Evaluating Pretrained Causal Language Models for Synonymy

Anonymous ACL submission

Abstract

The scaling of causal language models in size and training data has enabled them to tackle increasingly complex tasks. Despite the development of sophisticated tests to reveal their new capabilities, the underlying basis of these complex skills remains unclear. This paper adopts the view that complex skills might be explained using simpler ones, and hypothesizes that linguistic skills may represent these simpler skills. As an initial step in exploring this hypothesis, we focus on one lexical-semantic concept, that of synonymy, laying the groundwork for future investigations into its relationship with more complex skills. We develop a comprehensive test suite to assess various aspects of synonymy under different conditions and evaluate primarily causal open-source models. We find that language models effectively recognize synonymy, but their understanding is significantly influenced by the formulation of the evaluation task.

1 Introduction

As language models have grown in size and have been trained on larger datasets, they have gained the ability to handle tasks that were previously out of reach. This advancement has led to the development of increasingly complex evaluation tasks to explore these new capabilities.

Recent benchmarks, like BIG-bench (Srivastava et al., 2023), introduce tasks that require a blend of different skills to solve effectively. However, it remains unclear what these specific skills are (Arora and Goyal, 2023) and how they are connected to the performance of a language model on a specific task.

In this study, we hypothesize that some of these skills may be characterized using linguistic concepts or their properties. As a first step, we aim to define these properties by examining different aspects of one fundamental linguistic concept, that of synonymy.

While a significant amount of work has focused on evaluating syntactic concepts, either through benchmarks (Warstadt et al., 2020) or studies on specific phenomena (Chowdhury and Zamparelli, 2018; Da Costa and Chaves, 2020; Hu et al., 2020; Wilcox et al., 2023), there has been comparatively less emphasis on evaluating lexical-semantic capabilities, particularly in the case of causal language models. As a result, this area remains relatively underexplored.

Therefore, our study aims to thoroughly examine synonymy recognition in language models by focusing on its distinct properties, including both implicit aspects, such as substituting a word with its synonym, and explicit aspects, such as expressing the relationship between two words.

To connect the data with the linguistic concepts we use minimal pairs, which are a standard method for analyzing linguistic phenomena. In contrast with the majority of previous studies that rely on gap-filling and prompting techniques, we evaluate the minimal pairs using sentence perplexity, which enables us to avoid adding any heuristics to the evaluation.

Our main contributions are the following:

- An evaluation dataset consisting of 1 600 items and 19 200 sentences
- A methodology for constructing evaluations under multiple conditions
- Evaluation results of open-source models along with their analysis
- An analysis of the role of tokenisation
- An investigation of cross-test consistency

2 Related work

Previous work focuses on masked language models like BERT and mBERT (Devlin et al., 2019)

and shows that, while they capture some lexical meaning, by successfully predicting standard lexical relations (Vulić et al., 2020), they struggle with less typical constructions (Espinosa Anke et al., 2021).

Focusing on synonyms, Garcia (2021) finds that BERT's embeddings correctly identify a synonym as closer in representation to the target word than an homonym 75% of the time but fail to distinguish synonyms from words with different meanings in the same context. Extending these findings, Jang et al. (2022) show that pre-trained-only masked LMs seem to perform rather poorly in the synonym/antonym recognition task (SAR), with performance approaching random guess in the case of BERT. On the other hand, in synonym masked word retrieval (MWR) the results seem strong, with an WHR (weighted hit rate) of only 1.4% (lower is better). These results show that the way the test is constructed can significantly affect performance. In line with these findings, our study aims to develop multiple tests that capture various aspects of the same concept for a comprehensive evaluation.

While most lexical-semantics research has been focused on masked language models, the recent prominence of causal language models has led to a growing number of studies dedicated to them. In this context, we mention the work of Truong et al. (2023), which probes causal models for their knowledge of synonyms and antonyms through both cloze-style tasks and prompting. The authors conclude that in most cases the models do not effectively capture these relationships. Additionally, the authors note that the WHR metric, which seemingly indicates strong model performance in cloze tasks, may not be a reliable indicator. Since this metric counts hits from a predetermined list of completions, it does not account for many possible irrelevant completions. In order to avoid this issue in our study, we use sentence perplexity rather than eliciting generation.

3 Experimental settings

In this section we outline the experimental settings, including the minimal pairs tests, the subjacent data and the models.

3.1 Minimal pairs

Our tests are based on minimal pairs, consisting of a correct example and one or more incorrect examples. In each test we vary the following conditions: the type of test, the type of incorrect item and the absence or presence of context. A summary of the tests with examples is given in Appendix A.

First dimension: Test type. We have defined three types of tests, based on properties of synonymy, that we have named `substitution`, `relation` and `reference`.

As a starting point in defining these properties, we consider the following operational definition, adapted from Cruse (1986):

> *Two words are synonyms if they share all or a part of their contexts of occurrence.*

The **substitution** property derives from the definition: if synonyms share their contexts of occurrence, then we can assume a word can be replaced by its synonym without dramatically changing the meaning of the sentence.

> She plays the **character** of a factory worker.
> She plays the role of a factory worker.

The second property comes from discourse. When writing, if a word needs to be repeated in the following sentence, the writer might prefer to use a synonym in order to avoid repetition. We have named this property **reference**, as it involves referencing the initial word by its synonym.

> She plays the **character** of a factory worker.
> This role [...]

Given the difficulty of manually creating a viable continuation after the reference for each item in our evaluation dataset, we have limited ourselves to truncated examples as illustrated above. We believe these provide a relevant approximation, even though they present a more difficult setting.

The third property comes from the ability that humans possess to verbalise the concept of synonymy. A human can identify that there is a specific relation between two words and put a name on this relation. We have called this property simply **relation**.

> She plays the **character** of a factory worker.
> **Character** is a synonym of role.

Second dimension: Negative element. Since we use minimal pairs, the comparison element must

be a single word. Additionally, because we use perplexity instead of gap-filling, it must be specified in advance.

We considered two types of negative elements: one controlled and one uncontrolled. For the controlled case, we deemed a relevant negative element to be **another synonym** of the target word, one that does not align with the context's intended sense. As a result, we limited our target words to those that are polysemous. We provide a more detailed explanation of the synonym selection process in section 3.2.

> She plays the **character** of a factory worker.
> She plays the role of a factory worker.
> She plays the quality of a factory worker.

The uncontrolled element, on the other hand, was represented by ten **randomly selected words**.

> She plays the **character** of a factory worker.
> She plays the role of a factory worker.
> She plays the yoga of a factory worker.

