Lexical Sensitivity in LLMs: A Benchmark for Simplification, Substitution and Domain-Specific Vocabulary

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are now starting to be integrated into digital communication, from everyday writing assistants to domain-specific support systems. Yet, their tendency to prioritize high-frequency lexical items often comes at the expense of nuance, leading to a form of semantic flattening. For tasks where lexical sensitivity and diversity matter—such as simplification, this bias risks narrowing expression meaning rather than making text more accessible. To address this, we introduce a benchmark for evaluating LLMs' lexical sensitivity focusing on synonym generation across both general and domain-specific contexts. We compare GPT-3.5 and LLaMA-3.1 against structured lexical resources (Merriam-Webster and Datamuse), applying ten metrics that capture lexical overlap, semantic similarity, and ranking consistency on two corpora: complex modern vocabulary and diachronic French gardening texts. Our results show that while LLMs generate semantically coherent synonyms, they consistently underperform on domain-rich and high-complexity terms: GPT aligns more closely with curated thesauri, whereas LLaMA produces broader but less structured alternatives. We release our open-access framework [1] as both an evaluation resource for lexical diversity and a diagnostic tool simplification.

1 Introduction

LLMs are increasingly embedded in everyday communication, shaping language across media, literature, and speech (Yakura et al., 2024), while also holding promise for language learning, writing support, and accessibility (Ye et al., 2025; Gubelmann et al., 2024). As their outputs proliferate, concerns arise about how model preferences shape vocabulary. When LLMs systematically favour frequent terms, less common or domain-specific vocabulary may fall out of use (Juzek and Ward, 2025; Guo et al., 2024), reducing linguistic richness and expressiveness. This risk is acute in text simplification, where lexical homogenization can lead to oversimplification and the loss of nuance, ultimately under-representing less frequent but pedagogically valuable terms (Li et al., 2023; Lee et al., 2024; Tu, 2024; Kirk et al., 2024). In education, this means that the very learners who most need simplification may be disadvantaged when nuance and variety are erased. Striking a balance between accessibility and lexical diversity remains a central challenge (Shardlow et al., 2024).

At the heart of this challenge lies synonym generation, a key mechanism for vocabulary development (Rothschild, 2025). In speech and language therapy, as well as in education, practising synonym production has long been used to foster lexical diversity and strengthen expressive skills (Al-Darayseh, 2014). Traditionally, thesauri have provided structured alternatives, but building them is resource-intensive. LLMs, by contrast, can generate dynamic, context-sensitive synonyms from broad training data (Gabín and Parapar, 2025; Xu et al., 2025; Ashizawa et al., 2024). Yet existing evaluation resources remain limited: they are context-dependent and rarely address domain-specific or diachronic challenges. This is particularly problematic for simplification pipelines, where models must replace complex words without flattening meaning (Saggion, 2017; North et al., 2025).

We address this gap with a benchmark that directly tests whether LLMs can provide nuanced synonym generation of the kind valued in therapy and education. Our framework isolates lexical sensitivity, extending evaluation beyond context-driven substitution to domain-specific and diachronic settings. It serves as both a diagnostic tool for simplification systems and a resource for assess-

[1] https://github.com/vs1rr/domain_specificity_benchmark

ing whether LLMs preserve lexical richness when substituting complex or domain-specific terms. Our contributions are fourfold: (i) we introduce a benchmark for evaluating lexical sensitivity in LLMs, spanning both complex and domain-specific vocabulary; (ii) we design a comprehensive evaluation protocol with ten metrics that capture lexical overlap, semantic similarity, and ranking consistency; (iii) we apply the benchmark to two contrasting corpora: (a) complex words from a modern readability dataset and (b) a diachronic corpus of historical French gardening texts; and (iv) we release our open-source framework and code as a modular tool for diagnosing lexical adaptability in LLMs.[2].

Our findings reveal that while LLMs excel at producing semantically coherent synonyms, they struggle with specialized vocabulary. GPT demonstrates higher precision and closer alignment with curated thesauri, whereas LLaMA generates a broader but less structured range of alternatives. Taken together, these results highlight the potential of hybrid approaches that combine the reliability of traditional lexical resources with the flexibility of dynamic modeling. Without such deliberate design, LLMs risk over-relying on frequent, standardized vocabulary at the expense of nuance and diversity—an issue particularly critical in simplification pipelines, where synonym variety directly determines text quality and accessibility.

2 Background

Relation to Lexical Substitution and Simplification Our benchmark builds on two established NLP tasks: lexical substitution and lexical simplification. Lexical substitution (Kremer et al., 2014) replaces a target word with a meaning-preserving alternative in context, while lexical simplification (Shardlow et al., 2024) focuses on generating replacements that improve accessibility. Recent work increasingly leverages LLMs for substitution or simplification through fine-tuning or prompting (Kew et al., 2023; Cripwell et al., 2023; Farajidizaji et al., 2024). Our work extends this line in three ways: (i) isolating lexical sensitivity outside sentential context; (ii) benchmarking against structured thesauri rather than only human annotations; and (iii) introducing domain-specific and diachronic evaluation. Together, these innovations provide diagnostic tools that highlight whether models can preserve lexical diversity and domain richness, capabilities especially relevant for simplification pipelines.

