

Lost in Sampling: Assessing Lexical Reachability in LLMs via the Word Coverage Score (WCS)

Samer Awad¹, Javier Conde¹, Carlos Arriaga¹, Tairan Fu²,
Javier Coronado-Blázquez³, Pedro Reviriego¹

¹IPTC, Universidad Politécnica de Madrid
Madrid, Spain

²Politecnico di Milano
Milano, Italy

³Banco de España
Madrid, Spain

Abstract

Modern Large Language Models (LLMs) are often criticized for producing repetitive and homogeneous text, despite possessing vast latent vocabularies. While previous research has focused on model knowledge and training data, we investigate the role of decoding mechanics in suppressing linguistic diversity. It is well-known that sampling filters cut the tail of the token probabilities, but the specific impact at the word level remains unexplored. In this paper, we introduce the Word Coverage Score (WCS), a metric that quantifies the extent to which contextually appropriate human vocabulary is mathematically pruned by standard sampling filters (e.g., Top- p , Top- k , and Min- p). Rather than assessing static knowledge, the WCS measures the lexical survival rate of words as a function of sampling parameters through a deterministic Forced-Path Audit. By auditing 13 open-weight models on 10,000 evaluation samples per model from the modern FineWeb corpus at four temperatures, we identify which human lexical choices are rendered unreachable by the decoder, even when they reside within the underlying probability space. To characterize word usage reduction without relying only on absolute erasure, we additionally introduce the Low-Coverage Word Rate (LCWR), the fraction of target words reachable in fewer than a specified proportion of contexts. In matched downstream generation experiments, greater WCS is associated with greater output-text lexical diversity: repeated-measures Spearman correlations are $\rho = .67$ for Type-Token Ratio (TTR), 95% Confidence Interval $[.50, .82]$ and $\rho = .66$ for Measure of Textual Lexical Diversity (MTLD), 95% Confidence Interval $[.51, .81]$. These results provide quantitative evidence regarding how inference configurations shape vocabulary compression, establishing WCS as a computationally efficient diagnostic tool for hyperparameter debugging and vocabulary calibration.

1 Introduction

Open-ended LLM generations can exhibit both intra-model repetition and inter-model homogeneity, with different models producing similar responses to prompts that admit many plausible answers (Jiang et al., 2025). On standardized creativity tasks, LLM responses are also substantially more similar to one another than human responses are (Wenger and Kenett, 2026). Related human-subject experiments find that access to generative AI can improve the average quality of individually written stories while reducing the diversity of the stories produced across participants (Doshi and Hauser, 2024). These studies document output homogenization, but they do not by themselves establish training-time mode collapse or identify a single cause. Evidence about post-training is more specific. Preference-tuned models, especially those tuned with reinforcement learning, can show lower lexical and syntactic diversity even while producing more high-quality outputs and therefore greater effective semantic diversity (Shypula et al., 2025). Structural tokens in chat templates can independently reduce semantic output diversity (Yun et al., 2025), and newer instruction-tuned models have been observed to produce less lexically and syntactically diverse news text than older non-instruction-tuned models (Gude et al., 2026). We therefore treat post-training as one possible contributor to output homogenization rather than assuming that every alignment stage has the same effect.

A relevant aspect of this homogeneity is how LLMs use vocabulary. In a controlled comparison of human-authored English news with matched output from six LLMs, the human texts had the highest standardized type-token ratio and MTLD; the size of the gap nevertheless varied by model family, and the largest LLaMA model was an exception to the general restricted-vocabulary pattern (Muñoz-Ortiz et al., 2024). A separate evaluation of ChatGPT

found that lexical diversity varied with the model version, task, presence penalty, and assigned role; within the range that produced valid text, temperature had only a small effect (Martínez et al., 2025). These output-level findings motivate a more direct test of whether sampling filters exclude contextually appropriate mid-to-low-frequency words. Human word frequencies are often approximated by Zipf’s law (Zipf, 1949). A recent preprint reports that a two-parameter Mandelbrot distribution fit LLM token rank frequencies better than Zipf in 35 of 36 evaluated model–domain combinations (Bogdan and de Valois-Franklin, 2026); however, that result does not show that LLM tails are steeper than human tails or identify decoding as its cause. Our forced-path audit instead measures the contribution of inference-time filtering directly.

Previous research has focused on model knowledge and training data; this work instead examines the impact of decoding mechanics on vocabulary access. Building upon the well-established reality that token samplers truncate the tail of probability distributions, we introduce the Word Coverage Score (WCS), a diagnostic metric designed to explicitly quantify how these token-level filters map to the exclusion of contextually appropriate human words. We utilize this framework to evaluate 13 open-weight LLMs on the same 10,000 FineWeb word-context samples per model, including matched Base/Instruct comparisons that hold the tokenizer constant within each family. The results show that standard sampling configurations impose significant reachability constraints on standard human vocabulary.

2 Token Sampling and LLM Homogeneity

For each token prediction, LLMs compute the logits for all tokens in their vocabulary, producing tens of thousands of potential candidates. Even if the majority of these candidates possess a negligible probability, the cumulative number of potential trajectories is vast. This configuration creates a literal "Garden of Forking Paths" mirroring Jorge Luis Borges’ famous short story, *"El jardín de senderos que se bifurcan"* (Borges, 1998). In this narrative, Borges describes an infinite, multi-dimensional labyrinth where every contextually appropriate future step is kept alive simultaneously, allowing parallel timelines to branch out and co-exist. Within an LLM, the presence of sophisticated vo-

cabulary in the latent probability space represents a web of these parallel, co-existing forks. However, the generation of a singular textual sequence requires the decoding system to collapse this multi-dimensional space into a linear path, effectively reverting to what Borges categorizes as traditional fiction where a character selects one alternative and eliminates all the rest.

The sampling filter acts as a truncation mechanism that eliminates these winding branches. Standard sampling filters such as Top- p , Top- k , and Min- p remove low-probability candidates and restrict final token selection to the surviving set. Nucleus sampling was introduced in response to two observed failure modes: repetitive text under maximization-based decoding and incoherent text when sampling indiscriminately from an unreliable probability tail (Holtzman et al., 2020). Truncation can address those failures, but by construction it also makes every excluded token unreachable. Our experiments quantify how often that token-level exclusion blocks contextually appropriate words.

This truncation exposes a fundamental dilemma inherent to the autoregressive generation paradigm: a practical trade-off between local text coherence and global linguistic diversity. To reduce the risk of low-quality or less coherent generation when sampling from unconstrained, long-tail distributions, modern token sampling filters explicitly prioritize high-probability predictability at the cost of reducing lexical diversity. Because current architectures rely heavily on token-by-token selection from bounded distributions to maintain sequential continuity, they may struggle to simultaneously protect the sophisticated nuances of human vocabulary. Consequently, this structural bottleneck imposes a predictable upper bound on the accessible vocabulary and generative diversity of modern LLMs.