The `substitution` and `relation` tests are valid with both types of negative elements. However, the `reference` test becomes problematic when the negative item is randomly selected, as the random word may accidentally represent a valid reference to another word in the sentence. Hence, for the `reference` test type, we consider only the controlled setting.

> *She plays the **character** of a factory worker.
> This employee

One could argue that the `substitution` test might face a similar issue, as the random selection of candidates is uncontrolled, potentially resulting in incorrect examples that are valid for reasons other than those intended. However, given the difficulty of placing a random word in context, we considered this less likely.

Third dimension: Context. The final aspect of our testing involves the inclusion or exclusion of **context**. `Substitution` and `reference` are meaningless without context, as disambiguation becomes impossible. The only type of test that holds in the absence of context is the `relation` test.

> **Character** is a synonym of role.

In this case, the differentiation between correct synonym and incorrect synonym no longer makes sense. In the absence of context, both words are equally synonyms of the target word. Our goal here will therefore be a different one: to determine whether both synonyms are recognized as synonyms by the language model. To achieve this, we can only compare each of the synonyms with randomly chosen words.

> **Character** is a synonym of role.
> **Character** is a synonym of yoga.

> **Character** is a synonym of quality.
> **Character** is a synonym of yoga.

Lastly, we examined the **term** that a language model prefers for conveying this relationship. The purpose of this experiment was to investigate whether a learner might recognize a connection between two words without necessarily identifying it as "synonymy". We compared the explicit expression of this concept with two different paraphrases for the relationship, as illustrated below.

> **Character** is a synonym of role.
> **Character** is the same as role.
> **Character** means role.

To summarize, we have constructed six tests, as follows: two `substitution` tests, one controlled with an incorrect synonym as negative element and one uncontrolled with ten random words as negative elements, one `reference` test, controlled, and three `relation` tests, one controlled and one uncontrolled with context present and one uncontrolled with context absent.

3.2 Resources

Our tests are based on two resources: Wordnet (Miller, 1994) and Semcor (Miller et al., 1994).

WordNet is a linguistic resource where synonyms are clustered together in groups called synsets and assigned an identifier. SemCor is a corpus consisting of sentences in which one or more words are manually annotated with a sense identifier from WordNet. This identifier permits

to connect the word to a group of synonyms, thus indicating its specific meaning within that context.

Based on these resources we created two preliminary data collections containing all the information needed for our tests. All subsequent evaluation tasks were developed from these collections.

For the first collection we extracted four elements: a target word, its context of appearance, a synonym that fits the context and a synonym that does not fit the context. The target words and contexts were extracted from SemCor. The correct synonym was selected from the synset in WordNet indicated by the sense identifier. For the incorrect synonym, we chose one from the other synsets for that word in WordNet. For the second data collection we extracted a list of random words, that were obtained from WordNet entries (more details on the extraction are provided in Section 3.3).

For example, in the sentence below, the target word is *character* and the associated sense identifier is *1:09:01*. The correct synonym *role* was chosen from the synset designated by the identifier (*role, theatrical role, part, persona*). The incorrect synonym *quality* was chosen among all synonyms in the other synsets (*quality, lineament, fiber, fibre*).

SemCor
She plays the character synset="1:09:01" of a factory worker.

WordNet
character:
1:09:00 quality, lineament (a characteristic property)
1:07:01 fiber, fibre (the complex of attributes that determines a persons morals)
1:09:01 role, theatrical role, part, persona (an actor's portrayal of someone in a play)

3.3 Collection construction

In this section we discuss briefly the extraction of our preliminary data collections, focusing on one key factor impacting their quality: the method for selecting correct and incorrect synonyms out of the respective synsets (for a detailed discussion please see Appendix B).

The inherent properties of SemCor and WordNet theoretically support automated selection of the synonyms. To determine if this selection method is sufficiently effective, we used it to create a collection of 185 items, which was used in turn to create a `substitution`-type test. Human performance on this test showed an accuracy of only 76%, which we judged insufficient.

Consequently, we curated the dataset by selecting the synonyms from the respective synsets manually. The accuracy of human subjects on the resulting `substitution`-type test of 200 items was of 89%, a 13% improvement. The annotators agreed on 83% of the cases, albeit with a Cohen's kappa of 0.46. Given that the human evaluation was a quality check and that the truth value did not come from the human annotations, no adjudication was performed and no items were excluded. The 200-items collection was subsequently used as a basis for all our tests.

3.4 Models

As part of future work we envisage to correlate the performance in lexical-semantic tasks with the characteristics of language models, such as the training data, tokenization and architecture. For this reason, we have evaluated primarily language models that are completely open source, where the training dataset was available, with the exact content and in the exact order as used in training, with open checkpoints and parameters.

Another criterion for the choice of the models was for them to be pre-trained only (no fine-tuning), as we were interested in how they learn directly from data, with no supervision. To be able to explain more easily the results, we added the constraint that models should be monolingual (English). The final condition was that the models needed to have a research paper or technical report attached, for transparency reasons.

We identified two causal models that met our conditions: OLMo (1B, 7B) (Groeneveld et al., 2024) and Amber (7B) (Liu et al., 2023). An inventory of their characteristics is presented in Appendix C.

To ensure these models perform comparably to more established models with similar characteristics, we ran a selection of our tests on the following models: Llama 2 (7B) (Touvron et al., 2023), Llama 3 (8B) (Dubey et al., 2024) and Mistral (7B) (Jiang et al., 2023), in their pre-trained and fine-tuned variants. The three models are monolingual and roughly the same size as our main models.

3.5 Metric

To evaluate an item from our tests with a language model, we measure which of the two examples in

