The Vocabulary Fingerprint of AI Rewriting: Common Words AI Language Models Prioritize

TextPulse Research · Working Paper

Abstract

Certain words have become informal markers of AI writing. Readers now treat “underscores”, “pivotal”, and “tapestry” as signs that a text came from an AI language model, and published studies confirm that these words have surged in the scientific literature after 2022. This study measures the vocabulary signature in a setting the published work does not cover, the rewriting of existing human text. We compared 60,786 human-written academic texts with rewrites of the same texts produced by eight model configurations across six model families, amounting to 40.4 million tokens in total. Candidates were the 1,000 most common content words in the corpus, 250 for each part of speech, and 57 widely publicized AI words tracked separately. Every candidate was scored with a regularized log-odds ratio of AI use against human use within the corpus. The signature is consistent. The strongest AI-leaning words are formal connectives and Latinate substitutions such as “thereby” (13 times the human usage rate), “consequently” (11 times), and “utilized” (7 times), while models suppress the plain words humans prefer, with “used” cut to one ninth of its human rate. The famous chatbot words behave differently. Terms such as “meticulously” (214 times), “pivotal” (15 times), and “underscores” (19 times) are strongly enriched, but “delve”, the most famous marker of all, is not enriched in rewriting at all. Word origin analysis shows the AI-leaning vocabulary is 44 percent Latinate against 10 percent for the human-leaning set, with nearly twice the syllable count. The signature varies sharply by model. The model that preserves citations best in a prior-work audit alters vocabulary the most. We release the full lexicon of 1,057 words with corpus statistics, WordNet synsets, and an AI score per word as an open, publicly available dataset, as well as a full-vocabulary lexicon that attempts to score the entire English vocabulary (i.e., every single-word WordNet lemma; 90,520 entries), of which 3,371 pass a two-signal evidence check and receive a score above zero.

1. Introduction

When a language model rewrites a sentence, it does not choose words the way the original author did. It chooses the words that its training and tuning made the most probable. Repeated across millions of users, these choices have produced a recognizable fingerprint. Readers describe it with informal labels such as “AI fluff”, and since 2024 the flood of low-effort machine text has been referred to as “AI slop”, which is a term that dictionaries selected as a word of the year in 2025. A specific set of common English words appears in AI output at several times, sometimes hundreds of times, the human usage rate.

The published evidence for this phenomenon comes almost entirely from generated text, meaning that text the model initially wrote from a prompt. Word-frequency studies of scientific abstracts and conference peer reviews show sharp post-2022 increases in a stable set of style words. What has not been measured yet is the vocabulary effect of the *transformation* case, where a model *rewrites text a human already wrote*. Rewriting is among the most common realistic uses of language models on academic text, and it is the case where vocabulary change matters most directly, since every changed word replaces a word the author actually chose.

We use a paired corpus of human-written texts and an AI-generated version of each. Since we compare each rewrite with the human text it originated from, the topic and content are constant, and any vocabulary change is caused by the model itself. This study uses that design at scale to answer four separate questions. Which common words do models inject and suppress when rewriting academic text? Do the famous chatbot words behave the same way in rewriting as in generation? How strongly does the signature differ across models on identical inputs? What kind of vocabulary is it, measured by word origin and complexity? With this analysis, we release a publicly available lexicon that tags each studied word with an AI score derived from corpus statistics.

2. Literature review and popular usage

Word-level evidence for an AI vocabulary fingerprint has emerged rapidly after ChatGPT's release. Kobak et al. (2025) analyzed more than 15 million biomedical abstracts from 2010 to 2024 using an excess-vocabulary method modeled on excess mortality, and found a sudden surge in a set of style words including "delve" and "underscore" after 2022. From the excess usage they estimate that at least 13.5 percent of 2024 abstracts were processed with a language model, rising to about 40 percent in some other corpora. Liang et al. (2024) applied a corpus-level maximum likelihood method to peer reviews at AI conferences and estimated that 6.5 to 16.9 percent of review text was significantly modified by language models, with the adjectives "commendable", "meticulous", and "intricate" rising 9.8, 34.7, and 11.2 times human use. Their finding that adjectives carry the most reliable signal aligns with a pattern in our results. A cross-lingual study documents the same lexical changes spreading through news writing in 34 languages (Juzek, 2026).

Juzek and Ward (2025) investigated why models overuse these specific words. Testing model families and pipeline stages, they conclude that the overrepresentation is not explained by architecture or training data alone, and point to reinforcement learning from human feedback (RLHF) as the most probable source. Subsequent work strengthens the case that feedback-based alignment changes lexical choice toward a limited set of prioritized vocabulary (Juzek, 2025). Section 3 explains this further.

In real-world use, the AI vocabulary is community knowledge. Detector vendors publish lists of AI-favored words and phrases. The largest is GPTZero's AI Vocabulary resource, built from 3.3 million texts, which reports phrases appearing 10 to more than 200 times as often in AI documents as in human documents. Popular writing guides circulate lists of words to remove so that text stops

sounding mechanical, robotic, or machine-written, and our own article on AI jargon in research writing collects the same vocabulary from an editing perspective. These informal labels are now part of the language itself. “AI fluff” is used to refer to the filler register. “AI slop”, a term that spread from online forums in 2022 and was seen as a general label in 2024, was named word of the year for 2025 by Merriam-Webster, the American Dialect Society, and the Macquarie Dictionary.

The commonly cited vocabulary can be grouped by lexical category. The following examples recur across the published studies and the online popular lists.

- Verbs: delve, underscore, foster, leverage, bolster, elucidate, showcase, harness, embark, navigate.
- Nouns: tapestry, realm, landscape, caveat, lens, testament, beacon, interplay, myriad, synergy.
- Adjectives: pivotal, crucial, multifaceted, intricate, meticulous, holistic, nuanced, paramount, commendable, transformative.
- Adverbs: meticulously, seamlessly, notably, invariably, effortlessly.
- Phrases: “it is important to note”, “plays a pivotal role”, “a testament to”, “in the realm of”, “rich tapestry”, “delve into”, “the smoking gun”, “one honest caveat”, “in today’s fast-paced world”, “not only … but also”.

