Exploring Synonymy Representation in Large Language Models: A Comparative Analysis with Referential and Use-Based Lexical Resources

Sara Besharati
Université du Québec à Montréal (UQAM)
Montréal, Québec, Canada
besharati.sara@courrier.uqam.ca

Abstract

This study investigates how synonymy is represented in Large Language Models (LLMs) and examines the extent to which their representations align with human intuitions about context-dependent and use-based synonymy, in comparison to traditional lexical resources like WordNet. Focusing exclusively on English verbs, we utilized sentences from the Concepts in Context (CoInCo) dataset to generate substitute candidates from three sources: LLM-generated synonyms (RoBERTa), gold use-based synonyms (CoInCo), and referential synonyms (WordNet). Substitutions were evaluated by calculating cosine similarity scores within the T5 model's embedding space to assess semantic alignment.

The results show that RoBERTa (mean 0.656) and CoInCo (mean 0.625) exhibited the highest mean similarity scores to the target verbs, indicating that contextual modeling is highly effective in capturing fine-grained, context-dependent aspects of verb meaning. Conversely, WordNet (mean 0.548) provided a more conservative distribution. Analysis of the overlap among the top four candidate synonyms from each source revealed minimal convergence. The agreement between WordNet and RoBERTa was particularly low, with an average overlap of only 4.94%. This lack of convergence demonstrates that the methods generate largely distinct sets and offer complementary rather than redundant lexical knowledge.

Furthermore, while RoBERTa demonstrates enhanced contextual flexibility, its similarity scores show greater variability (standard deviation 0.133) compared to CoInCo and WordNet. This variability signals potential semantic drift, contrasting with WordNet's more stable but less contextually dynamic synonym sets.

Keywords: Verb synonymy, Large Language Models, RoBERTa, T5, WordNet, CoInCo, context-dependent synonymy, use-based synonymy, cosine similarity, lexical resources, word embeddings

1 Introduction

Synonymy refers to a paradigmatic lexical relation between two or more linguistic expressions that share the same or nearly the same meaning in certain contexts (Maienborn et al., 2011).

Among various types of synonymy, one well-known type consists of synonym pairs that share the same referent that refer to the same entity, event, or property in the world. These are what we term **referential synonyms** in this work. Lexical databases like WordNet (Miller, 1995) offer a structured means of identifying such referential synonymy by organizing words into synsets — sets of words that are interchangeable in at least one context because they share a common referent or denotation. Importantly, referential meaning is often the only kind of meaning considered in formal semantics and formal pragmatics. In these traditions, meaning is typically equated with truth-conditional content, where the central concern is whether expressions refer to the same entities or have the same extensions in possible worlds.

Although lexicographers group such words with similar meaning as synonyms, they may not always function interchangeably in any context. For example, the clear synonymy relation between the words "fiddle" and "violin" depends on the formality or informality of the context in which they appear; therefore, the degree of interchangeability between them depends on the context. In a more detailed example, if we search for a synonym for the word "kill", we may find "murder" which is not an absolute synonymy for this word, but it can be a near synonymy which is different in terms of intentionality (Cruse, 1986).

Synonymy judgment is constrained by context (Murphy, 2003). For example, the words "prize" and "award", which are referential synonyms of each other, can be perfect synonyms in a context like (a). They might still be considered

synonyms in a neutral context like (b). However, their meanings might not be similar enough to be considered synonyms in context (c).

(a) Joe won the prize (award) for the best drawing.
(b) what is a synonym for prize? award
(c) The planintiff received a hefty award (≠ prize) in the lawsuit.

The importance of context in synonymy is expressed in several approaches to semantics. Harris's distributional hypothesis (Harris, 1954) posits that words that occur in similar contexts tend to have similar meanings. This principle underlies the foundation of distributional semantics, where word meanings are derived from patterns of co-occurrence in large corpora. Moreover, words possess a flexible and generative nature, highlighting the context-dependent meaning of words (Pustejovsky, 1998).

Context is also essential in how computational models like transformer-based Large Language Models (LLMs) (Vaswani, 2017; Lin et al., 2022) represent natural language expressions. These models use token and positional embeddings, and attention mechanisms to capture relationships between words in a sentence (Vaswani, 2017). Thus, LLMs provide a powerful framework for analyzing synonymy by capturing context-sensitive relationships between words.

