CALIBRATION-GUIDED STATIC VOCABULARY HEAD PRUNING FOR EDGE LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern language models often employ a large vocabulary to cover multiple languages and modalities, which makes their deployment expensive. To address this challenge, we study *static* vocabulary-head pruning that permanently restricts the head to a survival subset $\mathcal{S} \subseteq \mathcal{V}$ chosen once, offline, with *no* post-pruning fine-tuning. The de-facto criterion for this choice is token *frequency*, and we first establish why frequency is the wrong signal along three axes: a participation-frequency mismatch, blindness to substitutability, and the loss of rare-but-critical tokens. We introduce PASS (**P**articipation **A**nd **S**ubstitution **S**election), a calibration-guided, training-free method built on two complementary per-token signals: *participation* which measures the rate at which a token enters the model's top-K at decode time; and *substitution cost* which is a mass-weighted logit deficit incurred by replacing a token with its nearest surviving neighbour. We then propose a two-stage selector that first uses participation to tag tokens as keep, prune or uncertain decisions, and then uses substitution cost to rank only the remaining uncertain tokens. PASS reduces the decode-time head cost from $\mathcal{O}(Vd)$ to $\mathcal{O}(|\mathcal{S}|d)$ per token and at a 50% keep fraction, retains nearly entire of full-head quality on multiple evaluation benchmarks as GSM8K, TrivaQA, MATH-500, *etc.*

1 INTRODUCTION

Every token generated from a large language model (LLM) requires one dense matrix-vector product against the full embedding matrix $W \in \mathbb{R}^{V \times d}$ (`LM-Head`). As modern vocabularies expand to accommodate multilingual tokenization, code, and multi-modal code-books, this component has become one of the dominant parameter blocks in the model. For edge hardware deployment with limited batch size inference, memory bandwidth and not compute becomes a key bottleneck dominating wall-clock time and power. Each decoded token requires memory read of full `LM-Head`, and this single matrix-vector product can compete with entire transformer backbone per-token cost.

The situation deteriorates for *large-vocabulary small models*. Recent models with multilingual tokenizers can push vocabulary into the 150K–300K range (Gemma Team, 2025; Xu et al., 2025; Yang et al., 2025; Tao et al., 2024), and a unified multimodal pipeline can further stack the text vocabulary with one or more visual codebooks within a single head (Wu et al., 2025a;b; Chameleon, 2024; Sun et al., 2024). These trends scale vocabulary independently of hidden dimension, and the output projection can consume upto *one-third* of total parameters and $\sim 20\%$ wall-clock per-token decode latency (Figure 1). We emphasize that tokens generated in practical edge settings are bounded by several factors such as user locale, applications and task served. Hence, the `LM-Head` is provisioned for a vocabulary far larger than the one the model actually exercises and a substantial fraction of rows never once enters the model's top-K at decode time. We call this as **Head Slack**: an unused capacity in W that is paid for on every generated token, regardless of whether the corresponding rows ever compete to be within top-K during decoding.

To address *Head Slack*, in this work we explore *static vocabulary-head pruning* which can permanently restrict the head to a survival set $\mathcal{S} \subseteq \mathcal{V}$ chosen offline once to ship a smaller head and pay $|\mathcal{S}|/|\mathcal{V}|$ of the original per-token bandwidth cost. Given that $\mathcal{S}$ is committed up front, the entire premise of our work reduces to a single question: ***which tokens end up in $\mathcal{S}$?*** A default choice widely adopted across prior works (Goel et al., 2025; Ushio et al., 2023a; Jean et al., 2015; Mi et al., 2016) is **frequency** that tokens are ranked by how often they occur in a representative calibration

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

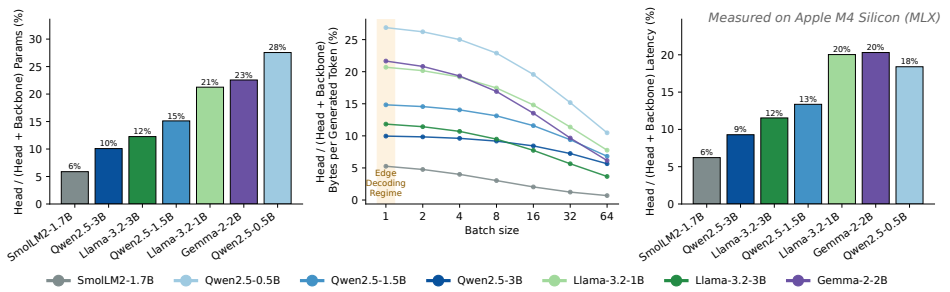

Figure 1: **Output Head Projection is Expensive for Edge Inference.** (a) *Parameter Share*: LM head parameter share across seven public models: the head consumes 6-28% of non-embedding parameters, with share growing as V outpaces d; (b) *Bandwidth Share*: An analytical estimate of head share of bytes moved per generated token vs. batch size (with S=2048). At batch=1 in conventional *edge decoding regime*, the head accounts for up to *27%* of memory traffic; (c) *Latency Share*: Measured decode-step latency share on Apple M4 Silicon (MLX, fp16, batch=1), the head consumes 6-20% of wall-clock time per token decoded.

corpus and the rare tail is dropped. We identified three issues associated with corpus frequency: ❶ *Participation-Frequency Mismatch*: frequency measures how often a token appears in text but Head Slack depends on how often the model competes to generate it; ❷ *Substitution blindness*: two tokens with identical corpus counts can differ significantly depending on whether the rest of the head can absorb their behavior; ❸ *Rare-but-critical tokens*: byte fallbacks, format markers, low-resource-language pieces, and code/math symbols are rare yet load-bearing which are blindly overlooked by raw frequency.