However, the literature does not contain a paired measurement. The published studies compare corpora across time or against reference sets, so they measure generation and adoption at the same time. None of them holds the source text constant and asks what a model does to an author’s vocabulary when it *rewrites*. This study aims to fill this gap, and the paired design also allows us to use reliable statistics to generate a word-level lexicon, which no published list provides with open per-word “AI scores”.

3. Why models favor these words

The signature is explained by three overlapping concepts. A language model writes by repeatedly choosing the most likely next token from a probability distribution. Everything the model knows about language is expressed in that distribution, and ordinary decoding samples from its upper region. Perplexity is the standard measure of how “surprising” a text is under such a distribution. Text that remains inside the model’s high-probability choices has low perplexity, which is why perplexity-based detectors treat easy predictability as a machine signal.

Tuning changes the distribution. After pretraining, models are aligned with reinforcement learning from human feedback, in which human raters reward outputs they preferred. Raters’ preferences encode a register. Polished, formal, slightly elevated wording is rewarded, and the tokens that express that register receive a constant probability boost. Once “utilized” is scored as safer than “used” and “pivotal” as safer than “important”, the aligned model treats the elevated word as the most likely next token in contexts where a human author would have simply written the plain one.

Juzek and Ward (2025) found that base models overuse the focal words far less than their feedback-tuned versions, and subsequent experiments reproduce similar polished-to-plain swaps when RLHF tuning is applied (Juzek, 2025).

Every rewrite that swaps a plain word for the preferred polished variant produces more text in the elevated register, some of which enters future training data. The vocabulary signature this study measures is the visible lexical trace of that probability concentration, and the reason why it generally clusters in specific words rather than spreading evenly across the English vocabulary.

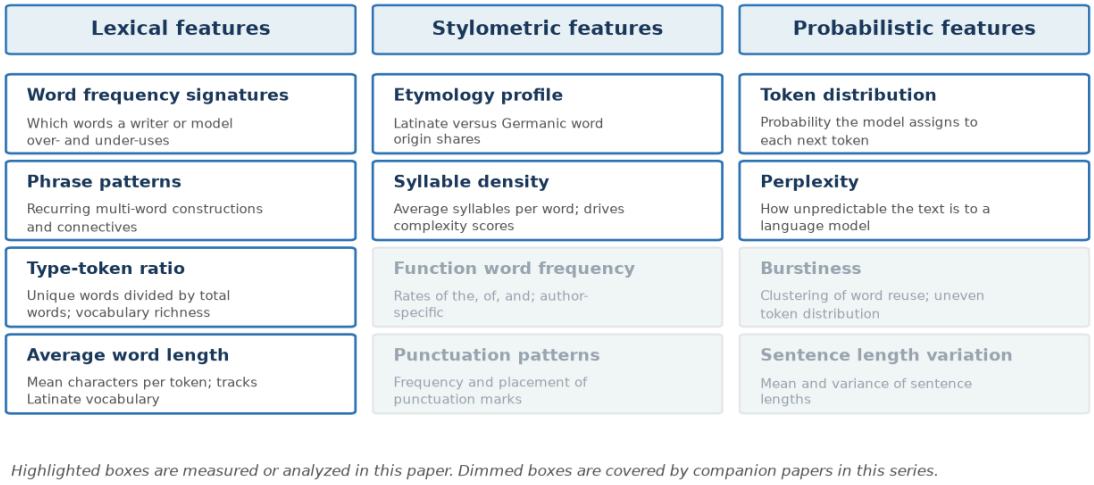

Figure F1. The feature families of AI text analysis. Highlighted families are measured in this paper.

4. Data and methods

4.1 Corpus

The audit uses the internal corpus of 60,786 human-AI paraphrase pairs built from open-access academic text that was also used in our citation fidelity study (TextPulse Research, 2026c). Each pair holds a source text of roughly 100 to 400 words from published scholarly writing across STEM, medicine, social science, economics, and humanities disciplines, and a rewrite of that text by one of eight model configurations. Corpus 1 contains 38,323 pairs from two configurations, a DeepSeek chat model and a Gemini Flash-class model, produced under a prompt designed for faithful rewriting. Corpus 2 contains 22,463 pairs from six configurations in the DeepSeek, Grok, Mistral, Qwen, Gemini, and OpenAI families under a common rephrase prompt. The paired sides amount to 40.4 million tokens. Model comparisons are reported within each corpus.

4.2 Candidates

Candidates are the most common content words in the corpus, not rare words. For preprocessing, we lowercased and tokenized both sides, excluded function words, and assigned each word its dominant WordNet part of speech. The 250 most frequent words for each of noun, verb, adjective, and adverb form the primary candidate set of 1,000 words. A second tier adds the publicized AI words from the studies and lists reviewed in Section 2. Most of them are already inside the frequency cutoff, and the 57 that are outside it are tracked with identical statistics regardless of rank, so that the famous vocabulary is measured rather than assumed. This gives 1,057 studied words. Dominant-POS assignment is a simplification for words that serve several parts of speech, and a few candidates such as “within” and “through” are prepositions that WordNet classes as adverbs. We retain them and note the labeling.

WordNet is a large lexical database of English where nouns, verbs, adjectives, and adverbs are grouped into ‘synsets’, or sets of synonyms that each express one distinct concept (Miller, 1995). Synsets are linked to one another and to their member words by semantic and lexical relations such as hypernymy, antonymy, and derivation, so the vocabulary of each lexical category forms a semantic network that can be traversed by meaning rather than by spelling (Fellbaum, 1998). We use WordNet to assign each candidate its part of speech, and to map every studied word to its synsets and synonym neighbors (see Section 5.6).

