

ARC ‘Challenge’ Is Not That Challenging

Lukasz Borchmann
Snowflake AI Research
lukasz.borchmann@snowflake.com

Abstract

ARC Challenge¹ appears more difficult than ARC Easy for modern LLMs primarily due to an evaluation setup that prevents direct comparison of answer choices rather than inherent complexity. Although some researchers have quietly shifted to a more appropriate evaluation scheme over the last year, the implications of this change have yet to be widely acknowledged. We highlight this overlooked shift, show how similar evaluation practices falsely imply reasoning deficits in other benchmarks, and demonstrate that fairer methods dramatically reduce performance gaps (e.g. on SIQA) and even yield superhuman results (OpenBookQA). In doing so, we reveal how evaluation shapes perceived difficulty and offer guidelines to ensure that multiple-choice evaluations accurately reflect actual model capabilities.

1 Introduction

A substantial set of benchmarks regularly employed in LLM testing consists of multiple-choice problems, commonly considered in a setup where each provided option is scored under the model, and the one with the highest likelihood is compared against the gold standard to determine accuracy. This refers, among others, to popular evaluators of MMLU (Hendrycks et al., 2021), ARC Easy and Challenge (Clark et al., 2018), BoolQ (Clark et al., 2019), RACE (Lai et al., 2017), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2019), SIQA (Sap et al., 2019), COPA (Gordon et al., 2011), and HellaSwag (Zellers et al., 2019).

Details of this setup differ but generally follow one of the two conventions. Under one convention, the model considers each candidate answer in **separation**, without alternative options displayed (Figure 2), while under the other, the model sees all candidate **options** together in the prompt (Figure 3). We argue that the first setup is commonly

¹Not to be confused with the recent ARC Prize.

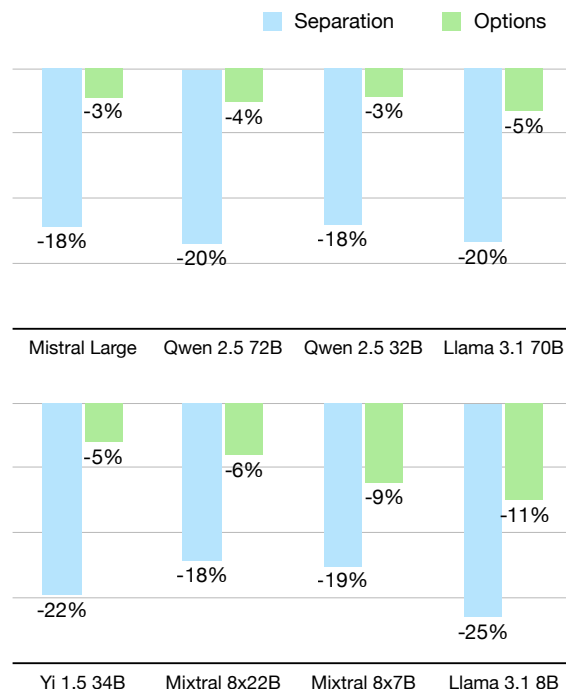


Figure 1: Difference between **ARC Challenge** and **ARC Easy** accuracies when considering each answer separately compared to seeing all options. The gap is vastly reduced, up to six times in this comparison.

overused and rarely preferred since it does not simulate the natural reasoning context in which multiple choices are compared directly. Importantly, it introduces a false notion of how challenging a particular problem is, as switching from the first to the second might result in a 35% improvement in model accuracy, as shown in Section 2 experiments.

1.1 Hardly answerable in separation

Consider the question, ‘Which of these items contains only a solution?’ Given the option ‘a jar of pickles,’ confronting a single item with a question and assessing whether pickles fulfill the definition of the solution suffices. They do not, so this option is incorrect. The question can be addressed under



Figure 2: Model considers particular choices in `separation` without knowing the alternative (prompt includes only the question). Because options may vary in length, it is a good practice to normalize them (Gao, 2021).