3 Methodology

The Word Coverage Score (WCS) quantifies the differences between a model’s latent lexical knowledge and its generative accessibility under specific decoding constraints. Rather than evaluating an LLM’s static capability to comprehend a word in isolation, the WCS operates as a dynamic behavioral metric that measures whether a contextually appropriate, human-authored vocabulary choice remains mathematically reachable during sequential autoregressive generation. By formalizing reachability as a product of step-wise token survival, the

WCS provides a rigorous framework to quantify how token sampling configurations impose reachability constraints on the accessible vocabulary space. The following subsections describe each of the elements of the proposed WCS which are illustrated in Figure 1.

3.1 The Forced-Path Audit

To evaluate the WCS, we employ a *Forced-Path Audit*. Given a human-authored reference sequence S from an evaluation corpus $\mathcal{X}$, we identify a target lexical unit or word w composed of n sub-word tokens $(t_1, t_2, \dots, t_n)$. We provide the model with the ground-truth prefix context C and force a deterministic traversal of the path w .

At each transition i , we extract the full probability distribution $P(\cdot \mid C, t_1, \dots, t_{i-1})$. Rather than sampling, we record the rank and scalar probability of the ground-truth token t_i . This allows us to determine if t_i would have survived the truncation of a given sampling algorithm.

3.2 Lexical Survival Functions

We define Reachability ($\mathcal{R}$) as a binary indicator of whether a token remains in the "active" vocabulary set after a sampling filter is applied. A multi-token word w is considered covered ($\mathcal{R} = 1$) if, and only if, every constituent token t_i survives the filter at its respective step, i.e. the token is among the candidates selected for sampling. In our evaluation we focus on established sampling techniques such as Top- k and Top- p , and also include Min- p , a dynamic truncation method proposed to improve the quality-diversity balance, especially at higher temperatures (Minh et al., 2025).

3.3 The Word Coverage Score (WCS)

The Word Coverage Score is the aggregate mean reachability across a target set of words W , each with a set of contexts C_w under a specific parameter configuration $\theta \in \{k, p, m\}$.

$$WCS(\theta) = \frac{1}{\sum_{w \in W} |C_w|} \sum_{w \in W} \sum_{c \in C_w} \mathcal{R}_{\theta}(w, c) \quad (1)$$

By calculating $WCS(\theta)$ across a continuous range of values, we generate a lexical decay curve. This function allows for the identification of how the sampler parameters impact the potential lexical richness.

The use of several contexts per word enables us to also do a per-word calculation of the WCS as follows:

$$WCS(\theta, w) = \frac{1}{|C_w|} \sum_{c \in C_w} \mathcal{R}_{\theta}(w, c) \quad (2)$$

which is of interest to analyze the reachability per word. For example, when a word is not reachable across any evaluated context, $WCS(\theta, w) = 0$, it indicates that the configuration has rendered that lexical entry operationally inaccessible.

Absolute operational inaccessibility, $WCS(\theta, w) = 0$, becomes less discriminative as the number of evaluated contexts grows because a word needs to survive only once to avoid being counted as inaccessible. We therefore define the Low-Coverage Word Rate at threshold τ :

$$LCWR_{\tau}(\theta) = \frac{1}{|W|} \sum_{w \in W} \mathbb{I}[WCS(\theta, w) < \tau]. \quad (3)$$

LCWR measures severe rather than absolute inaccessibility; higher values indicate that more target words are reachable in only a small fraction of contexts. We use $\tau = 0.10$ as the primary word-level threshold. With 50 contexts per word, $WCS(\theta, w) < 0.10$ means that the word survives in at most four contexts. We retain context-level WCS as the primary continuous aggregate and absolute erasure as an extreme endpoint.

3.4 Frequency-Based Lexical Selection

We use a band-limited random sampling protocol focusing on words that are neither common nor extremely rare. We establish the candidate lexical set using the Google Web Trillion Word Corpus as compiled by (Norvig, 2009). From this list, we retain lowercase entries of at least three characters ranked between 10,000 and 40,000 in total frequency. The eligible entries are shuffled with seed 13. We inspect a candidate pool four times larger than the requested sample and retain the first 200 words for which 50 coherence-checked FineWeb contexts can be collected.

3.5 Context Selection

To evaluate lexical reachability in a contemporary web context, we utilize the `sample-10BT` configuration of the FineWeb dataset (Penedo et al., 2024). FineWeb provides a large, filtered collection of contemporary web text spanning multiple registers.

Each word in W is then mapped back to N_c naturalistic contexts within the *FineWeb* corpus for

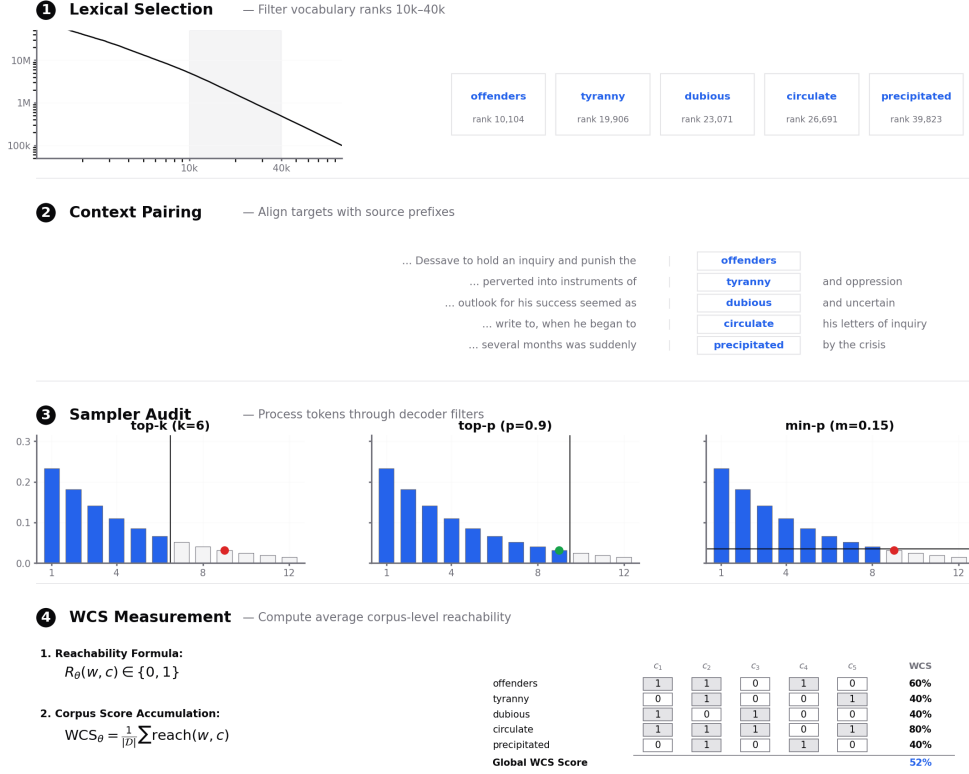


Figure 1: Four-stage methodology for calculating the Word Coverage Score (WCS): select frequency-bounded target words, pair them with human-written contexts, audit token survival under sampling filters, and aggregate reachability into WCS.

evaluation, ensuring that the WCS measures the reachability of legitimate human expressions. The following procedure is used to generate each of the N_c contexts:

- Seeded streaming collection:** We stream the FineWeb training split with a shuffle buffer of 10,000 documents and seed 13. For each candidate word, we retain up to 80 distinct lowercase occurrences with sufficient preceding context.
- Contextual extraction:** For every occurrence, we retain the preceding passage and use the final $L = 256$ model tokens as the audit prefix C .
- Context verification:** To ensure semantic integrity, each candidate context is evaluated for coherence using the gemini-2.5-flash-lite model (Google DeepMind, 2025a) in a zero-shot binary classification setup. Contexts classified as non-coherent, containing artifacts such as table-of-contents or index fragments, or lacking sufficient linguistic information are discarded. The first 50 accepted contexts are

retained for each word; candidates with fewer than 50 accepted contexts are skipped.

