its own items is in bold.

labels of answers can drastically impact the model's accuracy (Wang et al., 2024a; Zheng et al., 2024). Wang et al. (2024b) found that consistency across different answer label assignments increased with model capability. Similarly, an LLM's consistency in answering its own generated questions may be a broad indicator of its abilities; this warrants further analysis with other tasks beyond COPA.

6.4 Composition Quality

Just like Orig-COPA items, valid Gen-COPA items each have an agreed-upon alternative that is considered the correct answer. However, even items with a correct answer may not meet all composition quality standards reflected in the original benchmark. As with most benchmarks, while there are some explicit authoring guidelines for COPA, many of the quality standards are only implicitly defined. Thus, we enlisted one of the authors of Orig-COPA items to assess the intrinsic authoring quality of the valid Gen-COPA items. A item rated as **high-quality** is one the rater deems they would accept as an additional item in the benchmark.

Table 10 shows the proportion of items categorized as high-quality among the set of valid Gen-COPA items produced by each LLM. Though the association is more moderate than that of validity, the rate of high-quality items is also correlated with answering accuracy on Orig-COPA ($r_s = .76, p = .007$). This again suggests that LLMs that correctly answer the original benchmark also have better authoring ability.

To illuminate the characteristics underlying these ratings, Table 11 describes and exemplifies the most common reasons that an item failed to receive a high-quality rating, while Table 12 lists examples

Model	High-Quality
BLOOM-7B	0.533
BLOOM-176B	0.500
FALCON-7B	0.500
FALCON-40B	0.569
LLAMA2-7B	0.448
LLAMA2-13B	0.541
LLAMA2-70B	0.616
MISTRAL-7B	0.663
MPT-7B	0.470
MPT-30B	0.567
PHI-2	0.629

Table 10: Rate of high-quality valid Gen-COPA items

of items that were marked as high-quality. Notably, in these high-quality items, the `lpa` is highly related to the premise, but its causal relation is unclear compared with the `mpa`. Consequently, these items are challenging to answer without discerning causal relatedness separately from coarse semantic relatedness. This is a prominent feature of Orig-COPA items and was a key reason the benchmark remained unbeaten for so long. All Gen-COPA items with their validity and quality annotations are publicly available for further analysis[11].

7 Related Work and Outlook

Researchers are increasingly turning to LLMs to replace human effort in developing and evaluating NLP systems. In particular, LLMs are being used to synthesize labeled data in order to train or fine-tune models (Choi et al., 2024; He et al., 2022; Li et al., 2023). LLMs are also being applied to score the output quality of other models (Chiang and Lee, 2023; Kocmi and Federmann, 2023; Wang et al., 2023). Our work aligns with the above endeavors

[11] huggingface.co/datasets/roemmele/Gen-COPA

Description	Example
The `lpa` is not plausible in any context regardless of the premise	**Premise**: The woman was arrested. What was the *cause* of this? [FALCON-40B] **mpa**: She stole something. **lpa**: She was a zombie.
The relation between the `mpa` and the premise is clear but very trivial	**Premise**: The students cleaned the beach. What happened as a *result*? [LLAMA2-70B] **mpa**: The beach was clean. **lpa**: The beach was polluted.
The `mpa` is more plausible than the `lpa`, but the relation is more temporal or semantic than causal	**Premise**: The sales associate spoke to the customer. What was the *cause* of this? [PHI-2] **mpa**: The customer made a purchase. **lpa**: The sales associate went on vacation.
The `lpa` is also a plausible causal relation, but the `mpa` is a closer temporal relation to the premise	**Premise**: I broke my nose. What happened as a *result*? [BLOOM-7B] **mpa**: My nose started bleeding. **lpa**: I took a week off work.
The relation between the `mpa` and premise is clear only with assumptions not defined in the premise	**Premise**: The landlord is sending eviction notices. What happened as a *result*? [PHI-2] **mpa**: The tenant had to leave. **lpa**: The landlord offered a lower rent.

Table 11: Common quality problems in valid Gen-COPA items

Premise: The girl gave her friend her lunch. What was the *cause* of this? [MPT-30B] **mpa**: She was concerned about her friend's lack of food. **lpa**: Her friend sat next to her at lunch.
Premise: The student x-rayed the patient's arm. What happened as a *result*? [MPT-30B] **mpa**: He discovered that the patient's arm was broken. **lpa**: He put the patient's arm in a cast.
Premise: The movie star was in seclusion. What was the *cause* of this? [FALCON-40B] **mpa**: The movie star needed a break from the spotlight. **lpa**: The movie star found a secret underground club.
Premise: The reader was puzzled by a joke. What happened as a *result*? [FALCON-40B] **mpa**: She looked up the explanation. **lpa**: She laughed at the joke.
Premise: The woman destroyed the artwork. What was the *cause* of this? [LLAMA2-70B] **mpa**: The artwork was an insult to her religion. **lpa**: The artwork was an advertisement for a store.