4.3 Statistics

For every candidate and model we compute the usage rate per 10,000 tokens on each side (human or AI) and a log-odds ratio with an informative Dirichlet prior, which is the standard method for comparing word use between corpora (Monroe et al., 2008), with a z-score on the regularized difference. For a word w, let y_AI and y_H be its token counts on the AI and human sides, n_AI and n_H the total token counts on each side, and p the word’s pooled rate of use across both sides. The regularized log-odds ratio is

$$\delta(w) = \log\left((y_{AI} + \alpha_0 \cdot p) / (n_{AI} + \alpha_0 - y_{AI} - \alpha_0 \cdot p) \right) - \log\left((y_H + \alpha_0 \cdot p) / (n_H + \alpha_0 - y_H - \alpha_0 \cdot p) \right),$$

with total prior mass $\alpha_0 = 500$. The prior acts as $\alpha_0 \cdot p$ pseudo-occurrences of the word on each side, which stabilizes the estimate for words with few occurrences. The z-score divides $\delta(w)$ by its estimated standard error, $\mathrm{sqrt}(1/(y_{AI} + \alpha_0 \cdot p) + 1/(y_H + \alpha_0 \cdot p))$. The AI score released in the lexicon maps this log-odds, pooled over all models, onto a 0 to 100 scale through the logistic (sigmoid) function σ:

$$\text{AI score}(w) = 100 \cdot \sigma(\delta(w)) = 100 / (1 + \exp(-\delta(w))),$$

so 50 means equal use on both sides, scores above 50 mean AI-leaning use, and scores below 50 mean human-leaning use. Document frequencies, the share of texts containing a word at least once, come from a full corpus scan. The same run counts a curated list of 58 publicized AI phrases in addition to the common “not only … but also” construction, and discovers high-contrast word sequences directly from the corpus. Word origin is classified Latinate or Germanic with a

documented suffix heuristic and curated core lists. Syllables are counted with a standard vowel-group method. All scripts are released.

4.4 The full-vocabulary lexicon

The 1,057 studied words cover the common core, but the same corpus supports scoring far more of the English vocabulary. We therefore release a second, full-vocabulary lexicon that attempts to score every single-word WordNet lemma, 90,520 lemma and part-of-speech entries in total, against the same 40.4 million tokens. Each surface form in the corpus is assigned its dominant WordNet part of speech and reduced to its base lemma with WordNet's morphological analyzer, so that "used", "using", and "uses" all count toward the verb "use".

Two corpus signals are computed for every lemma. The token signal δ_tok is the Dirichlet-regularized log-odds ratio of AI against human token rate defined in Section 4.3, computed on lemma counts. The document signal is the Jeffreys-smoothed log-odds ratio of document presence: with d_AI and d_H the number of AI rewrites and human sources containing the lemma at least once, and D_AI and D_H the total number of documents on each side,

$$\delta_{doc} = \log\left((d_{AI} + 0.5) / (D_{AI} - d_{AI} + 0.5) \right) - \log\left((d_H + 0.5) / (D_H - d_H + 0.5) \right).$$

Requiring both signals prevents a lemma from scoring on heavy repetition inside a small handful of texts. The assumption is that a true AI-preferred word is used more often overall and appears in more documents. A lemma receives a score above zero only when it passes an evidence check, which is at least 25 combined occurrences, and a z-score of at least 3.29 ($p < 0.001$) on each signal independently, in the AI direction. For a lemma that passes the check, the AI score maps the mean of the two log-odds, $\delta = (\delta_{tok} + \delta_{doc}) / 2$, through a one-sided logistic transform:

$$\text{AI score} = 100 \cdot (2 \cdot \sigma(\delta) - 1) = 100 \cdot (2 / (1 + \exp(-\delta)) - 1),$$

so the score runs from just above 0 (barely detectable overuse) to 100 (extreme overuse). A 95 percent confidence interval is obtained by mapping $\delta \pm 1.96 \cdot s$ through the same transform, where $s = (1/2) \cdot \sqrt{v_{tok} + v_{doc}}$ combines the variance estimates of the two signals. Every other lemma is scored 0, whether the corpus contains too little evidence, the usage does not differ enough, or the preference is toward human writing. Human-leaning direction and the complete underlying statistics are released for every observed lemma, so any alternative cut can be recomputed. Note that the two lexicons use intentionally different scales: the 1,057-word lexicon is centered, with 50 meaning equal use, while the full-vocabulary score is one-sided, with 0 meaning no observable AI preference. The two releases are related as subset and whole: the 1,057-word studied lexicon is a small subset containing the most prominent AI-preferred words of the common vocabulary, while the full-vocabulary lexicon scores the entire English vocabulary as represented in WordNet.

5. Results

5.1 The rewrite signature

Models replace plain vocabulary with more formal alternatives. The strongest AI-favored words in the primary candidate set are connectives and Latinate verbs. “Thereby” appears at 10.1 per 10,000 tokens in rewrites against 0.77 in the human sources, and appears in 27 percent of all rewrites against 2 percent of human texts. “Consequently” rises from 0.96 to 11.0 per 10,000. “Employed” rises from 1.25 to 7.8, “utilized” from 0.84 to 5.7, and “constitutes” from 0.24 to 6.6. The preposition “within” rises more than fivefold from 7.5 to 40.9 per 10,000 and appears in two thirds of all AI rewrites.

“Used”, the most human-leaning word in the study, falls from 16.9 per 10,000 in human text to 1.8 in AI rewrites. “Different” falls from 13.8 to 3.5, “important” from 8.7 to 1.8, “people” from 8.7 to 2.6, and “way” from 4.3 to 0.5. The pattern is one substitution repeated across the vocabulary. The author’s ordinary word is discarded, and the model’s elevated equivalent replaces it.