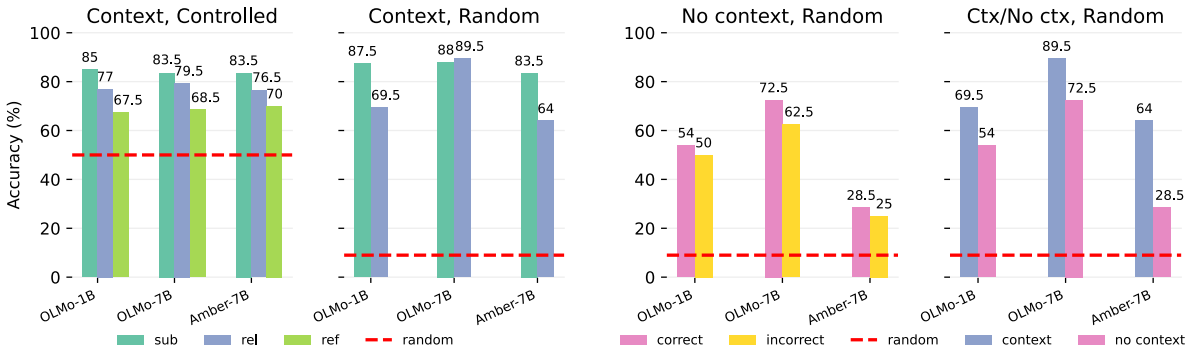

Figure 1: **Left**: Accuracy on the `substitution` (sub), `relation` (rel) and `reference` (ref) tasks with context, under two conditions: **(i)** binary controlled with incorrect synonym (*Controlled*) and **(ii)** ten randomly chosen words (*Random*) **Right**: **(iii)** Accuracy on the `relation` task without context (*No context*), comparing the *correct* and the *incorrect* synonym with random words and **(iv)** Accuracy on the `relation` task, with and without context (*Context/ No context*)

the minimal pair is more likely according to the model. To this end, we use sentence perplexity, which is the exponentiated average log-likelihood of each token in the sentence, conditioned on the previous tokens.

We measure the sentence perplexity of the correct example and that of the incorrect example and, if the sentence containing the correct synonym has the lowest perplexity, than the item is considered correct. For random tests, where a correct example is compared to ten incorrect examples with random words, the item is considered resolved correctly if the correct example has the lowest perplexity among the ten. Accuracy on a test is then calculated by dividing the number of correct items by the total number of items in the test.

4 Main results

We summarize the main results in Figure 1[1]. The first two graphs present the results for tests that include context (`substitution`, `reference`, and `relation` with context): on the left the results under the controlled condition, against the incorrect synonym (Context, Controlled), while on the right those under the random condition (Context, Random) (as detailed in Section 3, the `reference` test is not valid under this setting). The third graph displays the results of the only test without context, namely the `relation` test. This test assesses whether the language model correctly identifies both synonyms, by comparing the performance of the model in recognizing the correct synonym versus random words, and the incorrect synonym versus random words (No context, Random). Finally, the fourth graph compares the results of the `relation` test with context with those of the same test without (Ctx/No ctx, Random). The results are also provided in table form in Appendix D.

The assertions regarding differences in test results are backed by significance tests, which are provided and discussed in Appendix E.

4.1 Synonymy recognition

We observe that language model performance in synonymy recognition tests is affected by all three dimensions we varied in the test design: the type of test, the nature of the negative element (a single incorrect synonym or multiple random words), and the presence or absence of context. We will first discuss the results on tests with context, and afterwards the results on the test without.

Test type. We find that the `substitution` test is the easiest for almost all models under both conditions (first two graphs in Figure 1). This result may be due to the format of the test, which makes it more difficult for incorrect words to fit since they need to be closely related to surrounding words. Inserting an incorrect word disrupts the conditional probabilities throughout the sentence, but when the incorrect word appears at the end, as in the `relation` test, its impact is much more limited. This hypothesis is supported by Figure 2, where we observe that in the `relation` test, the perplexity differences between the correct and incorrect examples are closer to 0, indicating that the model is hesitant. In contrast, the `substitution` test shows a flatter distribution, indicating more pronounced differences between the two examples in the pair.

[1]The code and data are available at https://anonymous.4open.science/r/bm-semlex-62CB/

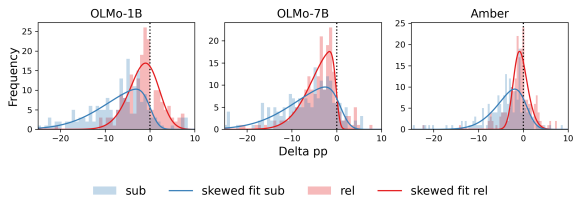

Figure 2: Distributions of perplexity differences for the `relation` and `substitution` tests in the random setting. The image represents a zoom on the fitted gaussian, the complete graph is present in Appendix F.

An exception seems to be the OLMo-7B model, for which the differences between `substitution` and `relation` are not significant. An explanation might be that this model, due to a combination of larger training data and size, manages to better encapsulate the synonymy relation than its competitors and to leverage it even when the synonym is far from the target word in context, thus narrowing the gap between the two tests.

Apart from OLMo-7B, the `relation` test shows a slight decrease in accuracy for all other models, likely due to the increased complexity of having the synonym and the context at a distance and having to connect them via the target word and the relation.

The `reference` test presents an even higher decrease in accuracy. This outcome is expected because the relation is not expressed and the only connector between the synonym and the context is the term 'This'.

Negative element. When the negative element is the *incorrect synonym*, all models have very similar performance (first graph in Figure 1). In contrast, in *random tests* (second graph), the differences between models become more pronounced, especially between OLMo-7B and the other models. This result indicates that random words might have different effects depending on the language models, while the controlled item triggers a more uniform result.

The `substitution` test becomes easier when the negative elements are random words rather than the incorrect synonym, however the differences are not significant. This similarity in performance between the two conditions might be because the context demands strong connections with surrounding words, and both the incorrect synonym and random words are unlikely to fit. For the `relation` test, results seem to have a slightly opposite tendency, namely they seem to be slightly better when the opposing item is the incorrect synonym (with the exception of OLMo-7B). Since the synonym appears at the end of the sentence in this case, its distance from context may allow random words to be more probable.

Context. As the `substitution` and `reference` tests become meaningless without context, only one test type allows us to vary this condition: the `relation` test. Additionally, the only possible negative candidates are random words, since, in the absence of context, both the correct and incorrect synonyms have a synonymy relation with the target word.

The `relation` test without context no longer opposes the correct and incorrect synonym, but permits us to test whether the model correctly identifies each of them as synonyms, by comparing them with random words (third graph in Figure 1). Additionally, it allows us to determine whether there is a bias towards one of the two.

We find that the results are surprisingly good even without context, indicating that, unlike some experiments with BERT-type models, lexical-semantic information is embedded in the word form and can be activated using the term "synonym". We also note that the models seem to demonstrate higher accuracy in recognizing the correct synonym than the incorrect one, with significant differences in the case of OLMo-7B. A probable explanation could be sense frequency: words tagged in SemCor may predominantly appear with their most frequent sense, making the contextual synonym one of the most frequent synonyms for that word. However, we observe that this advantage does not automatically translate in a better performance in the controlled tests with respect to the random tests, as `substitution` still seems to work better with random words.