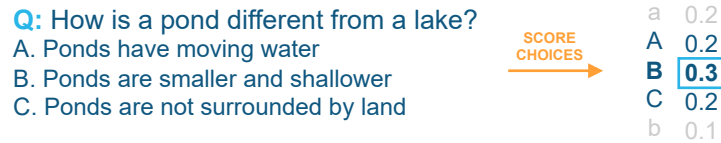


Figure 3: Model sees the context of all possible `options` in the prompt. Because all of the options are single letters (likely single tokens), scores require no normalization.

both evaluation setups because it does not require the availability of other options, such as ‘a can of mixed fruit.’

In contrast, some questions inherently demand comparative evaluation: let us think about ‘Which of these most likely has the *greatest* mass?’ and the option ‘puppy.’ This question’s answer cannot be determined without comparing the mass of the ‘puppy’ to the masses of all other provided options. It is the greatest compared to ‘chicken’ or ‘lizard’ but not in the context of ‘horse’ or ‘elephant.’ Though it can work to some extent, relying on the likelihood assigned in `separation` to each of the animals is an unreasonable way of determining the heaviest one. It feels natural to provide the model with the `options` to choose from instead because it allows the model to directly compare and contextualize choices, reflecting a more authentic reasoning process. This aspect, however, is commonly overlooked.

Importantly, such ‘hardly answerable in separation’ questions are prevalent in ARC datasets, constituting 21% of ARC Easy and 31% of ARC Challenge (see Appendix B). Despite this fact, it is widespread to evaluate them without seeing all of the options simultaneously (Touvron et al., 2023a,b; Jiang et al., 2023; Peng et al., 2023; 01. AI et al., 2024; Gemma Team et al., 2024b, *inter alia*).

2 Impact on evaluation results

Figure 4 shows the difference in model accuracy when options are presented in isolation versus all at once. Not surprisingly, different setups hugely change the evaluation results, partly because of the vast presence of ‘hardly answerable in separa-

tion’ questions and partially because such a setup, equivalent to what human test takers see, doesn’t introduce unnecessary obstacles.

For example, switching from `separation` to `options` improves the Llama 3.1 70B ARC Challenge accuracy from 64% to 93%, rendering this ARC subset significantly less challenging. Moreover, since the procedure change has a much higher impact on ARC Challenge than on ARC Easy, switching reduces the accuracy gap between these subsets as much as six-fold (Figure 1). These findings suggest that the previously perceived difficulty was primarily an artifact of the evaluation method rather than the tasks’ complexity.

The difference seems somewhat known in the LLM community, but not broadly, and needs to be stated explicitly. E.g. concerning the Llama family, authors seem to silently switch from `separation` to `options` between Llama 2 and Llama 3, similar to Mistral between Mixtral 8x7B and Mixtral 8x22B, or DeepSeek before their V2 (detailed assessment available in Appendix A).

3 Are other benchmarks affected?

Yes. Analogous changes in evaluation procedures would vastly improve OpenBookQA scores. Concerning Llama 3.1 70B, one can achieve improvement from 48% to 89% (see Figure 5). For some reason, most authors who switched from `separation` to `options` in ARC evaluation did not follow on some other multi-choice problems.

If they switched, they could notice that OpenBookQA is essentially solved, as current models achieve scores above human performance (92% compared to 96% of Qwen 2.5 72B).

