Evaluating Pretrained Causal Language Models for Synonymy

Ioana Ivan **Carlos Ramisch** **Alexis Nasr**
Aix Marseille University, CNRS, LIS, Marseille, France
{ioana.ivan, carlos.ramisch, alexis.nasr}@lis-lab.fr,

Abstract

The scaling of causal language models in size and training data 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. We argue that complex skills might be explained using simpler ones, represented by linguistic concepts. As an initial step in exploring this hypothesis, we focus on the lexical-semantic concept of synonymy, laying the groundwork for research into its relationship with more complex skills. We develop a comprehensive test suite to assess various aspects of synonymy under different conditions, and evaluate causal open-source models ranging up to 10 billion parameters. We find that these models effectively recognize synonymy but struggle to generate synonyms when prompted with relevant context.

1 Introduction

As language models (LMs) 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.

While complex datasets can demonstrate LM overall progress (Wei et al., 2022), they assess multiple implicit skills simultaneously, often without offering a clear picture of what precise skills contribute to performance (Arora and Goyal, 2023). For instance, a complex skill such as creative writing involves a combination of subjacent skills, including temporal coherence to ensure logical progression, commonsense reasoning to create plausible scenes, and the ability to use synonyms to avoid repetition and enrich the vocabulary. The interest in the connection between complex skills and their subjacent ones is twofold. First, Chen et al. (2023) propose that training first on subjacent skills such as synonymy may be more effective than training on complex skills like creative writing directly. Second, if performance on subjacent tasks correlates with performance on complex ones, this could help explain why language models sometimes appear to acquire complex abilities suddenly, suggesting that the necessary building blocks were already in place.

These subjacent skills might be defined in different ways, such as in terms of performance-improving datasets (Chen et al., 2023) or logical reasoning and rhetoric (Yu et al., 2024). Another way to define them could be via linguistic concepts. Although this approach might seem trivial, the majority of benchmarks do not test linguistic concepts, and the training of LMs does not particularly take them into account. Given that the skills of LMs are not yet well defined, we propose linguistic concepts and their properties as a starting point to structure the exploration of LM skills.

While a significant amount of work focused on syntax, either through benchmarks such as BLiMP (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 for causal LMs. As a result, this area remains relatively underexplored. Our study aims to examine the lexical-semantic concept of synonymy, by focusing on its different properties, including both implicit aspects, such as substituting a word with its synonym, and explicit aspects, such as expressing the relation between two words.

In line with previous benchmarks on linguistic knowledge such as BLiMP (Warstadt et al., 2020) and SyntaxGym (Gauthier et al., 2020), we create minimal pairs that we evaluate using sentence perplexity. The advantage of this method is that it does not require a prompt in the form of an instruction and can be performed on small, pre-trained only

Findings of the Association for Computational Linguistics: ACL 2025, pages 12533–12551

models. However, while this method does indicate whether a model tends to choose the correct synonym over an alternative, it does not determine whether a synonym will actually be generated when the context constrains the model to do so. Therefore, we develop a second set of evaluations, based on gap filling, to complete the study.

Our main contributions are:

- A minimal pairs dataset consisting of 1,600 items and 19,200 sentences and a gap-filling dataset consisting of 400 items,[1]
- A methodology for constructing evaluation datasets under multiple conditions,
- Evaluation of LMs and analyses of the role of tokenization and of cross-test consistency.

2 Related work

Previous work focuses on masked language models like BERT (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 focuses on masked LMs, the recent prominence of causal LMs led to a growing number of studies dedicated to them. Truong et al. (2023) probe causal models for their knowledge of antonyms through both cloze-style tasks and prompting. They conclude that in most cases the models do not effectively capture this relationship, underlining that the WHR metric, which seemingly indicates strong model performance in cloze tasks, may not be a reliable indicator. Since it counts hits from a predetermined list of completions, it does not account for many possible irrelevant completions. To avoid this issue in our gap-filling experiments, we used a list of correct completions instead of incorrect ones.

3 Experimental settings

This section outlines the experimental settings, including minimal pairs and gap-filling tests, subjacent data, and models.

3.1 Minimal-pairs tests

The tests based on minimal pairs consist of a correct example and one or more incorrect ones. We vary the type of test, the type of negative element, and the presence of context (examples in Appendix A).