In more details, we compare the cosine similarities between the T5 large language model (Raffel et al., 2020)'s representation of a target verb in a sentence with that of the following groups: (1) the recommended substitutions by human annotators for the target words in specific sentences (**gold use-based synonyms**), (2) the suggested synonyms by WordNet for the target words (**referential synonyms**), (3) the synonyms generated by LLMs for the target words (**LLMs synonyms**), and (4) a random group including unrelated substitutions with no lexical connection to the original target words. Given that **referential** synonyms do not take context into account, while **use-based** synonyms and LLM-generated alternatives do, we want to investigate how close referential synonyms are to the others in meaning, and how much their overlaps differ. In this study, we investigate what cosine similarity scores within LLMs' embedding space reveal about the degree of alignment among these different sources of synonymy. To ensure a robust and unbiased evaluation of these different synonym types, we calculate similarity scores within the embedding space of a T5 model, chosen for its strong performance in capturing sentence-level meaning and to avoid the potential bias of using our synonym-generating model (RoBERTa) for evaluation.

2 Literature Review

Advancements in contextualized language models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) have driven significant progress in understanding the syntactic and semantic relationships in language. Probing studies such as (Rogers et al., 2021) have highlighted BERT's capacity to encode rich linguistic knowledge, including entity types, semantic roles, and proto-roles, which supports its application in semantic analysis. These studies establish that BERT's architecture allows it to differentiate word meanings in various contexts effectively. Similarly, (Staliūnaitė and Iacobacci, 2020) showed that RoBERTa can handle lexical relations such as antonymy, highlighting the appropriateness of its contextualized embeddings for semantic analysis. This aligns with (Vulić et al., 2020), who revealed RoBERTa's strength in predicting lexical relations, particularly homonymy and synonymy and (Nair et al., 2020), which highlighted that BERT's alignment with human judgments on word sense relatedness is strong for distinctly separated senses (homonymous).

Lexical substitution (LS) (McCarthy and Navigli, 2007) is the task of replacing a target word in context with a semantically similar alternative while preserving the meaning of the original sentence. Although LS was initially proposed as a method for evaluating word sense disambiguation systems (McCarthy, 2002), subsequent research has focused on identifying the most suitable substitutions to enhance performance in various natural language processing tasks, including paraphrase generation (Fu et al., 2019), text simplification (Štajner et al., 2022), machine translation (Agrawal and Carpuat, 2019), and style transfer (Helbig et al., 2020), among others. In this research, we build on the idea of LS to compare and analyze the similarity levels of words selected with contextual information during substitution with those chosen using thesaurus-based resources.

In another related study, a new benchmark called SWORDS (Lee et al., 2021) was proposed, asking Candidates to first generate substitution using a context-free resource such as a thesaurus, and

then evaluated by humans for their contextual fit. This method is grounded in the intuition that it is cognitively easier for humans to evaluate a given list than to retrieve suitable words from scratch like CoInCo which is limited to human recall. Based on their results, Compared to CoInCo, SWORDS provides 4.1 times more substitutes per target word while maintaining contextual appropriateness by obtaining 1.5 times more appropriate than those found in previous datasets for the same number of candidates. Moreover, the limitations of applying BERT naively to lexical substitution is addressed in (Zhou et al., 2019) by explaining that masking the target word and sampling predictions often yields semantically inappropriate substitutes. Therfore, they proporsed a novel embedding dropout technique which helps BERT to generate substitute candidates that are sensitive to context but not overly biased toward the original word, and tried to increase the diversity and relevance of proposed substitutes. Moreover, to ensure substitute quality, the proposed approach incorporates a validation step that measures the semantic consistency between the original and substituted sentences using BERT's contextualized sentence representations.

As mentioned, one of the essential criteria of lexical substitution is preserving the meaning of the original sentence. While using pre-trained language models (PLMs) ensures contextual relevance, they might produce words that are semantically distant from the original target. In another recent work (Vladika et al., 2025), by concatenating the masked sentence with the original sentence, the authors introduced CONCAT, an augmented lexical substitution approach designed to enhance the contextual information available to the model during masked token prediction, thus improving the generation of contextually appropriate and grammatically coherent substitutes.