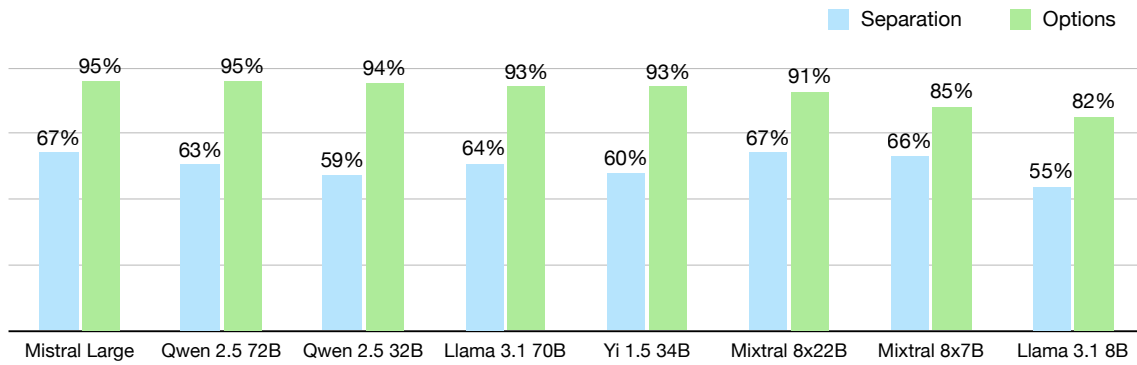


Figure 4: **ARC Challenge** evaluation results depending on whether the model sees other options or considers each answer separately. Differences reach up to 35%, and assumed setup impacts model rankings.

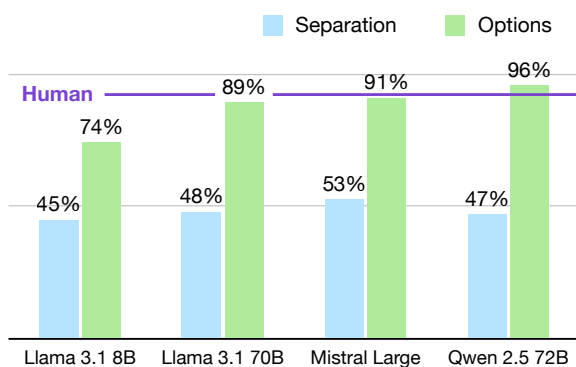


Figure 5: **OpenBookQA** evaluation results depending on whether the model sees other options or considers each answer separately. In a setup with options, current models outperform human test takers.

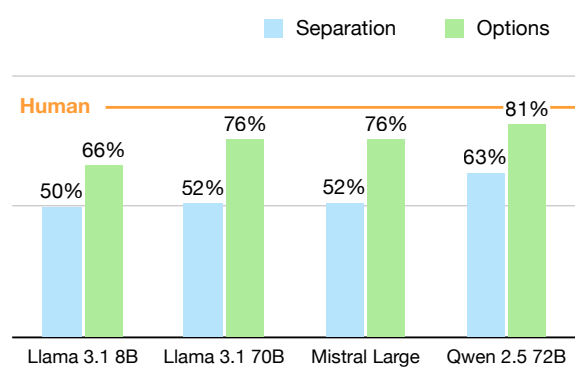


Figure 6: **SIQA** evaluation results depending on whether the model sees other options or considers each answer separately. Reformulation leads to up to 24% improvement.

In the case of SIQA, reformulation leads to a 24% increase in Llama 70B accuracy. However, the best models perform 5% below the human baseline, suggesting room for improvement (Figure 6).

These dramatic increases, however, call into question previous interpretations of model capability on both SIQA and OpenBookQA.

4 Why does it matter?

We argue that the benchmark’s challenge should result from the inherent complexity of the knowledge or reasoning required, not its formulation or evaluation procedure.

The **separation** setup is unnecessarily complicated and not consistent with how humans would approach the multi-choice problem, leading to existing assessments of human performance being incompatible. For example, the fact that strong LLMs perform 30% worse than humans on SIQA doesn’t mean they are deficient in commonsense reasoning about social situations if under **options**

the difference largely disappears. This mismatch can falsely suggest deficits in reasoning capabilities that are not truly present.

Notably, the gap between LLMs and humans on SIQA has been previously used to argue that LLMs might lack social intelligence and social commonsense (Sap et al., 2022).

5 Suggestions for multi-choice eval

There are many arguments for using the **options** for the evaluation of multi-choice QA problems. We have already described a few, including the presence of ‘hardly answerable in separation’ questions and the fact it is consistent with the usual approach to assessing human performance, as humans naturally consider all choices in a single context.