We observe that when comparing the results on the `relation` test with and without context, context has a significant positive impact for all three models, with the highest increase for the Amber model, from 28.5% to 64% (fourth graph in Figure 1). This outcome is expected, as language models rely primarily on context in their predictions. When context is absent, they may favor words that frequently co-occur with the target word, even if those words are not synonyms.

As mentioned in Section 3.4, in order to verify if the Amber and OLMo models are representative of the broader language model family, we conducted a subset of tests on more well known models like Llama and Mistral. Given that they

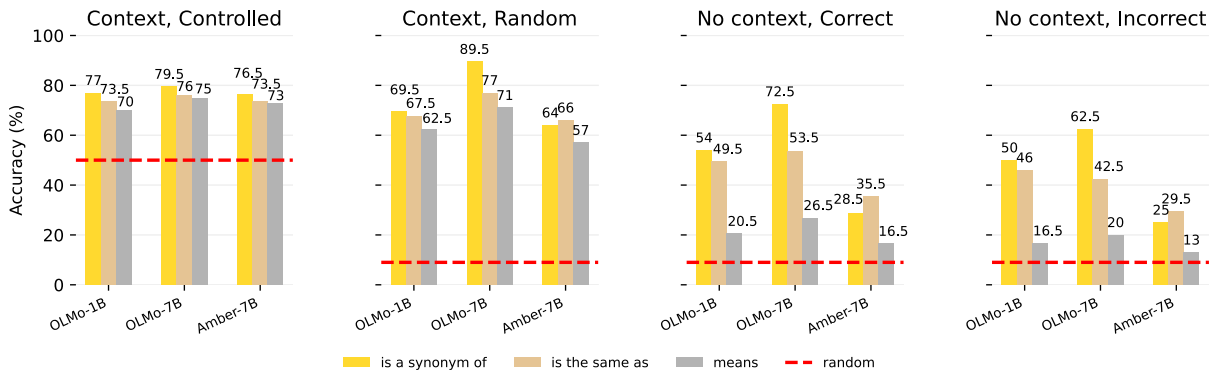

Figure 3: Results on different variants of expressing the synonymy relation: **(i)** Accuracy on the `relation` task with *controlled* context **(ii)** and with uncontrolled, *random* candidates **(iii)** Accuracy on the `relation` task *without context*, for the *correct* synonym against random words and **(iv)** for the *incorrect* synonym against random words.

produced similar results, we were able to conclude that the models are indeed representative, and that their performance accurately reflects that of similar models (see Appendix G for the results).

4.2 Preferred expression

To identify the preferred expression for synonymy, we analyzed the `relation` test, where the synonymy relation is made explicit. We examined both versions with context, involving controlled or random negative candidates (the first two graphs in Figure 3), as well as the version without context (the last two graphs in Figure 3). We compared the standard expression "is a synonym of" with two paraphrases: "is the same as" and "means".

We initially expected the paraphrases to be easier for the language models, assuming the term "synonymy" would appear infrequently in their training data. However, the results show the opposite: the phrase "is a synonym of" yields better performance, at least for the two OLMo models. This may be because paraphrases are more ambiguous and can be associated with other relations, whereas the phrase "is a synonym of" is unequivocal. The exception is the Amber model, that seems to prefer the expression "is the same as". This discrepancy might be due to differences in the composition of the training data of the two models.

We also observe that when context is present, the differences between expressions are modest, but when context is absent, they become more significant. This aligns with expectations, as in the absence of context, the term describing the relation becomes the main distinguishing factor besides the target word.

For all three models, the least preferred expression is "means", likely due to its single-word composition and inherent ambiguity. This is particularly visible in the tests without context, where performance using this expression approaches random guess in certain conditions.

4.3 Impact of tokenization

In linguistic studies, minimal pairs refer to pairs of examples that differ by a single word, which is the relevant unit for human evaluation. For language models, the relevant unit is the token, which can be a word but can also be part of a word. Since our minimal pair examples differ at the word level, they may be tokenized into different numbers of tokens. This discrepancy in token length might affect the conditional probabilities in the sentence perplexity computation and ultimately, the accuracy.

To verify if the accuracy on our tests was affected by the difference in token length, we computed the accuracy for each difference in token length. If the accuracy would remain stable across all possible differences in length, we could conclude that tokenisation does not have a big impact. More precisely, we considered the items belonging to the three types of controlled tests (`substitution`, `relation` and `reference`) given that in these tests we oppose the correct synonym to the incorrect synonym (600 items in total). For each item, we computed the difference in token length between the correct synonym and the incorrect synonym, which ranged from -2 to 2 tokens. Finally, for each value of this difference, we computed how many of the items were correctly predicted (Table 1).

We found that the accuracy changed only slightly when conditioned on the difference in token length. Notably, when the delta is 1, -1 or -2, the accuracy is higher. While a decrease in accuracy is observed with a delta of 2, this result is based on a small

Delta	-2	-1	0	1	2
Accuracy %	78	78	75	79	67
Count	14	117	214	100	14

Table 1: Accuracy conditioned on the difference in token length between the correct and the incorrect synonym. A delta of -2 means that the correct synonym is 2 tokens shorter that the incorrect synonym.

number of items. The Pearson correlation coefficient points in the same direction, with a value of -0.015, indicating virtually no correlation between the difference in tokens and the accuracy.

We therefore conclude that the difference in token length between the synonyms does not affect accuracy in our tests.

4.4 Consistency across tests

In this section we aim to measure the consistency of language models across different tests. We examine whether a minimal pair that is accurately predicted under a set of specific test conditions, continues to be accurately predicted when varying the conditions. To achieve this, we compute the conditional probability of being correct (or wrong) in the second test, knowing we were correct (or wrong) in the first test. We provide the complete counts and calculations in Appendix H.

Test type. We start by examining whether the results on the `substitution` test are coherent with those on the `relation` and `reference` tests, when keeping the same conditions on the negative element (incorrect synonym) and context (present). We find that a correct item in the `substitution` test is also correct in the `relation` test in 83% of the cases. However, if the item is incorrect, the probability for it to also be incorrectly predicted in the `relation` test is only 57%. For the `reference` test, the prediction based on `substitution` is even worse: the correct items of the `reference` test are predicted with a lower probability (of 71%), while the incorrect items are at random guess accuracy.

Negative example. Next, we look at whether the controlled condition can be used in order to predict the random one, when the test type stays the same. We find that in the `substitution` test, if an item is correctly predicted in the controlled test, there is a high likelihood (91%) it will also be correctly predicted in the random test. However, only a small portion of the errors in the controlled test carry over in the random one (33%).