Our Proposed Method: We propose **PASS** (**P**articipation **A**nd **S**ubstitution **S**election), a calibration-guided, training-free static pruning method built on two complementary per-token signals combined via a two-stage selector mechanism (Figure 5). First, the **Participation Signal** denoted as π_i (Section 4.1) is the fraction of calibration positions at which token i enters the top-K and measures decode-time competition across tokens rather than surface frequency. Second, the **Substitution cost** denoted as $\mathrm{cost}(i)$ (Section 4.2) is the mass-weighted logit deficit of replacing token i with its nearest surviving neighbor which favors tokens having high decode-time mass and lack a close substitute. We combine these two via a two-stage selector (Section 4.3) where the computationally cheap signal (π_i) is used to identifying confident ends of the vocabulary followed by second expensive signal to rank ambiguous middle range for per-token keep/prune decision. Our key contributions are:

- **Diagnostic Foundation:** We formalize *Head Slack* as a key inference bottleneck of large-vocabulary small models and empirically diagnose the three failure concerns with frequency-based static pruning (Section 3) across multiple model families.
- **Method Innovation:** We introduce PASS, a training-free static pruning framework built on two calibration-derived signals (participation and substitution cost) combined by a two-stage selector that generates per-token provenance (auto-keep, mid-keep, mid-prune, auto-prune) with a single calibration pass (Section 4).
- **Theoretical Innovation:** We prove that any token with zero calibration participation (π_i=0) admits an exact greedy-invariance certificate (Proposition 1) on the calibration distribution. It implies removing the entire zero-participation set leaves the greedy decode unchanged at every calibration position with no fine-tuning requirements.
- **Empirical Evidence:** Across three models (Qwen2.5-0.5B, 1.5B and Gemma2-2B), we empirically establish that:❶ both participation (π) and substitution ($\mathrm{cost}(i)$) signals outperform conventional frequency signal; ❷ π_i and $\mathrm{cost}(i)$ provide complementary information which benefits PASS; ❸ PASS can achieve $\sim 2\times$ (B200 GPU, Apple Silicon) and $\sim$ 6-8$\times$ (CPU) speedup on `LM-Head` with 50% compression; ❹ PASS significantly outperforms frequency-based baselines on multiple downstream benchmarks (GSM8K, MATH, HumanEval, etc.).

2 BACKGROUND AND NOTATION

Output projection in LLMs: Let $\mathcal{V} = \{1, \ldots, V\}$ be the vocabulary and d the hidden dimension. Let $W \in \mathbb{R}^{V \times d}$ be the output projection with i-th row $\boldsymbol{w}_i$. In this work, all checkpoints evaluated have zero output bias, so for a hidden state $\boldsymbol{h} \in \mathbb{R}^d$ the logit and next-token distribution is given as:

$$\ell_i(\boldsymbol{h}) = \boldsymbol{w}_i^\top \boldsymbol{h}, \qquad p^\star(\boldsymbol{h})_i = \frac{\exp \ell_i(\boldsymbol{h})}{\sum_{k=1}^{V} \exp \ell_k(\boldsymbol{h})}, \quad i \in \mathcal{V}. \tag{1}$$

We write Top-$K(W\boldsymbol{h})$ for the K indices of largest logit at $\boldsymbol{h}$ and $i^\star(\boldsymbol{h}) = \arg\max_i \ell_i(\boldsymbol{h})$ for the argmax value. Given a survival set $\mathcal{S} \subseteq \mathcal{V}$ with keep fraction $f = |\mathcal{S}|/|\mathcal{V}|$ after pruning, the pruned head renormalizes over $\mathcal{S}$ as follows:

$$p^{\mathcal{S}}(\boldsymbol{h})_i = \begin{cases} \dfrac{\exp \ell_i(\boldsymbol{h})}{\sum_{k \in \mathcal{S}} \exp \ell_k(\boldsymbol{h})} & i \in \mathcal{S}, \\ 0 & i \notin \mathcal{S}. \end{cases} \tag{2}$$

Calibration corpus and Hidden-state set: We assume access to a calibration corpus of C tokens drawn from the domain representative of the target deployment. Streaming this corpus through the model produces a set of final-layer hidden states $\mathcal{H} = \{\boldsymbol{h}_j\}_{j=1}^{N}$. Since each corpus token contributes one hidden state ($C = N$), we write C and N interchangeably and each signal within this work: corpus frequency (Section 3), participation frequency (Section 4.1), and substitution cost (Section 4.2), is a statistic estimated over this corpus.