Unlike most lexical substitution (LS) tasks, our objective is not to identify the best single substitute for a target word. Instead, we aim to examine how different substitution strategies influence the contextual representation of the target word. By using various lexical resources to generate candidate substitutes—such as human-annotated datasets, thesaurus-based lists, or embeddings-derived candidates, we investigate how the choice of substitution method affects the semantic similarity between the original and modified sentences. This allows us to evaluate not only the appropriateness of substitutes, but also the sensitivity of contextual word meaning to different substitution sources.

In totall, the goal of this work is to examine to what extent LLM-generated synonyms align with both referential and use-based substitutes in context. Specifically, this study evaluates the degree of similarity between LLM-generated synonyms and those provided by CoInCo, a resource reflecting use-based synonymy derived from human judgments, and WordNet, which encodes referential synonymy through structured lexical relations. By focusing on verb synonymy and utilizing cosine similarity to compare embeddings, this work adds to the field by directly assessing whether LLMs encode synonymy in ways that align more closely with context-dependent human judgments or traditional lexical resources. In addition, looking at the overlap between different synonym sets not only reveals how much lexical information is shared across them, but also identifies the unique contributions of each source, offering insights into whether LLMs capture synonymy in ways that converge with context-dependent human judgments or remain closer to static, dictionary-based relations. Taken together, these complementary analyses provide a multidimensional evaluation framework that deepens our understanding of LLMs' representational capabilities.

3 Method

As the LLMs from the BERT family have shown their effectiveness in natural language understanding (Zhang et al., 2020; Staliūnaitė and Iacobacci, 2020), we selected the RoBERTa [1] model (Liu et al., 2019) for generating the LLM-based synonyms—Compared to other LLMs in the BERT family, RoBERTa showed the ability to learn more about lexical features (Staliūnaitė and Iacobacci, 2020).

Additionally, we used the Text-to-Text Transfer Transformer (T5) model (Raffel et al., 2020) to represent both the original sentences and their paraphrases in order to avoid potential bias that could arise from using RoBERTa for both synonym generation and embedding-based evaluation. In other words, using the same model for generating substitutes and for measuring their semantic similarity could risk inflating alignment scores due to model-specific representational artifacts. Therefore, by employing T5 we aimed to reduce this

[1] https://huggingface.co/docs/transformers/en/model_doc/roberta

confound and ensure that the evaluation reflects model-independent semantic similarity.

Moreover, as the effect of context on lexical meaning was a key factor in our investigation, we selected the CoInCo (Concepts in Context) corpus (Kremer et al., 2014) for our experiments, which provides access to 35,000 target words and at least six synonyms for each word recommended by human annotators based on context.

3.1 Data Processing

This study focused exclusively on English verbs, as verbs play a central role in encoding actions and relational meanings in a sentence. We used 4,081 sentences from CoInCo with a target verb and its synonyms, but we only retained sentences where the target verbs had a synonym available in WordNet. Furthermore, we imposed a vocabulary constraint by requiring that the target verbs and all its synonym candidates existed within the T5 model's vocabulary, ensuring accurate word representations in the model's embedding space. After applying these filters, 4,039 target sentences remained, forming the final dataset.

The alternative sentences for each target verb were created by replacing the target verb in the sentence with synonyms from four different sources, resulting in four groups. The first group utilized CoInCo's gold use-based synonyms substitutions, which reflected use-based (contextual) synonymy. For each target word, CoInCo provides several synonyms and a label representing the frequency of each synonym. We selected the substitute with the highest frequency based on the label "freq" in CoInCo. The second group relied on synonyms extracted from WordNet, we used NLTK[2] library in Python to collect all the WordNet synonyms. To determine the intended sense of each verb in context, we employed the classic Lesk algorithm (Lesk, 1986) for word sense disambiguation (WSD). The Lesk algorithm is a knowledge-based WSD method that selects the most appropriate WordNet synset for a word by measuring overlap between the context of the target word and the dictionary definitions (glosses) of its possible senses. Specifically, the algorithm compares the words surrounding the target word in the sentence with the definitions and example usages of its candidate synsets in WordNet, selecting the sense with the greatest lexical overlap.