Test type Based on properties of synonymy, we define three types of tests: `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 contexts, then we 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 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.

[1] Code and datasets used in this study can be found at https://github.com/ioana-ivan/bm-semlex.

The third property comes from the ability that humans possess to verbalize 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**. This test is similar to the MWR test from Jang et al. (2022), with the difference that in our case it can also include context.

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

Negative element Unlike gap filling, in minimal pairs the compared element must be specified in advance, and must be a single word for the pair to be minimal. We considered two types of negative elements: controlled (C) and uncontrolled (U). 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 (details in § 3.3). As a result, we limit our target words to those that are polysemous. 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.
> (C) She plays the quality of a factory worker.
> (U) 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 is problematic when the negative item is randomly selected, as the random word may accidentally make a valid reference to another word in the sentence. Hence, for the `reference` test, we consider only the controlled setting.[2]

Context This aspect 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 `relation`.

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

In this case, distinguishing correct and incorrect synonyms 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 test suites: 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 two uncontrolled (with/without context).

3.2 Gap-filling tests

Causal language models only take into account left context in generation, requiring gap-filling tests to place the gap at the end. This effectively means giving the model a prompt and obtaining its response.

We attempted to match every minimal pairs test with a corresponding gap-filling test, however not all test types are suitable. In the `substitution` test, the position of the synonym in the sentence is variable and there is nothing in the left context to trigger the generation of a word in particular (or their synonym). For example, the prompt *She plays the* can be validly completed by the word *flute* and not the intended words *character* or *role*. The `reference` test faces the same issue: *She plays the*

[2]The `substitution` test might face a similar issue, as random candidate selection may coincidentally result in incorrect examples that are valid. However, given the difficulty of placing a random word in context, we considered this unlikely.

character of a factory worker. This is nice.

This problem does not manifest for the `relation` test. The presence of the lexical relation name (*synonym*) and of the original word (*character*) constrains the output, making it unlikely to have a valid continuation that is not a synonym.

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

Thus, we considered only the `relation` test type for gap-filling and, due to space constraints, conducted tests only for `relation with context`. We used a single prompt, the same as in the perplexity tests. The set of correct elements was either the curated correct synonym used in perplexity tests or the correct synset from WordNet. To support evaluation of models not fine-tuned on instructions, we aimed to keep the prompt as simple as possible.

LMs generate text as tokens instead of words, which brings about the challenge of determining the number of tokens needed to capture the target output. Given that target synonyms might be multi-token, that there might be orthographic variation (e.g. quotes around the predicted word) and meta-linguistic content ("is a synonym of the word [...]"), we settled on a number of 10 tokens, high enough to permit some variability in the form of the answer, but small enough to keep the synonym in proximity to the prompt. We varied the number of predicted sequences and of correct completions (synonyms).

Number of predictions LMs can generate the top-most probable output (greedy sampling), but this setting is highly restrictive, raising the question of whether the correct response might be reached by considering a broader part of the next-token distribution. Therefore, we established a strategy to generate probable multi-token sequences: we decided on sampling from the distribution by setting the top-k parameter (with $k = 10$). In this way, we explore a broader range of top tokens in the distribution and also follow a common practice in generation. We considered two scenarios: (1) a single prediction of a ten-token sequence in a greedy manner, and (2) ten predictions of ten-token sequences by sampling from the top-k tokens.

Number of synonyms We considered two settings: (1) comparing the prediction with the curated synonym (used in perplexity-based tests), and (2) comparing it with the entire synset.

3.3 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 connects the word to a synset specifying its meaning in context. Based on these resources we created three preliminary data collections containing all the information needed to build our tests. All subsequent evaluation tasks were developed from these collections.

For the first collection we extracted four elements: a target word, its context in SemCor, a synonym that fits the context and a synonym that does not, both from WordNet. The correct synonym was selected from the synset indicated by the sense identifier, while the incorrect synonym was chosen from the other synsets of the word. The second collection contained the same elements, but instead of selecting correct or incorrect synonyms, we extracted the entire synsets. For the third data collection we extracted a list of random words, that were obtained from WordNet entries (more details on the extraction are provided in Sec. 3.4).