Other benefits include enabling compatible evaluation in a likelihood and generative manner, allowing one to obtain comparable scores with LLMs behind closed and limited APIs. Moreover, it eliminates the need to decide which normalization

method to use when aggregating scores from several option tokens, which is, to some extent, arbitrary and impacts model ranking.

Nevertheless, it is not preferred in all cases commonly considered under the likelihood-scoring evaluation scheme.

5.1 Why likelihood scoring in the first place?

Likelihood-based scoring is a natural choice for problems from pure language modeling, Winograd schemas, or fill-in-the-gap setups such as encountered in LAMBADA (Paperno et al., 2016), HellaSwag (Bisk et al., 2019), or WindoGrande (Sakaguchi et al., 2019) datasets.

For other problems, it is effectively a variant of constrained decoding, that is, the model is restricted to selecting from given candidate options rather than generating open-ended text. It guarantees that models will not emit CoTs before answering the question and removes the need for output postprocessing, such as extracting the letter associated with the selected option and normalizing its casing. Moreover, it allows us to obtain meaningful results with base models, e.g. intermediate checkpoints from LLMs’ self-supervised pretraining since we are constraining the output to one of the most probable options under the model.

5.2 To show, or not to show options

Suppose the options are of equal length, and it is not helpful to consider them simultaneously. This is the case when we deal with a straightforward yes/no response, and no comparative reasoning is necessary, as in the BoolQ dataset (Clark et al., 2019). In similar scenarios, there are no arguments for dropping separation in favor of options.

We are in the position that the options variant is preferred if there is a risk of a ‘hardly answerable in separation’ question presence (Section 1.1) or it simply makes it easier to consider all of the options at once because it ensures the model can leverage direct comparisons. This seems to be the case for virtually all other multi-choice QA problems, such as MMLU (Hendrycks et al., 2021), ARC (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), PIQA (Bisk et al., 2019), or SIQA (Sap et al., 2019). In fact, most similar problems are already being evaluated in the options scheme.

Arguably, options has some possible or actual disadvantages: the order of the presented choices might impact the evaluation results (e.g. models might bias toward the first listed choice), it might be

easier to exploit pattern recognition, and the setup requires slightly more compute. Nevertheless, we consider these to be outweighed by benefits and recommend broad use of the options scheme.

Limitations

Despite our recommendations and their benefits, several limitations and uncertainties exist in identifying precisely which evaluation methods were employed in previously reported results.

Firstly, because authors of LLMs’ technical reports rarely or never report such details, our assessment of which of the options and separation they employed is based on attempting to replicate reported accuracy scores under both setups and observing which condition aligns best (Appendix A). Fortunately, given the magnitude of the observed differences, the performance gap is so large that one can differentiate between the mentioned approaches with a high degree of certainty.

Secondly, in a search for the separation overuse candidates, we mainly relied on intuition. We considered the most popular benchmarks due to their widespread use and availability of performance data, later confirming the intuition experimentally. Though it was a tale of ARC, OpenBookQA, and SIQA, many other widely used benchmarks may benefit from revisiting their evaluation setup.

6 Summary

We draw the community’s attention to shifting from evaluating answers in isolation to evaluating them alongside all other options. Over the last year, such a change happened in the reported ARC Challenge and ARC Easy scores, vastly impacting their evaluation results. After discussing the implications, we considered whether other popular benchmarks might undergo similar reformulation, identifying OpenBookQA and SIQA as candidates. In the former, recent models outperform humans, even though there is a room of 40% between humans and LLMs in the widespread setup. The fact that the gap drastically narrows under the all-options evaluation method highlights how the testing format can distort perceived difficulty.

We concluded with a guideline for evaluating multi-choice problems, arguing that the setup where the model sees all options is preferred over considering each answer separately, except for causal or masked language modeling problems.