The seeded streaming and shuffle-buffer procedure makes the collection reproducible while sampling occurrences from different locations in the streamed FineWeb documents.

3.6 Continuous Metrics and Future Outlook

The binary formulation of the WCS provides an interpretable, behavioral diagnostic of absolute word exclusion; however, it discretizes the underlying probability space. To capture continuous probability distribution changes across the generation path, we define a continuous, probability-weighted variant of the metric (WCS_{prob}):

$$WCS_{\text{prob}}(\theta) = \frac{1}{\sum_{w \in W} |C_w|} \sum_{w \in W} \sum_{c \in C_w} \prod_{i=1}^n (P(t_i | c, t_1, \dots, t_{i-1}) \cdot \mathbb{I}(t_i \in \mathcal{V}_\theta)) \quad (4)$$

where $\mathcal{V}_\theta$ denotes the truncated candidate token set preserved by the sampler under configuration θ , t_i represents the constituent sub-word tokens

of target word w , and $\mathbb{I}$ is an indicator function tracking token survival.

This continuous formulation offers a metric capable of tracking shifts in probability mass allocation that do not trigger a hard binary state change. In the rest of the paper, we focus on absolute word reachability and leave the empirical characterization of the continuous WCS_{prob} decay curves and their direct correlation with human-perceived lexical diversity metrics for future work.

4 Experiments

The experimental framework is designed to empirically trace how a word navigates from a model’s latent internal distribution to its final generation sequence. To achieve this, our methodology is segmented into three logical phases: first, we select a comprehensive suite of base and chat-oriented (commonly denoted as ‘-it’) models; second, we formalize a rigorous tracking protocol that audits the step-by-step token survivability of target human vocabulary; and third, we systematically evaluate word reachability under common truncation thresholds and temperature scales.

4.1 Model Selection

To isolate the impact of alignment and distillation on lexical diversity, we evaluate both the Base (raw) and Chat-oriented (commonly denoted as ‘-it’ or ‘-Instruct’) variants of selected computationally tractable open-weight architectures. This selection shown in Table 1 ensures a broad cross-section of the current LLM landscape while maintaining computational accessibility for independent researchers. For each architecture, the Base version represents the model’s fundamental linguistic distribution, while the Chat-oriented version represents the distribution post-instruction tuning and preference alignment (e.g., via Supervised Fine-Tuning (SFT) and Reinforcement Learning from Human Feedback (RLHF)). For models like DeepSeek-R1-Distill-Qwen-14B, which do not have an equivalent base, we use the corresponding Qwen-14B-Base as the reference point.

4.2 Audit Protocol

For each model, we evaluate the same $N_w = 200$ words and $N_c = 50$ contexts per word. We define a trial as a "success" (Reachability $\mathcal{R} = 1$) only if the complete multi-token sequence of the target word w remains within the valid sampling set $\mathcal{V}_\theta$ for a given parameter θ .

4.3 Sampling Sweep

We systematically sweep the sampling decoders to identify the impact on word coverage:

1. **Nucleus (p):** 0.50–0.95 in increments of 0.05, plus 0.99.
2. **Top- k :** integers from 1 through 20, followed by increments of 5 from 25 through 80.
3. **Min- p :** 0.01–0.10 in increments of 0.01.

Additionally, we conduct experiments at four temperatures, $T = 0.6, 0.7, 1.0, 1.5$, spanning lower-temperature conditions, the unscaled $T = 1.0$ distribution, and an aggressive high-temperature condition. Prior work on narrative generation finds that increasing temperature is only weakly associated with novelty and is also correlated with reduced coherence, suggesting a trade-off rather than a simple creativity control (Peepkorn et al., 2024).

5 Results

The code and results are available at <https://github.com/WordsGPT/WCS>; an interactive companion visualiser is available at <https://wordsgpt.github.io/WCS/>. The visualiser allows the WCS curves to be inspected across models, samplers, parameters, temperatures, and aggregation levels. We present our results by plotting WCS and LCWR against the sampler parameters. At the word level, we focus on $LCWR_{0.10}$, the fraction of words reachable in fewer than five of their $N_c = 50$ contexts. This distinguishes persistent low coverage from the more extreme condition $WCS(\theta, w) = 0$, which means that the word is unreachable in all fifty evaluated contexts under that configuration.

It is informative to examine reachability at the individual word-context level. Figure 2 shows the fraction of all 10,000 word-context pairs that remain reachable under each sampling algorithm at $T = 1.0$, where temperature leaves the model log-its unscaled. Substantial word-context reachability constraints remain under this neutral-temperature comparison. To summarize the low-coverage tail of the per-word distribution while retaining absolute operational inaccessibility as its extreme endpoint, subsequent analyses use LCWR.

The results for Top- p sampling with a temperature of $T = 0.7$ are shown in Figure 3. At $p = 0.95$, mean context-level WCS across the 13 models is

Family	Base Variant	Aligned Variant
Llama	Llama-3.1-8B (Meta AI, 2024a)	Llama-3.1-8B-Instruct (Meta AI, 2024b)
Mistral	Mistral-7B-v0.3 (Mistral AI, 2024b)	Mistral-7B-Instruct-v0.3 (Mistral AI, 2024a)
Qwen (Small)	Qwen3.5-9B-Base (Qwen Team, 2026b)	Qwen3.5-9B (Qwen Team, 2026a)
Qwen (Mid)	Qwen2.5-14B (Qwen Team, 2024a)	Qwen2.5-14B-Instruct (Qwen Team, 2024b)
Gemma-3	Gemma-3-12B-pt (Google DeepMind, 2025c)	Gemma-3-12B-it (Google DeepMind, 2025b)
Gemma-4	Gemma-4-E4B (Google DeepMind, 2026a)	Gemma-4-E4B-it (Google DeepMind, 2026b)
DeepSeek	Qwen2.5-14B (Qwen Team, 2024a)	DeepSeek-R1-Distill-Qwen-14B (DeepSeek-AI, 2025)

Table 1: Selected model pairs for evaluating the Word Coverage Score (WCS).