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 of a factory worker.
>
> **WordNet** character:
> 1:09:00 quality, lineament
> 1:07:01 fiber, fibre
> 1:09:01 role, theatrical role, part, persona

We restricted our choice of target word to non-compound, polysemous nouns. While focusing exclusively on polysemous words may introduce bias, we considered them to present a more difficult case. For monosemous words, the word form alone may suffice to identify its unique sense and, therefore, its synonym. In contrast, for polysemous words, although the form may help identify a set of possible synonyms, context is required to disam-

biguate between them, which is an added difficulty. Regarding part of speech, we focused on nouns, as more aspects of synonymy were more straightforward to test on them. For instance, the substitution test is difficult to adapt to verbs without manual intervention, and the reference test in its current form does not apply to verbs at all. Finally, we selected non-compound words to simplify inflection.

3.4 Collection construction

The inherent properties of SemCor and WordNet theoretically support automated synonym selection. To determine whether 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 (discussion in Appendix B).

Consequently, we curated the dataset by selecting the synonyms from the respective synsets manually. The accuracy of humans on the resulting `substitution`-type test of 200 items was of 89%, a 13% improvement. Annotators agreed on 83% of the cases, albeit with a Cohen's kappa of 0.46. Given that human evaluation was a quality check and that the truth value did not come from 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.

3.5 Models

We envisage correlating the performance in lexical-semantic tests with the characteristics of LMs in the future. Therefore, our evaluation favors open-source LMs with available training datasets, checkpoints, and hyperparameters. Another criterion for the choice of models was pre-trained only models (no fine-tuning), as we were interested in how they learn directly from data, with no supervision. To simplify result interpretation, we added the constraint that models should be monolingual (English). The final condition was that the models must be described in a research paper or technical report.

Two causal models met our conditions: OLMo (1B, 7B) (Groeneveld et al., 2024) and Amber (7B) (Liu et al., 2023), described in Appendix C. To ensure these models perform comparably to more established ones with similar characteristics, we ran a selection of our tests on Llama 2 (7B) (Touvron et al., 2023), Llama 3 (8B) (Dubey et al., 2024) and Mistral (7B) (Jiang et al., 2023), which are monolingual and roughly the same size as our models.

3.6 Metrics

To evaluate the minimal pairs, we measure which of the two examples is more likely according to the LM using sentence perplexity. Each example is fed to the model independently during inference, and sentence perplexity is computed in the standard way, as the average log-likelihood of each token, conditioned on the preceding tokens. If the sentence containing the correct synonym has the lowest perplexity, then the item is considered correct. For random tests, where a correct example is compared to ten incorrect examples, the item is considered correct if the correct example has the lowest perplexity among all.

To evaluate gap filling, we verify whether the generated sequence contains a synonym. In the single-prediction scenario, synonyms in the synset are tokenized and compared to the predicted token sequence. If the token sequence of a synonym is produced, the item is correct. With multiple predictions, we consider the item correct if a synonym appears in at least one of the 10 predictions.

4 Results

Figure 1 summarizes our main results. 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). The third graph displays the results of the only test without context, the `relation` test. This test assesses whether the language model correctly identifies both correct and incorrect synonyms as synonyms over 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). See Appendix D for results in table form and Appendix E for details on significance paired two-tailed t-tests.

4.1 Minimal-pairs tests

LM performance in synonymy recognition is affected by the type of test, the nature of the negative element (a single incorrect synonym or multiple random words), and the presence of context.