Existing Benchmarks for Lexical Diversity Table 1 summarizes leading synonym-related and lexical benchmarks, detailing their task focus, evaluation metrics, and distinguishing features. Existing benchmarks for lexical diversity span from frameworks for lexical simplification, i.e. BenchLS (Paetzold and Specia, 2016) and entity-level synonym set generation, i.e. EnSynFields (Huang et al., 2025), to paraphrase disambiguation, i.e. PAWS (Zhang et al., 2019), word sense disambiguation, i.e. WiC (Pilehvar and Camacho-Collados, 2019) and general natural language generation (NLG) with broad benchmarks such as GLGE (Liu et al., 2021) and GEM (Gehrmann et al., 2021). Evaluation metrics vary accordingly, including traditional measures like precision, recall, and F1-score, as well as BLEU, ROUGE, METEOR (Liu et al., 2021), and Spearman's (Jurgens et al., 2012). However, few benchmarks address the semantic breadth, ranking consistency, or synonym diversity central to our framework.

LLMs' Role in Corpus Linguistics Beyond lexical diversity, researchers have examined LLMs' roles in corpus linguistics, debating whether these models complement or challenge traditional analysis. Studies investigating how LLMs interact with lexicographic resources (Lew, 2024, 2023) suggest that AI tools can enhance dictionary compilation rather than replace them. Unlike static resources, LLMs generate dynamic outputs. However, while evaluations exist for dictionary integration (Lew, 2024), comparable evaluations for thesauri remain underdeveloped. To our knowledge, no existing benchmark systematically targets synonym diversity in domain-specific and diachronic contexts, nor directly connects to simplification tasks.

LLMs vs. Human Lexical Diversity Recent studies have compared the lexical diversity and richness of LLMs' outputs to human-generated texts (Martínez et al., 2024; Kendro et al., 2024; Herbold et al., 2023). Findings indicate that while LLMs often exhibit lower diversity and use fewer unique words than humans, newer models (e.g., *ChatGPT-4*) can match or even exceed human performance in certain contexts. For example, Martínez et al. (2024) highlight that although *ChatGPT-3.5* uses fewer distinct words,

[2] https://github.com/vs1rr/domain_specificity_benchmark

ChatGPT-4 demonstrates higher lexical diversity, sometimes surpassing human benchmarks. Similarly, Kurt Pehlivanoğlu et al. (2024); Sahib et al. (2023) explore paraphrasing capabilities, showing that *ChatGPT-4* maintains comparable complexity to human writing, unlike its predecessor, which tends to simplify language. Nonetheless, the need for a benchmark persists, especially in synonym generation tasks where human-annotated gold standards are lacking.

Benchmark	**Task**	**Evaluation Metrics**
BenchLS	LS	PR, RE
EnSynFields	ESD	PR, RE, F1
GLGE	NLG	BLEU, ROUGE, METEOR
GEM	NLG	BLEU, ROUGE, Humans
SemEval-2012 T2	LSE	Spearman's ρ
WiC	WSD	AC
PAWS	PD	AC, F1

Table 1: Overview of synonym-related and lexical semantic benchmarks. Tasks include simplification, paraphrasing, entailment, and NLG. Abbreviations: LS = Lexical Simplification, ESD = Entity Synonym Detection, LSE = Lexical Semantic Entailment, NLG = Natural Language Generation, WSD = Word Sense Disambiguation, and PD = Paraphrase Disambiguation. Evaluation metrics include standard classification scores (PR = Precision, Re = Recall, AC = Accuracy, $F1$), generation metrics (BLEU, ROUGE, METEOR), and semantic similarity (ρ, Spearman's).

Defining Lexical Diversity Lexical diversity reflects the variability and richness of word usage and is often linked to linguistic sophistication. However, as Jarvis (2013) note, there is no consensus on how to define or measure it. Common metrics like *Type-Token Ratio* (Hess et al., 1984; Richards, 1987) and *Moving-Average Type-Token Ratio* (Bestgen, 2025) are simple but sensitive to text length (Martínez et al., 2024; Bestgen, 2024). Length-invariant alternatives such as *Measure of Textual Lexical Diversity (MTLD)* (McCarthy and Jarvis, 2010), Weighted *MTLD* (Vidal and Jarvis, 2020; Kyle et al., 2021), *D Measure from Voc-D Model*, and *Hypergeometric Distribution Diversity* (McCarthy and Jarvis, 2007) offer improvements but introduce trade-offs in interpretability and efficiency. In this study, we adopt a task-specific approach to evaluating lexical diversity and synonym quality, grounded in metrics tailored to LLM-generated synonym sets. Unlike traditional corpus-level measures that assess global vocabulary richness, our framework use a comprehensive set of metrics to evaluate Lexical Overlap, Semantic Similarity and Ranking Consistency especially in setting where ground truth is not present.This aligns with recent computational studies that examine synonym variety and contextual appropriateness in LLM outputs (Kendro et al., 2024).

3 Benchmark Design

Figure 1 illustrates the methodology adopted in a schematic way; the subsections that follow describe our design choices in more detail.

3.1 Data

We chose complex, domain-specific corpora to test models on domain sensitivity, historical variation, and lexical precision, areas likely underrepresented in training. Though focused on two domains, our goal is to assess generalization to specialized, unfamiliar contexts. Both corpora are freely available on our GitHub.