Context. Finally, we investigate whether the results of the `relation` test without context could be predicted from the same test with context. We found that this is indeed the case for 74% of the correct items, but the incorrect items differed between the two tests.

We conclude that while language models seem to be relatively consistent in the correct items they identify in different tests, the errors they commit vary with each test. However, the number of incorrect items might be too small in order to be sure that our results are relevant. We hope to be able to address this limitation in future work.

5 Conclusions

In this study, we evaluated pre-trained causal language models on the linguistic concept of synonymy. We developed a comprehensive test suite designed to assess various aspects of synonymy, both in controlled and uncontrolled environments, and validated our datasets with human subjects.

We found that the performance of language models depends on the test conditions. Overall, the `substitution` test is easier, followed by the `relation` and `reference` tests. Synonymy tests are generally easier when context is provided: `substitution` tests are performed better with random words, while `relation` tests are easier when the opposing item is an incorrect synonym, indicating that the position of the word might play a significant role. Even in the absence of context, although accuracy decreases as expected, models like OLMo-7B still achieve relatively high accuracy. We also observed a bias favoring the correct synonym over the incorrect one, likely due to a sense frequency bias in SemCor.

The preferred expressions for conveying synonymy vary by language model: "is a synonym of" for the OLMo models and "is the same as" for the Amber model. We attribute this discrepancy to differences in training data content.

Finally, while language models consistently identify the same items correctly, the errors they make are different depending on the test.

For future work, we aim to investigate how training data affects model performance on synonymy tests. We plan to use the evaluation framework developed in this study to examine the progression of language models during training, analyze differences in preferred expressions for synonymy, and explore performance variations between models.

6 Limitations

Resources. SemCor is an outdated dataset that contains inappropriate language (from the authors' observations). Although we attempted to mitigate this issue through manual selection, using newer datasets with similar annotations may be more advisable.

Dataset. With respect to the dataset construction, we focused exclusively on nouns, non-compound words and polysemous words (on account of our methodology for the latter). However, the synonymy phenomenon extends to other grammatical categories. Furthermore, the manual selection of synonyms by the authors may have introduced bias, as the chosen synonyms were selected to facilitate easier substitution, potentially making the `substitution` test easier.

Another point to consider is that we did not evaluate whether the synonym pairs we selected are frequent enough and representative of the synonymy phenomenon in English.

Methodology. Given that we were primarily interested in the accuracy that models show on the tests, we do not directly use the perplexity values of the correct/incorrect sentences in our study, but the sign of the difference (except for one analysis). We believe that these values might provide valuable insight.

Regarding the test conditions, we believe that the conditions we varied in our dataset are not exhaustive; other properties of synonymy, such as commutativity, could lead to the development of new test types. Additionally, our two negative candidates are unequal in number, and further insights might be gained by comparing an incorrect synonym to a single random word instead of ten, as well as comparing ten random words to ten synonyms.

Results. Concerning the results, while we conducted some tests on well-known models, we did not run the entire test suite and therefore our conclusion regarding their relevance is limited to the scope of the tests that we were able to do. Furthermore, all models we did test were of roughly similar sizes; larger models may yield different insights. Additionally, we only compared the standard expression of synonymy to two paraphrases that we subjectively deemed relevant. More extensive testing could reveal a more favored expression. In discussing the results, we occasionally shared our intuitions or hypotheses about why certain outcomes occurred, even though these insights were not validated through experiments.

We only examined the impact of tokenization in the controlled setting and cannot rule out the possibility of an effect being present in the random setting. When measuring the consistency models display across tests, we varied a limited range of conditions, and we generally conditioned on the test having better accuracy. Some of the analysis was based on a relatively low number of examples (around 30), which might be insufficient to make a relevant claim.

7 Ethical considerations

B2. We downloaded SemCor and WordNet from `nltk` (version `3.8.1`)[2]. The licence mentioned (WordNet copyright and license agreement) grants permission to use, modify and distribute the resource. The models we tested were downloaded from Hugging Face[3]. From our understanding, all models licenses (Apache license 2.0 for OLMo and Amber, Mistral AI Research License for Mistral, Llama 2 Community License Agreement for Llama 2 and Llama 3.2 Community License Agreement for Llama 3) permit use of the models. The license for the code and data that we release are present in the repository url we provide in Section 4.

C1. The experiments were performed on the existing local computing infrastructure.

D1. The text of instructions given to annotators (mentioned in section 3.3) is available in the repository.

D2. The two annotators were recruited among authors' acquaintances and were not paid.

E1. Code writing was assisted by Copilot[4] and occasionally ChatGPT[5].

References

Sanjeev Arora and Anirudh Goyal. 2023. A theory for emergence of complex skills in language models. *Preprint*, arXiv:2307.15936.

Shammur Absar Chowdhury and Roberto Zamparelli. 2018. RNN simulations of grammaticality judgments on long-distance dependencies. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 133–144, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

[2] https://www.nltk.org/nltk_data/
[3] https://huggingface.co/models
[4] https://github.com/features/copilot
[5] https://chatgpt.com/

D. A. Cruse. 1986. *Lexical Semantics*. Cambridge University Press, New York.

Jillian Da Costa and Rui Chaves. 2020. Assessing the ability of transformer-based neural models to represent structurally unbounded dependencies. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 12–21, New York, New York. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Abhimanyu Dubey et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Luis Espinosa Anke, Joan Codina-Filba, and Leo Wanner. 2021. Evaluating language models for the retrieval and categorization of lexical collocations. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1406–1417, Online. Association for Computational Linguistics.

Marcos Garcia. 2021. Exploring the representation of word meanings in context: A case study on homonymy and synonymy. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3625–3640, Online. Association for Computational Linguistics.

Dirk Groeneveld et al. 2024. Olmo: Accelerating the science of language models. *Preprint*, arXiv:2402.00838.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Myeongjun Jang, Frank Mtumbuka, and Thomas Lukasiewicz. 2022. Beyond distributional hypothesis: Let language models learn meaning-text correspondence. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2030–2042, Seattle, United States. Association for Computational Linguistics.

Albert Q. Jiang et al. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Zhengzhong Liu et al. 2023. Llm360: Towards fully transparent open-source llms. *Preprint*, arXiv:2312.06550.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

George A. Miller, Martin Chodorow, Shari Landes, Claudia Leacock, and Robert G. Thomas. 1994. Using a semantic concordance for sense identification. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Aarohi Srivastava et al. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Hugo Touvron et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Thinh Hung Truong, Timothy Baldwin, Karin Verspoor, and Trevor Cohn. 2023. Language models are not naysayers: an analysis of language models on negation benchmarks. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 101–114, Toronto, Canada. Association for Computational Linguistics.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240, Online. Association for Computational Linguistics.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. 2023. Using Computational Models to Test Syntactic Learnability. *Linguistic Inquiry*, pages 1–44.