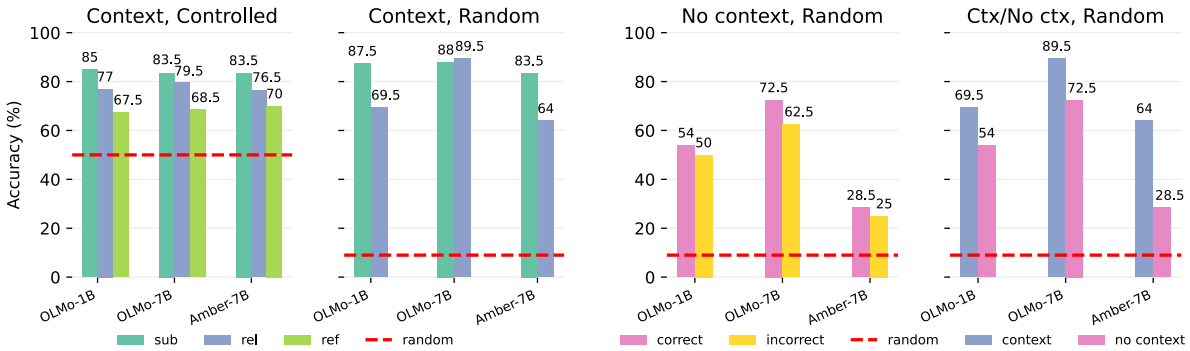

Figure 1: **Left**: Accuracy on `substitution` (sub), `relation` (rel) and `reference` (ref) tasks with context, under two conditions: **(i)** binary controlled with incorrect synonym (*Controlled*) and **(ii)** ten randomly chosen words (*Random*) **Right**: **(iii)** Accuracy on `relation` task without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

Test type The `substitution` test yields the best scores for almost all models under both conditions (first two graphs in Figure 1). This 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, as opposed to the `relation` test, where they are at the end. An exception seems to be OLMo-7B, for which the differences between `substitution` and `relation` are not significant. We hypothesize that, due to its larger training data and size, this model better encapsulates synonymy than its competitors, and leverages 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 models, potentially due to the increased complexity of having the synonym and the context at a distance, having to connect them via the target word and the relation. The `reference` test presents an even higher decrease in accuracy. This is expected as 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 suggests that random words might have different effects depending on the language models, while the controlled item triggers a more uniform result. Garcia (2021) shows that BERT models fail to distinguish synonyms from words with different meanings in the same context, but our study shows that this differentiation does take place for the models we tested. However, our models are not only architecturally different, but also at least ten times larger in terms of parameters.

The `substitution` test yields better scores when the negative elements are random words, than the incorrect synonym, but the differences are not significant. For the `relation` test, results seem to follow an opposite trend: they seem slightly better when the opposing item is the incorrect synonym (except for OLMo-7B). Since the synonym appears at the end of the sentence, its distance from context may allow random words to be more probable.

Context As the `substitution` and `reference` tests become meaningless without context, only the `relation` test allows us to vary this condition. Additionally, the only possible negative candidates are random words, since, in the absence of context, both the correct and incorrect synonyms are valid. The `relation` test without context no longer opposes correct and incorrect synonyms, allowing us to test whether the model correctly identifies them as synonyms when compared to random words (Figure 1, third graph).

Results are surprisingly good, indicating that lexical-semantic information is present in our models and can be elicited with the term *synonym*. Models seem to demonstrate higher accuracy in recognizing the "correct" synonym than the "incorrect" one, with significant differences for OLMo-7B. A probable explanation could be sense frequency: words in SemCor tend to occur with their most frequent sense, making the contextual synonym one of the most frequent ones for that word. How-

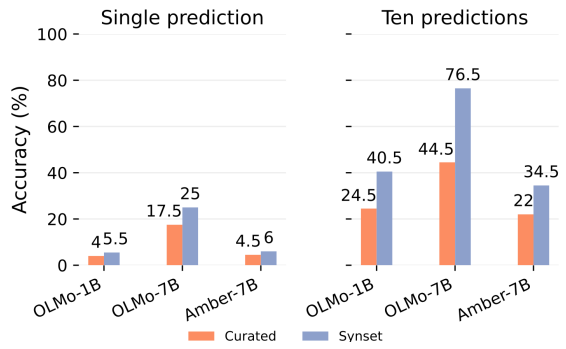

Figure 2: Accuracy on the `relation with context` test in **Left**: single-prediction and **Right**: ten predictions (top-k, with $k = 10$) settings.

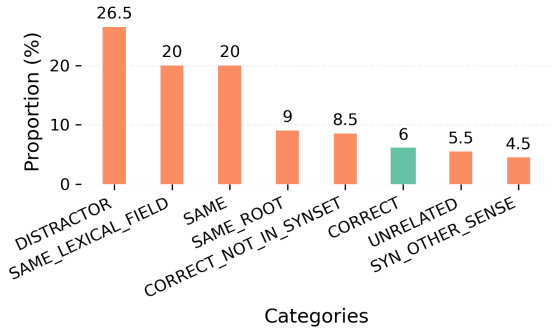

Figure 3: Proportion of examples per category.

ever, this advantage does not translate into better performance in controlled tests with respect to random tests, as `substitution` still seems to work better with random words.