References

01. AI et al. 2024. [Yi: Open Foundation Models by 01.AI](#).
- Jinze Bai et al. 2023. [Qwen Technical Report](#).
- Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. 2019. [PIQA: Reasoning about Physical Commonsense in Natural Language](#). *CoRR*, abs/1911.11641.
- Tom B. Brown et al. 2020. [Language Models are Few-Shot Learners](#). *CoRR*, abs/2005.14165.
- Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. [BoolQ: Exploring the Surprising Difficulty of Natural Yes/No Questions](#).
- Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. [Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge](#).
- DeepSeek AI et al. 2024a. [DeepSeek LLM: Scaling Open-Source Language Models with Longtermism](#).
- DeepSeek AI et al. 2024b. [DeepSeek-V2: A Strong, Economical, and Efficient Mixture-of-Experts Language Model](#).
- Leo Gao. 2021. [Multiple Choice Normalization in LM Evaluation](#).
- Leo Gao et al. 2024. [A framework for few-shot language model evaluation](#).
- Gemma Team et al. 2024a. [Gemma 2: Improving Open Language Models at a Practical Size](#).
- Gemma Team et al. 2024b. [Gemma: Open Models Based on Gemini Research and Technology](#).
- Andrew S. Gordon, Zornitsa Kozareva, and Melissa Roemmele. 2011. [Choice of plausible alternatives: An evaluation of commonsense causal reasoning](#). In *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning*.
- Aaron Grattafiori et al. 2024. [The Llama 3 Herd of Models](#).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring Massive Multitask Language Understanding](#).
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7B](#).
- Albert Q. Jiang et al. 2024. [Mixtral of Experts](#).
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding Comprehension Dataset From Examinations](#).
- Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. [Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering](#).
- Mistral AI. 2024. [Cheaper, Better, Faster, Stronger](#).
- Denis Paperno, Germ  n Kruszewski, Angeliki Lazaridou, Quan Ngoc Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fern  ndez. 2016. [The LAMBADA dataset: Word prediction requiring a broad discourse context](#).
- Bo Peng et al. 2023. [RWKV: Reinventing RNNs for the Transformer Era](#).
- Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. [Winogrande: An adversarial winograd schema challenge at scale](#).
- Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. [Neural theory-of-mind? on the limits of social intelligence in large LMs](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Maarten Sap, Hannah Rashkin, Derek Chen, Ronan LeBras, and Yejin Choi. 2019. [SocialQA: Commonsense reasoning about social interactions](#).
- Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timoth  e Lacroix, Baptiste Rozi  re, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. [LLaMA: Open and Efficient Foundation Language Models](#).
- Hugo Touvron et al. 2023b. [Llama 2: Open Foundation and Fine-Tuned Chat Models](#).
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a Machine Really Finish Your Sentence?](#)

A Claiming setup used by other authors

As authors of LLM technical reports rarely or never provide such details, we redo their evaluations in **options** and **separation** setups. If the reported score is in the same ballpark as one of these, and visibly distant from the other one, we claim they used the first. This assessment is backed by the notion that no other change in the prompt could cause a 20%+ improvement in ARC Challenge scores. The exception could be using a generative setup with CoT for some heavy reasoners, but we do not suspect authors to use CoT if they are not reporting this because it would be a serious flaw.

The results of this analysis for the ARC Challenge are presented in Table 1. Prompt ‘reverse engineering’ becomes more troublesome in the context of SIQA and OpenBookQA datasets (Table 2-3) as some authors do not directly report scores but average them with other commonsense reasoning problems. We’re not claiming any setup for these.

Finally, some authors tackling OpenBookQA followed [Brown et al. \(2020\)](#) in normalizing the likelihood by the likelihood of the completion given ‘Answer:’ as context. To address this possibility, we introduce two additional variants referred to as **separation_b** and **options_b**.

B Estimating number of questions hardly answerable in separation

To determine whether questions are answerable given a single option or require the context of other options, we process them in batches of 20 using gpt-4o-2024-11-20 model and the following prompt with few-shot examples:

Consider the question, "Which of these items
→ contains only a solution?" Given the
→ option "a jar of pickles," confronting a
→ single item with a question and assessing
→ whether pickles fulfill the definition
→ of the solution suffices. They do not, so
→ this option is incorrect.