Table 12: Examples of high-quality Gen-COPA items

in exploring the use of LLMs as an alternative to human authoring of evaluation items.

Our particular goal in using LLMs for this assessment authoring is to better understand their capabilities, rather than to produce a novel benchmark that is valuable for evaluating future LLMs. However, there are some emerging demonstrations of using LLMs to facilitate the creation of evaluation data that is unique in its design and scope. For instance, Anthropic (2023) described prompting an LLM to derive multiple-choice Q&A sets from short sections of text, which are then used to evaluate the same LLM's ability to answer those questions when the relevant information is provided in one long document. As another example, Tian et al. (2024) utilized an LLM as an interactive tool combined with human feedback in the authoring process for a novel challenge set focused on physical problem-solving.

Given the short lifecycle of benchmarks, rapid creation of new ones is pivotal to capturing further progress in NLP. Additionally, because they do not serve a broader purpose outside of research, the authoring standards for benchmarks are not always clearly understood by the human authors tasked with their creation, and these standards are not necessarily well-anchored to other real-world writing tasks. The current paradigm of LLM interaction via few-shot prompting suggests that LLMs can perform sophisticated authoring tasks in the absence of explicit guidelines just by observing representative examples. As current benchmarks quickly age and new benchmarks become more complex in what they aim to measure, LLMs are likely to take on a sanctioned role in their development.

8 Conclusion

This paper looks at a notable assessment of commonsense reasoning, COPA, as an authoring task performed by LLMs. The results indicate that models that answer COPA items correctly (both LLM-authored and human-authored ones) are also better at writing COPA items. In future work, we will investigate how this extends to other benchmarks.

Our work bears upon the trend of widening generalizability of NLP models across tasks. Previously, models designed for text generation were not considered directly applicable to commonsense question answering benchmarks, because this answering task was presumed to require a distinct model architecture supporting answer label prediction. This is no longer a constraint in the new paradigm of LLMs. Increasingly, we are observing models demonstrate the same core knowledge across highly diverse task representations.

Limitations

The current paradigm of LLMs has some key limitations that are reflected in this work. First, there has been limited transparency into the development process behind most LLMs, even the open-weight models used here. With exception to BLOOM which had transparency as an explicit goal of its development, not all details of the LLMs in this work regarding their training data, model architecture, and optimization techniques have been clearly documented. Given this as well as the complexity of these details, it is not easy to interpret why exactly different types of LLMs perform differently on the same benchmarks. So while we conclude there is a positive association between answering and generating COPA, we do not propose an explanation for why certain models (e.g. the LLAMA2 family) excel on these tasks more than other others (e.g. the BLOOM family).

Second, LLM behavior is highly sensitive to prompt design, such that different prompts representing the same task and input features can yield different outputs. As a result, prompt optimization has become a significant focus of utilizing LLMs. Finding tractable solutions to this optimization process is an ongoing research endeavor (e.g. Zhou et al., 2023). Meanwhile, this is typically done manually without any guarantee of optimality. In particular, researchers often write a few different prompt variations based on their knowledge of the task and select the one that performs best on the task evaluation. In this work, we took a principled rather than data-driven approach to selecting the prompts for answering and generating COPA, in which the prompt design matches the conceptual design of the benchmark defined in Roemmele et al. (2011). It is possible that varying the language of the prompt as well as the number of exemplars (shots) would result in better performance in terms of answering accuracy or Gen-COPA validity for some LLMs, which could yield a different view of how they compare to one another.

Ethical Considerations

The business of data annotation has grown dramatically with the expansion of NLP systems and LLMs in particular. There are serious concerns about the ethics of this business when it comes to fair compensation of workers (e.g. Perrigo, 2023). For the annotation in this work, the expert rater was compensated as part of their normal job role. The external raters we employed on Prolific were paid at the rate of $12/hour, which meets Prolific's recommended universal standard for fair pay (Prolific, 2023).

LLMs have a well-known risk of generating offensive content (e.g. Gehman et al., 2020). We anticipated this risk would emerge in some of the Gen-COPA items and sought to reduce exposure to these items to people outside our internal team. As described in Section 6.2, the expert rater consented to conduct an initial review of all Gen-COPA items. They assigned a "content warning" to items they deemed potentially offensive or harmful, and these items were consequently not shown to external raters on Prolific. While we cannot guarantee that the external raters were not offended by items that were not flagged in the initial review, we did not receive any report of this from raters.

Acknowledgements

(Gordon) The project or effort depicted was or is sponsored by the U.S. Army Research Laboratory (ARL) under contract number W911NF-14-D-0005, and that the content of the information does not necessarily reflect the position or the policy of the Government, and no official endorsement should be inferred.

References

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. The falcon series of open language models. *Preprint*, arXiv:2311.16867.