Held-out evaluation and Distributional cost: In this work, we evaluate any survival set $\mathcal{S}$ on a disjoint held-out corpus $\mathcal{H}_{\text{test}}$ drawn from the same domain as $\mathcal{H}$. The distributional cost of pruning to $\mathcal{S}$ at $\boldsymbol{h} \in \mathcal{H}_{\text{test}}$ is the negative log of the probability mass $\mathcal{S}$ retains,

$$\Delta(\boldsymbol{h}; \mathcal{S}) \;=\; -\log Z_{\mathcal{S}}(\boldsymbol{h}), \qquad Z_{\mathcal{S}}(\boldsymbol{h}) \;:=\; \sum_{i \in \mathcal{S}} p^\star(\boldsymbol{h})_i. \tag{3}$$

Note that Δ is zero when nothing is lost and grows as $\mathcal{S}$ discards mass due to pruning. We choose to summarize Δ over $\mathcal{H}_{\text{test}}$ by percentiles and track P_{99} since a simple averaging mask out the rare positions at which pruning notably changes the output.

3 BEYOND FREQUENCY: WHY FREQUENCY-BASED PRUNING FAILS

As we discussed in Section 1, the key question we attempt to answer in this work remains "which tokens can be dropped?". A classical de-facto recipe is domain-dependent corpus frequency of tokens that keep the frequent tokens and drop the cold tail. We formally define this baseline as:

Given a calibration corpus of C tokens, the *corpus frequency* of token i is given by:

$$\text{freq}_i(C) \;=\; \frac{1}{C} \sum_{t=1}^{C} \mathbb{1}[\, x_t = i \,] \;\in\; [0, 1], \tag{4}$$

Intutively, it is the fraction of corpus positions at which i appears as the observed next-token label. With k_f as the Keep-Fraction, frequency-based static pruning baseline keeps the Top-$(k_f \cdot V)$ tokens ranked by $\text{freq}_i(C)$ and prunes the rest (rank 1 = most frequent).

► **[Failure Mode I] The Participation-Frequency Mismatch:** *Does how often a token appears in text predict how often the model competes to generate it?*

Frequency pruning ranks tokens by their count in the calibration corpus. However, cost of `LM-Head` vector-matrix multiplication associated with V tokens is paid at decode time where what matters remain if a token is a serious next-token *candidate* and if it can enter the top-K competition for the argmax, termed as participation (π and defined formally in Section 4.1). We argue that these two notions significantly diverge and a token can be frequent as *input* yet rarely *generated* in the model contexts and vice-versa. For example, in Qwen2.5-1.5B model, the end-of-text piece `<|endoftext|>` token has corpus frequency near zero but consistently remain in the Top-20 (3.65% of all positions in $\mathcal{H}_{\text{test}}$) of any token. Greedy tokenizers further compound the gap, since a whole-word piece and its stem (e.g. `describe` vs. `_describ`) are distinct tokens with distinct decode-time roles, but the surface-level frequency counts don't capture which one model prioritizes.

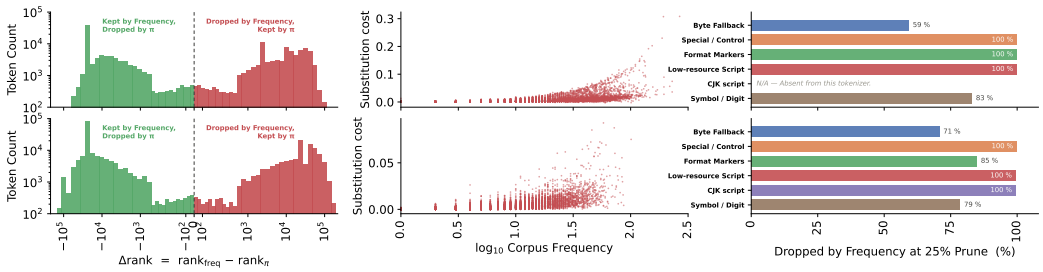

Figure 2: **Three Failure Cases of Frequency Pruning:** Top row for Qwen2.5-1.5B (V=151,936) and Bottom row for Gemma-2-2B (V=256,000) calibrated with 20M tokens. (a) *Participation mismatch:* Distribution of $\Delta\text{rank} = \text{rank}_{\text{freq}} - \text{rank}_{\pi}$ across the vocabulary with two signals showing significant disagreement (heavy bimodal tails). (b) *Substitution blindness:* Per-token substitution cost (Eq. 8) against corpus frequency have wide vertical spread at fixed frequency showing two tokens with the same corpus count can differ by more than $100\times$ in the cost of removing them. (c) **Rare-but-Critical Tokens:** Per-category share of rare-but-critical tokens deleted with 25% frequency-based removal.

Figure 2(a) plots the per-token i's rank disagreement $\Delta\text{rank} = \text{rank}_{\text{freq}(i)} - \text{rank}_{\pi(i)}$ over the whole vocabulary estimated on $\mathcal{H}_{\text{test}}$ using Equations 4 and 5. If frequency can be faithful proxy for decode-time competition across tokens, the mass should concentrate at $\Delta\text{rank} \approx 0$ but we instead found both signals exhibit heavy, bimodal distribution. It can be observed that left side (kept by frequency, dropped by π) collects tokens that are common in text but never enter the Top-20; and the right side (dropped by frequency, kept by π) collects tokens that are rare in $\mathcal{H}_{\text{test}}$ tokens distribution but highly competitive during the decoding, indicating the disagreement between these two signals.