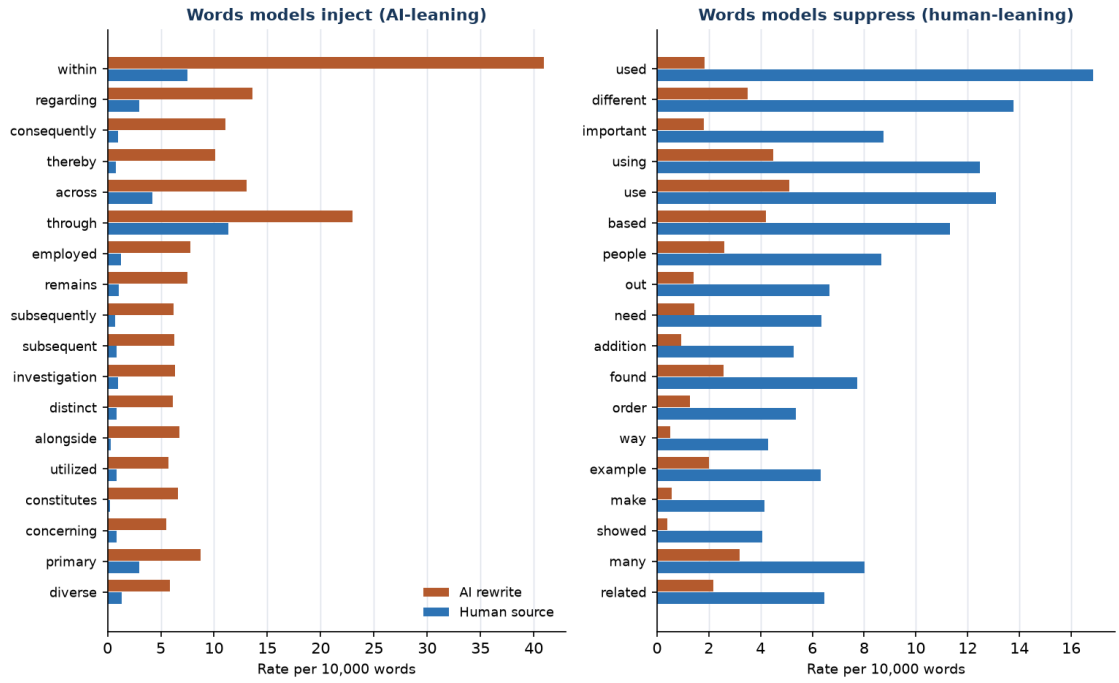

Figure F2. The words models inject and suppress when rewriting, with rates per 10,000 tokens.

5.2 Differences between models

Signature strength varies widely on the same inputs. Counting candidates whose usage changed strongly (z of at least 10), the Gemini Flash configuration of corpus 1 changed 422 of the 1,057 studied words, Grok 403, Mistral 317, and DeepSeek 295, while OpenAI changed 132 and Qwen only 76. Individual models also have identifiable habits. Grok is responsible for the most extreme single-word behavior in the corpus, writing “thereby” at 16.4 per 10,000 tokens against 0.63 in its sources and “delineates” at 5.3 per 10,000 against 0.004, a word its source texts rarely use. The corpus 2 Gemini configuration writes “meticulously” at 9.3 per 10,000 in texts whose human versions do not contain the word at all.

Comparing these results with our citation fidelity audit of the same corpus leads to a caution for tool evaluation. Grok preserved citations better than any other configuration in that study, corrupting 0.30 percent of citations, yet it rewrites vocabulary second most aggressively. Qwen showed the weakest citation fidelity and the weakest vocabulary change. Faithfulness to a text's references and faithfulness to its voice are different properties, and as shown here, a model can excel at one while failing the other.

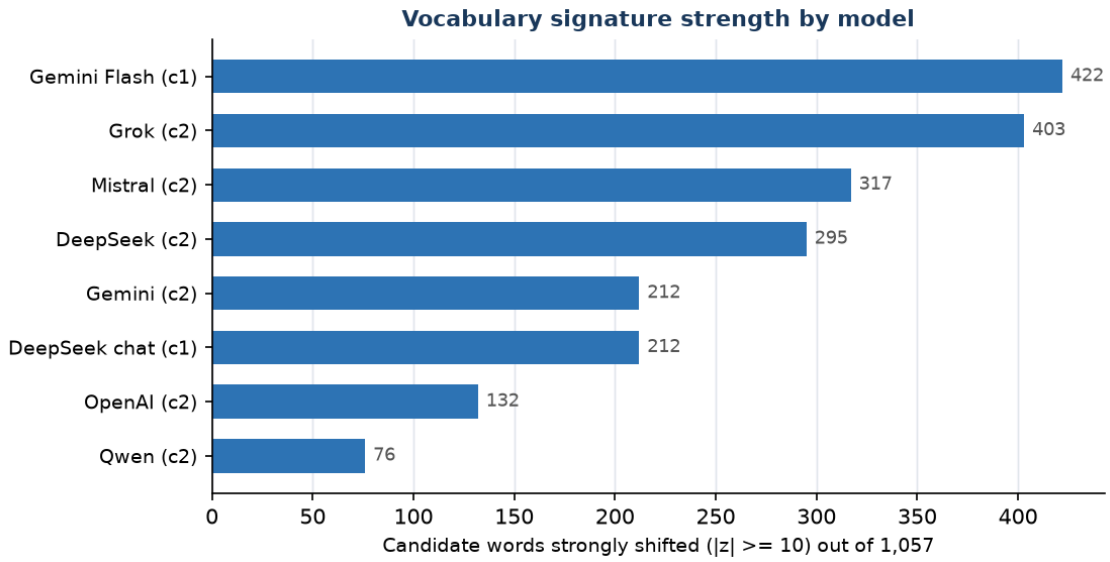

Figure F3. Number of candidate words strongly changed by each model configuration.

5.3 "Underscoring" the publicized AI words in rewriting

The famous chatbot vocabulary is obviously enriched in rewriting, but not uniformly, and the differences are informative. "Meticulously" is at 214 times its human rate, "meticulous" at 30 times, "bolster" at 21, "underscores" at 19, "commendable" at 18, "paramount" at 16, "pivotal" at 15, and "intricate" at 15. These match the adjective-heavy signature Liang et al. (2024) found in their study on peer reviews.

The surprise is "delve". The single most famous AI word appears in rewrites at almost exactly its human rate, 0.033 per 10,000 tokens on both sides, and "delve into" as a phrase is similarly equally used. "Tapestry" is negligible on both sides, and "caveat" and "showcase" show no enrichment. The famous list describes free chat generation, where the model opens with "let us delve into" because it is composing from nothing. In rewriting, however, the model works with the human author's sentence as a reference, and its signature concentrates in substitutions and connectives instead. Public AI-word lists therefore transfer only partially to the rewriting case, which is one reason a task-specific lexicon (generation from scratch, and rewriting from a source text) is worth releasing.