Complex Case Study We chose the corpus collected by Maddela and Xu (2018) to analyse complexity, which contains around $15,000$ words. Following their methodology, lexical complexity was measured using a regression-based readability model that combines multiple features, including word length, inverse frequency, and syntactic depth. Words were then selected based on their complexity scores, using the 10th and 90th percentiles as thresholds to identify the simplest and most complex words in the dataset. This approach balanced the analysis across diverse distributions while excluding outliers such as Optical Character Recognition (OCR) errors or misclassified tokens.

Domain-Specific Case Study To evaluate how LLMs handle domain-specific vocabulary, we relied on the GOM corpus, a collection of 21 French gardening manuals published between 1802 and 1918, introduced in Colliaux and van Trijp (2024). This corpus, containing approximately $8,000$ words, reflects specialized knowledge on cultivation practices and offers a rich lexical window into historical, thematic language use. To quantify domain specificity, each word was assigned a keyness score based on the log ratio of its frequency in the GOM corpus to its frequency in *FRANTEXT*,a large reference corpus of over $4,000$ digitized French texts spanning the 16th century to

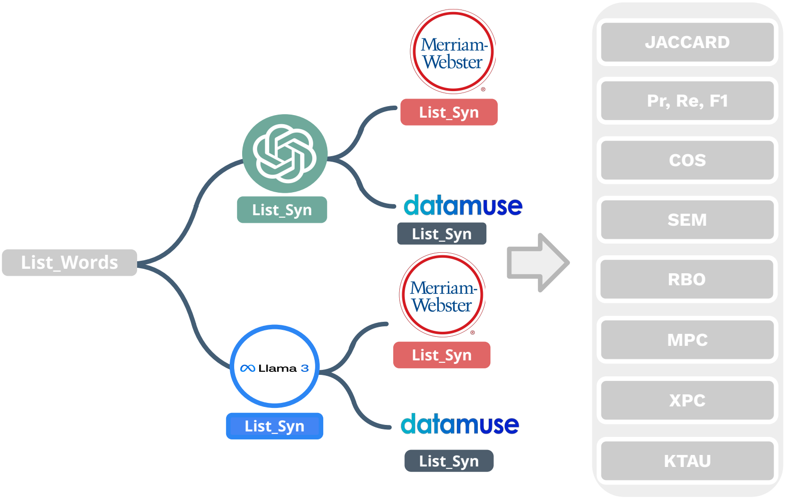

Figure 1: We generate synonyms using two models, then compare their outputs to two thesauri, assessing performance with ten metrics that measure overlap, semantic similarity, and ranking consistency. For abbreviations of the evaluation metrics, see Section 3.3.

the present (Bernard et al., 2002). The keyness of each term was calculated as $k = log\frac{f_{GOM}}{f_{REF}}$ where f_{GOM} is the frequency of occurrences in our corpus and f_{REF} is the frequency of occurrences in a reference corpus. We then applied a percentile-based thresholding approach: words in the top 10th percentile of keyness scores were classified as *high-keyness*, indicating strong domain anchoring, while those in the bottom 10th percentile were marked as *low-keyness*, suggesting more general usage. For example, the verb "arroser" (*"to water"*) is highly relevant in the context of gardening and thus receives a high keyness score due to its elevated frequency in GOM compared to FRANTEXT. In contrast, words with near-zero keyness scores appear with similar frequency across both corpora.

Data Preprocessing To enhance the quality and comparability of the synonym datasets, several preprocessing steps were applied. First, all words were converted to lowercase to mitigate discrepancies due to capitalization. Common stop-words were then removed to ensure focus on meaningful synonyms. Single-character words, often representing noise, were filtered out. Next, words were lemmatized to their base forms, allowing for more consistent comparisons. To further refine the dataset, scores were aggregated for unique lemmas by computing the mean of their occurrences (n), ensuring fair representation as in $A = \frac{1}{n}\sum_{i=1}^{n} \text{score}_i$

3.2 Models

Large Language Models We evaluate synonym generation across different architectures and computational constraints, balancing performance, efficiency, and applicability in real-world educational and linguistic tasks. First, a GPT-based model was included for its advanced language generation capabilities. We used `GPT-3.5-Turbo` model via the `OpenAI API`. The model operates with max_tokens = 50, $temperature$ = 0.7, and top_p = 0.9 to ensure a balance between diversity and precision[3]. Additionally, a LLaMA model were incorporated as strong open-weight alternatives known for their robust performance in natural language tasks. More precisely, we used `Meta-LLaMA-3.1-8B-Instruct` model, a locally hosted LLaMA 3.1 variant with 8 billion parameters, executed via the `LLaMA_cpp` library. The model operates with max_tokens = 50, $temperature$ = 0.7, top_k = 50, and top_p = 0.9.

A single prompt was used in this study for both LLMs to ensure consistency. The system prompt defines the model's role as a helpful assistant providing synonyms for words, while the user prompt explicitly instructs the model to return a plain, comma-separated list of synonyms without additional formatting, such as special characters, as

[3] Our aim isn't to benchmark the latest models, but to highlight persistent structural challenges in synonym generation. GPT-4 showed only minor improvements and still struggled with domain-specific and rare terms, underscoring the task's inherent complexity

it can be seen in Figure 2. Finally, the chosen models represent distinct paradigms, proprietary vs. open-source, allowing us to probe structural differences in synonym generation behavior without conflating architectural or training scale variation. Likewise, the thesauri contrast expert-curated and API-based resources. Our goal was not to exhaustively benchmark all systems, but to establish a diagnostic framework for evaluating lexical sensitivity in controlled conditions. Moreover, given the rapid pace of LLMs development, it is not feasible to evaluate every available model.