► **[Failure Mode II] The Substitution Blindness:** *Among two equally droppable-looking tokens, is one safer to remove than the other?*

Corpus frequency assigns a single scalar per token and cannot express how the rest of the `LM-Head` would absorb removal of a token. Two tokens with identical (low) frequency can differ sharply in impact they can create once pruned. If a token's head row has a near-identical surviving neighbor, the model can degrade marginally when it is pruned with the neighbor reproducing similar logit for every hidden state. For example, `sofa` and `couch` are two tokens as near neighbors and if `sofa` $\in \mathcal{S}$, `couch` can be pruned with marginal degradation. However, if the token row is directionally isolated, no survivor can replace it without notable difference in logit and pruning it distorts the output distribution wherever it carried mass. Equation 4 is blind to this substitution behavior where given two tokens with exactly same frequency count, it can't differentiate which token is droppable.

Figure 2(b) plots per-token substitution cost (Equation 8 defined formally in Section 4.2) against $\log_{10}$ corpus frequency. It can be clearly observed that at any fixed frequency the scatter points span a wide vertical range and two tokens with the same corpus count can differ by more than $100\times$ in the cost of removing them. Therefore, a simple frequency based thresholding cannot effectively rank tokens by removal impact with neighboring substitution tokens. Note that substitution cost (Equation 8) is a property of `LM-Head` geometry rather than of counts, it also varies substantially at fixed participation (Equation 5) along with corpus frequency (Sections 4.2 & 4.3).

► **[Failure Mode III] Rare-but-Critical Tokens:** *Are all rare tokens safe to drop?*

Some tokens are rare but special. Examples include: the byte-fallback pieces are required for tokenizer correctness, format and control markers, image-codebook entries in a unified multimodal decoding head (Chameleon, 2024; Sun et al., 2024), and low-resource-language fragments. Tokens belonging to these example categories are infrequent and sensitive to frequency-based pruning but removing byte fallbacks can break tokenizer completeness and removing a format marker can corrupt the structured generation.

Figure 2(c) quantifies the exposure and illustrate tokens deleted by a 25% frequency-based pruning on Qwen2.5-1.5B model and Gemma-2-2B tokenizer. Note that Equation 4 removes every special/control marker, low-resource-script pieces, and CJK-script pieces across two models. Note that even 59–71% of byte-fallback pieces and 79–83% of symbol/digit pieces are removed which can significantly hurt downstream performance and requires appropriate handling before deployment.

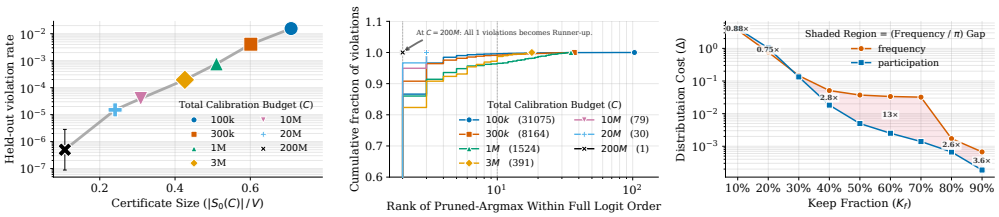

Figure 3: **Investigating Participation Cost** (Qwen2.5-1.5B, V=151,936; Nemotron-CC-v2 calibration; $|\mathcal{H}_{\text{test}}|$=20M disjoint held-out positions; K=20). **(a)** Certificate size $|\mathcal{S}_0(C)|/V$ and Held-out violation rate as the calibration tokens increases C=$10^5 \rightarrow 2\times10^8$ tokens. **(b)** Empirical CDF over violations caused by pruned `LM-Head` with K_f=0.5 and corresponding ranking with full logit order; **(c)** Held-out P_{99} of Δ (Eq. 3) for π-only vs. Frequency-only pruning at $f \in \{10, \ldots, 90\}\%$.

4 PASS: PARTICIPATION AND SUBSTITUTION SELECTION

In Section 3, we discussed three failure modes of frequency-based pruning. In this section, we introduce PASS, a calibration-guided static pruning framework that directly addresses these gaps. PASS rests on two per-token signals computed in a single calibration pass: ❶ *participation frequency* π_i (Section 4.1), which measures decode-time competition, and ❷ *substitution cost* $\text{cost}(i)$ (Section 4.2), which measures how much of a token's role the surviving head can inherit. Both signals are combined via a two-stage selector (Section 4.3) that identifies confident tokens using the computationally cheap signal π_i and reserves the more expensive $\text{cost}(i)$ signal for remaining tokens. Figure 5 summarizes the pipeline. PASS's keep/prune decision is associated with a provenance label (**auto-keep**, **mid-keep**, **mid-prune**, **auto-prune**) allowing per-token interpretability of the selection.