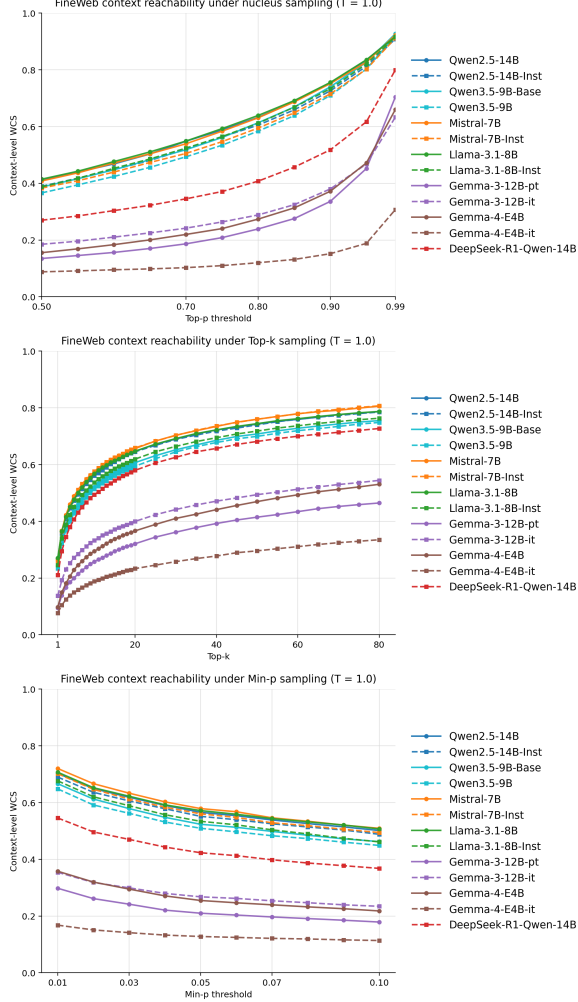


Figure 2: Fraction of the 10,000 FineWeb word-context pairs (200 words, each with 50 contexts) that are reachable, $WCS(\theta)$, when using Top- p (top), Top- k (middle), and Min- p (bottom) sampling at $T = 1.0$. Instruct models use dashed lines and their Base counterparts use solid lines.

46.1%, so 53.9% of the 10,000 human-attested word-context paths per model are excluded. Although 98.9% of targets remain reachable in at least one of their 50 contexts, $LCWR_{0.10} = 8.3\%$, so only 215 of the 2,600 model/word combinations are accessible in fewer than five contexts. At $p = 0.99$, mean context-level WCS rises to 59.0%, word-level reachability to 99.5%, and LCWR decreases to 4.4%. When comparing Base models (solid lines) and their chat-oriented (Instruct/it) counterparts (dashed lines), five of the six matched families show lower context-level reachability after

post-training, with Gemma-3-12B as the exception.

As a threshold-sensitivity check at $T = 0.7$ and $p = 0.95$, the aggregate low-coverage rate is 4.8% at $\tau = 0.05$, 8.3% at $\tau = 0.10$, and 18.8% at $\tau = 0.20$. The expected change in magnitude motivates reporting the threshold explicitly; the primary analysis uses $\tau = 0.10$ throughout.

Across models, the Gemma family has non-uniform patterns between generations. At $T = 0.7$ and $p = 0.95$, Gemma-3 increases from a context-level WCS of 21.7% for the Base checkpoint to 27.5% for the it checkpoint, while LCWR decreases from 23.0% to 11.5%. Gemma-4 moves in the opposite direction: context-level WCS decreases from 26.2% to 12.1%, while LCWR increases from 13.0% to 46.0%. Qwen, Llama, and Mistral show smaller Base-to-Instruct reductions at the same setting. DeepSeek-R1-Distill-Qwen-14B obtains a WCS of 41.0% and LCWR of 5.0%, compared with 60.8% and 1.0% for Qwen2.5-14B Base.

Averaging across the six matched Base/Instruct pairs over all FineWeb temperatures, samplers, and parameter settings, mean context-level WCS decreases from 49.9% for Base models to 47.7% for Instruct models. This aggregate trend is not universal, with Gemma-3-12B-it improving over its Base counterpart.

Temperature has a strong but sampler-dependent effect. At Top- $p = 0.95$, mean WCS increases from 40.2% at $T = 0.6$ to 46.1% at $T = 0.7$, 67.4% at $T = 1.0$, and 92.1% at $T = 1.5$; LCWR correspondingly decreases from 11.1% to 8.3%, 2.1%, and less than 0.1%. At Min- $p = 0.05$, mean WCS values are 34.3%, 36.6%, 43.7%, and 54.8%, while LCWR values are 14.0%, 12.8%, 9.0%, and 5.2%. Pure Top- $k = 20$ remains effectively constant at a WCS of 53.7% and an LCWR of approximately 3.6% across temperatures, as expected because temperature rescaling preserves token rank.

Table 2 reports word-level reachability, LCWR, and WCS under documented default or recommended decoding settings for representative evaluated models. These values complement the one-dimensional sampler sweeps: Figure 3 isolates the

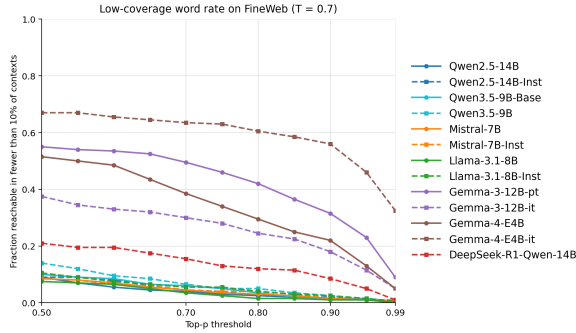


Figure 3: Low-Coverage Word Rate, $LCWR_{0.10}$: fraction of the 200 FineWeb target words reachable in fewer than five of 50 contexts under Nucleus Sampling (p) at $T = 0.7$. Instruct models use dashed lines and their Base counterparts use solid lines.

effect of varying Top- p , whereas practical decoding configurations may combine Top- p , Top- k , and temperature. Absolute erasure ranges from 0.5% to 4.0% under these settings, but LCWR ranges from 4.0% to 30.0%, revealing substantially more persistent unreachability. Context-level WCS ranges from 16.3% to 43.4%, corresponding to 1,627–4,340 reachable word-context pairs out of 10,000. The lowest result is for Gemma-4-E4B-it: 96.0% of target words are reachable at least once, yet 30.0% survive in fewer than five contexts and only 16.3% of all evaluated word-context paths remain reachable.

The results for Top- k and Min- p are shown in Figures 4,5. Word coverage can be seen to improve for most models, but there are still models that lose a significant fraction of word-context paths even under aggressive settings. The Gemma family shows again a different pattern across generations: Gemma-3-12B-pt and Gemma-4-E4B behave broadly like the other Base models, while Gemma-4-E4B-it retains substantially lower context-level reachability than its Base counterpart. This gap is also clear under Gemma-4’s default settings, Top- $p = 0.95$, Top- $k = 64$, and $T = 1.0$, as shown in Table 2. Gemma-4-E4B has a WCS of 36.4% and LCWR of 6.5%, whereas Gemma-4-E4B-it falls to a WCS of 16.3% and rises to an LCWR of 30.0%. Because these checkpoints use the same tokenizer, this within-family difference is not attributable to tokenizer segmentation.