Now let us think about "Which of these most
→ likely has the greatest mass?" and the
→ option "puppy." It can be considered only
→ with other options because it is the
→ greatest compared to "chicken" or "lizard"
→ " but not in the context of "horse" or "
→ elephant".

These questions represent two classes of
→ questions: "answerable without other
→ options" and "unanswerable without other
→ options".

Other examples of "answerable without other
→ options" are:

- Kerry made a simple flashlight. She recorded
→ the following statements in her lab book.
→ Which statement is an inference? (
→ Answerable, because it suffices to
→ compare options against the definition of
→ inference)
- A scientist on a field trip discovered a new
→ organism. She examined its cells under a
→ microscope and observed several different
→ structures, including a nucleus, a cell
→ wall, and some chloroplasts. This
→ organism would correctly be classified in
→ which of the following kingdoms? (
→ Answerable, because it can be answered by
→ deciding if the kingdom provided in the
→ option can be associated with having a
→ nucleus, a cell wall, and chloroplasts)
- Many types of motion occur in our solar system.
→ Which type of motion describes one Earth
→ year? (Answerable, because it suffices
→ to validate if the motion describes one
→ year or not)
- When trees develop leaves in the spring, 10
→ changes occur on the forest floor. Why
→ does the development of leaves cause
→ changes on the forest floor? (Answerable,
→ because it is enough to verify if a
→ particular option described the possible
→ cause of change)
- Using a softball bat to hit a softball is an
→ example of using which simple machine? (
→ Answerable, because all one needs to do
→ is to check if the described simple
→ machine is the explanation of how a
→ softball bat works)
- Which is a statement about climate? (
→ Answerable, because it is possible to
→ verify a single option against the
→ climate definition)
- How do word processors on computers benefit
→ most students? (Answerable, because it
→ can be answered in separation whether
→ most students benefit from this feature
→ of the word processor)
- Photosynthesis occurs in which of these
→ organisms? (Answerable because it
→ suffices to check if the organism
→ mentioned in the option performs
→ photosynthesis)
- Which two theories of Moon formation propose
→ that much or all of the material
→ comprising the Moon came from Earth? (
→ Because it suffices to validate if both
→ theories mentioned in a single option
→ describe the Moon as formed from Earth
→ material)
- Plants and animals are composed of organic
→ compounds. Which of the following are the
→ common elements found in organic
→ compounds? (Answerable, because it
→ suffices to check if the option consists
→ of compounds appearing in both plants and
→ animals)

Other examples of "unanswerable without other
→ options" are:

- A ball is dropped from different heights. When
→ the ball is dropped from the highest

Model	Reported	Measured	s / o	Assessment
Llama 65B (Touvron et al., 2023a)	56.0	55.6 / 70.2		separation
Llama 2 70B (Touvron et al., 2023b)	57.4	57.4 / 79.6		separation
Llama 3 70B (Grattafiori et al., 2024)	92.9	64.2 / 91.3		options
Mistral 7B (Jiang et al., 2023)	55.5	54.1 / 74.6		separation
Mixtral 8x7B (Jiang et al., 2024)	59.7	59.9 / 83.3		separation
Mixtral 8x22B (Mistral AI, 2024)	91.3 [†]	70.7 / 91.8		options
DeepSeek 67B (DeepSeek AI et al., 2024a)	59.0	60.1 / 84.6		options
DeepSeek V2 (DeepSeek AI et al., 2024b)	92.4 [†]	70.3 / 92.2		options
Qwen 14B (Bai et al., 2023)	84.4	47.3 / 86.6		options
Yi 6B (01. AI et al., 2024)	50.3 [†]	55.7 / 80.5		separation
Gemma 7B (Gemma Team et al., 2024b)	53.2	53.2 / 79.0		separation
Gemma 2 27B (Gemma Team et al., 2024a)	71.4	65.8 / 90.0		separation

Table 1: Measured and reported **ARC Challenge** scores with our assessment of the setup used by authors. The 25-shot prompting used in contrast to the 0-shot is denoted by [†] (in the case authors use such a setup in their report).