4.1 SIGNAL 1: PARTICIPATION FREQUENCY (π)

Corpus frequency (Eq. 4) counts how often a token appears in the calibration corpus but we propose *participation* to count how often the model *competes to decode* it. For an hyperparameter K (default Top-20 in this work), the *participation frequency* of token i at calibration budget C is given by:

$$\pi_i(C) \;=\; \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[\, i \in \text{Top-}K(W\boldsymbol{h}_j)\,] \;\in\; [0,1], \tag{5}$$

Formally, it is the fraction of calibration positions at which i enters the top-K candidates for the argmax. Note that frequency records observed raw labels, participation records the model's own top-K competition, so the two signals occupy the same range ([0,1]-valued per-token statistics over $\mathcal{H}$) but measure different aspects. We also define *zero-participation set* at budget C as:

$$\mathcal{S}_0(C) \;=\; \{\, i \in \mathcal{V} \;:\; \pi_i(C) = 0 \,\}, \tag{6}$$

$\mathcal{S}_0(C)$ collects tokens which the model never selected as a top-K candidate over calibration corpus.

Proposition 1 (Zero-participation greedy invariance)**.** *Under the pruned head $\mathcal{S} = \mathcal{V} \setminus \mathcal{S}_0(C)$ with renormalization (Eq. 2), the greedy decode is unchanged at every calibration position $\boldsymbol{h}_j \in \mathcal{H}$:* $\arg\max_{k\in\mathcal{S}} p^{\mathcal{S}}(\boldsymbol{h}_j)_k \;=\; \arg\max_{i\in\mathcal{V}} p^{\star}(\boldsymbol{h}_j)_i$. *Proof in Appendix C.1.*

We refer to $\mathcal{S}_0(C)$ as a *certificate* of prunability, and its normalized cardinality $|\mathcal{S}_0(C)|/V$ as the *certificate size*. Note that this guarantee is restricted to the calibration positions in $\mathcal{H}$), so whether it reflects on unseen text ($\mathcal{H}_{\text{test}}$) is an empirical argument we evaluate next. Figure 3 attempts to address this off-calibration question on Qwen2.5-1.5B where we calibrated on Nemotron-CC-v2 and evaluated on 2M disjoint held-out positions. A held-out position is a *violation* when the greedy token $i^{\star}(\boldsymbol{h})$ lies in $\mathcal{S}_0(C)$ and the *violation rate* is the fraction of token positions it happens within $\mathcal{H}_{\text{test}}$.

☛ **More calibration ensures safety and reduces violation:** Figure 3(a) presents the certificate size $|\mathcal{S}_0(C)|/V$ and the held-out violation rate as the calibration budget C increases from 10^5 to 2×10^8 tokens on Qwen2.5-1.5B. It can be observed that certificate size shrinks from roughly 65% down to 13% as rare tokens progressively escape $\mathcal{S}_0$ by appearing in some top-K increasing safety from dropping incautiously. Also the violation rate simultaneously falls significantly from 1.5×10^{-2} to 5×10^{-7} providing empirical support on held-out data and confirming practicality of static pruning.

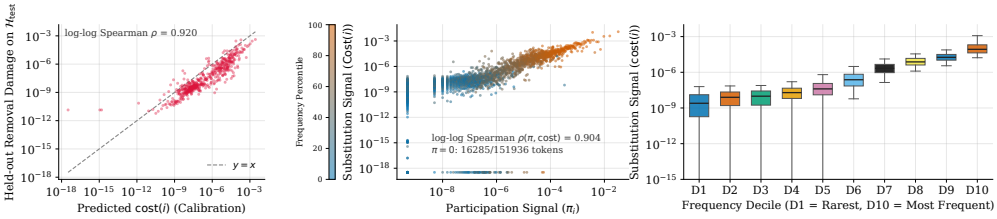

Figure 4: **Investigating Substitution Cost** (Qwen2.5-1.5B, V=151,936; Nemotron-CC-v2 calibration; $|\mathcal{H}_{\text{test}}|$=20M disjoint held-out positions; K=20). **(a)** Predicted $\text{cost}(i)$ vs. measured held-out cost of removing i alone; **(b)** Co-relation between π_i vs. $\text{cost}(i)$ colored by frequency percentile (dark = common tokens); **(c)** Substitution cost(i) segregated by rank-based frequency decile.

☛ **If certificate fails, the model falls back on its runner-up:** A low failure rate is not sufficient on its own because a handful of catastrophic substitutions can notably corrupt generation. To characterize the nature of failures, for each violation we identify which token the pruned `LM-Head` generates in place of the removed argmax and its rank r in the original ranking (necessarily $r \geq 2$, since r=1 is the removed token). Figure 3(b) reports the empirical CDF of these ranks across calibration budgets. Across every setting we examine, the surrogates concentrate at the very top of the ranking. At C=3×10^6, over 95% of the 391 violations resolve at r=2 with none exceeding r=10, and even at the very limited calibration budget of C=10^5 where 31,075 violations occur, over 80% still resolve at rank 2. Failures under the π=0 certificate are therefore significantly graceful in the model's true ranking and the pruned-argmax effect falls to a close neighbor rather than a distant token.

