Layer-wise Decodability of Word Frequency in Decoder-Only Language Models

Anonymous EMNLP Submission

Abstract

We investigate how word frequency is represented across layers in decoder-only language models. Using frozen hidden states, we probe lexical decision, direct frequency classification, and continuous-frequency regression in a controlled evaluation centered on matched 3B models and extended to a nine-model sweep across the Llama, Qwen, and Smol families. Frequency information is strongly decodable across families and scales: binary high-versus-low probes reach high held-out accuracy, three-way probes support finer low/mid/high discrimination, and regression analyses recover a strong graded relation to continuous log frequency. Across the full sweep, one layer-wise pattern is especially consistent: the *relative* advantage of high-frequency over low-frequency words is largest in the earliest layers and decreases with depth, while frequency remains strongly decodable throughout the network. Token-restricted, reduced-confound, pooling, multi-seed, and probe-selectivity analyses support this interpretation, but do not justify a causal claim about how models use frequency during computation. Overall, the results show that decoder-only language models encode a robust and graded representation of word frequency, with the clearest early-layer signature appearing as stronger relative separability for high-frequency words.

1 Introduction

While decoder-only models dominate NLP, high performance does not reveal what linguistic information their internal states encode (Brown et al., 2020; Belinkov, 2022). Determining if hidden representations preserve variables shaping human language processing, like word frequency, clarifies how lexical information is organized across model depth (Ettinger, 2020; Oh and Schuler, 2023). Frequency is a robust psycholinguistic variable; high-frequency words are processed faster and more accurately, making them an ideal test case for representation analysis (Balota et al., 2004; Brysbaert et al., 2011).

Probing requires caution, as decodability does not establish functional use (Alain and Bengio, 2016; Hewitt and Manning, 2019). We therefore utilize controls and selectivity tests to ensure frequency effects are not merely artifacts of orthography or tokenization (Hewitt and Liang, 2019; Voita and Titov, 2020). While most layer-wise studies focus on encoder-based models and syntax (Jawahar et al., 2019; Rogers et al., 2020), a controlled account of word frequency in contemporary decoder-only models is currently missing.

We address this gap by probing three matched 3B models i.e Llama-3.2 (Touvron et al., 2023), Qwen2.5 (Hui et al., 2024), and SmolLM3 (Bakouch et al., 2025) alongside a nine-model sweep. From frozen hidden states, we evaluate lexical decision and frequency probing across several formulations. We further validate results through tokenization controls, confound matching, and probe selectivity checks.

Our contributions are threefold:

- We provide a controlled layer-wise analysis of frequency in matched decoder-only models.
- We demonstrate that frequency is strongly decodable, with relative separability peaking in early layers and decreasing with depth.
- We show this pattern remains robust under multiple configurations and continuous regression.

These results provide a precise account of how a central psycholinguistic variable is distributed across contemporary model architectures.

2 Related Work

Probing as representation analysis. Probing is a standard tool for testing what information is re-

coverable from neural language-model representations (Alain and Bengio, 2016; Hewitt and Manning, 2019). At the same time, probing studies have repeatedly emphasized that recoverability is not the same as functional use. Probe performance can reflect classifier capacity, data geometry, and task design, which is why control tasks, selectivity analyses, and restrained interpretation are essential (Hewitt and Liang, 2019; Voita and Titov, 2020; Belinkov, 2022). Recent interpretability work in the LLM era has reinforced this point, arguing that probing is most informative when it targets a clearly defined property and limits alternative explanations (Calderon and Reichart, 2025).

Layer-wise organization in pretrained language models. A large literature has examined how linguistic information is distributed across transformer layers. In encoder-based models, lower layers often capture more local or surface-oriented information, while higher layers support more abstract structure (Jawahar et al., 2019; Tenney et al., 2019; Rogers et al., 2020). Related work has shown that lexical-semantic information can remain distributed across depth rather than localizing to a single stage (Vulić et al., 2020), and that transformer feed-forward blocks can store lexical and factual associations in a memory-like form (Geva et al., 2021). Recent studies suggest that some classical layer-wise patterns remain relevant in newer language models (Li and Subramani, 2025). However, this literature has rarely addressed word frequency directly, and much of it remains centered on encoder architectures rather than decoder-only models.

Psycholinguistic evaluation of language models. A complementary line of work evaluates language models with psycholinguistic diagnostics instead of downstream benchmark scores. Ettinger (2020) showed that strong aggregate performance can coexist with clear weaknesses on targeted psycholinguistic tests. Subsequent work has examined the relation between model predictions and human reading behavior, showing that larger or stronger models do not necessarily provide better accounts of human processing measures (Oh and Schuler, 2023). Other studies have used language models to support psycholinguistic resource construction and analysis, further linking NLP models with cognitively motivated evaluation (Trott, 2024; Clark et al., 2026). Together, this literature motivates focused analysis of specific variables rather than assuming that human-like sensitivity follows from general model quality.

Word frequency as a psycholinguistic variable. Word frequency is one of the most established variables in lexical processing. It affects recognition speed and accuracy, including in lexical decision, and remains central to psycholinguistic accounts of word recognition (Forster and Chambers, 1973; Balota et al., 2004, 2007; Brysbaert et al., 2011). For representation analysis, it is attractive because it is both theoretically important and empirically robust, while still being distinguishable from lexicality itself. Despite this, prior NLP work has not provided a systematic layer-wise account of how word frequency is represented in contemporary decoder-only language models under controlled probing conditions.

Position of the present study. Our study sits at the intersection of these strands. We adopt the methodological caution of the probing literature, the layer-wise perspective of representation studies, and the targeted focus of psycholinguistic evaluation, and apply them to word frequency in decoder-only models. Concretely, we analyze matched 3B decoder-only models with layer-wise probes, direct frequency classification on real-word stimuli, lexical decision as a baseline task, and robustness checks designed to address common concerns about probe-based claims. This yields a more focused account of frequency representations in contemporary decoder-only language models than is currently available.

3 Methodology

3.1 Study design

We test whether word-frequency information is encoded in decoder-only hidden representations, and how that encoding varies with depth. The main paper centers on three matched 3B models, *Llama-3.2-3B*, *Qwen2.5-3B*, and *SmolLM3-3B-Base*. To assess whether the same layer-wise pattern appears beyond this matched comparison, we also run a nine-model sweep spanning Llama (*3.2-1B*, *3.2-3B*, *3.1-8B*), Qwen (*2.5-1.5B*, *2.5-3B*, *3-4B-Base*, *3-8B-Base*), and Smol (*SmolLM2-360M*, *SmolLM3-3B-Base*).

For each model, we extract frozen hidden states from every layer and train lightweight probes on top of them. The analysis consists of lexical decision, direct frequency probing on real-word stim-

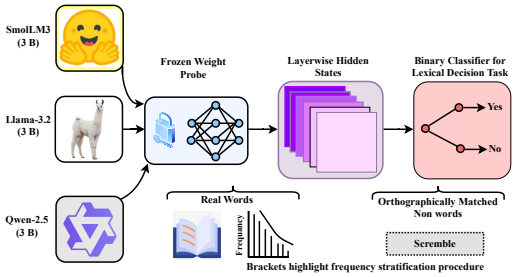

Figure 1: Overview of the experimental pipeline. For each decoder-only model, hidden representations are extracted from every layer, followed by layer-wise probing for lexical decision and frequency prediction. The main analysis is complemented by tokenization controls, reduced-confound matching, pooling ablation, selectivity checks, multi-seed stability analysis, and continuous-frequency regression.

Algorithm 1 Layer-wise frequency probing.
The algorithm trains probe classifiers $\phi^{(\ell)}$ at each layer ℓ to measure the frequency effect $\hat{\Delta}^{(\ell)}$ between higher-frequency (H) and lower-frequency (L) words using test accuracies $\hat{\pi}_H^{(\ell)}$ and $\hat{\pi}_L^{(\ell)}$.

Input: model $\mathcal{M}$, dataset $\mathcal{D}$
Phase 1: Representation extraction
$\mathbf{H}^{(0)}, \ldots, \mathbf{H}^{(L-1)} \leftarrow \textsc{ForwardPass}(\mathcal{M}, \mathcal{D})$
Phase 2: Layer-wise probing
for $\ell = 0$ to $L-1$ **do**
  $\phi^{(\ell)} \leftarrow \textsc{TrainProbe}(\mathbf{H}^{(\ell)})$
  $\hat{\pi}_H^{(\ell)}, \hat{\pi}_L^{(\ell)} \leftarrow \textsc{EvaluateFreqGroups}(\phi^{(\ell)})$
  $\hat{\Delta}^{(\ell)} \leftarrow \hat{\pi}_H^{(\ell)} - \hat{\pi}_L^{(\ell)}$
end for
Phase 3: Statistical analysis
for $\ell = 0$ to $L-1$ **do**
  $h^{(\ell)} \leftarrow 2\left(\arcsin\sqrt{\hat{\pi}_H^{(\ell)}} - \arcsin\sqrt{\hat{\pi}_L^{(\ell)}}\right)$
  $p^{(\ell)} \leftarrow \textsc{PropTest}(\hat{\pi}_H^{(\ell)}, \hat{\pi}_L^{(\ell)})$
end for
$\{p_{\text{FDR}}^{(0)}, \ldots, p_{\text{FDR}}^{(L-1)}\} \leftarrow \text{BH-FDR}(\{p^{(0)}, \ldots, p^{(L-1)}\})$
Return: $\{(\hat{\Delta}^{(\ell)}, h^{(\ell)}, p_{\text{FDR}}^{(\ell)})\}_{\ell=0}^{L-1}$

uli, and robustness analyses covering tokenization controls, reduced-confound matching, pooling ablation, multi-seed stability, probe selectivity, and continuous-frequency regression. Figure 1 summarizes the pipeline, and Algorithm 1 summarizes the main computation.