We used the implementation available in the NLTK library, which applies the simplified Lesk variant introduced by (Banerjee and Pedersen, 2002), considering context windows and WordNet relations such as glosses and examples.
The third group consisted of synonyms generated by RoBERTa, which provided a contextual representation of the target word. To extract the synonyms suggested by RoBERTa, we used a masked language modeling approach. Specifically, for each target word in a sentence, we replaced it with a mask token (<mask>), allowing the RoBERTa model provided by the Transformers [3] library in Python to predict the most likely word to fill in the blank. Among the model's predictions, we selected the most probable word that was different from the original target word, assuming it to be the best synonym candidate in the given context.

Lastly, a random group was developed, which included unrelated substitutions with no lexical connection to the original target words. In more detail, first, we used NLTK to extract all English verbs in the *omw-1.4* (Open Multilingual WordNet) corpus. Then, for each target verb, we selected a verb in the corpus that was neither the target verb nor among its WordNet synonyms. An example of a target sentence and its substitutions in each group is shown in Table 1, where the verb "end" is replaced by different alternatives.

Target	A mission to **end** a war
CoInCo	A mission to **stop** a war
RoBERTa	A mission to **stop** a war
WordNet	A mission to **finish** a war
Random	A mission to **take_a_dare** a war

Table 1: Comparison of Different Groups

3.2 Model

The main analysis in this study was based on the evaluation of the cosine similarity between the T5 model's representations of the target words and their substitutions from four sources: CoInCo (gold use-based synonyms), WordNet (referential synonyms), RoBERTa (LLM-generated synonyms), and Random (unrelated substitutions).

We used the T5 model as a contextual encoder to obtain sentence-level representations. T5 treats every NLP task as a text-to-text problem, making

[2]https://www.nltk.org/

[3]https://huggingface.co/docs/transformers/en/index

it highly adaptable for contextual semantic tasks. In our setup, we encode sentences containing a target verb using a pre-trained T5 model (e.g., T5-base), generating contextual embeddings that reflect the influence of surrounding lexical material on the target word. These representations serve as a semantic basis for comparing substituted and original sentences, allowing us to assess how well different models preserve or shift meaning when proposing substitutes —Our choice of the T5 model for evaluation was a deliberate methodological decision. First, to ensure an unbiased assessment, we used a different model for evaluation than for synonym generation (RoBERTa). This separation prevents the risk of inflated similarity scores that can arise from model-specific artifacts. Second, T5's encoder-decoder architecture is distinct from masked language models like RoBERTa and is particularly adept at capturing the broad contextual dependencies and subtle semantic shifts essential for our analysis, as it excels at tasks involving sentence-level meaning.

Tokenization: To represent the sentences in each group and extract the embedding of the target words, we first tokenized the input sentence using T5's tokenizer. Next, we lemmatized the target word and the T5's tokens using spaCy library in Python (Honnibal and Montani, 2017) to obtain their base forms for accurate comparison. Then we compared the target words with the T5 tokens (base forms) to identify the index of the token that matched the target word.

Representation: We passed the tokenized sentences through the T5 model to generate contextual embeddings for all tokens. Next, using the identified index, we extracted the high-dimensional embedding corresponding to the target word from the model's last hidden state.

3.3 Evaluation

Similarity Measurement: After collecting the embeddings of the four substitutes for the target words within the sentence contexts, for each sentence, we calculated the cosine similarity between each substitute and the target word, using the SciPy library in Python[4], to analyze the semantic alignment between the groups—Semantic alignment refers to the cosine similarity score between the contextual embedding of a substitute word and the target word within the T5 model's embedding space. A higher score indicates that the substitute is semantically closer to the original word in that specific context. For example, in *A mission to end a war*, the substitute "stop" would demonstrate a high semantic alignment, while the random substitute "tak_a_dare" would show a very low alignment.