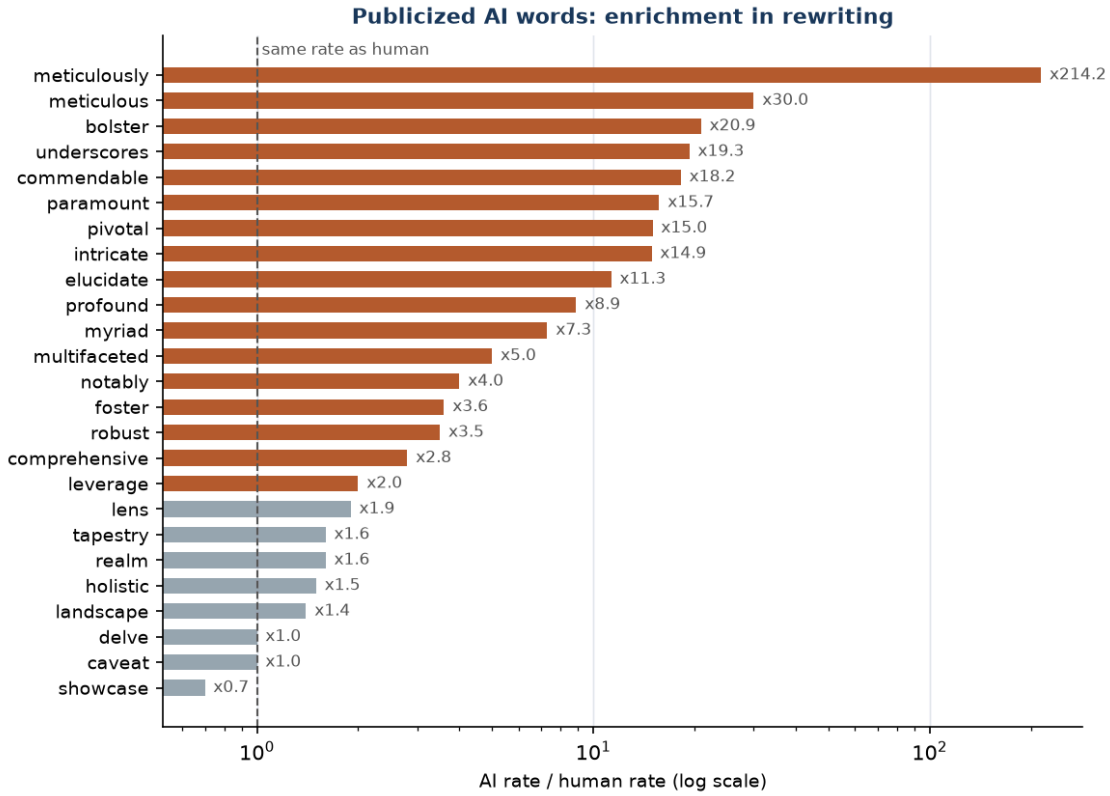

Figure F5. Enrichment of publicized AI words in rewriting. Grey bars are not meaningfully enriched.

5.4 Phrases

Curated phrases show the same two-level pattern. “Serves as a” is at 8 times its human rate with 2,842 occurrences, “stands as a” at 5 times, “plays a pivotal role” at 4 times, and “underscores the importance” and “delves into” at 3 times, while the chat-generation phrases “in today’s fast-paced world” and “rich tapestry” rarely occur in either side. The discovered sequences are more extreme than anything on the curated lists. “Delineates the” appears 1,285 times in rewrites and zero times in 60,786 human texts. “Thereby facilitating” appears 906 times against zero, “necessitated by” 959 times against once, and “ascertained that” 584 times against zero. Sequences of this type are close to deterministic markers of machine rewriting in this corpus, since the human baseline is empty.

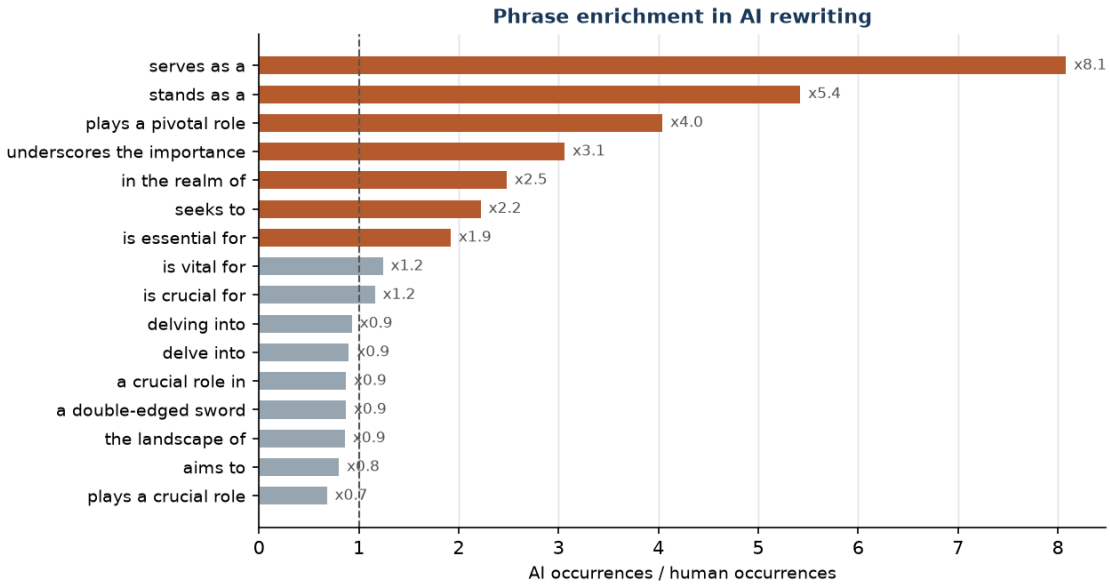

Figure F6. Phrase enrichment in AI rewriting.

5.5 Latinate vocabulary and word complexity

The character of the AI signature is consistent. Among the 100 most AI-preferred candidates, 44 percent are Latinate by our classification, the mean word length is 9.1 characters, and the mean syllable count is 3.3. Among the 100 most human-leaning candidates the Latinate share is 10 percent, the mean length is 5.5 characters, and the mean syllable count is 1.8. Models used through common online interfaces change academic writing toward longer, Latin-derived vocabulary, while the human authors of the original source texts relied on shorter, largely Germanic words. Since word length and syllable load drive readability formulas, this change alone makes rewritten text harder to read in general, independent of any change in content.