☛ **Participation outperforms frequency across the practical keep-fraction (k_f) range:** Beyond isolating $\mathcal{S}_0(C)$, π_i can rank the entire vocabulary and function as a standalone selection signal. Figure 3(c) compares π-ordered pruning against frequency-ordered pruning (Equation 4) across the keep-fraction $k_f \in [10\%, 90\%]$, scored by distributional cost (Δ). It can be observed that participation (π) standalone dominates across the practically relevant regime ($k_f > 0.3$ where non-affordable tokens are not bound to be discarded) with gap reaches 13× at k_f=50% and stays between 2.8× and 13× across $k_f \in [40\%, 70\%]$. Motivated by this consistent dominance for $k_f \geq 30\%$, we adopt π_i as PASS's primary ranking signal.

4.2 SIGNAL 2: SUBSTITUTION COST

Participation Signal (π) identifies tokens the model never select during decoding, but among the many tokens that do participate, not all are equally important to retain. Two participating tokens can carry very different consequences where if one has a near-duplicate row in the surviving `LM-Head`, its role can be effectively inherited with negligible distributional change, while another may be directionally isolated and can impact the output distribution wherever it carried mass. We propose *substitution cost* that aim to quantify *how expensive it is to lose token i from the `LM-Head`, given the survival head.* We first formally define the notion of a substitute, and then combine it with the mass a token actually carries at decode time.

Definition 1 (Nearest neighbour)**.** For a candidate set $\mathcal{R} \subseteq \{1, \ldots, V\}$, the nearest neighbour of token i in $\mathcal{R}$ (excluding i itself) under raw ℓ_2 distance on unembedding rows is given as:

$$\text{nn}_{\mathcal{R}}(i) \;=\; \arg\min_{k \in \mathcal{R} \setminus \{i\}} \|\boldsymbol{w}_i - \boldsymbol{w}_k\|_2. \tag{7}$$

The nearest neighbour identifies the surviving row that is geometrically closest to $\boldsymbol{w}_i$ in the output projection. However, if that neighbor is actually a *good* substitute at decode time depends jointly on how often token i is invoked and how well its role can be reproduced by $\text{nn}_{\mathcal{R}}(i)$ at the given hidden states. We combine both effects into a single per-token estimator as follows:

Definition 2 (Substitution cost)**.** The substitution cost of token i against candidate set $\mathcal{R}$ is

$$\text{cost}(i \mid \mathcal{R}) \;=\; \frac{1}{N} \sum_{j=1}^{N} p^{\star}(\boldsymbol{h}_j)_i \cdot \left(\boldsymbol{w}_i^{\top} \boldsymbol{h}_j - \boldsymbol{w}_{\text{nn}_{\mathcal{R}}(i)}^{\top} \boldsymbol{h}_j\right). \tag{8}$$

At each calibration position, $\text{cost}(i \mid \mathcal{R})$ multiplies two per-token quantities: ❶ the mass $p^{\star}(\boldsymbol{h}_j)_i$ that token i carried at $\boldsymbol{h}_j$ under the full head; ❷ the logit gap $\boldsymbol{w}_i^{\top} \boldsymbol{h}_j - \boldsymbol{w}_{\text{nn}_{\mathcal{R}}(i)}^{\top} \boldsymbol{h}_j$ that removing i would incur by rerouting that mass to its nearest surviving neighbour. A token scores high only

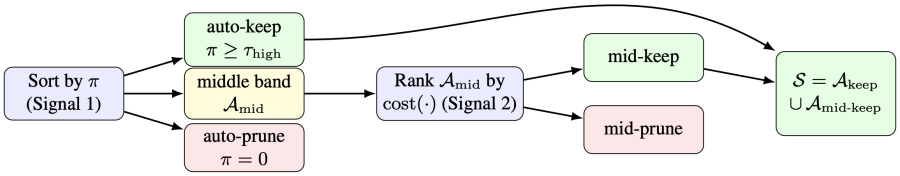

Figure 5: **PASS two-stage selection:** Stage 1 partitions by participation into auto-keep, middle band, and auto-prune. Stage 2 ranks only the middle band by substitution cost into mid-keep and mid-prune.

when it *both* carries decode-time mass *and* lacks a close substitute. Therefore, tokens that are rarely activated remain cheaply prunable through the $p^\star$ weight, and tokens with near-duplicate neighbours remain cheaply prunable through the logit gap. This estimator directly addresses the limitation of corpus frequency that can't distinguish two tokens with identical counts.

Choice of Nearest-Neighbour Metric: In PASS, we choose to use ℓ_2 in raw W-row space in place of cosine similarity. Using Cauchy-Schwarz, we have $|(\boldsymbol{w}_i - \boldsymbol{w}_k)^\top \boldsymbol{h}| \leq \|\boldsymbol{w}_i - \boldsymbol{w}_k\|_2 \|\boldsymbol{h}\|_2$ and note that $\|\boldsymbol{h}\|_2$ is same for every candidate w_k for a given $\boldsymbol{h}$. It implies that ℓ_2-nearest neighbor the surviving row that minimizes this upper bound on the induced logit gap. On the other hand, cosine similarity normalizes away the magnitude of $\boldsymbol{w}$ and can pair a high-norm row with a low-norm partner. Empirically, we also identified ℓ_2 performance superior than cosine as nn-metric.