Intersection Statistics To assess the degree of lexical similarity and agreement between the substitute word sets generated by CoInCo, RoBERTa, and WordNet, we analyzed the overlap of multiple top candidate synonyms rather than focusing solely on the best single substitute from each method. For each target word in a sentence, we extracted the top 4 substitute candidates from each source —CoInCo, RoBERTa, and WordNet — forming three sets of potential synonyms. We then computed pairwise intersections between these sets to determine the proportion of shared substitute words between each pair of methods. Specifically, for every sentence, we calculated the pairwise intersection between CoInCo and RoBERTa, CoInCo and WordNet, and RoBERTa and WordNet among their respective top candidates, expressed as a percentage. These overlap percentages were computed individually for each sentence to preserve context-specific lexical variation, and then averaged across the entire dataset to obtain an overall measure of alignment and divergence between methods. This approach captures not only agreement on the single most likely synonym, but also the broader semantic neighborhood suggested by each method, providing richer insight into how these lexical resources and models complement or diverge from each other in proposing substitute words. A high average overlap indicates that different methods tend to propose similar sets of plausible synonyms in context, whereas a low overlap reflects distinctive lexical suggestions unique to each method.

Statistical Significance The variables used in the statistical analysis were the cosine similarity values, calculated using T5 embeddings to represent both the original sentence and its paraphrases. Furthermore, the statistical methods applied to the data included Welch's ANOVA and the Games-Howell post-hoc test.

To assess the assumptions for parametric testing, the Levene test and Bartlett test were first performed using the SciPy library in Python. Both tests confirmed significant differences in group variances (Levene's test: Statistic = 277.167, p =

[4]https://scipy.org/

0.000; Bartlett test: Statistic = 894.552, p = 0.000), indicating a violation of the homogeneity of variances assumption. Additionally, normality checks showed that the overall distribution (pooled from all groups) was approximately normal, with skewness = –0.084 and excess kurtosis = 0.556, both of which fall within commonly accepted thresholds for normality.

Given these results, Welch's ANOVA was selected as it is robust to unequal variances and performs reliably even under moderate deviations from normality, particularly when the sample size is sufficiently large. This test revealed statistically significant differences across groups.

To determine which group means differed significantly, the Games-Howell test was used for pairwise comparisons. This post-hoc test is appropriate when variances are unequal and does not require equal group sizes, making it suitable for our analysis.

4 Result

The average overlap in the top four substitutes per target verb varied considerably across the three resources. The comparison between CoInCo and RoBERTa yielded the lowest agreement, with an average overlap of only **5.03%**, indicating that these two methods rarely propose the same candidate substitutions in context. A higher, yet still limited, degree of overlap was observed between CoInCo and WordNet at **12.60%**, suggesting slightly greater alignment between gold use-based synonyms substitutes and referential synonyms. The overlap between WordNet and RoBERTa was also low, at **4.94%**, demonstrating that the distributional predictions of the language model differ substantially from the lexical entries in WordNet. Overall, these results indicate that the three methods generate largely distinct sets of candidate substitutes, with minimal convergence beyond chance levels.

While cosine similarity metrics indicated some semantic alignment between methods, the overlap analysis reveals minimal agreement across the broader candidate synonym sets. The particularly low overlap values suggest that even when methods identify similar top-ranked substitutes, they tend to propose largely distinct alternatives overall. This pattern reflects complementary rather than redundant lexical knowledge among CoInCo, WordNet, and RoBERTa's substitute proposals.

In addition, statistical analyses demonstrated significant differences in the degree to which the four synonym groups aligned semantically with their target words. Welch's ANOVA revealed a highly significant effect of synonym source on mean cosine similarity scores ($F(3, \cdot) = 7510.70$, $p < .001$), with a large effect size ($\eta_p^2 = 0.489$), indicating that nearly half of the variance in similarity can be attributed to differences between groups.

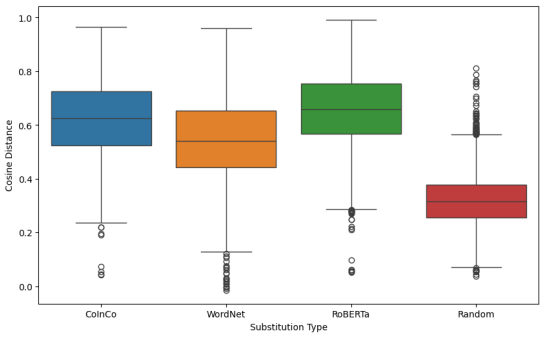

Figure 1: Cosine Similarities by Synonym Types

Post-hoc pairwise comparisons using the Games-Howell test (Table 3, Appendix A.2) confirmed that all group differences were statistically significant ($p < .001$). The largest effects were observed between CoInCo vs. Random and RoBERTa vs. Random, while moderate to strong effects were found in comparisons involving WordNet (e.g., RoBERTa vs. WordNet, CoInCo vs. WordNet). These results suggest that synonym sets derived from contextualized (LLM-based) and use-based resources exhibit stronger semantic alignment with target words than those generated from random selection or static lexical databases.