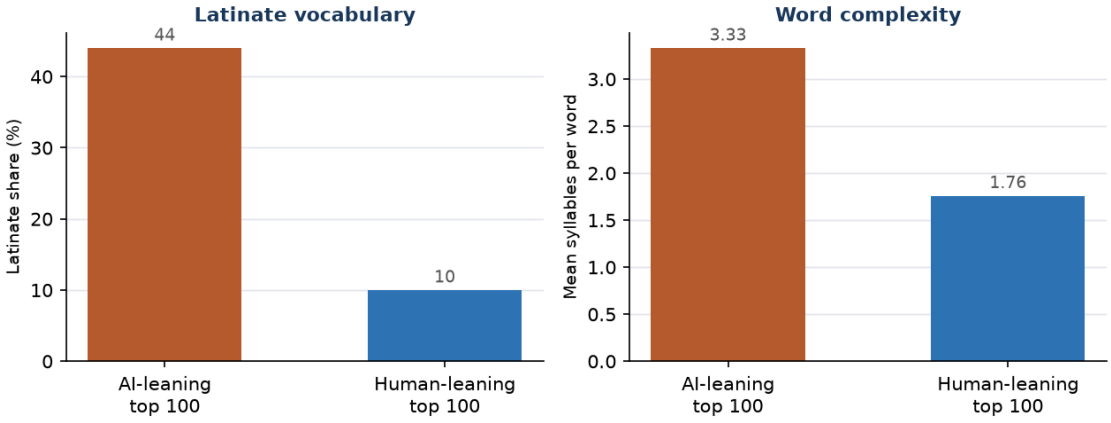

Figure F4. Word origin and complexity of the AI-leaning and human-leaning vocabularies.

5.6 The lexicon and the WordNet view

The released lexicon covers all 1,057 studied words. Each entry carries the word, its part of speech and tier, rates per 10,000 tokens on both sides, the log-odds ratio and z-score, an AI score from 0 to 100, WordNet synsets and lemmas, and a short gloss. By the score, 286 words lean AI (60 or above) and 213 lean human (40 or below). The extremes are intuitive, with “meticulously” at 99.5 and “besides” at 4.4.

Mapping the words onto WordNet synonym sets makes this substitution behavior more apparent. Within one synset, models and humans consistently choose different lemmas. In the set around “use”, the human choices “used” and “using” score 10 and 26 while “utilizing” and “employing” score 90 and 92. In the set around “show”, “shows” scores 7 while “demonstrated” scores 79 and “exhibited” 90. “Help” scores 11 while its synonym “facilitate” scores 81. The same divergence appears for “main” against “primary”, “basically” against “fundamentally”, and “always” against “invariably”. The signature is a systematic preference for one member of each synonym set, almost always the more formal and more Latinate member.

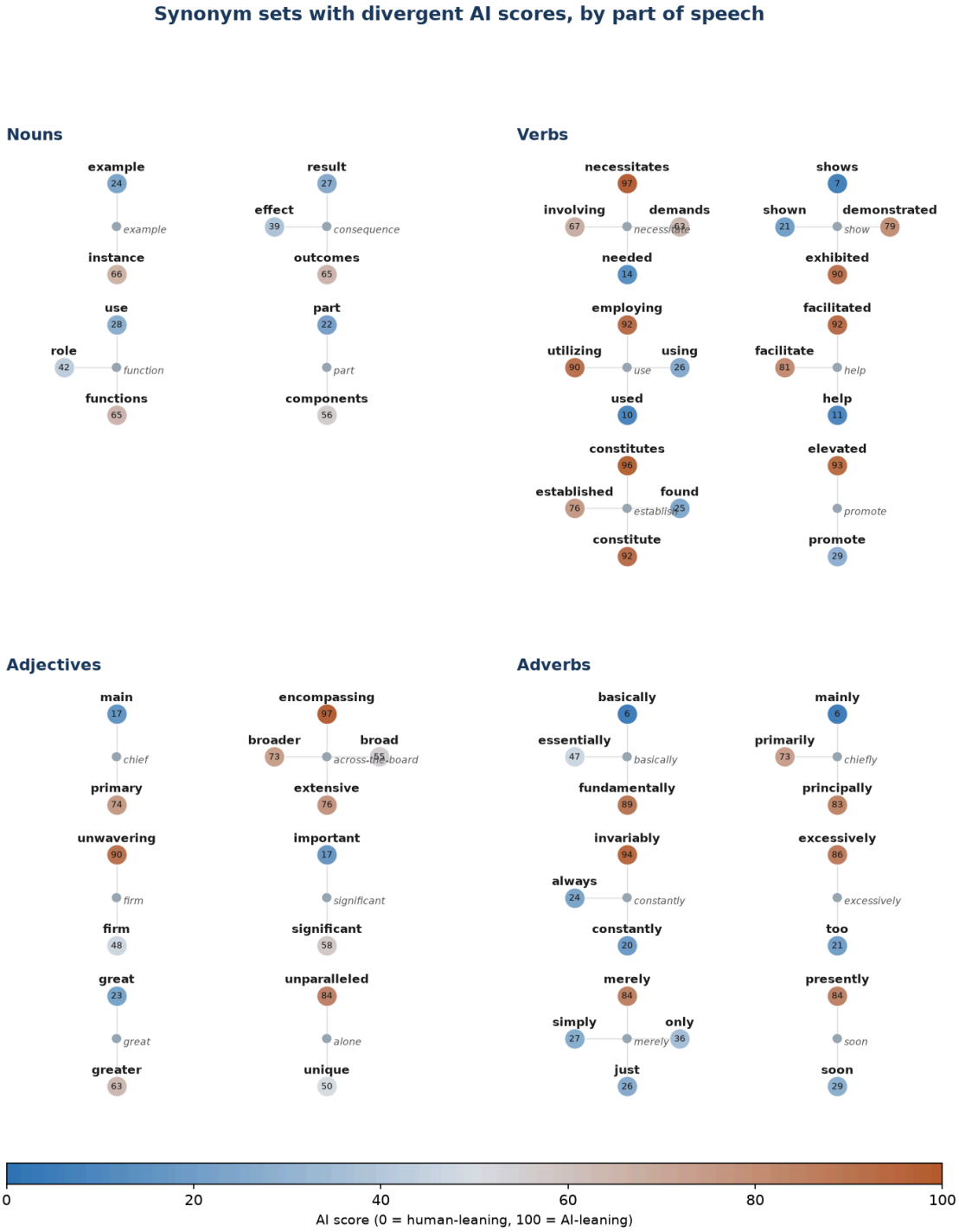

Figure F7. Synonym sets with divergent AI scores. Each word is colored by its corpus-derived AI score.