☛ **Cost reliably ranks Held-out removal damage:** To validate if $\mathrm{cost}(i)$ can meaningfully capture per-token removal impact, Figure 4(a) plots, for each token i, the calibration prediction $\mathrm{cost}(i \mid \mathcal{V})$ from Equation 8 on the x-axis against the ground-truth held-out removal damage $\mathbb{E}_{\boldsymbol{h} \in \mathcal{H}_{\text{test}}}\big[\Delta(\boldsymbol{h}; \mathcal{V} \setminus \{i\})\big]$ on the y-axis which is the mean distributional cost (Equation 3) of pruning i alone and renormalizing over the remaining $V-1$ tokens. Across the $144{,}011$ tokens strictly positive on both axes, we observe a strong monotonic relationship with log-log Spearman ρ=0.920. For clarification, Equation 3 and 8 can differ in magnitude since the x-axis multiplies mass by a logit gap while the y-axis measures the actual distributional damage on held-out text. However for a rank-based selector, ordering matters and the clear monotonic alignment in Figure 4(a) confirms that $\mathrm{cost}(i)$ as a valid estimator to rank tokens based on removal damage.

☛ **Cost carries complementary information to Participation:** Figure 4(b) plots $\pi_i(C_{\max})$ against $\mathrm{cost}(i)$ and it can be observed that despite two signals strongly correlated ($\rho(\pi, \mathrm{cost})$=0.904), the cost signal is not redundant. Given a fixed π, cost can significantly vary upto roughly two orders of magnitude. We exploit this residual spread for token bands that π treats as uniform to separate tokens by expected removal impact.

☛ **Cost is more informative where frequency is least informative:** Figure 4(c) reveals two complementary trends in the cost distribution across the frequency axis. Grouping all vocabulary tokens across rare (D_1–D_3), mid (D_4–D_7), and common (D_8–D_{10}) frequency bins (details in Appendix B), we find that the median cost rises monotonically from rare to common, which is consistent with the intuition that typical common tokens carry more decode-time mass and are individually more expensive to remove than typical rare ones. However, the within-stratum *spread* moves in the opposite direction with $15\times$ magnitude for rare bins, shrinking across mid bins, and then notably small in the common bins. This asymmetry suggests that rare tail is where true discardble tokens (small mass, close substitute) sit alongside rare-but-load-bearing tokens (meaningful mass, no substitute), while common tokens are more uniformly consequential.

4.3 TWO-STAGE SELECTION WITH AUTO-KEEP AND MID-REFINEMENT

Participation empirically outranks frequency as a whole vocabulary ranker at practical keep fractions while substitution can be interpreted not primarily as a whole-vocabulary ranker but a within-stratum discriminator. In PASS, we therefore decide to exploit each signal consecutively controlled by a knob $\alpha \in [0, 1]$ to capture their strengths, rather than combining them uniformly across the vocabulary. Given a keep token budget $V' = |\mathcal{S}|$, the two-stage selector proceeds as follows:

- **Stage 1 [Participation-based commitment]:** Let τ_{high} be the $\lceil \alpha V' \rceil$-th largest value of π among tokens with $\pi > 0$. We partition the vocabulary $\mathcal{V}$ across three bands:

$$\mathcal{A}_{\text{keep}} = \{i : \pi_i \geq \tau_{\text{high}}\}, \quad \mathcal{A}_{\text{prune}} = \{i : \pi_i = 0\}, \quad \mathcal{A}_{\text{mid}} = \mathcal{V} \setminus (\mathcal{A}_{\text{keep}} \cup \mathcal{A}_{\text{prune}}). \tag{9}$$