5.1 Tokenizer-Stratified Reachability

Figure 6 stratifies FineWeb WCS by the number of subword tokens in each target word. Because tokenizers assign different words to each token-count stratum, comparisons should be made within

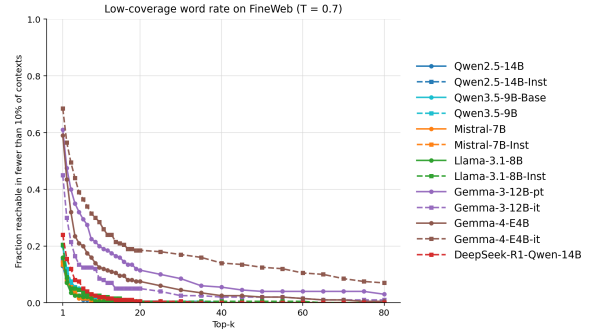


Figure 4: $LCWR_{0.10}$: fraction of the 200 FineWeb target words reachable in fewer than five of 50 contexts under Top- k Sampling at $T = 0.7$. Instruct models use dashed lines and their Base counterparts use solid lines.

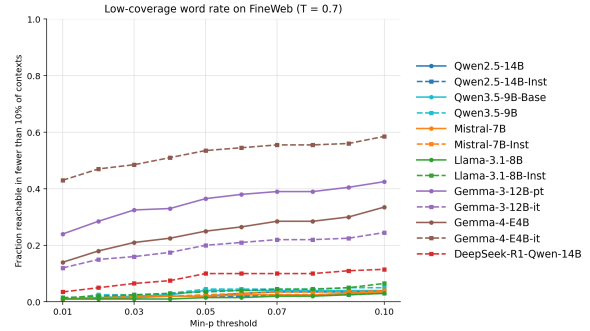


Figure 5: $LCWR_{0.10}$: fraction of the 200 FineWeb target words reachable in fewer than five of 50 contexts under Min- p Sampling at $T = 0.7$. Instruct models use dashed lines and their Base counterparts use solid lines.

model panels rather than by pooling strata across models. Token count affects WCS, but its direction is not universal: reachability decreases with token count for Gemma, remains approximately stable for Llama, and increases from the one-token stratum for Mistral. In any case, it should be noted that the number of words for some data points is low and further experiments would be needed to validate these observations.

6 Discussion

This section discusses the potential applications of the WCS as a diagnostic tool for auditing decoding configurations, as well as the implications of our findings for LLM text generation.

6.1 Interpretation of the Word Coverage Score and Potential Applications

When analyzing our experimental findings, it is important to discuss the information captured by the Word Coverage Score (WCS). The WCS fundamentally operates as a metric of binary exclusion: it isolates an extreme manifestation of word reachability limitations by detecting instances where a target word becomes mathematically impossible

Model / family	Default setting	Words reachable	$LCWR_{0.10}$	WCS
Qwen2.5-14B-Instruct ¹	$p = 0.8, k = 20, T = 0.7$	99.5%	4.0%	40.1%
Qwen3.5-9B ²	$p = 0.8, k = 20, T = 0.7$	99.0%	7.0%	36.7%
Llama-3.1-8B-Instruct ³	$p = 0.9, T = 0.6$	99.5%	4.0%	43.4%
Gemma-4-E4B ⁴	$p = 0.95, k = 64, T = 1.0$	99.5%	6.5%	36.4%
Gemma-4-E4B-it	$p = 0.95, k = 64, T = 1.0$	96.0%	30.0%	16.3%
DeepSeek-R1-Distill-Qwen-14B ⁵	$p = 0.95, T = 0.6$	98.5%	8.5%	36.4%

Table 2: FineWeb reachability under documented default or recommended decoding settings. Words reachable is the fraction of 200 target words reachable in at least one of 50 contexts. $LCWR_{0.10}$ is the fraction reachable in fewer than five contexts. WCS is the aggregate over all 10,000 evaluated word-context pairs. Rows with both Top- p and Top- k use the sequential combined filter. Mistral-7B-Instruct-v0.3 is omitted because its model card does not specify a single default Top- p value.

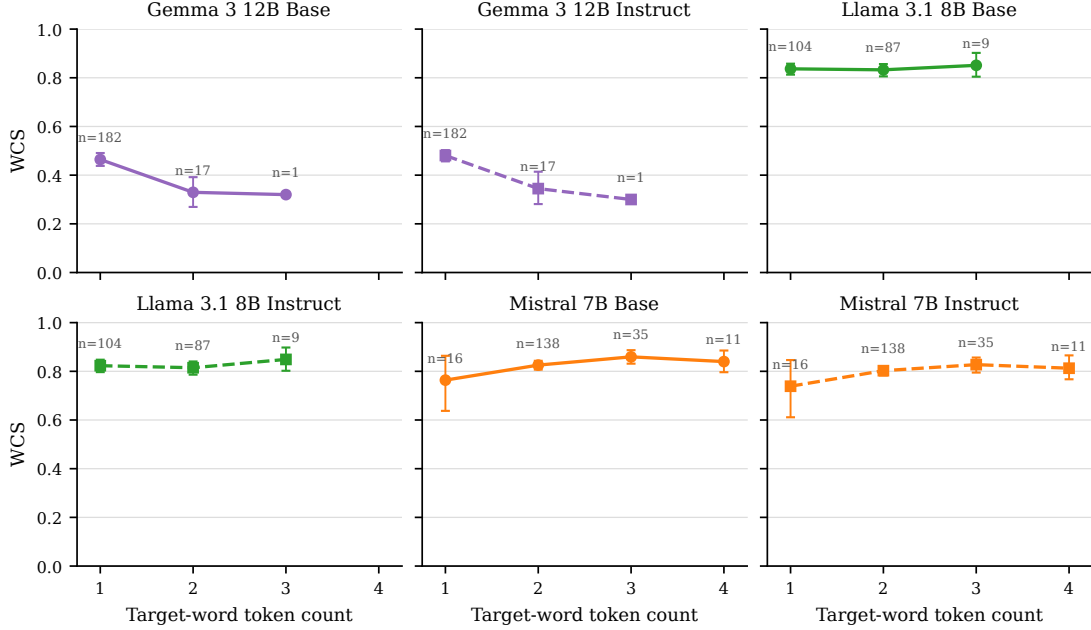


Figure 6: FineWeb WCS stratified by within-model target-word token count for six models under Top- $p = 0.95$ at $T = 1.0$. Each target word contributes 50 contexts. Error bars are 95% percentile confidence intervals from 2,000 target-word cluster-bootstrap draws; n denotes the number of target words in each stratum. Stratum membership differs across tokenizers, so panels support within-model rather than tokenizer-independent model comparisons.