Model	Reported	Measured	s / o	Assessment
Llama 65B (Touvron et al., 2023a)	52.3	50.3 / 60.1		separation
Llama 2 70B (Touvron et al., 2023b)	50.7	50.8 / 66.9		separation
Llama 3 70B (Grattafiori et al., 2024)	52.2	51.2 / 72.9		separation
Mistral 7B (Jiang et al., 2023)	— [◊]	50.9 / 62.4		—
Mixtral 8x7B (Jiang et al., 2024)	— [◊]	49.4 / 65.1		—
Mixtral 8x22B (Mistral AI, 2024)	—	51.1 / 67.3		—
DeepSeek 67B (DeepSeek AI et al., 2024a)	—	51.6 / 61.6		—
DeepSeek V2 (DeepSeek AI et al., 2024b)	—	52.2 / 70.0		—
Qwen 14B (Bai et al., 2023)	77.9	56.2 / 78.6		options
Yi 6B (01. AI et al., 2024)	—	52.5 / 71.0		—
Gemma 7B (Gemma Team et al., 2024b)	51.8	51.8 / 60.0		separation
Gemma 2 27B (Gemma Team et al., 2024a)	53.7	58.3 / 70.0		separation

Table 2: Measured and reported **SIQA** scores with our assessment of the setup used by authors. Some authors do not directly report scores but average them with other commonsense reasoning problems (denoted by [◊]), making our assessment unlikely to succeed.

↳ height, it makes the greatest noise or ↳ vibration when it lands on the ground. ↳ What is the best explanation for the ball ↳ making the greatest noise? (Unanswerable ↳ , because in order to choose the best ↳ explanation, one needs to consider ↳ several explanations) - If an experiment results in data that do not ↳ support the hypothesis, what is the most ↳ likely step to take next? (Unanswerable, ↳ because in order to choose the most ↳ likely step, one needs to consider the ↳ less likely alternative) - When an igneous intrusion comes into contact ↳ with surrounding rock, the surrounding ↳ rock will (Unanswerable, because one can ↳ easily verify if an option describes the ↳ possible outcome of contact with ↳ surrounding rock) - A research scientist writes a paper on the ↳ initial regrowth of a forest after a fire ↳ has damaged the entire ecosystem. Which ↳ title would be best for the paper? (↳ Unanswerable, because it is impossible to ↳ decide the best title without comparing ↳ it to other titles) - Jessica wants to see cells in an oak tree leaf. ↳ Which tool is best for Jessica to use to ↳ see the cells? (Unanswerable, because ↳ choosing the best tool depends on the set ↳ of tools considered and is ambiguous ↳ without a complete list of options ↳ considered) - Which factor is most likely to cause the ↳ number of rabbits living in an area to ↳ increase? (Unanswerable, because choosing ↳ the most likely case requires checking ↳ all of the causes under consideration) Now classify the following statements either as ↳ "unanswerable" or "answerable" in ↳ separation. Answer in a form of JSONL file containing " ↳ question", "category", and "explanation" ↳ keys.
---	---

Model	Reported	Measured	s / o / s _b / o _b	Assessment
Llama 65B (Touvron et al., 2023a)	60.2	47.0 / 59.0 / 60.2 / 56.2		separation _b
Llama 2 70B (Touvron et al., 2023b)	60.2	48.8 / 73.0 / 60.0 / 65.8		separation _b
Llama 3 70B (Grattafiori et al., 2024)	47.6	48.6 / 88.4 / 59.4 / 88.5		separation
Mistral 7B (Jiang et al., 2023)	— [◊]	44.2 / 71.6 / 55.0 / 57.8		—
Mixtral 8x7B (Jiang et al., 2024)	— [◊]	47.0 / 80.2 / 55.2 / 78.0		—
Mixtral 8x22B (Mistral AI, 2024)	—	49.6 / 81.6 / 61.2 / 78.4		—
DeepSeek 67B (DeepSeek AI et al., 2024a)	60.2	47.6 / 76.6 / 62.0 / 76.2		separation _b
DeepSeek V2 (DeepSeek AI et al., 2024b)	—	38.6 / 82.8 / 62.4 / 84.2		—
Qwen 14B (Bai et al., 2023)	—	43.8 / 87.0 / 54.6 / 79.8		—
Yi 6B (01. AI et al., 2024)	— [◊]	40.4 / 68.2 / 53.6 / 67.6		—
Gemma 7B (Gemma Team et al., 2024b)	—	44.8 / 65.2 / 58.2 / 65.8		—
Gemma 2 27B (Gemma Team et al., 2024a)	—	47.6 / 83.0 / 59.8 / 81.4		—