3.2 Stimuli and frequency labels

Lexical decision stimuli. Real-word and non-word items are drawn from the English Lexicon Project (ELP) (Balota et al., 2007). We retain real words with valid `Log_Freq_HAL` values and lexical decision norms. Non-words come from the same lexical decision resource and are used only for the baseline word-versus-non-word task (Balota et al., 2004). All strings are lowercased before tokenization.

Frequency labels. Frequency analyses use real words only. Following standard psycholinguistic practice, we divide words into tertiles using the 33rd and 66th percentiles of log-HAL frequency (Brysbaert et al., 2011), yielding *low*, *mid*, and *high* frequency groups. The main direct-frequency task uses the high and low groups only, while the three-way auxiliary task uses all three groups.

Controlled subsets. Because frequency can correlate with lexical form and tokenization, we construct three additional subsets. The first is a *single-token* subset, defined separately for each tokenizer, that restricts the analysis to words represented by a single token. The second is a *token-count-reduced lexical-decision subset*, which narrows token-count differences between high- and low-frequency groups. The third is a *reduced-confound matched* subset for direct frequency probing. This narrows lexical-form differences, but is not intended to claim perfect equality across all correlated lexical variables.

3.3 Layer-wise representation extraction

All models are run in evaluation mode with frozen parameters. For each input item x_i, tokenized to length m_i, we extract hidden states from every transformer layer (ℓ):

$$\mathbf{H}_i^{(\ell)} \in \mathbb{R}^{m_i \times d_\ell}, \qquad \ell \in \{0, \ldots, L-1\}.$$

Because decoder-only models do not provide a dedicated classification token, our default item representation is the hidden state at the final non-padding position:

$$\mathbf{z}_i^{(\ell)} = \mathbf{H}_i^{(\ell)}[m_i].$$

This yields one vector per item per layer and is consistent with the model's causal structure. As a robustness check, we also evaluate mean pooling over non-padding positions. Before probe training, we verify that extracted representations are non-degenerate by checking for non-trivial variance and the absence of all-zero vectors.

3.4 Probe models and training

For each layer ℓ, we train a probe on frozen representations. For classification, the default probe is a lightweight multilayer perceptron with two hidden layers of width 512 and 256. Hidden layers use

Kaiming initialization (He et al., 2015); the output layer uses Xavier initialization (Glorot and Bengio, 2010). Batch normalization (Ioffe and Szegedy, 2015) and dropout with rate 0.3 are applied in the hidden layers.

For a classification task with label set $\mathcal{Y}$, the probe is

$$f_\theta^{(\ell)} : \mathbb{R}^{d_\ell} \to \mathbb{R}^{|\mathcal{Y}|}.$$

We optimize Focal Loss (Lin et al., 2017) often called weighted cross-entropy when focusing parameter$\gamma = 0$ where Focal loss is stated as $\mathrm{FL}(p_t) = -\alpha_t(1-p_t)^\gamma \log(p_t)$ and p_t is the model's estimated probability for the ground-truth class (a higher p_t means the model is more confident and correct)

$$\mathcal{L}_{\text{cls}} = -\frac{1}{N}\sum_{i=1}^{N} w_{y_i} \log p_\theta(y_i \mid \mathbf{z}_i^{(\ell)}),$$

where w_{y_i} is the class weight for label y_i. Optimization uses AdamW (Loshchilov and Hutter, 2017) with learning rate 10^{-3} and weight decay 10^{-2}.

Each layer-task pair is split into train, validation, and test partitions in a 70/15/15 ratio. Classification splits are stratified by label. Feature standardization is performed with `StandardScaler` fit on training data only and then applied to validation and test data, preventing leakage (Rogers et al., 2020). Training uses early stopping with patience 10, a `ReduceLROnPlateau` scheduler with factor 0.5, gradient clipping with maximum norm 1.0, and class-balanced mini-batches through weighted sampling.

For continuous-frequency analysis, we fit a scalar-output probe to predict log-HAL frequency from the same frozen representations and evaluate it with R^2 and Spearman correlation.

3.5 Primary tasks

Lexical decision. The lexical decision task tests whether a representation separates real words from non-words:

$$y_i^{\text{lex}} \in \{0, 1\},$$

where 1 denotes a real word and 0 denotes a non-word.

Direct binary frequency probing. The main analysis tests whether frequency is decodable independently of lexicality. Restricting the data to real words only, we define

$$y_i^{\text{freq}} \in \{0, 1\},$$

where the labels correspond to low- and high-frequency groups.

Three-way frequency probing. As an auxiliary analysis, we also train a three-way probe over low-, mid-, and high-frequency words. This tests whether the same representations support finer frequency distinctions beyond the binary high-versus-low contrast.

Continuous-frequency regression. To test whether the encoded signal is graded rather than purely categorical, we fit a regression probe from hidden states to continuous log-HAL frequency:

$$r_i = \log \mathrm{HAL}(x_i).$$

3.6 Layer-wise frequency effect

To quantify how the *relative* advantage of high-frequency words changes with depth, we analyze subgroup performance within lexical decision.

For each layer ℓ, let $\hat{\pi}_H^{(\ell)}$ and $\hat{\pi}_L^{(\ell)}$ denote held-out lexical-decision accuracy for high- and low-frequency real words. We define the layer-wise frequency gap as

$$\hat{\Delta}^{(\ell)} = \hat{\pi}_H^{(\ell)} - \hat{\pi}_L^{(\ell)}. \tag{1}$$

Positive values indicate that high-frequency words are classified more accurately than low-frequency words at layer ℓ.

To quantify the same effect independently of raw accuracy, we compute Cohen's h:

$$h^{(\ell)} = 2\left(\arcsin\sqrt{\hat{\pi}_H^{(\ell)}} - \arcsin\sqrt{\hat{\pi}_L^{(\ell)}}\right), \tag{2}$$

following Cohen (2013) and Lakens (2013).

3.7 Robustness analyses

We perform six robustness analyses.

Single-token control. We repeat the frequency analysis on words represented as a single token by a model's tokenizer.

Token-count-reduced lexical-decision control. We repeat lexical decision on a subset that narrows token-count differences between high- and low-frequency words.

Reduced-confound matched probing. We repeat direct frequency probing on a subset constructed to narrow lexical-form differences between groups. This is treated as a reduced-confound control rather than a fully matched design.

Pooling ablation. We compare the default last-token representation with mean pooling over non-padding positions.

Multi-seed stability. We rerun selected probe configurations across five random seeds and summarize mean and standard deviation. These experiments are conducted on sampled anchor layers spanning early, middle, and late depth rather than every layer.

Probe selectivity. To test whether the main result depends on a highly expressive nonlinear probe, we compare the default MLP with a linear probe and a shuffled-label control. If $m^{(\ell)}_{\text{true}}$ is the performance of a probe trained on true labels and $m^{(\ell)}_{\text{shuf}}$ is the corresponding shuffled-label performance, then layer-wise selectivity is

$$\text{Sel}^{(\ell)} = m^{(\ell)}_{\text{true}} - m^{(\ell)}_{\text{shuf}}.$$

As with the seed analysis, selectivity is evaluated on sampled anchor layers rather than across the full layer sweep.

3.8 Evaluation metrics and statistical testing

For lexical decision and binary frequency probing, we report held-out accuracy and ROC-AUC. For the three-way frequency task, we report macro-F1. For continuous regression, we report coefficient of determination (R^2) and Spearman rank correlation (ρ). Multi-seed experiments are summarized as mean $\pm$ standard deviation across five seeds.

To test whether the subgroup difference in Eq. 1 is reliable at each layer, we use the two-proportion z-test with Yates correction (Agresti, 2013). Because these tests are repeated across layers, we control the false discovery rate with the Benjamini-Hochberg procedure (Benjamini and Hochberg, 1995). The corrected values are denoted

$$p^{(\ell)}_{\text{FDR}}, \qquad \ell = 0, \ldots, L-1.$$

We report statistical reliability and effect magnitude separately, using the two-proportion test for the former and Cohen's h for the latter.

4 Results

We reported a plot for the three different architectures of 3B parameters (Llama, Qwen, SmolLM) in the main text for clear understanding of the readers, as shown in Figure 2 and reserve the full sweep of the nine-models, as well as individual plots, exhaustive per-layer tables for the supplementary material §9. The figures therefore show all model variants, while the text summarizes the empirical patterns that recur across families, scales, and control settings.