Thesauri Creating a definitive gold standard for synonyms generation, especially in specialized domains, is challenging. Lacking one, we use a human-derived corpus as a reference. We selected the *Merriam-Webster Thesaurus* and *Datamuse* as representative thesauri, accessible via API. The *Merriam-Webster Thesaurus* is an expert-curated database that provides structured synonym and antonym lists with contextual examples. *Datamuse* is a web-based lexical resource that provides definitions, synonyms, antonyms, example usages, and other word-related data through a programmable API. Unlike large language models, Datamuse relies on structured, human-curated lexical databases, primarily based on authoritative sources like WordNet. Both the thesauri were accessed through their APIs. In our work, we do not consider the thesauri as absolute gold standards. Instead, they act as comparative baselines for systematic evaluation. Their limitations, including sparse domain-specific coverage and the absence of contextual adaptation, underscore the need for complementary human judgment and dynamic benchmarks discussed in Section 6 and Section 5. Other thesauri could be readily integrated in place of or alongside the current ones. The structure and evaluation script are designed to support such modular substitution, facilitating adaptation to diverse linguistic and licensing contexts.

English Prompt

You are a helpful assistant who provides synonyms for words. Provide a plain, clean list of synonyms for the word 'word'. Do not include any newlines, bullet points, special characters, or formatting. Just output the synonyms as a simple, comma-separated list of words without any extra spaces or symbols.

French Prompt

Vous êtes un assistant utile qui fournit des synonymes pour les mots. Donnez une liste simple et propre de synonymes pour le mot 'word'. N'incluez pas de sauts de ligne, de points, de caractères spéciaux ou de formatage. Donnez simplement les synonymes sous forme de liste de mots séparés par des virgules sans espaces ou symboles supplémentaires.

Figure 2: Prompts used to generate synonyms for a given word using LLMs in both English and French.

To address this limitation, we explicitly track instances where the thesauri return no synonyms, thereby accounting for their incompleteness. When the Merriam-Webster API fails to provide synonyms for a given word, we log the instance and record it instead of leaving the entry blank. Additionally, we track the proportion of words without synonyms to assess thesaurus coverage. This ensures that missing data is systematically incorporated into our evaluation rather than ignored, allowing for a more accurate comparison of lexical resource reliability. For the domain-specific scenario, we rely solely on Datamuse, as Merriam-Webster yielded many empty results.

3.3 Automatic Evaluation Metrics

To evaluate synonym set similarity between LLMs and traditional thesauri, we use ten metrics capturing lexical overlap, semantic similarity, and ranking consistency as it can be seen in Table 2. We used multiple metrics to balance strict correctness with domain fit, allowing flexible interpretation based on application needs. While we avoid ranking metrics, we note that precision may be especially relevant for synonym selection and will clarify this in the revision. Lexical overlap is assessed with Jaccard Similarity, JS, Precision (P), Recall(R), and F1-Score($F1$), while Cosine Similarity measures word co-occurrence patterns. Semantic similarity is computed and further refined with Mean Pairwise Cosine Similarity and Max Pairwise Cosine Similarity($MeanPCS$/$MaxPCS$), ensuring a nuanced assessment of coherence. Finally, ranking consistency is evaluated using Rank-Biased Overlap(RBO) and Kendall's Tau(KT), which prioritize top-ranked words and normalize correlation measures for interpretability. All similarity metrics are systematically computed across datasets, using the same embedding model,i.e. `SentenceTransformer('all-MiniLM-L6-v2')`.

Metric	Type	Strengths	Weaknesses	Applications
JS	S	Simple, interpretable	Sensitive to set size	Set overlap evaluation
P	S	Measures correctness	Ignores coverage	Synonym accuracy
R	S	Measures coverage	Ignores correctness	Synonym completeness
$F1$	S	Balances precision and recall	Requires interpretation	Overall synonym quality
CS	E	Captures semantic similarity	May misalign embeddings	Vector comparison
SS	E	Context-aware	Embedding quality-dependent	Contextual synonym relevance
RBO	R	Weighs top ranks heavily	Interpretation complexity	Ranked synonym comparison
$MeanPC$	E	Measures overall coherence	Hides outliers	Semantic consistency
$MaxPC$	E	Identifies closest match	Ignores distribution	Closest synonym similarity
KT	R	Captures rank order	Sensitive to ties	Rank correlation

Table 2: Evaluation metrics are categorized into set-based (S), embedding-based (E), and rank-based (R) approaches. The abbreviations used are: JS: Jaccard Similarity, P: Precision, R: Recall, $F1$: F1 Score, CS: Cosine Similarity, SS: Semantic Similarity, RBO: Rank-Biased Overlap, $MeanPC$: Mean Pairwise Cosine, $MaxPC$: Max Pairwise Cosine, and KT: Kendall's Tau.