Descriptive statistics further elucidate these differences (Table 4, Figure 1). RoBERTa and CoInCo exhibit the highest mean similarity scores (0.656 and 0.625, respectively), indicating that their substitutes generally preserve semantic similarity with the target word better than those from WordNet (mean = 0.548) or Random (mean = 0.321). Median values follow the same pattern, with RoBERTa's median (0.659) notably higher than the other methods, reflecting a consistent tendency to produce contextually relevant substitutes.

However, RoBERTa's similarity scores show greater variability (standard deviation = 0.133) compared to CoInCo and WordNet, with a wider range spanning from a minimum of 0.051 to a maximum near 0.99. This suggests that although RoBERTa frequently generates semantically close synonyms, it can also produce less similar or con-

textually inappropriate substitutes, highlighting the challenges LLMs face in fully capturing nuanced or rare word senses.

In contrast, WordNet displays a more conservative similarity distribution, with lower mean and median scores and narrower variability, indicative of a lexically constrained but stable synonym set that is less sensitive to context. The broader distributions for CoInCo and RoBERTa likely reflect their enhanced contextual adaptability, with RoBERTa especially able to generate a wider semantic spectrum of substitutes.

Table 2 presents a detailed comparison of the top substitute candidates generated by CoInCo, RoBERTa, and WordNet for the target verb "begin" in a specific context. This example illustrates the key findings regarding overlap and semantic strategy divergence.

Target Sentence	The company will begin mailing materials to shareholders.
CoInCo	start, commence, initiate
RoBERTa	start, initiate, shipping, sending, posting
WordNet	start, commence, set_out, start_out, get_down

Table 2: Examples of Different Group Substitutes

Demonstrating Complementary Lexical Knowledge (Low Overlap) This example illustrates the finding that the candidate sets are largely distinct and complementary, rather than redundant, as reflected by the overall low average overlap (e.g., WordNet vs. RoBERTa at 4.94%).

- CoInCo focuses on traditional initiation verbs: start, commence, initiate.
- WordNet includes similar terms (start, commence), but also phrasal verbs that may be less contextually appropriate, such as set_out or get_down.
- RoBERTa provides a mix of precise synonyms (initiate) and contextually derived words that describe the subsequent action (e.g., shipping, sending, posting).

The inclusion of verbs related to the direct object (e.g., shipping materials) by RoBERTa, alongside abstract synonyms (initiate), demonstrates a distinct, context-sensitive approach that differs substantially from the sets generated by CoInCo and WordNet, supporting the conclusion of minimal convergence across the broader candidate sets.

Demonstrating RoBERTa's Contextual Success and Flexibility This example highlights RoBERTa's ability to capture the fine-grained, context-dependent aspects of verb meaning and its greater contextual adaptability.

For the target verb "begin" in the context of mailing materials:

- RoBERTa proposes several substitutes that are highly specific to the context of distributing physical items: shipping, sending, and posting. These terms are not general synonyms for "begin" but are contextually functional substitutes that preserve the intended meaning of the action being initiated. This use of highly specific verbs validates the finding that RoBERTa excels in generating contextually relevant substitutes, contributing to its highest overall mean similarity score (0.656).

Demonstrating RoBERTa's Variability vs. Referential Stability This comparison addresses the finding of a trade-off between semantic precision and contextual flexibility.

- RoBERTa's Variability: RoBERTa shows high contextual flexibility by including shipping and posting. While contextually useful, these are semantically distant from the target verb's core "initiation" meaning, contributing to the model's greater variability (Std Dev = 0.133). The model risks semantic drift by moving too far into the related action space.
- WordNet's Stability: WordNet exclusively provides terms focused on the referential meaning of initiation (start, commence, set_out). This results in a conservative similarity distribution, demonstrating a stable but less contextually dynamic set.