($\mathcal{R} = 0$) to generate under a given context and sampling configuration. However, lexical accessibility is not an all-or-nothing phenomenon; it can also manifest as a probability dampening or elevation. In less severe scenarios, a sampling filter may allow a sophisticated word to remain within the truncated candidate pool ($\mathcal{V}_\theta$), yet reduce its assigned probability mass significantly below the threshold observed in natural human discourse. Such a word remains technically reachable, but its generational frequency is reduced. Because the WCS focuses strictly on the total elimination of linguistic paths, it does not account for this soft probabilistic reduction. Consequently, our empirical results must be interpreted as a conservative lower bound on the true scale of decoding-induced vocabulary constriction. The continuous WCS introduced in section 3.6 provides a tool that can detect both increments and decrements in the probability of a given word,

potentially providing finer detail to analyze and compare how LLMs use specific words.

Beyond our initial experiments, the WCS and its continuous variant provide useful tools that can be used to understand the impact of sampling algorithms and their hyperparameters on specific word generation probabilities. In fact, the continuous variant could also be used to provide feedback during the training or post-training of the models, which could potentially be used to align word generation probabilities with different target objectives. The WCS can also be used to analyze how a model completes texts generated by other models to evaluate if word reachability is better or worse than for human texts. These are some potential uses of the WCS, and we expect other researchers to find imaginative use cases and applications of the WCS, making it a valuable general-purpose tool for the NLP community.

6.2 The Zero-Sum Game of Sampling: Coherence vs Diversity

Our empirical results suggest a conflicting relationship between lexical reachability and structural text coherence. To recover sophisticated, human-utilized words lost in the 10k–40k frequency band, generative configurations often require elevating p near unity or increasing the temperature beyond typical defaults. However, this is not a complete solution: even at $T = 1.5$, our results show that many words remain unreachable in some contexts. This aligns with evidence that higher temperature can modestly increase novelty while correlating with reduced coherence (Peeperkorn et al., 2024). Consequently, practical decoding involves a trade-off: less restrictive settings can improve lexical reachability, while more conservative settings preserve coherence but can also remove viable human vocabulary from the candidate set.

6.3 Impact on Generative Diversity and Downstream Validation

A critical question is whether reduced word reachability translates into observable generative diversity decay. A specific word being rendered unreachable in a given context does not automatically mean the resulting text is semantically impoverished, as models can rely on synonyms and structurally equivalent alternatives. Our WCS is primarily a precise attribution tool for sampling-induced vocabulary loss, rather than a measure of text quality. In the matched generation experiment, greater forced-path reachability is generally associated with greater realized lexical diversity, with the strongest and most consistent evidence among Base models. Specifically, repeated-measures Spearman correlations with WCS are $\rho = .67$ for Type-Token Ratio (TTR) and $\rho = .66$ for Measure of Textual Lexical Diversity (MTLD); full details of these downstream evaluations are available in the public repository.

6.4 Mitigating Word Reachability Limitations

The most direct path to reduce word reachability limitations is to explore alternative sampling schemes and also the optimization of the sampler hyperparameters. However, we believe that mitigating word reachability issues demands a more fundamental shift in decoding pipelines. Current models treat sampling as a layer completely decoupled from the model’s underlying semantic under-

standing. Future research could explore semantic-guided decoders that can actively differentiate between low-probability data noise from contextually rich, low-frequency human vocabulary.

Another possible direction is to modify the training objectives of Large Language Models by incorporating explicit lexical diversity and reachability penalties directly into the loss function. Furthermore, preference optimization frameworks could be modified to reward stylistic and lexical variation rather than purely optimizing for flat, conversational uniformity. By treating vocabulary reachability as an optimization metric, models may be able to better preserve the vocabulary.

7 Conclusion

In this work, we have investigated the role of token sampling in vocabulary reachability within LLM generation. We introduced the Word Coverage Score (WCS), which provides an operational framework to isolate the direct impact of inference-time token truncation on word accessibility.

Our empirical results across several open-weight model families demonstrate that typical decoding filters act as deterministic exclusion boundaries that constrain a model’s operational vocabulary. In the contextually rich 10k–40k word frequency band, samplers with default parameters systematically impose reachability constraints, making some contextually appropriate words mathematically impossible to generate. This finding identifies a possible risk under recursive training rather than demonstrating long-term language change: prior work shows that indiscriminately training successive generative models on model-generated data can cause the tails of the original distribution to disappear (Martínez et al., 2024; Shumailov et al., 2024). Our experiments do not test recursive training, corpus evolution, or effects on human language.

These results illustrate the diagnostic capabilities of the WCS, which can be utilized as a general-purpose auditing framework for validating specific sampling algorithms and optimizing their hyperparameters during inference engineering. In fact, with the addition of the continuous variant, we have formalized a tool that provides the finer probabilistic detail needed to provide continuous feedback during model training or post-training. We expect the WCS and its probabilistic extension to serve as valuable evaluation metrics for auditing how LLMs use specific words across varying contexts.

Limitations

Several limitations must be acknowledged to better frame our results and guide future extensions of this work.

Model Architecture and Parameter Constraints:

Our empirical evaluation is confined to the selected computationally tractable open-weight checkpoints listed in Table 1, leaving larger frontier-scale models unexplored. Furthermore, because our methodology relies fundamentally on extracting complete, step-wise logit probability distributions at each token transition, this protocol cannot be extended to proprietary, closed-source models whose API endpoints only yield compiled text outputs or highly restricted top- k logit values.

Linguistic and Web Bias: The entire analytical pipeline was optimized for and restricted to the English language. The underlying corpora utilized for frequency estimation and verification are English-centric. Additionally, our context selection is derived from the FineWeb dataset, which consists exclusively of Web texts.

Corpus and Target-Sample Dependence: Absolute WCS depends on corpus register, target-word selection, and context sampling. Although the results presented are only for FineWeb, additional results for PG-19 are also available in the public repository.

LCWR Threshold Choice: LCWR converts the per-word WCS distribution into an interpretable tail rate but requires a threshold τ . We select $\tau = 0.10$ to identify words that survive in fewer than one in ten evaluated contexts and report sensitivity at $\tau \in \{0.05, 0.10, 0.20\}$. Context-level WCS remains the threshold-free primary aggregate.

Boundary Bounds of Frequency Bands: Our frequency-based token selection isolates a specific band bounded between the 10,000 and 40,000 rank thresholds. Although this specific window effectively avoids ubiquitous functional words and extreme low-frequency noise, it means that the impact on words outside this range is not explored.

Binary Nature of the WCS: A limitation of the WCS is its binary formulation of reachability ($\mathcal{R} \in \{0, 1\}$). The metric does not capture cases of soft probability suppression, where a word remains technically inside the valid sampling set $\mathcal{V}_\theta$ but has

its probability mass significantly reduced. The continuous WCS proposed in Section 3.6 could address this limitation, but it is not empirically evaluated in this work.

Static Context and Prefix Length Constraints:

The Forced-Path Audit operates within a localized window bounded by a maximum prefix length of $L = 256$ tokens. This provides context for localized syntactic evaluation but does not assess how long-range contextual dependencies interact with token probabilities.