For each target word, we compute similarity metrics between the list of synonyms generated by an LLM and those retrieved from the thesaurus. These metrics are computed on a per-word basis and subsequently aggregated across the dataset to yield model-level averages. We deliberately excluded sentential context in the synonym generation task to ensure a fair and controlled comparison with traditional lexical resources such as thesauri, which inherently lack contextual grounding. Including contextual cues would risk introducing evaluation asymmetries, as LLMs can leverage context to generate dynamic, situational synonyms, whereas thesauri offer static, context-agnostic alternatives. Our goal is not to assess the pragmatic adaptability of models, but rather to isolate and measure their lexical sensitivity in the absence of external cues, enabling a direct comparison of lexical choices under the same constraints.

4 Results and Discussion

Complexity Scenario For simplification, systems need to balance precision, avoiding misleading or unnatural substitutions, with recall, ensuring enough lexical variety to prevent repetitive or impoverished output. Our comparison of synonym sets generated by GPT and LLaMA across complexity levels (Table 3) highlights this trade-off.

For simple words, GPT shows stronger alignment with Merriam-Webster Thesaurus (MWT), achieving higher lexical overlap ($JS = 0.22$) and recall ($R = 0.32$), while LLaMA yields slightly lower overlap but greater ranking consistency ($KT = 0.64$). Both models maintain high precision with Datamuse (GPT: 0.86, LLaMA: 0.67). For complex words, GPT again aligns more closely with MWT ($JS = 0.18$, $P = 0.33$), whereas LLaMA exhibits better rank agreement ($RBO = 0.03$ vs. GPT's 0.02). Across both complexity levels, semantic similarity scores remain consistently high ($SS = 0.98$–0.99), suggesting that even when exact matches are missing, generated synonyms are semantically close. These results indicate that GPT is generally more reliable for complex word substitutions due to higher precision, whereas LLaMA offers broader but less structured alternatives that may increase lexical variety at the risk of inconsistency. For simplification, this means GPT may be safer in educational or accessibility-oriented settings, while LLaMA might be better suited when diversity is prioritized over strict fidelity.

We also note a consistent pattern across thesauri: precision is higher with Datamuse, while recall is higher with MWT. This reflects structural differences: Datamuse produces shorter, algorithmically curated lists, favoring precision, whereas MWT provides broader expert-curated synonym sets, improving recall but reducing precision due to larger candidate pools. This trade-off illustrates how reference resource choice shapes evaluation and reinforces the importance of multiple benchmarks for lexical sensitivity.

Domain-specific Scenario Domain-specific evaluation (Table 4) reveals that LLMs particularly struggle with **high-keyness modifiers**, especially adjectives and adverbs. In these cases, lexical overlap scores collapse ($JS = 0.00$, $P = 0.00$ for

Table 3: Comparison of metrics across different complexities. Abbreviations are used as in Section 3.3. DM stands for Datamuse thesaurus, and MWT for Merriam-Webster Thesaurus. Top scores per model/thesaurus are highlighted; overall bests are **bolded.** GPT excels in precision with simple terms; LLaMA shows diversity but less alignment

Metric	**GPT**				**LLaMA**			
	Simple		**Complex**		**Simple**		**Complex**	
	MWT	**DM**	**MWT**	**DM**	**MWT**	**DM**	**MWT**	**DM**
JS	**0.22**	0.11	0.18	0.06	0.18	0.13	0.13	0.08
P	0.41	**0.86**	0.33	0.73	0.27	0.67	0.18	0.50
R	0.32	0.11	0.29	0.06	**0.35**	0.14	0.33	0.09
$F1$	**0.36**	0.20	0.31	0.11	0.31	0.23	0.23	0.15
CS	0.42	**0.64**	0.05	0.49	0.36	0.51	0.10	0.48
SS	0.99	0.99	0.98	0.98	0.99	**1.00**	0.98	0.99
RBO	0.08	0.17	0.02	**0.32**	0.01	0.16	0.03	0.01
$MeanPCS$	**0.22**	**0.22**	0.19	0.19	**0.22**	0.21	0.19	0.18
$MaxPCS$	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
KT	**0.71**	0.64	0.69	0.63	0.64	0.63	0.67	0.62

Table 4: Comparison of results across different keyness levels. Abbreviations for the metrics are used as in Table 3. ADJ stands for adjectives, ADV for adverbs. The highest metric for model is highlighted, while the overall best values are in bold. Both models collapse on high-keyness adverbs, underscoring the difficulty of domain-rich substitution.

Metric	**GPT**						**LLaMA**					
	Low-Keyness			**High-Keyness**			**Low-Keyness**			**High-Keyness**		
	Adj	**Noun**	**Adv**	**Adj**	**Noun**	**Adv**	**Adj**	**Noun**	**Adv**	**Adj**	**Noun**	**Adv**
JS	0.01	0.02	0.01	0.01	0.01	0	0.01	**0.03**	0.01	0.01	0.01	0.0
P	0.01	0.02	0.02	0.01	0.01	0	0.02	**0.03**	0.02	0.01	0.02	0.0
R	0.14	0.26	0.02	0.07	0.15	0	0.12	**0.27**	0.01	0.07	0.12	0.0
$F1$	0.02	0.04	**0.06**	0.02	0.02	0	0.03	**0.06**	0.02	0.02	0.03	0
CS	0.08	**0.15**	0.08	0.04	0.08	0	0.09	**0.15**	0.08	0.04	0.09	0.0
SS	**0.97**	**0.97**	0.89	0.93	**0.97**	0.74	**0.97**	**0.97**	0.89	0.93	0.96	0.85
RBO	0.00	0.00	0.00	0.00	0.00	0	0.00	0.00	0.00	0.00	0.00	0.0
$MeanPCS$	0.22	0.21	0.22	0.21	0.21	0.22	0.22	0.21	0.22	0.21	0.21	**0.23**
$MaxPCS$	**1**	**1**	**1**	**1**	**1**	0.76	**1**	**1**	**1**	**1**	**1**	0.85
KT	0.67	0.62	0.36	**0.73**	0.59	0.05	0.55	0.19	0.68	0.68	0.58	0.50