This divergence illustrates that LLMs capture a broader semantic spectrum, sometimes sacrificing general semantic proximity for high contextual relevance (e.g., shipping vs. start), whereas static resources maintain referential consistency regardless of the specific surrounding context (the mailing materials clause). Overall, these findings highlight a trade-off between semantic precision and contextual flexibility: CoInCo and RoBERTa

generate more semantically aligned and context-sensitive synonyms on average, while RoBERTa's broader distribution signals occasional semantic drift. WordNet provides more stable but less contextually dynamic synonym sets, and Random substitutions unsurprisingly yield the lowest semantic alignment.

5 Conclusion

The closer semantic proximity between LLM-generated substitutes and use-based ones showed that LLMs produce word substitutes that are closer to use-based approaches like CoInCo. Furthermore, the closer proximity of LLM-generated and use-based substitutes to target words underscores the effectiveness of contextual modeling in capturing fine-grained, context-dependent aspects of verb meaning.

Nevertheless, the limited overlap across the full sets of candidate synonyms produced by different methods reveals that each approach explores distinct lexical neighborhoods, offering complementary perspectives on lexical semantics rather than converging on identical substitutions. This divergence highlights the complex and gradient nature of synonymy, where semantic equivalence varies depending on pragmatic and usage contexts.

Moreover, while LLMs exhibit greater contextual flexibility and adaptability, this strength is accompanied by increased variability in synonym quality, occasionally leading to semantic drift. In contrast, traditional lexical resources provide more stable but less context-sensitive synonym sets. These contrasting properties suggest that no single resource suffices to comprehensively model the full spectrum of verb synonymy.

Together, these findings reinforce the effectiveness of context-aware approaches, especially those powered by LLMs in capturing subtle nuances of verb similarity, while also underscoring that methods may agree semantically without converging on the same lexical choices. This divergence highlights the potential benefits of combining multiple approaches to broaden coverage and enhance robustness in synonym substitution tasks.

6 Limitations

While this study provides valuable insights into the representation of verb synonymy in LLMs and their comparison with lexical resources, several limitations must be acknowledged. First, the analysis was limited to verbs extracted from the CoInCo dataset, which may constrain the generalizability of findings to other parts of speech or less commonly used verbs. Second, this study relied on cosine similarity as the sole metric of alignment. This metric may not capture all semantic subtleties, particularly in cases involving polysemy or context ambiguity.

Additionally, while WordNet and CoInCo served as strong lexical benchmarks, incorporating other lexical databases or expert-annotated datasets could further strengthen the external validity of the findings. Future work could also include (i) asking human annotators to rate the synonyms generated by WordNet and RoBERTa as an additional metric for comparison, and (ii) extending the analysis beyond English to examine cross-linguistic patterns. Moreover, although this study mitigated potential bias by using a different model for evaluation, it did not systematically explore how variations in LLM architectures (e.g., GPT, XLNet, BERT variants) might influence synonym generation and semantic alignment. Addressing these gaps by covering a broader range of linguistic categories, employing diverse evaluation methods, and performing cross-architecture comparisons would provide deeper insights into model-specific behaviors in capturing context-aware synonymy.

Statement: *During the preparation of this work the authors used ChatGPT-4o to help with text editing, after which, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.*

References

Sweta Agrawal and Marine Carpuat. 2019. Controlling text complexity in neural machine translation. *arXiv preprint arXiv:1911.00835*.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *International conference on intelligent text processing and computational linguistics*, pages 136–145. Springer.

D.A Cruse. 1986. *Lexical Semantics*. Cambridge University Press.

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.

Yao Fu, Yansong Feng, and John P Cunningham. 2019. Paraphrase generation with latent bag of words. *Advances in Neural Information Processing Systems*, 32.

Zellig S Harris. 1954. Distributional structure. *Word*, 10(2-3):146–162.

David Helbig, Enrica Troiano, and Roman Klinger. 2020. Challenges in emotion style transfer: An exploration with a lexical substitution pipeline. *arXiv preprint arXiv:2005.07617*.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing.

Gerhard Kremer, Katrin Erk, Sebastian Padó, and Stefan Thater. 2014. What substitutes tell us-analysis of an "all-words" lexical substitution corpus. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 540–549.

Mina Lee, Chris Donahue, Robin Jia, Alexander Iyabor, and Percy Liang. 2021. Swords: A benchmark for lexical substitution with improved data coverage and quality. *arXiv preprint arXiv:2106.04102*.