As for the `relation` test, context has a significant positive impact for all models, with the highest increase for Amber, from 28.5% to 64% (fourth graph in Figure 1). This outcome is expected, as LMs 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.

Preferred expression We compared the standard expression *is a synonym of* with two paraphrases: *is the same as* and *means* in the `relation` test, where it is explicit. For all three models, the least preferred expression is *means*, possibly due to its single-word composition and inherent ambiguity. The phrase *is a synonym of* yields the best 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 (results and discussion in Appendix G).

Other models As mentioned in Sec. 3.5, we verify if Amber and OLMo are representative of the broader language model family, by conducting a subset of minimal pairs tests on more well known models like Llama and Mistral. Given that they produce similar results, we conclude that the models are indeed representative, and that their performance accurately reflects that of similar models in this task (see Appendix F for the results).

4.2 Gap-filling tests

Figure 2 shows that the results for the gap-filling tests are generally poor, indicating that the models tend to not generate a synonym when prompted to do so.

Curated vs. synset As expected, having more than one correct prediction leads to an overall improvement, but at varying degrees. In the single prediction scenario, the broader range of correct answers makes little difference for the OLMo-1B and Amber models. However, OLMo-7B, which has a significantly higher score, shows a more notable gain. The effect is much more visible when considering multiple predictions. Both OLMo-1B and Amber show a 15% improvement, and the accuracy of OLMo-7B nearly doubles.

Predictions The right-hand graph in Figure 2 shows that sampling from the top-10 predictions offers a better chance of generating a correct synonym compared to the single greedy prediction. This holds true both for the curated synonym, where models perform 2 to 5 times better, and for the entire synset setting, where the improvement is even greater (up to 8 times for OLMo-1B). This suggests that both the curated synonym and other synonyms from the synset are present near the top of the distribution, even if they are not ranked first.

Error analysis To gain a better understanding of the models' performance on this task, we conducted an error analysis by manually annotating the predictions from one experiment: single prediction, synset, using OLMo-1B (Figure 3). The categories we defined classify the predicted word based on the following characteristics: **correct**, belonging to the target synset, **same root**, i.e. a prefix of the reference, **same** word, identical to the reference, **correct not in synset**, for correct synonyms not

appearing in the target synset, **same lexical field** of the reference (e.g. hyponym, cohyponym, hypernym), **distractor**, that is, another word of the context or its synonym or inflected form, **synonym other sense** synonym of the reference, but not the correct sense, and **unrelated** to the context.

Most common errors, representing roughly half of the cases, are labeled *distractor*, *same*, and *same root*, indicating that the model tends to repeat a word from context when prompted for synonymy.

In approximately 30% of the cases, the model does not repeat the word but instead samples from the lexical field, although incorrectly. The *same lexical field* category represents 1 in 5 cases: the model generates a word that is part of the lexical field of the reference, but misses the lexical relation. As for *synonym other sense* cases, the model gets even closer to the correct answer by predicting a synonym, but not one that fits the context.

In 8.5% of the cases, labeled *correct not in synset*, the predictions might be erroneously classified as incorrect: the prediction is not in the target synset, but it is considered correct by the authors, hinting at a WordNet coverage issue.

4.3 Impact of tokenization

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).

Delta	-2	-1	0	1	2
Accuracy %	78	78	75	79	67
Count	18	150	285	126	21

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.

Accuracy changes slightly when conditioned on the difference in token length. When the delta is 1, -1 or -2, accuracy is higher. While a decrease in accuracy is observed with a delta of 2, this result is based on a small number of items. In both cases, differences are not significant (Appendix E).

4.4 Consistency across 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 (details in Appendix H).

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). 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.

Next, we look at whether the controlled condition can be used to predict the random one, when the test type stays the same. 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 to the random one (33%).

Finally, we investigate whether the results of the `relation` test without context could be predicted from the same test with context. This is indeed the case for 74% of the correct items, but the incorrect items differ between tests. We conclude that while LMs 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

We evaluated pre-trained causal LMs on the linguistic concept of synonymy. We developed a com-

prehensive test suite, validated by human subjects, designed to assess various aspects of synonymy, both in controlled and uncontrolled environments.