several categories), despite semantic similarity remaining high (SS up to 0.97). Nouns, by contrast, achieve higher precision and recall across all keyness levels, with LLaMA slightly outperforming GPT in low-keyness noun retrieval ($P = 0.03$, $R = 0.27$). GPT, however, demonstrates greater stability overall, with higher semantic similarity ($SS = 0.97$) and ranking consistency ($KT = 0.73$ for high-keyness adjectives). This uneven performance has direct consequences for simplification:

modifiers are often the carriers of nuance in specialized domains (e.g., technical adjectives and adverbs). When models fail to retrieve appropriate synonyms for these terms, simplified texts risk becoming oversimplified or distorted, erasing crucial distinctions. Conversely, both models are relatively stronger with nouns, which may make them more dependable for substituting central content words.

An apparent contradiction emerges in this scenario: surface-level overlap metrics (e.g., $F1$) fall near zero for high-keyness terms, while embedding-based measures (e.g., SS) remain high(often > 0.95). This apparent contradiction reflects the limitations of surface-level overlap metrics (like F1) when applied to highly specialized vocabularies. For domain-specific terms, traditional thesauri often lack relevant synonyms entirely or provide outdated lexical mappings. Consequently, LLMs may generate contextually appropriate alternatives that are absent from the reference lists, resulting in low Jaccard overlap and precision/recall. However, embedding-based measures such as SS, which are derived from mean sentence embeddings, still capture the overall semantic coherence between the sets, explaining the high scores. This suggests that while reference coverage is sparse, the embedding space retains some generalizability even for rare or historical terms. Nevertheless, for simplification, this gap poses a risk: models may appear semantically adequate by vector metrics but fail to provide substitutions that are pedagogically useful or lexically appropriate. In practice, this could lead to simplified outputs that preserve meaning but lack naturalness, usability, or alignment with learner needs. Addressing this requires not only better benchmarks but also future integration of human evaluation to ensure that substitutions enhance rather than flatten simplified text.

Key Implications for Simplification Our findings show that both GPT and LLaMA struggle with high-keyness terms, particularly adjectives and adverbs, where lexical overlap scores approach zero despite high semantic similarity values. This pattern has direct consequences for text simplification. Simplification systems often rely on synonym substitution to replace complex or domain-specific terms with more accessible alternatives. If LLMs fail to retrieve appropriate synonyms for high-keyness vocabulary, simplification pipelines risk either omitting critical concepts or substituting them with overly generic words. The result is a form of *semantic flattening*: domain-specific nuance is lost, and the simplified text may no longer adequately convey the original meaning. Our benchmark therefore highlights an important diagnostic signal for simplification research: ensuring that models preserve lexical richness when simplifying specialized or contextually anchored texts. Our benchmark provides an important diagnostic signal: ensuring that synonym substitution in simplification pipelines does not merely preserve semantics in vector space, but also maintains lexical richness, usability, and domain-appropriate nuance.

5 Conclusion & Future Work

Our benchmark advances evaluation resources for lexical simplification and substitution, providing diagnostic insights for simplification pipelines. By systematically comparing LLMs to traditional thesauri, we show that while models generate semantically coherent synonyms, they continue to struggle with domain-specific and high-complexity vocabulary. Precision and recall remain limited when measured against curated lexical resources, yet models consistently achieve high semantic similarity (e.g., $SS \approx 0.97$), indicating that they capture meaning-driven relationships even when exact lexical matches are absent. These findings highlight the risk of semantic flattening in simplification: when high-keyness terms are poorly substituted, critical nuance may be lost and outputs risk becoming oversimplified or misleading. Our open-access framework thus offers both a benchmark for lexical sensitivity and a tool for diagnosing where simplification systems compromise expressiveness.

Future work should strengthen LLMs' handling of specialized language through richer domain-specific datasets, hybrid AI–lexicographic approaches, and context-aware learning strategies. Beyond automated metrics, incorporating human-centered evaluation will be essential for assessing the pragmatic adequacy of LLM-generated synonyms in simplification. A promising next step is small-scale human studies in which annotators judge synonym sets for contextual appropriateness, readability, and usefulness. This will help bridge the gap between automatic benchmarks and real-world applications, supporting simplification systems that preserve both accessibility and lexical richness.