Michael Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26.

Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. 2022. A survey of transformers. *AI open*, 3:111–132.

Y Liu, M Ott, N Goyal, J Du, M Joshi, D Chen, O Levy, M Lewis, L Zettlemoyer, and V Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 364.

Claudia Maienborn, Klaus von Heusinger, and Paul Portner. 2011. *Semantics: An international handbook of natural language meaning*, volume 1. Walter de Gruyter.

Diana McCarthy. 2002. Lexical substitution as a task for wsd evaluation. In *Proceedings of the ACL-02 workshop on Word sense disambiguation: recent successes and future directions*, pages 089–115.

Diana McCarthy and Roberto Navigli. 2007. Semeval-2007 task 10: English lexical substitution task. In *Proceedings of the fourth international workshop on semantic evaluations (SemEval-2007)*, pages 48–53.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

MLynne Murphy. 2003. *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. *arXiv preprint arXiv:2010.13057*.

James Pustejovsky. 1998. *The generative lexicon*. MIT press.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Sanja Štajner, Daniel Ferrés, Matthew Shardlow, Kai North, Marcos Zampieri, and Horacio Saggion. 2022. Lexical simplification benchmarks for english, portuguese, and spanish. *Frontiers in artificial intelligence*, 5:991242.

Ieva Staliūnaitė and Ignacio Iacobacci. 2020. Compositional and lexical semantics in roberta, bert and distilbert: A case study on coqa. *arXiv preprint arXiv:2009.08257*.

A Vaswani. 2017. Attention is all you need. *Advances in Neural Information Processing Systems*.

Juraj Vladika, Stephen Meisenbacher, and Florian Matthes. 2025. Lexical substitution is not synonym substitution: On the importance of producing contextually relevant word substitutes. *arXiv preprint arXiv:2502.04173*.

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.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 9628–9635.

Wangchunshu Zhou, Tao Ge, Ke Xu, Furu Wei, and Ming Zhou. 2019. Bert-based lexical substitution. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3368–3373.

7 Appendix

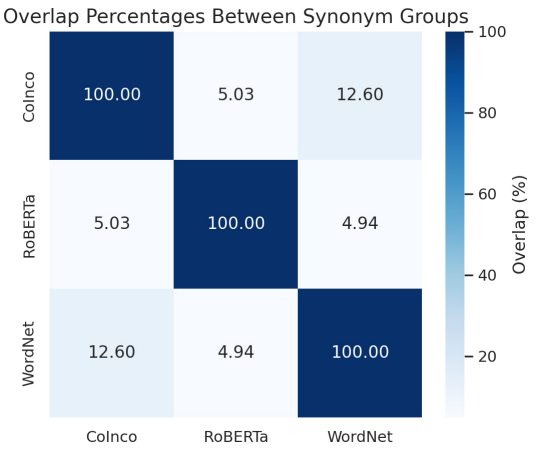

Figure 2: A.1: Overlap Percentages Between Synonym Groups

Group Comparison	**Mean Difference**	**p-value**	**Hedge's g**
CoInCo vs Random	0.3044	< 0.001	2.5167
CoInCo vs RoBERTa	-0.0307	< 0.001	-0.2245
CoInCo vs WordNet	0.0769	< 0.001	0.5155
Random vs RoBERTa	-0.3351	< 0.001	-2.8629
Random vs WordNet	-0.2275	< 0.001	-1.7304
RoBERTa vs WordNet	0.1076	< 0.001	0.7367

Table 3: Appendix A.2: Games-Howell Post-Hoc Test Results

Statistic	**CoInCo**	**RoBERTa**	**WordNet**	**Random**
Count	4039	4039	4039	4039
Mean	0.6251	0.6557	0.5481	0.3206
Std	0.1400	0.1332	0.1578	0.0983
Min	0.0424	0.0512	-0.0141	0.0384
25%	0.5248	0.5664	0.4418	0.2552
Median	0.6255	0.6589	0.5418	0.3160
75%	0.7263	0.7535	0.6535	0.3792
Max	0.9648	0.9922	0.9601	0.8109

Table 4: Appendix A.3: Descriptive Statistics of the Cosine Similarities