4.1 Lexical decision saturates quickly across the full model sweep

Figure 3 in §9 shows lexical decision accuracy across normalized depth for all nine variants. In every model, accuracy rises sharply in early layers and then remains high.

Peak lexical decision accuracy ranges from 0.893 for *SmolLM2-360M* to 0.986 for *Llama-3.1-8B*. Eight of the nine variants reach at least 0.949. Within the Llama family, peak accuracy increases with scale, from 0.965 for *Llama-3.2-1B* to 0.978 for *Llama-3.2-3B* and 0.986 for *Llama-3.1-8B*. The Qwen variants cluster more tightly, with peak accuracy between 0.949 and 0.963. In the Smol family, *SmolLM3-3B-Base* peaks at 0.962, whereas *SmolLM2-360M* peaks at 0.893.

4.2 Word frequency is strongly decodable across models and scales

We next evaluate direct probing of word frequency using real-word representations only.

Binary high-versus-low frequency probing is strong across the full sweep. Peak accuracy ranges from 0.921 for *SmolLM2-360M* to 0.975 for *Llama-3.1-8B*. First-layer performance ranges from 0.664 to 0.891. Final-layer binary accuracy remains high for every model, ranging from 0.885 to 0.968.

The same pattern appears under the harder three-way formulation. Peak macro-F1 ranges from 0.693 for *SmolLM2-360M* to 0.799 for *Llama-3.1-8B*. The strongest non-Llama three-way result is 0.796 for *Qwen3-4B-Base*. Among the three primary 3B models, peak three-way macro-F1 is 0.761 for *Llama-3.2-3B*, 0.750 for *Qwen2.5-3B*, and 0.763 for *SmolLM3-3B-Base*.

Table 1 summarizes the main task-level results for all nine variants. *Llama-3.1-8B* attains the highest peak result on lexical decision, binary frequency probing, three-way probing, and regression. Within the Qwen family, model ordering differs by task, and *Qwen3-4B-Base* slightly exceeds *Qwen3-8B-Base* on three-way probing.

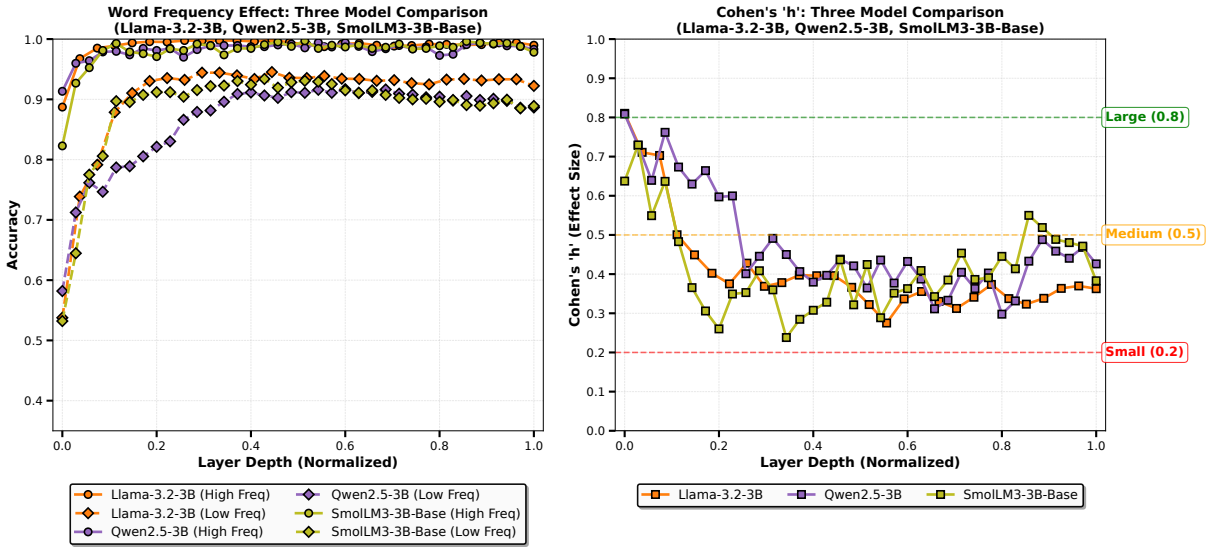

Figure 2: Layer-wise frequency effects for three different architectures (Size: 3B). Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive. Moreover, comparison of nine models and individual models is shown in section §9.

Table 1: Compact summary of the main task-level results for all nine model variants. Full per-layer tables are provided in the supplementary material.

Model	Lexical decision peak acc.	Binary frequency peak acc.	Three-way peak macro-F1	Regression peak R^2
Llama-3.2-1B	0.965	0.966	0.769	0.837
Llama-3.2-3B	0.978	0.970	0.761	0.862
Llama-3.1-8B	**0.986**	**0.975**	**0.799**	**0.873**
Qwen2.5-1.5B	0.949	0.959	0.745	0.778
Qwen2.5-3B	0.961	0.957	0.750	0.811
Qwen3-4B-Base	0.960	0.958	0.796	0.789
Qwen3-8B-Base	0.963	0.954	0.772	0.805
SmolLM2-360M	0.893	0.921	0.693	0.673
SmolLM3-3B-Base	0.962	0.961	0.763	0.829

4.3 Relative frequency separability is strongest in early layers

We next analyze how the high-frequency versus low-frequency subgroup difference changes with depth. Figure 13 shows subgroup accuracy trajectories and Cohen's h across layers.

The high-minus-low accuracy gap is maximal in layer 0 or layer 1 for every variant. Cohen's h peaks in layer 0 or layer 1 for eight variants, and at layer 5 for *SmolLM2-360M*.

The magnitude of the early effect varies across models. At the first layer, the high-minus-low accuracy gap ranges from 0.102 for *SmolLM2-360M* to 0.440 for *Llama-3.1-8B*. Corresponding early Cohen's h values range from 0.637 to 1.008. Both measures decrease with depth, but remain positive at the final layer. Final-layer high-minus-low gaps range from 0.053 to 0.199, and final-layer Cohen's h ranges from 0.317 to 0.599. All layer-wise high-versus-low differences remain significant after Benjamini-Hochberg correction.

4.4 Continuous regression shows graded frequency encoding

We then evaluate continuous-frequency regression. Peak R^2 ranges from 0.673 for *SmolLM2-360M* to 0.873 for *Llama-3.1-8B*. Peak Spearman correlation ranges from 0.821 to 0.935. Final-layer performance also remains strong, with final-layer R^2 values from 0.587 to 0.827 and final-layer Spearman correlation from 0.767 to 0.911. For all three Llama variants, peak R^2 exceeds 0.83. Among the two non-Llama primary 3B models, peak R^2 is 0.811 for *Qwen2.5-3B* and 0.829 for *SmolLM3-3B-Base*.

4.5 Robustness analyses preserve the main empirical pattern

We next test whether the main pattern persists under alternative task formulations and control settings.

The token-count-reduced lexical-decision subset preserves strong performance across the full sweep. Peak accuracy ranges from 0.896 to 0.984, and the maximum high-minus-low gap ranges from 0.167 to 0.278. The reduced-confound matched-pair probe also remains strong, with peak accuracy between 0.920 and 0.979. The balancing diagnostics show that this control narrows, but does not eliminate, residual lexical differences: after matching, residual differences remain in mean word length (about 1.75 characters), orthographic neighborhood size (about 1.28), and token count (about 0.82).

The single-token subset provides additional support for decodability. Peak single-token accuracy ranges from 0.920 to 0.991 for the nine variants present in that our workbook.

The pooling ablation shows similar results under mean pooling and last-token extraction. Switching to mean pooling changes peak lexical-decision accuracy by at most 0.004 across the nine variants, and the largest frequency gap still appears at or near the start of the network in every family.

The multi-seed stability analysis, conducted on sampled anchor layers, shows best mean F1 values between 0.886 and 0.986 across variants, with a maximum observed accuracy standard deviation of 0.0066.

The selectivity analyses also preserve the pattern. Linear probes remain strong, with best accuracy between 0.893 and 0.982 across the nine variants. Shuffled-label probes range from 0.487 to 0.634 on the sampled layers used for selectivity. Peak selectivity ranges from 0.484 to 0.565 for every variant.

Across these controls, frequency decodability remains strong, and the largest relative high-versus-low separation continues to appear in early layers.

5 Discussion

5.1 An early-layer frequency signature that generalizes across model variants

The main empirical result is not only that word frequency is decodable, but that its *relative* separability is strongest in the earliest layers and then decreases with depth. This pattern holds across the full nine-model sweep, spanning multiple families and scales. That consistency suggests that the early-layer dominance of frequency-related separability is not an artifact of a single architecture or model size, but a recurrent property of contemporary decoder-only language models.

One plausible interpretation is that earlier layers preserve stronger sensitivity to lexical-statistical information, whereas deeper layers transform representations in ways that reduce the relative advantage of high-frequency words without removing frequency information altogether. This is broadly consistent with prior layer-wise work showing that lower transformer layers often emphasize more local or surface-oriented properties, while later layers support more abstract structure (Jawahar et al., 2019; Tenney et al., 2019; Vulić et al., 2020). Our results extend that perspective to a psycholinguistically central variable in decoder-only models.