6 Limitations

This study has three main limitations. First, we evaluate only two LLMs (GPT-3.5 and LLaMA-3.1), which do not capture the full range of current or emerging systems. Our goal is to present a framework and methodology rather than a definitive ranking, but future work should extend evaluation to additional models, including those explicitly trained for *text simplification*. Second, our benchmark relies exclusively on structured thesauri and automated metrics. While this enables reproducibility and scalability, it does not replace the need for human judgment. A complementary study with linguists, educators, or end users rating synonym sets for adequacy in simplification scenarios would provide valuable insights into the practical usefulness of model outputs. Third, we intentionally abstract away from sentential context to isolate lexical sensitivity. This allows for direct comparison with static lexical resources but does not capture the full complexity of contextual substitution. Incorporating sentence-level evaluations is a critical next step for simplification pipelines. Beyond synonym retrieval, future work should also extend to related aspects of lexical competence, such as antonym recognition, paraphrasing, and diachronic semantic shifts, in order to broaden the diagnostic value of our benchmark and strengthen its applicability to real-world simplification tasks.

References

Al Al-Darayseh. 2014. The impact of using explicit/implicit vocabulary teaching strategies on improving students' vocabulary and reading comprehension. *Theory & Practice in Language Studies (TPLS)*, 4(6).

Arisa Ashizawa, Ryota Mibayashi, and Hiroaki Ohshima. 2024. Query expansion in food review search with synonymous phrase generation by llm. In *International Conference on Database Systems for Advanced Applications*, pages 252–260. Springer.

Pascale Bernard, Josette Lecomte, Jacques Dendien, and Jean-Marie Pierrel. 2002. Computerized linguistic resources of the research laboratory atilf for lexical and textual analysis: Frantext, tlfi, and the software stella. In *LREC*. Citeseer.

Yves Bestgen. 2024. Measuring lexical diversity in texts: The twofold length problem. *Language Learning*, 74(3):638–671.

Yves Bestgen. 2025. Estimating lexical diversity using the moving average type-token ratio (mattr): Pros and cons. *Research Methods in Applied Linguistics*, 4(1):100168.

David Colliaux and Remi van Trijp. 2024. The discourse of the french method: making old knowledge on market gardening accessible to machines and humans. In *CHR*.

Liam Cripwell, Joël Legrand, and Claire Gardent. 2023. Document-level planning for text simplification. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 993–1006, Dubrovnik, Croatia. Association for Computational Linguistics.

Asma Farajidizaji, Vatsal Raina, and Mark Gales. 2024. Is it possible to modify text to a target readability level? an initial investigation using zero-shot large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9325–9339, Torino, Italia. ELRA and ICCL.

Jorge Gabín and Javier Parapar. 2025. Leveraging retrieval-augmented generation for keyphrase synonym suggestion. In *European Conference on Information Retrieval*, pages 311–327. Springer.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clinciu, Dipanjan Das, Kaustubh Dhole, et al. 2021. The gem benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)*, pages 96–120.

Reto Gubelmann, Michael Burkhard, Rositsa V Ivanova, Christina Niklaus, Bernhard Bermeitinger, and Siegfried Handschuh. 2024. Exploring the usefulness of open and proprietary llms in argumentative writing support. In *International Conference on Artificial Intelligence in Education*, pages 175–182. Springer.

Yanzhu Guo, Guokan Shang, Michalis Vazirgiannis, and Chloé Clavel. 2024. The curious decline of linguistic diversity: Training language models on synthetic text. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3589–3604, Mexico City, Mexico. Association for Computational Linguistics.

Steffen Herbold, Annette Hautli-Janisz, Ute Heuer, Zlata Kikteva, and Alexander Trautsch. 2023. A large-scale comparison of human-written versus chatgpt-generated essays. *Scientific reports*, 13(1):18617.

Carla W Hess, Kelley P Ritchie, and Richard G Landry. 1984. The type-token ratio and vocabulary performance. *Psychological Reports*, 55(1):51–57.

Subin Huang, Daoyu Li, Chengzhen Yu, Junjie Chen, Qing Zhou, and Sanmin Liu. 2025. Empowering entity synonym set generation using flexible perceptual field and multi-layer contextual information. *PloS one*, 20(4):e0321381.

Scott Jarvis. 2013. Defining and measuring lexical diversity. In *Vocabulary knowledge*, pages 13–44. John Benjamins Publishing Company.

David Jurgens, Saif Mohammad, Peter Turney, and Keith Holyoak. 2012. Semeval-2012 task 2: Measuring degrees of relational similarity. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 356–364.

Tom S Juzek and Zina B. Ward. 2025. Why does ChatGPT “delve” so much? exploring the sources of lexical overrepresentation in large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 6397–6411, Abu Dhabi, UAE. Association for Computational Linguistics.

Kelly Kendro, Jeffrey Maloney, and Scott Jarvis. 2024. Lexical diversity in human-and llm-generated text. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 46.

Tannon Kew, Alison Chi, Laura Vásquez-Rodríguez, Sweta Agrawal, Dennis Aumiller, Fernando Alva-Manchego, and Matthew Shardlow. 2023. Bless: Benchmarking large language models on sentence simplification. *arXiv preprint arXiv:2310.15773*.

Hannah Rose Kirk, Bertie Vidgen, Paul Röttger, and Scott A Hale. 2024. The benefits, risks and bounds of personalizing the alignment of large language models to individuals. *Nature Machine Intelligence*, pages 1–10.

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.