Tokenizer Sub-Word Segmentation: The models evaluated utilize highly diverse sub-word tokenization vocabularies and therefore the same target word w can be split into a differing number of sub-word fragments across different model families. The number of tokens determines the number of filtering steps that must be passed for the word to be reachable. This means that results across models are not fully comparable for some words.

Ethical Considerations

Beyond these limitations, we briefly discuss downstream risks from effects that our study does not test.

Restrictive sampling creates configuration-specific lexical reachability limitations; it does not permanently remove words from a model or from human language. A downstream risk could arise if filtered model output becomes a substantial part of future training corpora. Recursive-training experiments have shown that indiscriminate reuse of generated data can cause early model collapse, in which the tails of the original distribution disappear (Martínez et al., 2024; Shumailov et al., 2024). That result supports taking distribution-tail preservation seriously, but it does not show that current decoding practices have already altered language evolution.

Classical accounts of linguistic relativity motivate broad questions about relationships among language, habitual thought, and categorization (Sapir, 1921; Whorf, 1956). They do not establish that exposure to LLM-generated text changes cognition, and our experiments contain no human-subject or longitudinal evidence bearing on that question. We therefore make no causal claim about effects on human thought; testing such effects would require a separate longitudinal research design.

Acknowledgments

This work was supported by the Agencia Estatal de Investigación (AEI) (doi:10.13039/501100011033) under Grant FUN4DATE (PID2022-136684OB-C22), by TUCAN6-CM (TEC-2024/COM460) funded by CM (ORDEN 5696/2024) and by European Union’s Horizon Europe research and innovation programme under project BRIDGE-AI (Grant 101299050). Access to the Gemini model was provided by the Google Cloud Research Credits program for the Gemini Academic Program. AI coding and writing assistants (e.g., Claude, Codex, Gemini) were used for code development and text polishing. The authors reviewed and edited all outputs and take full responsibility for the contents of this work.

References

- Alex Bogdan and Adrian de Valois-Franklin. 2026. [The surprising universality of llm outputs: A real-time verification primitive](#). *arXiv preprint arXiv:2604.25634*.
- Jorge Luis Borges. 1998. The garden of forking paths. In *Collected Fictions*, pages 119–128. Viking, New York. Original work published 1941.
- DeepSeek-AI. 2025. DeepSeek-R1-Distill-Qwen-14B. <https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B>. Hugging Face model checkpoint.
- Anil R. Doshi and Oliver P. Hauser. 2024. [Generative artificial intelligence enhances creativity but reduces the diversity of novel content](#). *Science Advances*, 10(28):eadn5290.
- Google DeepMind. 2025a. [Gemini 2.5 Flash-Lite](#). Model documentation; model code: gemini-2.5-flash-lite.
- Google DeepMind. 2025b. Gemma-3-12B-it. <https://huggingface.co/google/gemma-3-12b-it>. Hugging Face model checkpoint.
- Google DeepMind. 2025c. Gemma-3-12B-pt. <https://huggingface.co/google/gemma-3-12b-pt>. Hugging Face model checkpoint.
- Google DeepMind. 2026a. Gemma-4-E4B. <https://huggingface.co/google/gemma-4-E4B>. Hugging Face model checkpoint.
- Google DeepMind. 2026b. Gemma-4-E4B-it. <https://huggingface.co/google/gemma-4-E4B-it>. Hugging Face model checkpoint.
- Adrián Gude, Roi Santos-Rios, Francis Bond, Dan Flickinger, Carlos Gómez-Rodríguez, and Olga Zamaraeva. 2026. [More aligned, less diverse? analyzing the grammar and lexicon of two generations of LLMs](#). In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 38900–38911, San Diego, California, United States. Association for Computational Linguistics.
- Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. [The curious case of neural text de-generation](#). In *International Conference on Learning Representations*.
- Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, and Yejin Choi. 2025. [Artificial hivemind: The open-ended homogeneity of language models \(and beyond\)](#). In *Advances in Neural Information Processing Systems*, volume 38.
- Gonzalo Martínez, José Alberto Hernández, Javier Conde, Pedro Reviriego, and Elena Merino-Gómez. 2025. [Beware of words: Evaluating the lexical diversity of conversational LLMs using ChatGPT as case study](#). *ACM Transactions on Intelligent Systems and Technology*, 16(6):1–15.
- Gonzalo Martínez, Lauren Watson, Pedro Reviriego, José Alberto Hernández, Marc Juárez, and Rik Sarkar. 2024. Towards understanding the interplay of generative artificial intelligence and the internet. In *International workshop on epistemic uncertainty in artificial intelligence*, pages 59–73. Springer.
- Meta AI. 2024a. Llama-3.1-8B. <https://huggingface.co/meta-llama/Llama-3.1-8B>. Hugging Face model checkpoint.
- Meta AI. 2024b. Llama-3.1-8B-Instruct. <https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct>. Hugging Face model checkpoint.
- Nguyen Nhat Minh, Andrew Baker, Clement Neo, Allen G. Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. 2025. [Turning up the heat: Min-p sampling for creative and coherent LLM outputs](#). In *The Thirteenth International Conference on Learning Representations*.
- Mistral AI. 2024a. Mistral-7B-Instruct-v0.3. <https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.3>. Hugging Face model checkpoint.
- Mistral AI. 2024b. Mistral-7B-v0.3. <https://huggingface.co/mistralai/Mistral-7B-v0.3>. Hugging Face model checkpoint.
- Alberto Muñoz-Ortiz, Carlos Gómez-Rodríguez, and David Vilares. 2024. [Contrasting linguistic patterns in human and LLM-generated news text](#). *Artificial Intelligence Review*, 57:265.

- Peter Norvig. 2009. [Natural language corpus data](#). In Toby Segaran and Jeff Hammerbacher, editors, *Beautiful Data: The Stories Behind Elegant Data Solutions*, chapter 14, pages 219–242. O’Reilly Media, Inc.
- Max Peeperkorn, Tom Kouwenhoven, Dan Brown, and Anna Jordanous. 2024. [Is temperature the creativity parameter of large language models?](#) In *Proceedings of the 15th International Conference on Computational Creativity*, pages 226–235, Jönköping, Sweden. Association for Computational Creativity.
- Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2024. [The FineWeb datasets: Decanting the web for the finest text data at scale](#). In *Advances in Neural Information Processing Systems*, volume 37.
- Qwen Team. 2024a. Qwen2.5-14B. <https://huggingface.co/Qwen/Qwen2.5-14B>. Hugging Face model checkpoint.
- Qwen Team. 2024b. Qwen2.5-14B-Instruct. <https://huggingface.co/Qwen/Qwen2.5-14B-Instruct>. Hugging Face model checkpoint.
- Qwen Team. 2026a. Qwen3.5-9B. <https://huggingface.co/Qwen/Qwen3.5-9B>. Hugging Face model checkpoint.
- Qwen Team. 2026b. Qwen3.5-9B-Base. <https://huggingface.co/Qwen/Qwen3.5-9B-Base>. Hugging Face model checkpoint.
- Edward Sapir. 1921. *Language: An introduction to the study of speech*. Harcourt, Brace and Company, New York.
- Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Nicolas Papernot, Ross Anderson, and Yarin Gal. 2024. [AI models collapse when trained on recursively generated data](#). *Nature*, 631:755–759.
- Alexander Shypula, Shuo Li, Botong Zhang, Vishakh Padmakumar, Kayo Yin, and Osbert Bastani. 2025. [Evaluating the diversity and quality of LLM generated content](#). In *Conference on Language Modeling*.
- Emily Wenger and Yoed N Kenett. 2026. [Large language models are homogeneously creative](#). *PNAS Nexus*, 5(3):pgag042.
- Benjamin Lee Whorf. 1956. *Language, thought, and reality: Selected writings*. Technology Press of Massachusetts Institute of Technology. Edited by John B. Carroll.
- Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. 2025. [The price of format: Diversity collapse in LLMs](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15454–15468, Suzhou, China. Association for Computational Linguistics.
- George Kingsley Zipf. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Addison-Wesley Press, Cambridge, Massachusetts.