5.2 Frequency remains available throughout the network

The layer-wise profile should be interpreted carefully. Our results do not show that frequency is confined to early layers, nor that later layers stop encoding it. Direct frequency probes remain strong throughout the network, and the regression results show a robust graded relation to continuous log frequency even in middle and late layers. What changes with depth is the *relative* high-versus-low separation observed in lexical decision, not the existence of frequency information itself.

This distinction is important for the paper's main claim. A model can show weaker relative subgroup separation in later layers while still retaining strong absolute decodability. The most accurate summary is therefore that frequency is expressed most strongly in relative terms early in the network, while remaining decodable across depth. That is narrower and more defensible than treating frequency as an exclusively early-layer property.

5.3 What the cross-model sweep adds

The broader model sweep adds two useful qualifications. First, the main trend is stable across families and scales, which strengthens the claim that the early-layer profile is a general representational pattern rather than an isolated behavior of the three primary 3B models. Second, scale alone does not fully determine the strength or shape of the effect. The Llama variants improve fairly consistently with size, but the Qwen variants do not follow a simple monotonic ordering on every task. This suggests that model family, training recipe,

and representational organization also matter.

The paper therefore does not support a claim that larger decoder-only models necessarily encode frequency better in every respect. It supports a more limited conclusion: across a diverse set of decoder-only models, frequency is strongly decodable, and the strongest relative high-versus-low separation tends to occur early.

5.4 What the controls support, and what they do not

The control analyses strengthen the main result, but they also define its scope. Strong performance under token-restricted, reduced-confound, multi-seed, linear-probe, pooling, and regression settings makes it less likely that the observed pattern is driven by a single tokenization artifact, a particular random seed, or an overly expressive nonlinear probe. At the same time, the controls are not equally diagnostic. The reduced-confound subset does not provide perfect lexical matching, the single-token results are available in fully symmetric form across all variants as shown in section 9, and some auxiliary analyses are limited to sampled anchor layers.

The correct interpretation is therefore robust decodability under multiple checks, not complete elimination of every alternative explanation. This is consistent with the broader probing literature, which has repeatedly emphasized that decodability should not be equated with functional use (Hewitt and Manning, 2019; Voita and Titov, 2020; Belinkov, 2022). Our results show that decoder-only representations contain substantial frequency-related information and that this information has a clear layer-wise profile. They do not by themselves establish how that signal is used during model computation.

5.5 Psycholinguistic relevance

From a psycholinguistic perspective, the findings are still informative. Word frequency is one of the most reliable predictors of lexical access and lexical decision performance in human studies (Forster and Chambers, 1973; Balota et al., 2004; Brysbaert et al., 2011). The fact that decoder-only models encode a strong, graded, and early-emerging frequency signal suggests that a basic lexical-statistical variable remains highly visible in their internal states.

At the same time, prior psycholinguistic evaluations have shown that strong model performance should not be taken as evidence of human-like processing without direct testing (Ettinger, 2020; Oh and Schuler, 2023). Our results should therefore be read as a representational finding: decoder-only models preserve frequency information in a structured and measurable way, but the relation between that signal and human lexical processing remains an open question.

6 Conclusion

We investigated how word frequency is represented across layers in decoder-only language models using a controlled probing framework over a nine-model sweep spanning the Llama, Qwen, and Smol families. Three findings emerge. First, frequency is strongly decodable from frozen hidden states across model families and scales. Second, the strongest *relative* separation between high- and low-frequency words appears in the earliest layers and decreases with depth. Third, frequency remains decodable throughout the network, as shown by strong direct classification and continuous regression results in middle and late layers. Together, these results show that decoder-only language models encode a robust and graded representation of word frequency, with the clearest early-layer signature appearing as a disproportionate advantage for high-frequency words.

The control analyses strengthen this interpretation while also defining its scope. The main pattern survives token-restricted, reduced-confound, multi-seed, linear-probe, pooling, and regression settings, which makes a narrow artifact-based account unlikely. At the same time, this remains a probing study. It shows what information is recoverable from representations, not how that information is functionally used during model computation. A natural next step is therefore to combine layer-wise probing with causal intervention methods, richer contextual stimuli, and multilingual settings. This would help determine whether the early-layer frequency signature identified here reflects a broader property of lexical representation in decoder-only models or a narrower feature of the specific models and stimuli studied here.

Our **`Code & Data`** is available here for research community.

7 Limitations

This study has four main limitations.

First, the analysis is based on probing. Probes

show what information is recoverable from frozen representations, but they do not establish how that information is used during model computation. Our results therefore support claims about representational accessibility, not about causal mechanisms inside the models.

Second, the experiments use isolated lexical stimuli drawn from English lexical decision resources. This design is appropriate for studying word frequency as a controlled lexical variable, but it does not capture the full contextual setting in which decoder-only language models are typically used. Frequency effects in sentence context may interact with syntactic, semantic, and discourse information in ways that are not measured here.

Third, although the model sweep is broader than the three primary 3B models emphasized in the paper, it is still limited in scope. We study nine decoder-only variants from three families and focus on English. The extent to which the same layer-wise pattern holds for other architectures, multilingual models, or substantially larger instruction-tuned systems remains an open question.

Fourth, the control analyses strengthen the main conclusion but do not remove every possible alternative explanation. The reduced-confound subset narrows lexical-form differences, but it does not produce perfect equality across all correlated variables. In addition, some auxiliary analyses are narrower than the main layer sweep: the multi-seed stability and selectivity analyses are computed on sampled anchor layers rather than every layer. For this reason, the strongest claim supported by the evidence is a consistent representational pattern across the available models and controls, not an exhaustive account of every source of variation.

These limitations do not change the central empirical result of the paper. They define its scope. Our findings show that word frequency is strongly decodable in decoder-only language models and that its strongest relative separation appears early in the network, but broader causal and cross-linguistic generalization remains for future work.

8 Use of Large Language Models (LLMs)

Researchers used LLM for refining and polishing the manuscript.

9 Reproducibility Statement

We provide source code, preprocessing pipelines, reconstruction scripts, evaluation workflows, fixed random seeds, dependency versions, and model hyperparameters to support reproducibility. Experiments were conducted with Python 3.12.9 on a single NVIDIA RTX 4090 GPU with 24GB memory and it taken 6hrs to complete one framework.

References

Alan Agresti. 2013. *Categorical data analysis*. John Wiley & Sons.

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, and 4 others. 2025. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3.

David A Balota, Michael J Cortese, Susan D Sergent-Marshall, Daniel H Spieler, and Melvin J Yap. 2004. Visual word recognition of single-syllable words. *Journal of experimental psychology: General*, 133(2):283.

David A Balota, Melvin J Yap, Keith A Hutchison, Michael J Cortese, Brett Kessler, Bjorn Loftis, James H Neely, Douglas L Nelson, Greg B Simpson, and Rebecca Treiman. 2007. The english lexicon project. *Behavior research methods*, 39(3):445–459.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Marc Brysbaert, Emmanuel Keuleers, and Boris New. 2011. Assessing the usefulness of google books' word frequencies for psycholinguistic research on word processing. *Frontiers in Psychology*, 2:27.

Nitay Calderon and Roi Reichart. 2025. On behalf of the stakeholders: Trends in nlp model interpretability in the era of llms. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human*

Language Technologies (Volume 1: Long Papers), pages 656–693.

Thomas Hikaru Clark, Carlos Arriaga, Javier Conde, Gonzalo Martínez, and Pedro Reviriego. 2026. To words and beyond: Probing large language models for sentence-level psycholinguistic norms of memorability and reading times. *arXiv preprint arXiv:2603.12105*.

Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences*. routledge.

Allyson Ettinger. 2020. What bert is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Kenneth I Forster and Susan M Chambers. 1973. Lexical access and naming time. *Journal of verbal learning and verbal behavior*, 12(6):627–635.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Xavier Glorot and Yoshua Bengio. 2010. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*, pages 1026–1034.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of EMNLP-IJCNLP*, pages 2733–2743. ACL.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word 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 4129–4138.

Binyuan Hui, Jian Yang, Zeyu Cui, Jiaxi Yang, Dayiheng Liu, Lei Zhang, Tianyu Liu, Jiajun Zhang, Bowen Yu, Keming Lu, and 1 others. 2024. Qwen2. 5-coder technical report. *arXiv preprint arXiv:2409.12186*.

Sergey Ioffe and Christian Szegedy. 2015. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International conference on machine learning*, pages 448–456. pmlr.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3651–3657.

Daniël Lakens. 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and anovas. *Frontiers in psychology*, 4:62627.

Michael Li and Nishant Subramani. 2025. Echoes of bert: Do modern language models rediscover the classical nlp pipeline? *arXiv preprint arXiv:2506.02132*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988.

Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

Byung-Doh Oh and William Schuler. 2023. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? *Transactions of the Association for Computational Linguistics*, 11:336–350.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Sean Trott. 2024. Can large language models help augment english psycholinguistic datasets? *Behavior Research Methods*, 56(6):6082–6100.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

Ivan Vulić, Edoardo Maria Ponti, Robert Litschko, Goran Glavaš, and Anna Korhonen. 2020. Probing pretrained language models for lexical semantics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7222–7240.