A Description of different tests with examples

test	condition	example
sub	context, incorrect syn	She plays the role of a factory worker. She plays the quality of a factory worker.
	context, random wds	She plays the role of a factory worker. She plays the random1 of a factory worker. ... She plays the random10 of a factory worker.
ref	context, incorrect syn	She plays the **character** of a factory worker. This role She plays the **character** of a factory worker. This quality
rel	context, incorrect syn	She plays the **character** of a factory worker. **Character** is a synonym of role. She plays the **character** of a factory worker. **Character** is a synonym of quality.
	context, random wds	She plays the **character** of a factory worker. **Character** is a synonym of role. She plays the **character** of a factory worker. **Character** is a synonym of random1. ... She plays the **character** of a factory worker. **Character** is a synonym of random10.
	no context, random wds	**Character** is a synonym of role. **Character** is a synonym of quality. **Character** is a synonym of random1. ... **Character** is a synonym of random10.

Table 2: Example of test items for each of the test types (`substitution`, `reference` and `relation`) under different conditions.

B Data collection construction

The inherent properties of SemCor and WordNet would theoretically guarantee that any word from the synset linked to the identifier would constitute a viable correct synonym, while any word from other synsets would function as a viable incorrect synonym. This supports the possibility of an automated selection process.

In order to test the viability of this method, we created a data collection of 185 sentences by selecting the synonyms in an automatic way according to the heuristic we perceived to be the most discriminatory (frequency count in WordNet). From this collection we constructed a `substitution`-type test (Table 2) that we evaluated on human subjects.

The human subjects achieved scores of approximately 76%, which we considered too low to ensure sufficient quality for further experiments. To enhance the dataset's quality, we conducted a brief qualitative analysis, which allowed us to identify several problematic cases:

- some SemCor examples lack sufficient discriminating context - making it difficult to distinguish between the correct synonym *sum* and the incorrect one *quantity*

  > amount - sum - quantity
  > Multiply the result obtained in item 3 above by the amount used for each State in item 1 above.

- broken substitution - replacing the target word *bit* with the synonym *spot* does not yield a grammatical output (as the target word is part of a collocation)

  > bit - spot - moment
  > But I'm not one damned bit sorry I went out to question the people I know.

- WordNet is not exhaustive - *hare* appears in WordNet only as the meat and not as the animal, leading it to appear in a different synset than *coney*, even tough they share the same sense of *animal*

> rabbit - coney - hare
> We come upon a **rabbit** that has been caught in one of the brutal traps in common use.

Given the poor performance we registered for human subjects, as well as the errors identified in the analysis, we decided to curate the dataset, by choosing the contexts and the synonyms manually ourselves. We constructed a dataset of 200 items, each composed of the 5 elements described at the end of section 3.2.

To measure the improvement brought about by this method, we conducted another `substitution` test on human subjects. We observed that the accuracy improved on average by 15% (Table 3). The two annotators agreed on 83% of the cases, with a Cohen's kappa of 0.46. Finally, each annotator found around 30 out of 200 (15%) items to be bizarre in some way (ambiguous, inadequate choice of triples etc.), agreeing on 7 items.

Measure	Annot 1	Annot 2	Agree
Accuracy %	88.2	89.5	**83.3**
Weird item %	15.5	14	13

Table 3: Accuracy of the two annotators on the `substitution` test. On the second row, count of perceived bizarre items. The accuracy result for the annotator 1 is out of 186 (instead of 200), because they did not annotate a subset of the items that they found bizarre.

Given that this experiment is a sanity check, the annotations described above are not part of the final dataset. The truth value does not come from the annotations, but from the WordNet synsets and the authors' manual choice. Therefore, no adjudication was needed, and we decided to not exclude any of the items labeled as bizarre.

As we could observe an improvement in accuracy on this curated dataset with respect to the automatically-generated one, we used it as a base for all the subsequent experiments. From the data collection of 200 items and the randomly selected nouns from a list of 10 000 we constructed datasets for each of our tests, amounting to a total of 19 200 sentences.

C Models parameters

Size	Train. Tokens	Vocab	Hidden Size	Attn. Heads	Context Length	Pos.
OLMo 1B	2 Trillion	50 304	2048	16	2048	RoPE
OLMo 7B	2.5 Trillion	50 304	4096	32	2048	RoPE
Amber 6.7B	1.2 Trillion	32 000	4096	32	2048	RoPE

Table 4: Specifications of OLMo and Amber models.

D Main results table

The results are provided in Table 5.

E Significance

To verify if the results on our tests, discussed in Sections 4.1 and 4.2 show significant differences, we conducted paired two-tailed t-tests each time we claim that differences are present. The results of these tests are provided below to support the validity of our assertions.

The t-tests are performed on distributions of 200 values of 0 and 1, corresponding to the 200 test items correctly or incorrectly evaluated by a model. The average of these distributions corresponds to the

Test	Cond	O-1B	O-7B	Amber
sub	syn	85	83.5	**83.5**
	rand	**87.5**	88	**83.5**
ref	syn	67.5	68.5	70
rel	syn	77	79.5	76.5
ctx	rand	69.5	**89.5**	64

Test	Sub-test	O-1B	O-7B	Amber
rel	correct	54	72.5	28.5
no ctx	incorrect	50	62.5	25
	both	33	52	14.5

Table 5: **Left:** Accuracy (%) on the `substitution`, `reference` and `relation` with context under different conditions. **Right:** Accuracy (%) on the `relation` tests without context.

task's accuracy. The t-test's null hypothesis is that these averages are identical. We consider the results significant when the p-value is inferior to 0.05.

1. *We find that the* `substitution` *test is the easiest for almost all models under both conditions.*

   We show in Table 6 that the differences between the `substitution` on one side and the `relation` or `reference` tests on the other are significant in all conditions for the OLMo-1B And Amber-7B models. For the OLMo-7B model, while the difference is significant between the `substitution` and the `reference` test, the evidence is insufficient in the case of the `substitution` and `relation` tests.