Meltem Kurt Pehlivanoğlu, Robera Tadesse Gobosho, Muhammad Abdan Syakura, Vimal Shanmuganathan, and Luis de-la Fuente-Valentín. 2024. Comparative analysis of paraphrasing performance of chatgpt, gpt-3, and t5 language models using a new chatgpt generated dataset: Paragpt. *Expert Systems*, 41(11):e13699.

Kristopher Kyle, Scott A Crossley, and Scott Jarvis. 2021. Assessing the validity of lexical diversity indices using direct judgements. *Language Assessment Quarterly*, 18(2):154–170.

Jinsook Lee, Yann Hicke, Renzhe Yu, Christopher Brooks, and René F Kizilcec. 2024. The life cycle of large language models in education: A framework for understanding sources of bias. *British Journal of Educational Technology*, 55(5):1982–2002.

Robert Lew. 2023. Chatgpt as a cobuild lexicographer. *Humanities and Social Sciences Communications*, 10(1):1–10.

Robert Lew. 2024. Dictionaries and lexicography in the ai era. *Humanities and Social Sciences Communications*, 11(1):1–8.

Qingyao Li, Lingyue Fu, Weiming Zhang, Xianyu Chen, Jingwei Yu, Wei Xia, Weinan Zhang, Ruiming Tang, and Yong Yu. 2023. Adapting large language models for education: Foundational capabilities, potentials, and challenges. *arXiv preprint arXiv:2401.08664*.

Dayiheng Liu, Yu Yan, Yeyun Gong, Weizhen Qi, Hang Zhang, Jian Jiao, Weizhu Chen, Jie Fu, Linjun Shou, Ming Gong, et al. 2021. Glge: A new general language generation evaluation benchmark. *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 408–420.

Mounica Maddela and Wei Xu. 2018. A word-complexity lexicon and a neural readability ranking model for lexical simplification. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino-Gómez. 2024. Beware of words: Evaluating the lexical diversity of conversational llms using chatgpt as case study. *ACM Transactions on Intelligent Systems and Technology*.

Philip M McCarthy and Scott Jarvis. 2007. vocd: A theoretical and empirical evaluation. *Language Testing*, 24(4):459–488.

Philip M McCarthy and Scott Jarvis. 2010. Mtld, vocd-d, and hd-d: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods*, 42(2):381–392.

Kai North, Tharindu Ranasinghe, Matthew Shardlow, and Marcos Zampieri. 2025. Deep learning approaches to lexical simplification: A survey. *Journal of Intelligent Information Systems*, 63(1):111–134.

Gustavo Paetzold and Lucia Specia. 2016. Benchmarking lexical simplification systems. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*, pages 3074–3080, Portorož, Slovenia. European Language Resources Association (ELRA).

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. Wic: the word-in-context dataset for evaluating context-sensitive meaning representations. 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 1267–1273.

Brian Richards. 1987. Type/token ratios: What do they really tell us? *Journal of child language*, 14(2):201–209.

Daniel Rothschild. 2025. Language and thought: The view from llms. *arXiv preprint arXiv:2505.13561*.

Horacio Saggion. 2017. *Lexical Simplification*, pages 21–31. Springer International Publishing, Cham.

Thaeer M Sahib, Osamah Mohammed Alyasiri, Hussain A Younis, Dua' Akhtom, Israa M Hayder, Sani Salisu, and Darmawati Muthmainnah Besse. 2023. A comparison between chatgpt-3.5 and chatgpt-4.0 as a tool for paraphrasing english paragraphs. In *Int. Applied Social Sciences (C-IASOS-2023) Congress*, pages 471–480.

Matthew Shardlow, Fernando Alva-Manchego, Riza Theresa Batista-Navarro, Stefan Bott, Saul Calderon-Ramirez, Rémi Cardon, Thomas François, Akio Hayakawa, Andrea Horbach, and Anna Huelsing. 2024. The bea 2024 shared task on the multilingual lexical simplification pipeline. In *Proceedings of the 19th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2024)*, pages 571–589.

Dinh Huynh Mai Tu. 2024. Navigating the convenience trap with chatgpt and google translate: The risks of homogenization in translation teaching for vietnamese students. *International Journal of Linguistics, Literature and Translation*, 7(8):244–255.

Karina Vidal and Scott Jarvis. 2020. Effects of english-medium instruction on spanish students' proficiency and lexical diversity in english. *Language Teaching Research*, 24(5):568–587.

Yan Xu, Tao Wang, Yang Yuan, Ziyue Huang, Xi Chen, Bo Zhang, Xiaorong Zhang, and Zehua Wang. 2025. Llm-enhanced framework for building domain-specific lexicon for urban power grid design. *Applied Sciences*, 15(8):4134.

Hiromu Yakura, Ezequiel Lopez-Lopez, Levin Brinkmann, Ignacio Serna, Prateek Gupta, and Iyad Rahwan. 2024. Empirical evidence of large language model's influence on human spoken communication. *arXiv preprint arXiv:2409.01754*.

Jingheng Ye, Shen Wang, Deqing Zou, Yibo Yan, Kun Wang, Hai-Tao Zheng, Zenglin Xu, Irwin King, Philip S Yu, and Qingsong Wen. 2025. Position: Llms can be good tutors in foreign language education. *arXiv preprint arXiv:2502.05467*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. 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 1298–1308.