5.7 Scoring the entire English vocabulary

We attempted to score the entire English vocabulary as represented in WordNet: 90,520 single-word lemma entries across the four lexical categories of noun, verb, adjective and adverb. From these, 40,636 actually occur in the corpus under study. This observed set is the English vocabulary that the models and the human authors actually used across 60,786 academic texts and their AI

rewrites, a base that is large enough to generalize to academic rewriting. In it, 18,937 lemmas pass the 25-occurrence evidence threshold, and after the two-signal check, a final 3,371 lemmas show an observable AI-preferred difference and receive an 'AI score' above zero: 1,502 nouns, 838 adjectives, 779 verbs, and 252 adverbs. A further 1,347 lemmas lean significantly toward the human side and are labeled accordingly. Every remaining lemma is scored 0 by default. We attempted to score all words, but only these showed a difference that qualified. The remainder are considered to be used roughly equally between AI and human writing. Each word of the English vocabulary, represented by a node in the WordNet semantic network, was labeled with an AI score, and can be interactively viewed at https://textpulse.ai/research/ai-vocabulary-explorer.

The scored tail over the 1,057 studied words has a strong signal as well. The highest full-vocabulary scores go to low-frequency Latinate words: "erudition" (99.3), "sundry" (99.2), "approbation" (98.8), "exigency" (98.4), "stratagem" (98.1), "pecuniary" (98.0), and "salubrious" (97.9) are all almost absent from the human-written sources but recur in AI rewrites. The substitution signature of Section 5.5 therefore extends deep into rare vocabulary. Models do not only prioritize the elevated member of common synonym pairs, they re-introduce vocabulary that academic authors have generally 'retired'.

Figure F8 shows one WordNet synset cluster per part of speech, with synsets, their member lemmas, and the semantic relations between them, colored by corpus lean and labeled with a corpus statistic-derived AI score. The pattern of Section 5.6 repeats at each level of the semantic network. In the "use" synset, "utilize" (76) and "employ" (75) against plain "use"; in the enterprise neighborhood, "endeavor" (89) against "task"; among the satellites of "crucial", "pivotal" (87); in the adverbs, "fundamentally" (78) against "basically".

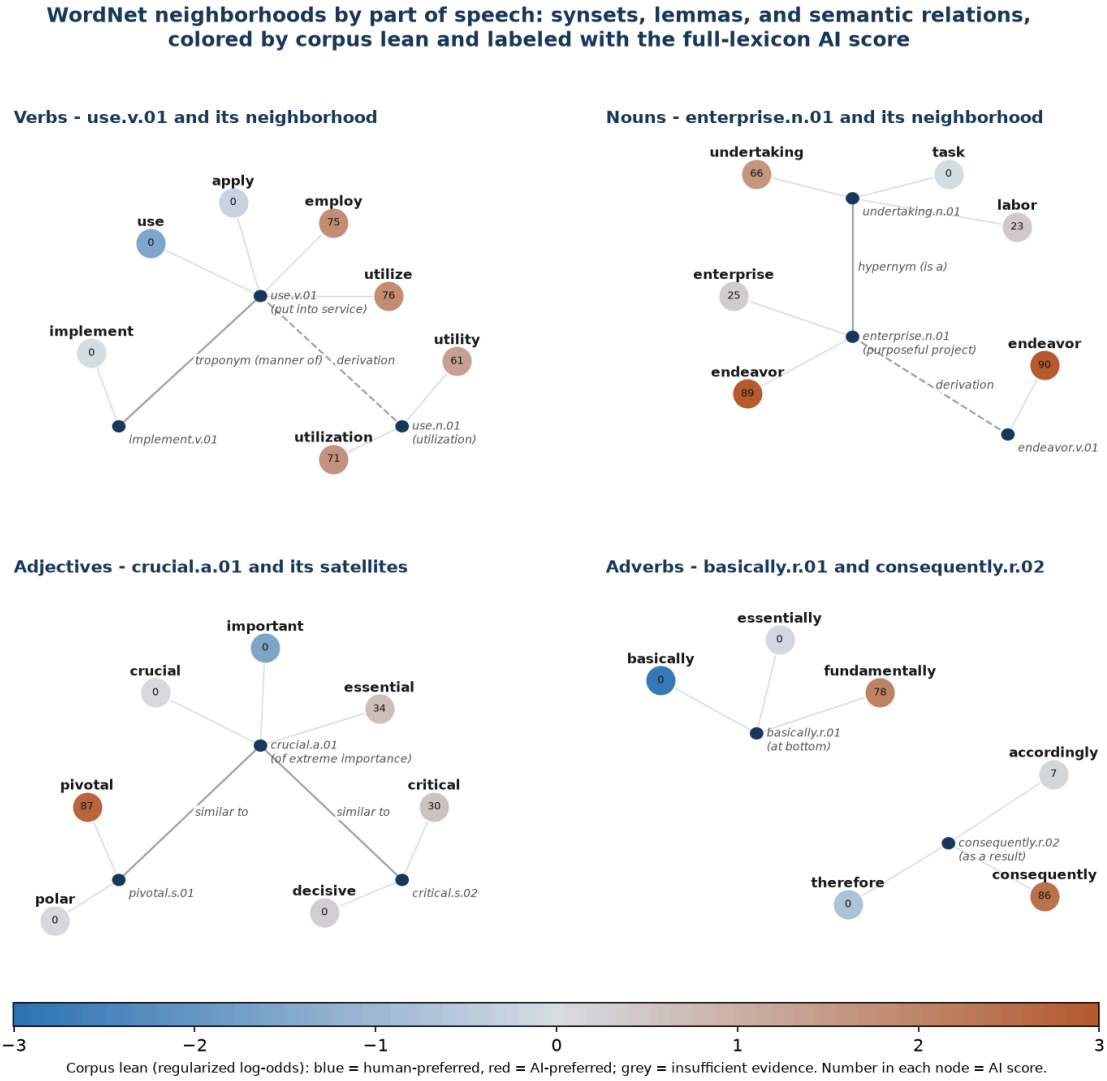

Figure F8. WordNet neighborhoods by part of speech: synsets, lemmas, and semantic relations, colored by corpus lean and labeled with the full-vocabulary AI score.