Supplementary Material

This appendix provides compact numerical summaries for the full nine-model sweep and the auxiliary control analyses. The main paper focuses on the most important cross-model patterns and the primary 3B comparisons, while the appendix

records the broader evidence used to assess robustness and interpretive scope.

All per-layer numerical outputs are provided in the companion supplementary spreadsheets: **Full metrics.csv**, **direct_freq_binary.csv**, **direct_freq_3way.csv**, **freq_regression.csv**, **single_token_ldt.csv**, **single_token_ldt.csv**, **token_matched_ldt.csv**, **confound_matched_probe.csv**, **stability_stats.csv**, **selectivity_linear_probe.csv**, **selectivity_shuffled_label.csv**, **selectivity_selectivity.csv**, **pooling_ablation.csv**, and **balance_table.csv**.

A. Full cross-model task summary

Table S1 reports the main task-level summary for all nine model variants. It complements the main text by making the full family- and scale-level comparisons explicit in one place.

Table S2 summarizes the continuous regression results. These values are included separately because they capture the graded nature of the frequency signal more directly than the classification tasks.

B. Layer-wise frequency-effect summary

Table S3 records the key layer-wise statistics behind the main claim of the paper. For each model, it reports the high-minus-low accuracy gap at the first and final layers, the layer at which that gap peaks, and the analogous summary for Cohen's h.

The full per-layer trajectories underlying Table S3 are contained in `Full metrics.xlsx`. The compact summary above is designed to make the early-layer concentration of the effect easy to verify without reproducing all per-layer tables in print.

C. Robustness and control analyses

Table S4 collects the main auxiliary controls across the nine-model sweep. These include single-token restriction, token-count-reduced lexical-decision controls, reduced-confound matched probing, multi-seed stability, and probe selectivity.

Two details are important for interpretation. First, the single-token workbook does not provide subgroup frequency-gap values for *Qwen3-4B-Base* or *Qwen3-8B-Base*, although it does provide single-token accuracy for those models. Second, *Llama-3.1-8B* is not present in the single-token workbook. For this reason, the single-token control is used as supportive evidence for decodability under tokenization restriction, but not as the primary basis for family-wide gap comparisons.

D. Pooling ablation

Table S5 compares last-token extraction, which is used in the main experiments, with mean pooling over non-padding positions. The differences are small across the full model sweep, and the main frequency result is not sensitive to this choice.

E. Balance diagnostics for the reduced-confound control

Table S6 reports compact diagnostics for the reduced-confound matching analysis. Matching equalizes group size within each model, but residual lexical differences remain after matching. This is why the matched analysis is treated as supportive rather than primary evidence in the main paper.

Across models, the post-matching residual gaps are similar in magnitude. Mean word-length differences remain about 1.75 characters, and mean token-count differences remain roughly 0.79 to 0.83 tokens. These diagnostics are included to make the scope of the matched control explicit.

F. Brief Representation of all tested models

Individual-level results across variants are shown for low and high-frequency items, with statistical validation. Figures 4 through 12 present results per model, while Figure 13 summarizes across all models.

F. Supplementary scope note

The appendix is designed to complement the main paper, not to restate it. The main paper reports the core nine-model trends and the primary 3B comparisons. The appendix records the broader numerical evidence behind those claims, including variant-level summaries and control-specific caveats. This division keeps the main paper compact while allowing reviewers to verify the full empirical basis of the study.

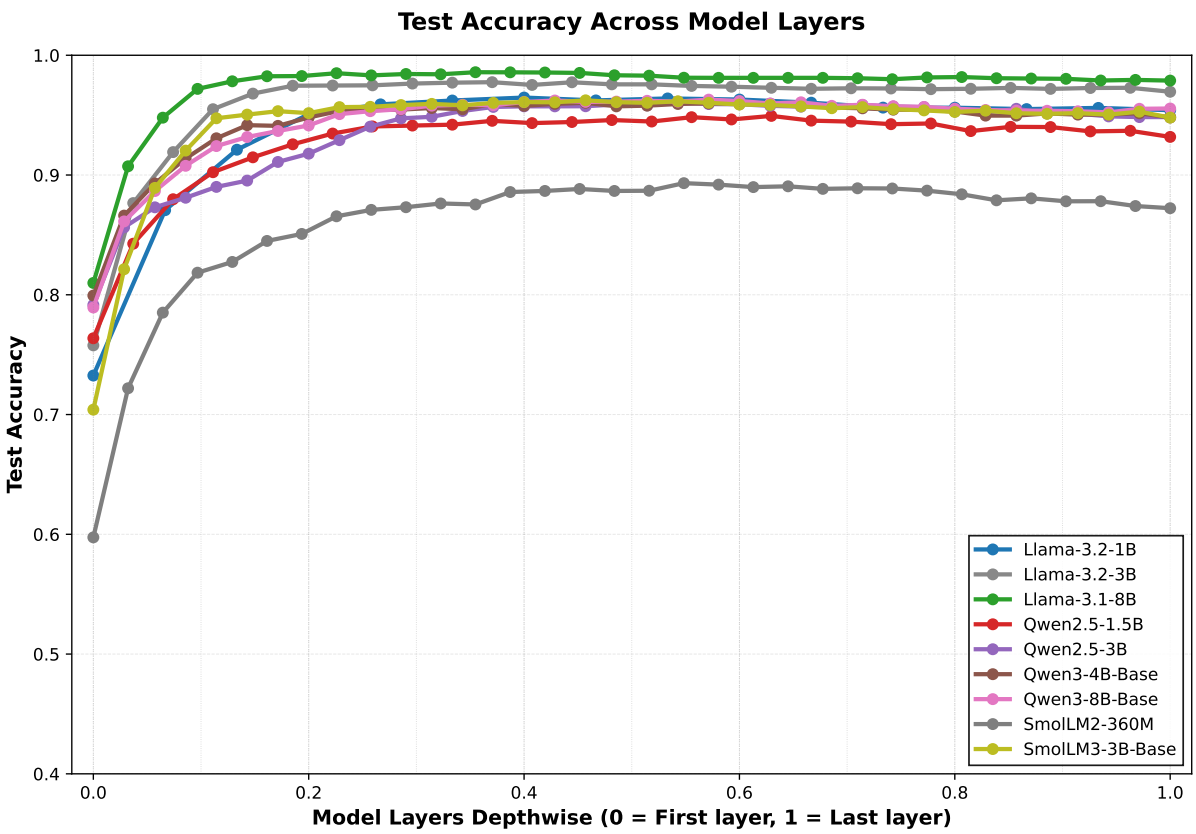

Figure 3: Held-out lexical decision accuracy across normalized depth for the full nine-model sweep. All variants show rapid early gains, followed by high and relatively stable accuracy in middle and late layers.

Table S1: Main task-level summary for the full nine-model sweep. Lex. denotes lexical decision, and Bin. denotes the binary high-versus-low frequency probe.

Model	Layers	Lex. peak layer	Lex. peak acc.	Lex. final acc.	Bin. peak layer	Bin. peak acc.	3-way peak layer	3-way peak macro-F1
Llama-3.2-1B	16	6	0.965	0.954	9	0.966	6	0.769
Llama-3.2-3B	28	10	0.978	0.970	7	0.970	8	0.761
Llama-3.1-8B	32	11	0.986	0.979	11	0.975	13	0.799
Qwen2.5-1.5B	28	17	0.949	0.932	17	0.959	15	0.745
Qwen2.5-3B	36	19	0.961	0.948	25	0.957	21	0.750
Qwen3-4B-Base	36	16	0.960	0.948	18	0.958	16	0.796
Qwen3-8B-Base	36	20	0.963	0.956	22	0.954	22	0.772
SmolLM2-360M	32	17	0.893	0.872	18	0.921	21	0.693
SmolLM3-3B-Base	36	16	0.962	0.948	22	0.961	19	0.763

Table S2: Continuous regression summary for the full nine-model sweep. The final column reports the number of lexical-decision layers for which the high-versus-low subgroup difference remains significant after Benjamini–Hochberg correction.

Model	Best R^2 layer	Best R^2	Best Spearman ρ	Significant lexical-decision layers
Llama-3.2-1B	7	0.837	0.918	16
Llama-3.2-3B	14	0.862	0.929	28
Llama-3.1-8B	18	0.873	0.935	32
Qwen2.5-1.5B	16	0.778	0.888	28
Qwen2.5-3B	19	0.811	0.902	36
Qwen3-4B-Base	20	0.789	0.895	36
Qwen3-8B-Base	20	0.805	0.906	36
SmolLM2-360M	21	0.673	0.821	32
SmolLM3-3B-Base	21	0.829	0.913	36

Table S3: Layer-wise frequency-effect summary across the nine model variants. Δ denotes the high-minus-low subgroup accuracy difference from the lexical decision setting, and h denotes Cohen's h.