We found that the performance of LMs of the order of 10 billion parameters in minimal-pairs tests depends on the test conditions. Overall, the `substitution` test has the highest scores, followed by the `relation` and `reference` tests. Synonymy tests are generally easier when context is provided. In this case, `substitution` tests yield better results with random words, while `relation` tests with the incorrect synonym, indicating that the position of the word might play a significant role. In the absence of context, although performance decreases, models like OLMo-7B still achieve relatively high accuracy. We also observed a bias favoring the correct synonym over the incorrect one, which we hypothesize is due to sense frequency bias in SemCor.

Gap-filling tests show that there is a low chance of generating a synonym when the models are directly prompted to do so. Furthermore, increasing the number of correct elements does not bring a definite improvement, except in the case of OLMo-7B. Additionally, tests with multiple predictions show that while the correct synonym might not be ranked first, correct synonyms are present near the top of the distribution.

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 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.

In the gap-filling tests, we arbitrarily selected a top-k value of 10. A grid-search exploring multiple values could further strengthen the results.

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`)[3]. 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[4]. 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 Sec. 4.

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

D1. The text of instructions given to annotators (mentioned in section 3.4) 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[5] and occasionally ChatGPT[6].

Acknowledgements

We thank Fiona Torbey and Hamish Short for their contributions to the dataset evaluation process and Xavier Alario for insightful comments on earlier versions. This work has been funded by the French Agence Nationale pour la Recherche, through the SELEXINI project (ANR-21-CE23-0033-01).

References

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

Mayee F. Chen, Nicholas Roberts, Kush Bhatia, Jue Wang, Ce Zhang, Frederic Sala, and Christopher Ré. 2023. Skill-it! a data-driven skills framework for understanding and training language models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

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.

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.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. SyntaxGym: An online platform for targeted evaluation of language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, 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.

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

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*.

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.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

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

Dingli Yu, Simran Kaur, Arushi Gupta, Jonah Brown-Cohen, Anirudh Goyal, and Sanjeev Arora. 2024. Skill-mix: a flexible and expandable family of evaluations for AI models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2023*.

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 3.3.

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.

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.

E Significance

To verify if the results on our tests, discussed in Sec. 4.1 and Sec. G 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 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).

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.

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.*

   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.

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

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.

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.

10. *However, OLMo-7B, which has a significantly higher score, shows a more notable gain.*

    When comparing the single-prediction test curated setting and synset setting accuracies for the OLMo-7B model, we obtain a t-value of -4.02 and a p-value inferior to 0.0001. We conclude that the differences in accuracies between the two tests are significant.

For the tokenisation claims (*In both cases, differences are not significant*) we performed an independent t-test, comparing (1) the accuracies when the difference in tokens is -2 with 0 and (2) when the difference is 0 with when it is 2. We obtained a t-value of 0.29 and p-value of 0.61 for the first and a t-value of -0.74 and p-value of 0.76 for the second, permitting us to draw the above conclusion.

F 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 4) 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 5).

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 (4). 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.

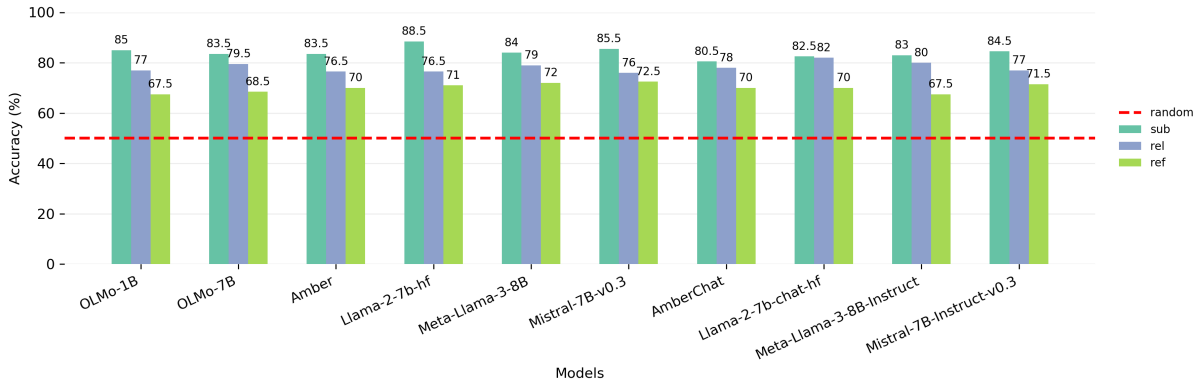

Figure 4: Accuracy on the `substitution`, `relation` and `reference` tasks (with context) for a diverse range of models.