6. Discussion

For writers and editors, the practical lists need an update. The words that expose machine involvement in AI-rewritten academic text are not primarily the famous chat words. They are ordinary formal substitutions, “thereby”, “consequently”, “employed”, “utilized”, “constitutes”, and the near-deterministic sequences built from them. A document that was rewritten by an AI tool with every “used” turned into “utilized” carries a stronger AI signature than one containing a single “delve”. The publicly released lexicon supports this use directly, since each word carries a score based on (human and AI) corpus statistics rather than a place on an informal list based on popularity.

For detection and its risks, vocabulary changes of this size are part of the reason that perplexity-based detectors work at all, and the discovered zero-baseline sequences show how strong the lexical

evidence can be. But the same evidence warns against reading any single word as proof. “Delve” is not enriched in rewriting, several publicized words are barely enriched, and human authors do use most of this vocabulary at some rate. Word-level signals support probabilistic judgment, and not accusation, a distinction our detector agreement study develops from the score side (TextPulse Research, 2026a).

For AI model builders, the cross-model distribution shows that the AI signature is not inevitable. On identical inputs, Qwen changed only 76 of 1,057 common words strongly while Gemini Flash changed 422. Vocabulary restraint is achievable and varies by vendor, but no benchmark reports it. The contrast with citation fidelity in the same corpus highlights this point, since the most reference-faithful model is among the least voice-faithful. Tools that advertise rewriting should be measured on both citation preservation and also rate of lexical (vocabulary) change.

This study has some limitations. It measures rewriting of English academic text, and the signature of free generation from scratch or other genres will differ, as the “delve” result shows. The two corpora used different one-shot prompts, so cross-model comparison is confined to corpus 2. Dominant-POS assignment and the Latinate heuristic are documented simplifications for the sake of the study. The human baseline is published academic writing, which is itself edited text, so the human rates describe a formal writing style rather than a casual/conversational one. Finally, the configurations are those captured in this corpus, and vendors update models continuously.

7. Conclusion

When language models rewrite academic text, they prioritize a specific vocabulary. The signature is measurable in every model tested, is dominated by formal connectives and Latinate synonym substitutions, reaches several hundred fold for individual words, and includes word sequences that do not occur in the human baseline at all. It differs from the publicized chat vocabulary online in ways that matter for anyone using word lists to judge the origin of a text as human or AI. Also, it varies enough across models that vocabulary restraint should be treated as a measurable property of rewriting tools, along with the citation fidelity we measured in prior work. The full lexicon, with per-word corpus evidence and AI scores, is made publicly available for future work.

Data availability

The studied lexicon (CSV and JSON, 1,057 entries with per-word statistics, WordNet mappings, and centered AI scores where 50 means equal use), the full-vocabulary lexicon (CSV and JSON, 90,520 lemma and part-of-speech entries with two-signal statistics, confidence intervals, evidence status, and one-sided AI scores where 0 means no observable AI preference), the candidate score tables, per-model tables, document frequencies, phrase and discovered-sequence tables, and all analysis and figure code are openly available on Zenodo at https://doi.org/10.5281/zenodo.22028377. Both lexicons can also be browsed interactively and

downloaded at the TextPulse vocabulary explorer at https://textpulse.ai/research/ai-vocabulary-explorer.

References

American Dialect Society. (2026). 2025 word of the year is “slop”. American Dialect Society.

Fellbaum, C. (Ed.). (1998). WordNet: An electronic lexical database. MIT Press.

GPTZero. (2024). AI vocabulary: The words and phrases AI uses most. https://gptzero.me/ai-vocabulary

Juzek, T. S. (2025). Word overuse and alignment in large language models: The influence of learning from human feedback. arXiv:2508.01930.

Juzek, T. S. (2026). AI-associated lexical shifts across 34 languages: Cross-lingual convergence and diachronic uptake in news writing. arXiv:2605.25358.

Juzek, T. S., & Ward, Z. B. (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 (COLING 2025), 6397-6411. https://aclanthology.org/2025.coling-main.426/

Kobak, D., González-Márquez, R., Horvát, E.-Á., & Lause, J. (2025). Delving into LLM-assisted writing in biomedical publications through excess vocabulary. Science Advances, 11(27), eadt3813. https://doi.org/10.1126/sciadv.adt3813

Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Liu, S., Huang, Z., McFarland, D. A., & Zou, J. Y. (2024). Monitoring AI-modified content at scale: A case study on the impact of ChatGPT on AI conference peer reviews. In Proceedings of the 41st International Conference on Machine Learning (ICML 2024), PMLR 235. arXiv:2403.07183.

Merriam-Webster. (2025). Word of the year 2025: Slop. Merriam-Webster.

Miller, G. A. (1995). WordNet: A lexical database for English. Communications of the ACM, 38(11), 39-41. https://doi.org/10.1145/219717.219748

Monroe, B. L., Colaresi, M. P., & Quinn, K. M. (2008). Fightin’ words: Lexical feature selection and evaluation for identifying the content of political conflict. Political Analysis, 16(4), 372-403. https://doi.org/10.1093/pan/mpn018

TextPulse Research. (2026a). Do AI detectors agree? An inter-rater reliability study of commercial AI text detectors on academic writing. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22003419

TextPulse Research. (2026b). Do AI models invent references? A verification audit of citations in AI-generated academic text. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22010511

TextPulse Research. (2026c). What happens to citations when AI rewrites academic text? A large-scale paired audit. TextPulse Working Paper. https://doi.org/10.5281/zenodo.22010530