Model	First Δ	Final Δ	Peak-Δ layer	Peak Δ	First h	Final h	Peak-h layer	Peak h
Llama-3.2-1B	0.349	0.092	0	0.349	0.801	0.450	0	0.801
Llama-3.2-3B	0.350	0.067	0	0.350	0.810	0.363	0	0.810
Llama-3.1-8B	0.440	0.053	0	0.440	1.008	0.317	0	1.008
Qwen2.5-1.5B	0.293	0.138	1	0.308	0.696	0.555	1	0.803
Qwen2.5-3B	0.331	0.097	0	0.331	0.809	0.426	0	0.809
Qwen3-4B-Base	0.389	0.131	0	0.389	0.874	0.480	0	0.874
Qwen3-8B-Base	0.360	0.106	0	0.360	0.819	0.436	0	0.819
SmolLM2-360M	0.102	0.199	1	0.303	0.206	0.599	5	0.709
SmolLM3-3B-Base	0.291	0.089	0	0.291	0.637	0.383	1	0.730

Table S4: Robustness and control summary for the nine-model sweep. Stability and selectivity are computed on sampled anchor layers rather than on the full layer sweep.

Model	Single-token peak acc.	Token-matched peak acc.	Token-matched Δ	Confound-matched peak acc.	Stability F1 $\pm$ std	Linear F1	Shuffle F1	Peak selectivity
Llama-3.2-1B	0.954	0.963	0.171	0.965	0.962 ± 0.002	0.959	0.488	0.493
Llama-3.2-3B	0.963	0.977	0.167	0.968	0.976 ± 0.001	0.965	0.506	0.504
Llama-3.1-8B	0.988	0.984	0.207	0.979	0.986 ± 0.001	0.982	0.624	0.545
Qwen2.5-1.5B	0.930	0.951	0.278	0.956	0.946 ± 0.001	0.944	0.644	0.558
Qwen2.5-3B	0.954	0.962	0.186	0.956	0.957 ± 0.002	0.954	0.520	0.532
Qwen3-4B-Base	0.991	0.956	0.185	0.958	0.959 ± 0.001	0.952	0.574	0.565
Qwen3-8B-Base	0.980	0.959	0.201	0.957	0.963 ± 0.002	0.949	0.599	0.537
SmolLM2-360M	0.920	0.896	0.181	0.920	0.886 ± 0.004	0.892	0.518	0.484
SmolLM3-3B-Base	0.939	0.961	0.187	0.963	0.962 ± 0.001	0.959	0.556	0.510

Table S5: Pooling ablation across the nine model variants. The main paper uses last-token extraction; mean pooling is reported here as a robustness check.

Model	Pooling	Best layer	Accuracy	F1	Δ	Cohen's h
Llama-3.2-1B	Mean pool	7	0.963	0.963	0.081	0.472
Llama-3.2-1B	Last token	6	0.965	0.965	0.080	0.485
Llama-3.2-3B	Mean pool	12	0.976	0.976	0.051	0.345
Llama-3.2-3B	Last token	10	0.978	0.977	0.058	0.397
Llama-3.1-8B	Mean pool	14	0.985	0.985	0.028	0.201
Llama-3.1-8B	Last token	11	0.986	0.986	0.039	0.279
Qwen2.5-1.5B	Mean pool	13	0.947	0.947	0.106	0.490
Qwen2.5-1.5B	Last token	17	0.949	0.949	0.120	0.521
Qwen2.5-3B	Mean pool	25	0.960	0.961	0.068	0.365
Qwen2.5-3B	Last token	19	0.961	0.961	0.078	0.436
Qwen3-4B-Base	Mean pool	22	0.962	0.962	0.095	0.484
Qwen3-4B-Base	Last token	16	0.960	0.960	0.113	0.491
Qwen3-8B-Base	Mean pool	23	0.962	0.962	0.082	0.391
Qwen3-8B-Base	Last token	20	0.963	0.963	0.080	0.370
SmolLM2-360M	Mean pool	15	0.897	0.896	0.207	0.689
SmolLM2-360M	Last token	17	0.893	0.892	0.183	0.580
SmolLM3-3B-Base	Mean pool	13	0.964	0.964	0.067	0.348
SmolLM3-3B-Base	Last token	16	0.962	0.962	0.076	0.437

Table S6: Compact balance diagnostics after reduced-confound matching. Group sizes are equalized within each model, but length and token-count differences remain, so this analysis should be interpreted as reduced-confound rather than fully matched.

Model	High n after	Low n after	High length mean	Low length mean	High token mean	Low token mean
Llama-3.2-1B	13391	13391	7.181	8.926	1.848	2.669
Llama-3.2-3B	13391	13391	7.181	8.926	1.848	2.669
Llama-3.1-8B	10393	10393	7.177	8.923	1.851	2.668
Qwen2.5-1.5B	13391	13391	7.181	8.926	1.851	2.677
Qwen2.5-3B	13391	13391	7.181	8.926	1.851	2.677
Qwen3-4B-Base	10393	10393	7.177	8.923	1.855	2.676
Qwen3-8B-Base	10393	10393	7.177	8.923	1.855	2.676
SmolLM2-360M	13391	13391	7.181	8.926	1.928	2.716
SmolLM3-3B-Base	13391	13391	7.181	8.926	1.848	2.669

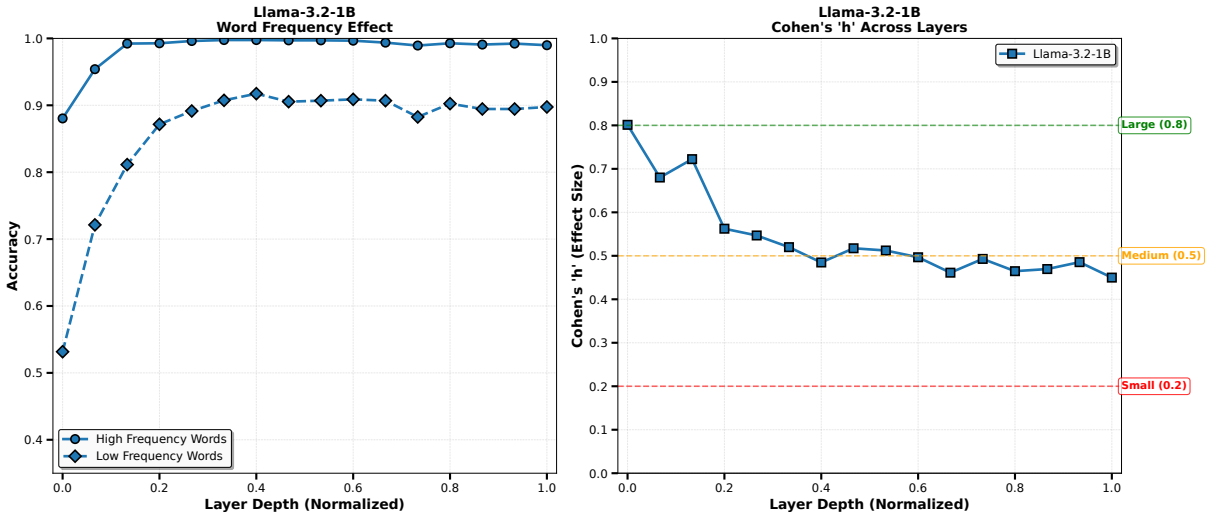

Figure 4: Layer-wise frequency effects for Llama-3.2-1B model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

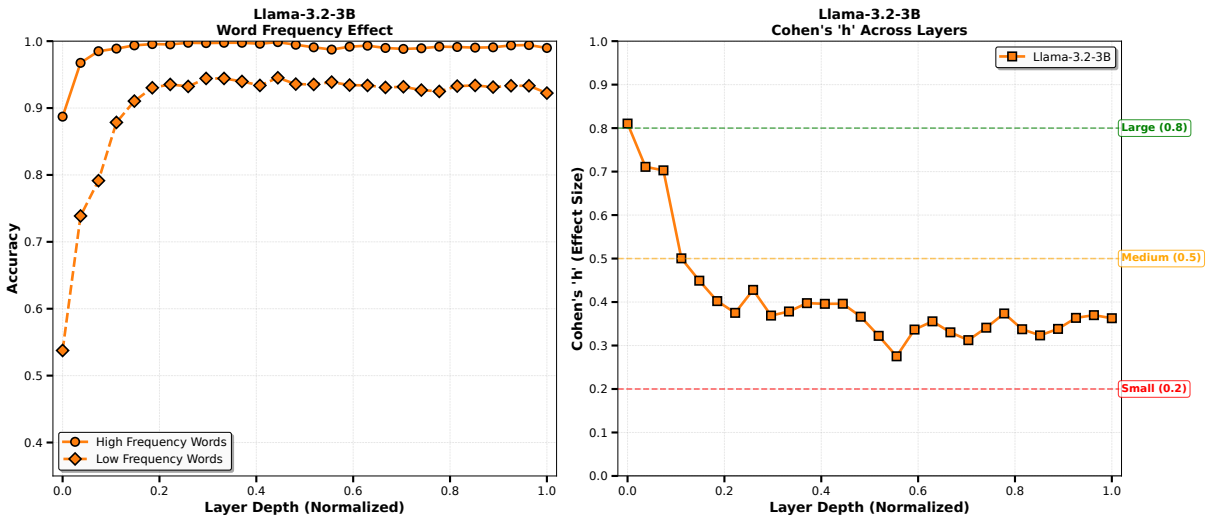

Figure 5: Layer-wise frequency effects for Llama-3.2-3B model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