Model	**sub vs. rel**				**sub vs. ref**	
	controlled		random		controlled	
	t	p	t	p	t	p
OLMo-1B	2.50	0.013	5.00	0.000	13.71	0.000
OLMo-7B	1.18	0.239	-0.52	0.603	10.92	0.000
Amber-7B	2.13	0.034	5.32	0.000	10.08	0.000

Table 6: Paired t-test between the results on `substitution` and `relation` or `substitution` and `reference` tests. P-values greater than 0.05 are highlighted in red.

2. *Apart from OLMo-7B, the* `relation` *test shows a slight decrease in accuracy for all other models.*

   Supported by the same evidence provided in Table 6.

3. *When the negative example is the incorrect synonym, all models have very similar performance. In contrast, in random tests, the differences between models become more pronounced.*

   We observe from Table 7 that, in the case of the controlled condition (incorrect synonym), although we cannot say that the performances are the same, at least there is insufficient evidence of the contrary.

Model 1	**Model 2**	sub		rel		ref	
		t	p	t	p	t	p
OLMo-1B	**OLMo-7B**	0.73	0.47	-1.15	0.25	-0.17	0.86
OLMo-1B	**Amber-7B**	0.60	0.55	0.18	0.86	-0.59	0.56
OLMo-7B	**Amber-7B**	0.00	1.00	1.13	0.26	-0.44	0.66

Table 7: Paired t-test between the models on different tests, in the **controlled** condition.

   With respect to the second claim, namely that differences are more pronounced in the random condition, this is only significant when comparing OLMo-1B and Amber with the OLMo-7B model in the `relation` test (Table 8).

4. *The* `substitution` *test becomes easier when the negative candidates are random words rather than the incorrect synonym. For the* `relation` *test, on the contrary, results are better when the opposing item is the incorrect synonym.*

Model 1	Model 2	sub		rel	
		t	p	t	p
OLMo-1B	**OLMo-7B**	-0.24	0.809	-6.31	0.000
OLMo-1B	**Amber**	1.52	0.131	1.52	0.131
OLMo-7B	**Amber**	1.74	0.083	7.50	0.000

Table 8: Paired t-test between the models on different tests, in the **random** condition.

In the case of `substitution`, we observe from Table 9 that while there seems to be a difference in favor of the random condition, the results are not significant. In the case of the `relation` test the results seem to lean towards a small advantage of the incorrect synonym, with the exception of OLMo-1B.

Model	**sub** random vs. controlled		**rel** random vs. controlled	
	t	p	t	p
OLMo-1B	0.84	0.399	-1.90	0.059
OLMo-7B	1.48	0.139	3.08	0.002
Amber-7B	0.00	1.000	-3.40	0.001

Table 9: **Left:** Paired t-test for the `substitution` test, between the controlled and the random condition, for each model. **Right:** Same for the `relation` test.

5. *We also note that the models seem to demonstrate higher accuracy in recognizing the correct synonym than the incorrect one.*

   We observe that the differences are only significant in the case of OLMo-7B (Table 10).

Model	correct vs. incorrect	
	t	p
OLMo-1B	0.92	0.360
OLMo-7B	2.58	0.011
Amber-7B	1.00	0.319

Table 10: Paired t-test for `relation` with context between the correct synonym and the incorrect synonym tests.

6. *We observe that when comparing the results on the* `relation` *test with and without context, context has a significant positive impact for all three models.*

   From Table 11, we observe that the differences are significant between the test with context and the same test without, for all models.

Model	relation context vs. no ctx	
	t	p
OLMo-1B	3.80	0.000
OLMo-7B	4.61	0.000
Amber-7B	9.01	0.000

Table 11: Paired t-test for the relation test, between the test with context and the test without.

7. *The phrase "is a synonym of" yields better performance, at least for the two OLMo models. [...] the exception is the Amber model, that seems to prefer the expression "is the same as".*

   As shown in Table 12, for the tests with context we observe a preference for the expression "is a synonym as". This preference is significant when compared to the expression "means", but less significant when compared to "is the same as". With respect to the `relation` test without context,

the preference for the "is a synonym of" expression is significant in almost all cases. One exception is the Amber model, that has a preference for the expression 'is the same as', which is significant in the case of the `relation` test without context.

Model	"syn of" vs "means"				"syn of" vs "same as"			
	controlled		random		controlled		random	
	t	p	t	p	t	p	t	p
OLMo-1B	2.91	0.004	2.43	0.016	1.61	0.108	0.71	0.481
OLMo-7B	2.08	0.039	6.32	0.000	1.82	0.071	4.90	0.000
Amber	1.40	0.162	2.36	0.019	1.74	0.083	-0.71	0.481

Table 12: Paired t-test between the expressions "is a synonym of" and "means" and between "is a synonym of" and "is the same as" for the `relation` with context test.

Model	"syn of" vs "means"				"syn of" vs "same as"			
	correct		incorrect		correct		incorrect	
	t	p	t	p	t	p	t	p
OLMo-1B	10.01	0.000	10.01	0.000	1.74	0.083	1.30	0.195
OLMo-7B	13.02	0.000	11.89	0.000	5.94	0.000	5.76	0.000
Amber	4.78	0.000	4.98	0.000	-2.69	0.008	-1.68	0.095

Table 13: Paired t-test results between the expressions "is a synonym of" and "means" and between "is a synonym of" and "is the same as" for the `relation` without context test.

8. *We also observe that when context is present, the differences between expressions are modest, but when context is absent, they become more significant.*

   This affirmation is supported by the results in Tables 12 and 13, where we observe that when context is present (Table 12) less tests have significant differences. On the contrary, when context is absent (Table 13), few of the tests do not have significant differences.

9. *For all three models, the least preferred expression is "means".*

   As we can see from the previously-mentioned tables, the differences between the preferred expression and "means" are almost in all cases significant.

F Histogram of delta perplexities

The histogram of differences in perplexity between the correct and incorrect synonym is provided in Figure 4.