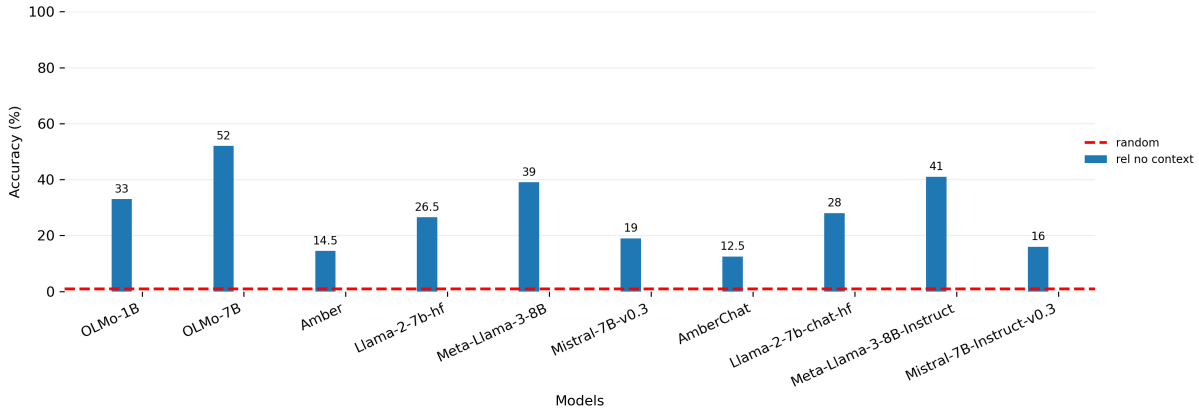

Figure 5: Accuracy on the `relation` task without context for a diverse range of models.

G 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 6), as well as the version without context (the last two graphs in Figure 6). We compared the standard expression "is a synonym of" with two paraphrases: "is the same as" and "means".

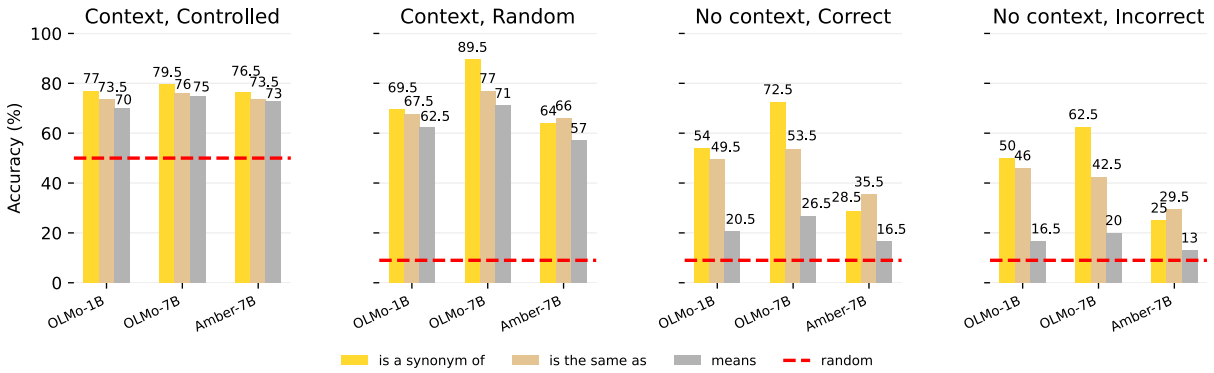

Figure 6: Results on different variants of expressing the synonymy relation: **(i)** Accuracy on the `relation` task with *controlled* context **(ii)** and with uncontrolled, *random* candidates **(iii)** Accuracy on the `relation` task *without context*, for the *correct* synonym against random words and **(iv)** for the *incorrect* synonym against random words.

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", possibly 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.

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.