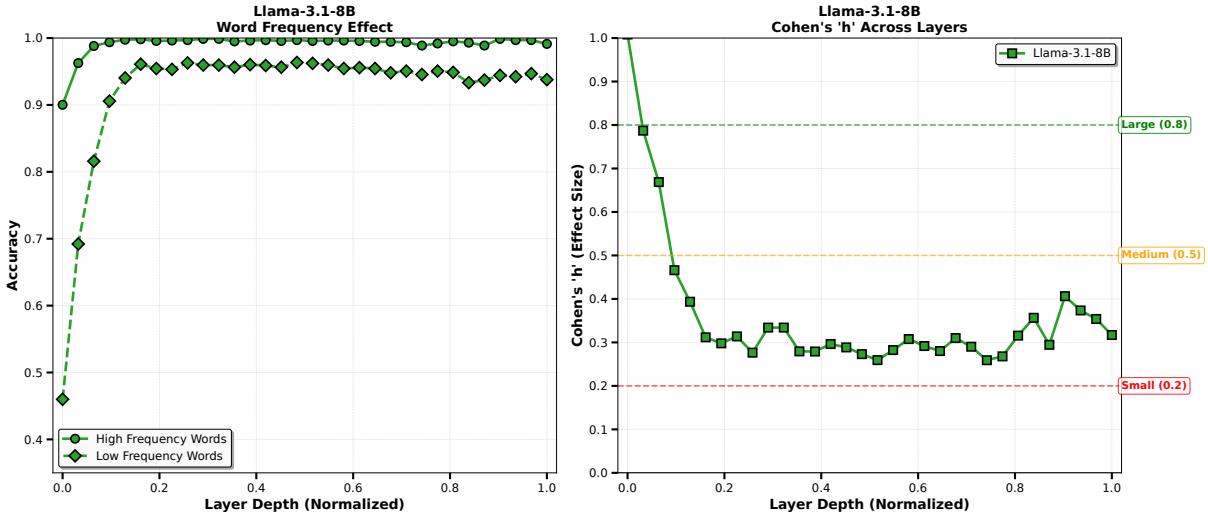

Figure 6: Layer-wise frequency effects for Llama-3.1-8B model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

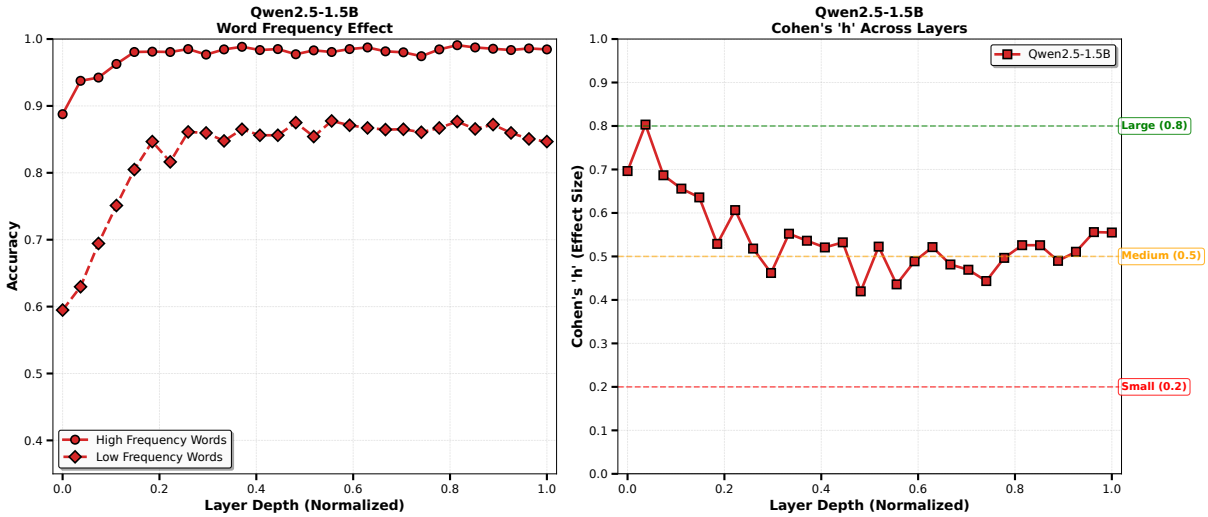

Figure 7: Layer-wise frequency effects for Qwen2.5-1.5B model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

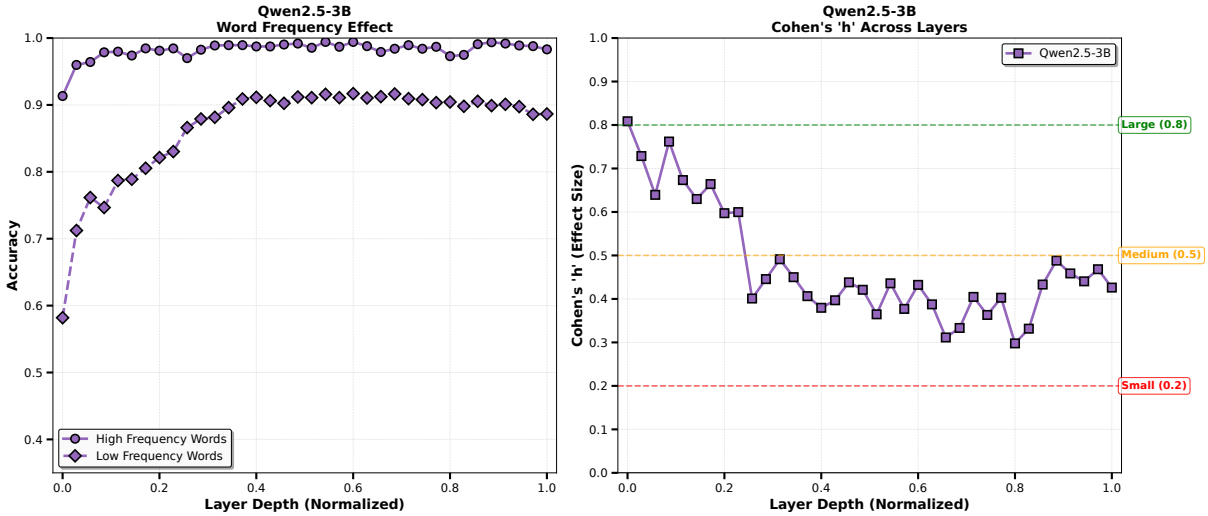

Figure 8: Layer-wise frequency effects for Qwen2.5-3B model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

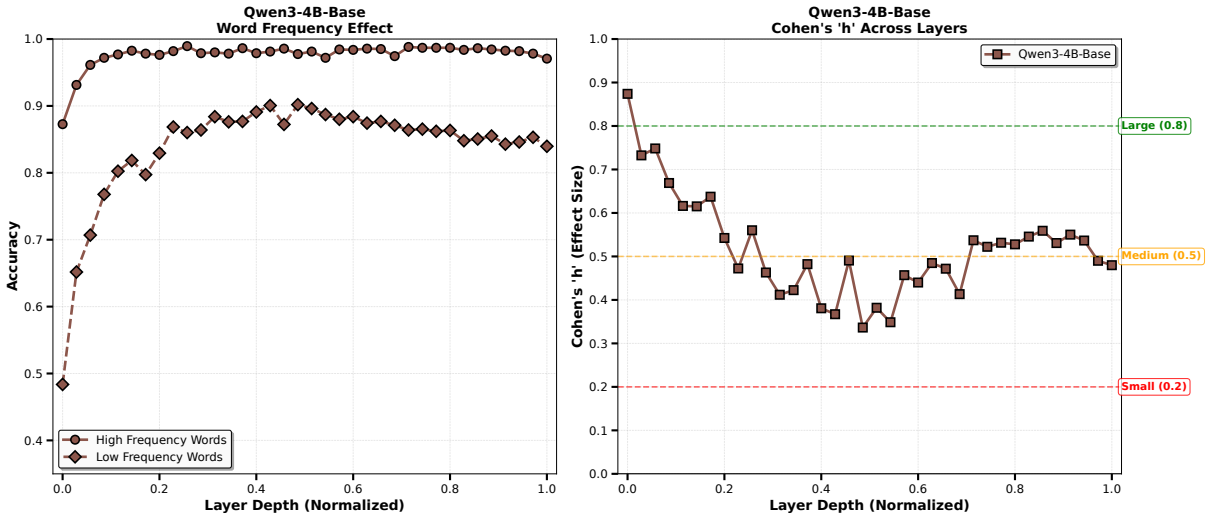

Figure 9: Layer-wise frequency effects for Qwen3-4B-Base model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