A Selected Word List and Aggregate Reachability

Table 3 lists the 200 target words sampled from the 10k–40k frequency-rank band. Each word was paired with 50 FineWeb contexts, yielding 10,000 forced-path audit samples. The count column reports the source frequency count in millions. More common selected words appear toward the beginning of the table, while rarer selected words appear toward the end.

Table 3: Selected target words from the 10k–40k frequency band. Frequency counts are shown in millions.

Word	Rank	Count	Word	Rank	Count	Word	Rank	Count
nonfiction	10019	5.04	phenomena	10146	4.94	preschool	10167	4.92
observers	10332	4.80	lining	10381	4.76	glove	10453	4.71
wastewater	10455	4.71	relaxed	10615	4.60	adhesive	10643	4.59
aperture	11127	4.27	intuitive	11181	4.24	coordinating	11191	4.23
borrow	11358	4.14	shocked	11678	3.95	diaper	11702	3.93
refuge	11736	3.91	contexts	11944	3.80	rename	12029	3.76
petrol	12158	3.69	lever	12422	3.57	webs	12569	3.50
honours	12636	3.47	abundant	12717	3.44	competitor	12862	3.38
encounters	12915	3.36	protests	13042	3.30	vaccines	13219	3.24
turbine	13229	3.24	palms	13514	3.13	textures	13560	3.11
microscope	13698	3.05	checksum	13713	3.05	worksheet	13776	3.02
brewing	13896	2.97	autonomy	13990	2.94	allergic	14062	2.91
addict	14072	2.91	juniors	14086	2.90	admits	14096	2.90
padding	14131	2.89	comprehension	14182	2.87	flea	14392	2.80
karate	14617	2.72	ceremonies	14652	2.71	invaluable	14736	2.68
undergo	14957	2.62	mirage	15383	2.49	wrought	15949	2.35
comfortably	16082	2.31	roadside	16139	2.30	cessation	16872	2.14
psychotherapy	16899	2.13	emblem	17091	2.09	dunes	17097	2.09
praised	17239	2.06	horseback	17690	1.97	bullying	17807	1.95
robotic	17964	1.92	crypt	18063	1.90	cubes	18162	1.89
sanity	18221	1.88	forefront	18597	1.82	teal	18672	1.80
frontiers	18775	1.79	drank	18821	1.78	watersports	18895	1.77
commute	18959	1.75	prognosis	19019	1.75	determinants	19034	1.74
phishing	19206	1.72	newsfeed	19224	1.72	ibn	19272	1.71
overdrive	19574	1.66	coffin	19591	1.66	moi	19720	1.64
ode	19721	1.64	conferred	19847	1.62	absorbing	19926	1.61
wilder	19964	1.61	sultan	20280	1.56	nah	20554	1.53
pinot	20923	1.48	deputies	20998	1.47	socioeconomic	21225	1.45
inherit	21257	1.44	surreal	21289	1.44	encrypt	21302	1.44
constable	21318	1.43	phono	21441	1.42	bps	21643	1.40
renters	21838	1.37	radiological	22019	1.35	motorway	22030	1.35
iodine	22050	1.35	decorators	22345	1.32	psoriasis	22352	1.32
motorists	22526	1.30	biodiesel	22722	1.28	presume	22857	1.27
dharma	23121	1.25	calligraphy	23210	1.24	starship	23323	1.23
livelihood	23358	1.23	drowned	23473	1.22	sweeps	23508	1.21
calves	23565	1.21	lacy	23573	1.21	walkthroughs	23742	1.19
deliverables	24026	1.17	inset	24086	1.16	whispers	24142	1.16
digitized	24337	1.14	palmetto	24668	1.12	pleasantly	24767	1.11
communicates	24788	1.11	wildfire	24965	1.09	fruity	24990	1.09
sonnet	25111	1.08	bong	25226	1.07	inadvertently	25274	1.07
reacted	25375	1.06	interplay	25570	1.05	azure	25682	1.04
pathogenesis	25721	1.04	workbench	25759	1.04	frontage	25909	1.03
marketer	26005	1.02	priori	26063	1.02	unison	26139	1.01
bassist	26352	1.00	waterfowl	26542	0.98	payback	26663	0.98
slipper	27106	0.95	governs	27153	0.94	extraordinarily	27253	0.94
antagonist	27459	0.93	cadre	27463	0.93	velour	27490	0.92
marketable	27623	0.92	machined	27858	0.90	liturgical	27888	0.90
cyclones	28459	0.87	lament	28523	0.87	terraces	28525	0.87
repo	28685	0.86	hassles	28814	0.85	mosques	29297	0.83
distractions	29515	0.82	tentatively	29692	0.81	conceptions	29762	0.81
cfg	29771	0.81	vigilance	29889	0.80	humankind	30114	0.79
forearm	30313	0.78	eerie	30535	0.77	electrochemical	30681	0.77
jetty	30713	0.77	ascertained	30857	0.76	varnish	31047	0.75
refinements	31111	0.75	repost	31114	0.75	streaks	31550	0.73
subtracted	32269	0.71	pathologist	32474	0.70	irradiated	32530	0.70
sari	32632	0.69	hydrophobic	32843	0.68	prim	33359	0.67
feudal	33531	0.66	receptacles	34096	0.64	socialization	34264	0.64
cosmological	34520	0.63	compensating	34618	0.63	guestrooms	34696	0.62
dendritic	35147	0.61	unskilled	35413	0.60	numerator	35512	0.60
biodegradable	35892	0.59	lagging	36202	0.58	radii	36358	0.58
overflowing	36438	0.57	croquet	36538	0.57	ceasefire	36610	0.57
creatinine	36635	0.57	landmines	37122	0.56	erections	37209	0.55
multifunctional	37364	0.55	revamp	37429	0.55	subdomain	37616	0.54
subspecies	37788	0.54	nuance	37939	0.54	freer	38009	0.53
yanks	38150	0.53	photojournalism	38241	0.53	pronoun	38652	0.52
subtropical	39006	0.51	unbeaten	39592	0.50	aloft	39597	0.50
encased	39739	0.49	credentialing	39746	0.49			