Anthropic. 2023. Prompt engineering for Claude's long context window. *anthropic.com*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Juhwan Choi, Eunju Lee, Kyohoon Jin, and YoungBin Kim. 2024. GPTs are multilingual annotators for sequence generation tasks. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 17–40, St. Julian's, Malta. Association for Computational Linguistics.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration

in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Andrew Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2012. SemEval-2012 task 7: Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 394–398, Montréal, Canada. Association for Computational Linguistics.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks are all you need. *Preprint*, arXiv:2306.11644.

Xuanli He, Islam Nassar, Jamie Kiros, Gholamreza Haffari, and Mohammad Norouzi. 2022. Generate, annotate, and learn: NLP with synthetic text. *Transactions of the Association for Computational Linguistics*, 10:826–842.

Mojan Javaheripi and Sébastien Bubeck. 2023. Phi-2: The surprising power of small language models. *https://www.microsoft.com/en-us/research/blog*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation*, pages 193–203, Tampere, Finland. European Association for Machine Translation.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gérard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, et al. 2022. The bigscience ROOTS corpus: A 1.6TB composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Zhuoyan Li, Hangxiao Zhu, Zhuoran Lu, and Ming Yin. 2023. Synthetic data generation with large language models for text classification: Potential and limitations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10443–10461, Singapore. Association for Computational Linguistics.

MosaicML NLP Team. 2023. Introducing MPT-7B: A new standard for open-source, commercially usable llms.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The RefinedWeb dataset for Falcon LLM: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

Billy Perrigo. 2023. Exclusive: OpenAI used Kenyan workers on less than $2 per hour to make ChatGPT less toxic. *Time*.

Prolific. 2023. Prolific's payment principles.

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *2011 AAAI Spring Symposium Series*.

Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. AI models collapse when trained on recursively generated data. *Nature*, 631(8022):755–759.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas Griffiths, and Faeze Brahman. 2024. MacGyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5324, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Haochun Wang, Sendong Zhao, Zewen Qiang, Bing Qin, and Ting Liu. 2024a. Beyond the answers: Reviewing the rationality of multiple choice question answering for the evaluation of large language models. *Preprint*, arXiv:2402.01349.

Jiaan Wang, Yunlong Liang, Fandong Meng, Zengkui Sun, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023. Is ChatGPT a good NLG evaluator? a preliminary study. In *Proceedings of The 4th New Frontiers in Summarization Workshop*, pages 1–11. Association for Computational Linguistics.

Xinpeng Wang, Bolei Ma, Chengzhi Hu, Leon Weber-Genzel, Paul Röttger, Frauke Kreuter, Dirk Hovy, and Barbara Plank. 2024b. "my answer is C": First-token probabilities do not match text answers in instruction-tuned language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 7407–7416, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large language models are human-level prompt engineers. In *The Eleventh International Conference on Learning Representations*.

A Novelty of Gen-COPA with regard to Orig-COPA

As described in Section 6, we verified that none of the Gen-COPA items used in our analyses are exact duplicates of Orig-COPA items. However, it's important to consider the overall similarity between Gen-COPA and Orig-COPA, to ascertain that the Gen-COPA items are not just trivial variations of Orig-COPA items. Table 13 reports two metrics pertaining to this for the passable Gen-COPA sets produced by each LLM. The first metric, **Common 3-grams**, indicates the overall proportion of trigrams in the Gen-COPA items that also appear in Orig-COPA. These percentages range from ≈8% to ≈12% for all LLMs, indicating that the majority of trigrams in the Gen-COPA items are not contained in Orig-COPA. The second metric pertains to ROUGE-3 F1, which is computed to determine the similarity of each Gen-COPA item to each of the Orig-COPA items. For a given Gen-COPA item, the maximum of these scores is selected, which corresponds to the Orig-COPA item that is most similar to that Gen-COPA item in terms of trigram overlap. **Max ROUGE-3** reports the mean of these maximum scores across the Gen-COPA items for each LLM. These scores range from 0.058 to 0.091 for all LLMs, indicating that on average a particular Gen-COPA item has only a minimal degree of trigram overlap with the Orig-COPA item it most closely resembles. Thus, the majority of Gen-COPA items are reasonably distinct from Orig-COPA items.

Model	Common 3-grams	Max ROUGE-3
BLOOM-7B	0.079	0.058
BLOOM-176B	0.079	0.063
FALCON-7B	0.091	0.067
FALCON-40B	0.090	0.070
LLAMA2-7B	0.089	0.063
LLAMA2-13B	0.093	0.063
LLAMA2-70B	0.096	0.067
MISTRAL-7B	0.085	0.059
MPT-7B	0.091	0.065
MPT-30B	0.123	0.091
PHI-2	0.096	0.071

Table 13: Redundancy of Gen-COPA with Orig-COPA