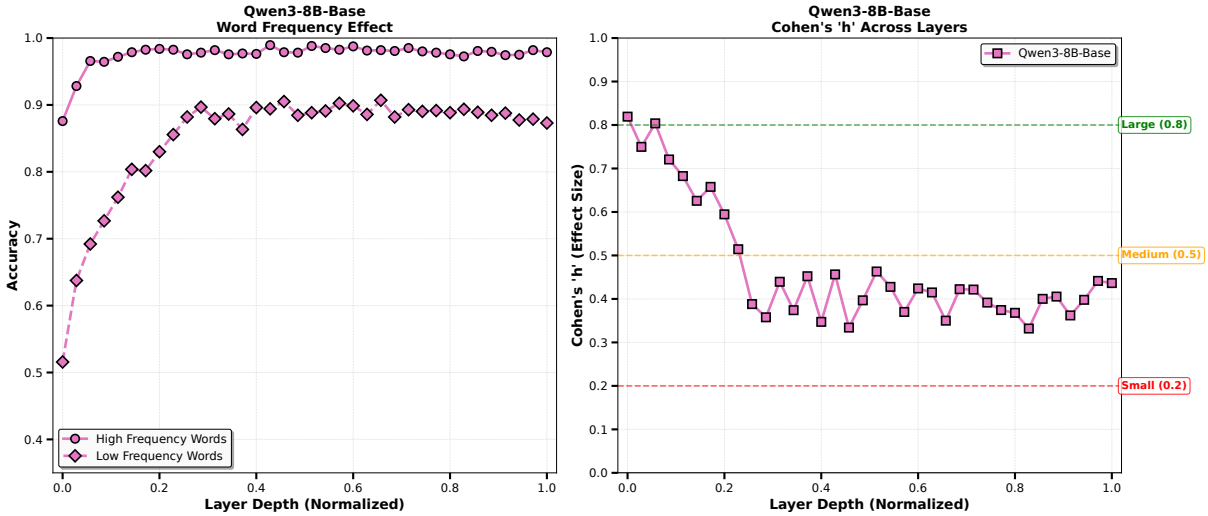

Figure 10: Layer-wise frequency effects for Qwen3-8B-Base model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

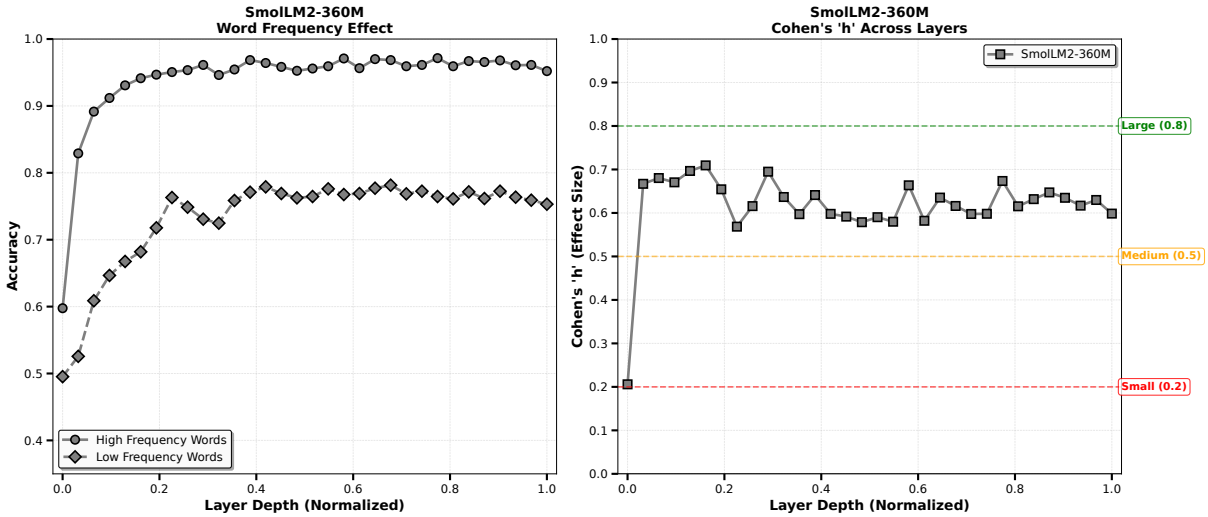

Figure 11: Layer-wise frequency effects for SmolLM2-360M model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

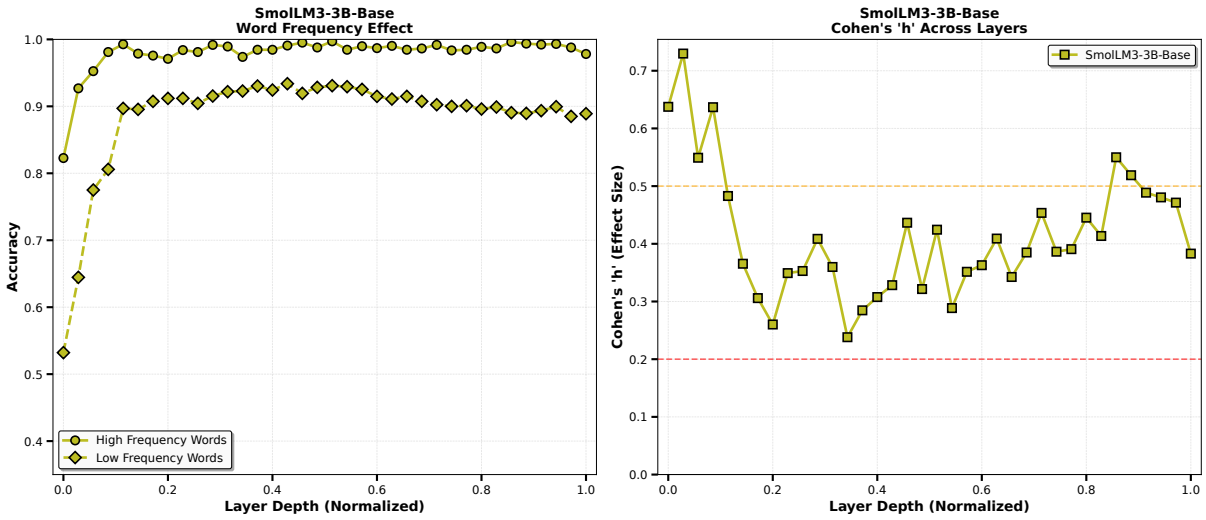

Figure 12: Layer-wise frequency effects for SmolLM3-3B-Base model. Left: subgroup accuracy trajectories across normalized depth.Solid line shows high-frequency words; dashed line shows low-frequency words across the models. Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.

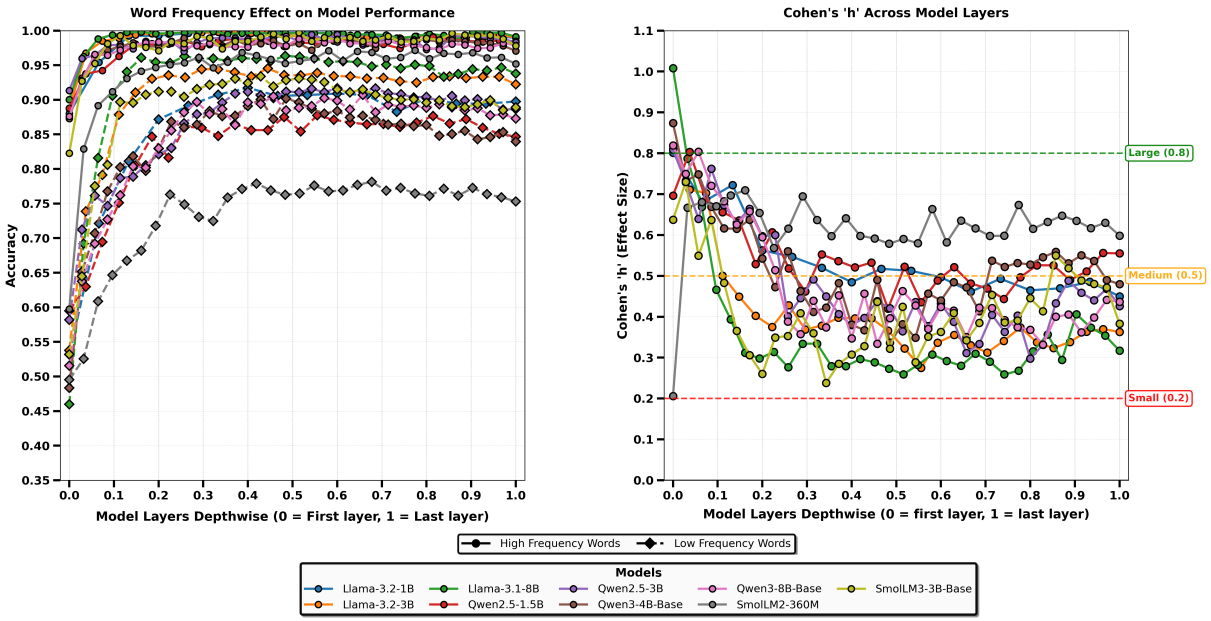

Figure 13: Layer-wise frequency effects for the full nine-model sweep. Left: subgroup accuracy trajectories across normalized depth.Solid Line shows high and Dashed line presents low frequency words across the models.Right: Cohen's h across layers with conventional reference lines. Across the full sweep, high-frequency words show their largest relative advantage in early layers, and that advantage decreases with depth while remaining positive.