Table 3: Measured and reported **OpenBookQA** scores with our assessment of the setup used by authors. Some authors do not directly report scores but average them with other commonsense reasoning problems (denoted by [◊]).

Questions to classify:

[List of 20 questions and choices]

The model returned batches of JSONL, such as:

```
{
  "question": "Which best describes the structure
    ↳ of an atom?", "category": "unanswerable
    ↳ ", "explanation": "Determining the best
    ↳ description requires comparing all
    ↳ options to identify the most accurate one
    ↳ ."
}
{"question": "Which is a statement about climate
    ↳ ?", "category": "answerable", "
    ↳ explanation": "It is possible to verify
    ↳ each option against the definition of
    ↳ climate to determine the correct answer
    ↳ ."
}
{"question": "During which activity should a
    ↳ student wear goggles?", "category": "
    ↳ answerable", "explanation": "It suffices
    ↳ to check if the activity described in the
    ↳ option requires goggles for safety."
}
{"question": "Which natural event occurs with
    ↳ the most frequency?", "category": "
    ↳ unanswerable", "explanation": "
    ↳ Determining the most frequent event
    ↳ requires comparing the frequency of all
    ↳ listed events."
}
```

During this procedure, we estimated the percentage as 21% for ARC Easy and 31% for ARC Challenge.

C Evaluation details

All evaluations were conducted using `lm_eval` 1170ef9 (Gao et al., 2024). We used HF implementations and base variants of models (exact versions in Table 4) with either default prompts and `acc_norm` metric or prompts outlined below.

Inferences were performed with bf16 precision, flash attention (whenever available), and dynamic batch size, using transformers 4.47.0 and torch 2.5.1 on eight NVIDIA H100 GPUs.

Model

```
huggyllama/llama-65b
meta-llama/Llama-2-70b-hf
meta-llama/Meta-Llama-3-70B
mistralai/Mistral-7B-v0.1
mistralai/Mixtral-8x7B-v0.1
mistralai/Mixtral-8x22B-v0.1
deepseek-ai/deepseek-llm-67b-base
deepseek-ai/DeepSeek-V2
Qwen/Qwen-14B
01-ai/Yi-6B
google/gemma-7b
google/gemma-2-27b
```

Table 4: Exact variants of models used for evaluation.

Concerning ARC Easy and Challenge datasets, for the `separation` setup, we follow the standard `lm_eval` configuration with:

```
doc_to_text: "Question: {{question}}\nAnswer:"
doc_to_target: "{{choices.label.index(answerKey)
    ↳ }}"
doc_to_choice: "{{choices.text}}"
```

In contrast, for the `options` setup, we use:

```
doc_to_text: !function arc_utils.doc_to_text
doc_to_target: "{{choices.label.index(answerKey)
    ↳ }}"
doc_to_choice: "{{choices.label}}"
```

with `doc_to_text()` defined as:

```
def doc_to_text(doc):
    prompt = "Question: " + doc["question"] + "\n
    ↳ nOptions:\n"
    for l, t in zip(doc["choices"]["label"], doc
    ↳ ["choices"]["text"]):
        prompt += l + '. ' + t + '\n'
    prompt += "Answer: "
    return prompt
```

Analogous changes were introduced to OpenBookQA and SIQA templates.