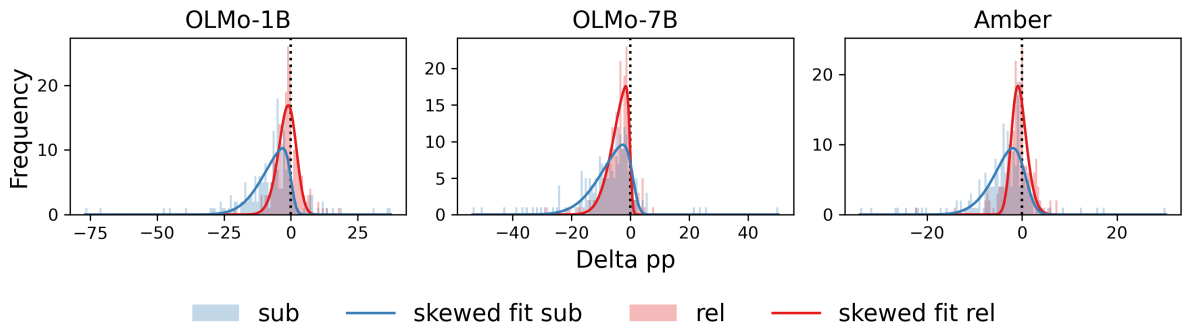

Figure 4: Histogram of delta perplexities

G Synonymy test results on more well-known models

To make sure that the open-source models we chose for the evaluation are not exceptions in the landscape of language models, we have performed some of the tests on other, more well-known language models. We performed two tests suites: (i) the three test types with the incorrect synonym (Figure 5) and (ii) the `relation` test without context, where both the correct and the incorrect synonym must be better than random words at the same time (Figure 6).

Looking at the results, we observed that other models show the same stability in performance, as well as the same accuracy range in the `substitution`, `relation` and `reference` tests with the incorrect synonym, as the OLMo and Amber models (5). For the `relation` test where context is absent, the differences between language models are more pronounced, but we notice that our three language models are representative of the accuracy values we encounter: OLMo-7B has the highest value, OLMo-1B approaches the mean, while Amber is the second lowest.

We concluded that the OLMo and Amber models are representative of their type and size within the language model family, and thus their results accurately reflect those of similar models.

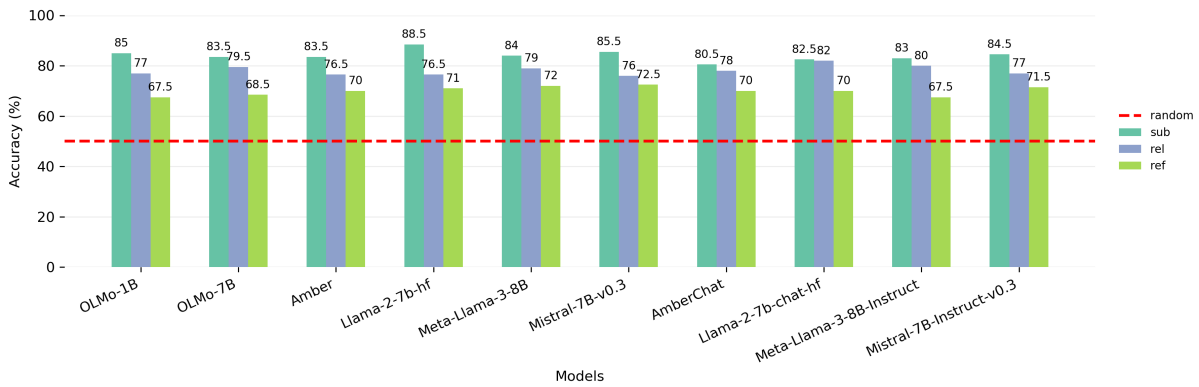

Figure 5: Accuracy on the `substitution`, `relation` and `reference` tasks (with context) for a diverse range of models.

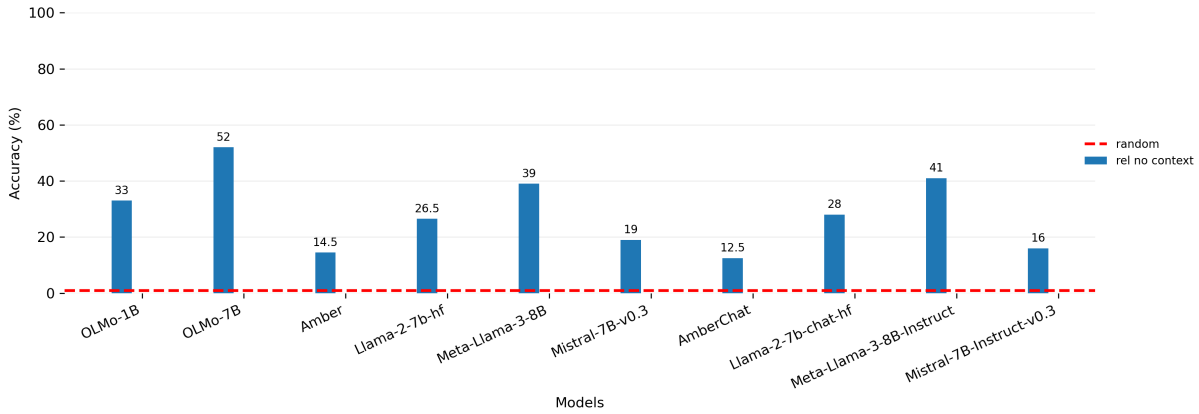

Figure 6: Accuracy on the `relation` task without context for a diverse range of models.

H Consistency across models, conditional probabilities

We provide the complete results of the conditional probabilities across different test settings in Tables 14-17 below. For example, the first line of Table 14 should be read as follows:

$\mathbb{P}(Rel = 0|Sub = 0) = 0.57$ (the probability that `relation` is incorrect (0) knowing `substitution` is incorrect is 57%)

$\#(Rel = 0 \cap Sub = 0) = 17$ (the number of incorrect items in both `relation` and `substitution` is 17)

Sub	Rel	$\mathbb{P}(Rel\|Sub)$	#
0	0	0.57	17
0	1	0.43	13
1	0	0.17	29
1	1	0.83	141

Table 14: Distribution of the correct (1)/incorrect(0) items of the `relation` test conditioned on the correct and incorrect items of the `substitution` test, OLMo-1B.

Sub	Ref	$\mathbb{P}(Ref\|Sub)$	#
0	0	0.57	16
0	1	0.43	14
1	0	0.29	49
1	1	0.71	121

Table 15: Distribution of the correct/incorrect items of the `reference` test conditioned on the correct and incorrect items of the `substitution` test, OLMo-1B.

Ctrl	Rand	$\mathbb{P}(Rand\|Ctrl)$	#
0	0	0.33	10
0	1	0.67	20
1	0	0.09	15
1	1	0.91	155

Table 16: Distribution of the correct/incorrect items of the random `substitution` test conditioned on the correct and incorrect items of the controlled `substitution` test, OLMo-1B.

No_ctx	Ctx	$\mathbb{P}(Ctx\|No_ctx)$	#
0	0	0.38	8
0	1	0.62	13
1	0	0.26	47
1	1	0.74	132

Table 17: Distribution of the correct/incorrect items of the `relation` test with context conditioned on the correct and incorrect items of the `relation` test without, OLMo-7B.