Task	Method	Qwen2.5-0.5B			Qwen2.5-1.5B			Gemma2-2B		
		K_f=0.5	K_f=0.3	K_f=0.1	K_f=0.5	K_f=0.3	K_f=0.1	K_f=0.5	K_f=0.3	K_f=0.1
GSM8K	Baseline		34.34			60.96			23.28	
	Frequency	32.42	0.00	0.00	61.03	0.00	0.00	20.12	20.44	0.00
	VocabTrim(Goel et al., 2025)	32.89	0.00	0.00	60.29	10.04	0.00	20.02	15.78	0.00
	VT(Ushio et al., 2023b)	31.97	0.00	0.00	58.49	0.00	0.00	17.59	0.00	0.00
	PASS [Ours]	34.34	33.66	26.54	60.88	60.50	54.44	23.20	23.12	8.00
MATH	Baseline		8.54			13.06			9.20	
	Frequency	7.54	4.54	0.00	11.06	11.08	2.44	9.22	8.98	8.80
	VocabTrim(Goel et al., 2025)	8.12	7.08	0.00	12.84	12.15	0.92	9.18	9.00	8.74
	VT(Ushio et al., 2023b)	6.84	2.17	0.00	10.92	5.08	0.00	8.86	6.42	0.00
	PASS [Ours]	8.54	8.54	0.92	13.06	13.08	0.46	9.10	9.08	8.80
MMLU	Baseline		47.57			60.96			52.98	
	Frequency	47.57	47.57	47.57	60.96	60.96	60.96	52.98	52.98	52.98
	VocabTrim(Goel et al., 2025)	47.57	47.57	47.57	60.96	60.96	60.96	52.98	52.98	52.98
	VT(Ushio et al., 2023b)	47.57	47.57	47.57	60.96	60.96	60.96	52.98	52.98	52.98
	PASS [Ours]	47.57	47.57	47.57	60.96	60.96	60.96	52.98	52.98	52.98
TriviaQA	Baseline		19.18			39.70			55.17	
	Frequency	15.14	15.37	11.06	34.66	33.73	21.58	53.10	51.70	33.31
	VocabTrim(Goel et al., 2025)	18.62	17.81	9.84	39.12	37.05	19.86	54.41	50.82	31.74
	VT(Ushio et al., 2023b)	17.21	15.34	6.58	37.18	31.64	14.27	52.43	46.17	24.85
	PASS [Ours]	19.18	18.41	10.92	39.67	37.77	21.25	54.98	51.37	34.40
NQ-Open	Baseline		6.68			14.68			18.75	
	Frequency	5.65	5.45	4.68	14.65	13.99	9.14	16.75	15.40	13.21
	VocabTrim(Goel et al., 2025)	6.47	6.18	4.72	14.56	13.85	8.97	18.21	16.98	12.88
	VT(Ushio et al., 2023b)	6.02	5.47	3.92	13.82	11.94	7.12	17.08	14.84	10.41
	PASS [Ours]	6.68	6.51	4.85	14.68	14.07	9.78	18.70	17.42	13.13
HumanEval	Baseline		29.27			35.98			19.51	
	Frequency	17.68	0.61	0.61	23.17	2.44	0.61	17.07	0.00	0.00
	VocabTrim(Goel et al., 2025)	21.95	10.37	3.66	31.71	15.24	4.27	18.42	11.58	4.88
	VT(Ushio et al., 2023b)	19.51	5.49	1.22	27.44	8.54	1.83	16.46	5.49	1.22
	PASS [Ours]	22.56	13.41	8.54	32.93	20.73	7.93	18.90	15.24	9.76
MBPP	Baseline		29.60			43.20			27.80	
	Frequency	17.20	0.00	0.00	24.40	2.40	0.00	22.60	0.00	0.00
	VocabTrim(Goel et al., 2025)	17.80	4.20	0.00	23.60	1.80	0.00	24.60	1.20	0.00
	VT(Ushio et al., 2023b)	17.40	1.40	0.00	19.80	5.00	0.00	23.20	0.40	0.00
	PASS [Ours]	19.80	0.20	0.00	29.00	6.60	0.00	24.80	0.00	0.00

Table 1: Performance comparison on downstream tasks across models and keep-fractions K_f.

Tokens in $\mathcal{A}_{\text{keep}}$ band are committed to survival as *auto-keep*, and tokens in $\mathcal{A}_{\text{prune}}$ are committed to removal as *auto-prune* (certified safe by Proposition 1).

- **Stage 2 [Cost-based Mid-Refinement].** For each $i \in \mathcal{A}_{\text{mid}}$, we compute $\text{cost}(i \mid \mathcal{R})$ with candidate set $\mathcal{R} = \mathcal{A}_{\text{keep}} \cup \mathcal{A}_{\text{mid}}$. Let $\mathcal{A}_{\text{mid-keep}}$ be the Top-$(V' - |\mathcal{A}_{\text{keep}}|)$ subset of $\mathcal{A}_{\text{mid}}$ ranked by cost (*mid-keep*), we prune the remaining as *mid-prune*. The final survival set is:

$$\mathcal{S} \;=\; \mathcal{A}_{\text{keep}} \cup \mathcal{A}_{\text{mid-keep}}, \tag{10}$$

Note that at $\alpha \to 1$ the whole budget is committed by π and cost is unused, while at $\alpha \to 0$ the entire budget is spent on the cost ranking of the middle band. Detailed pseudocode (Algorithm 1) and design-choices (candidate-set construction and coverage-undershoot handling) are in Appendix B.1.

5 Performance Comparison and Ablation Studies

➤ **Perplexity Retention with Vocabulary Pruning:** Figure 6 reports the fraction of full-head log-likelihood that the pruned head retains on held-out set ($\mathcal{H}_{\text{test}}$). At k_f=0.5, PASS can retain ∼99% on both Qwen models against ∼96% for frequency suggesting halving the head costs PASS under 1% of likelihood. On Gemma-2-2B (V=256K), PASS outperforms by ∼7% at k_f=0.5 (96.6 vs. 89.7) and ∼5 points at k_f=0.3. PASS and Frequency-based pruning tends to converge at both extremes: at k_f=0.9 nearly every high-mass token survives either ranking, while at k_f=0.1 both lose 14-22% of PPL because total remaining tokens falls well below set of tokens the model actually uses during decoding.

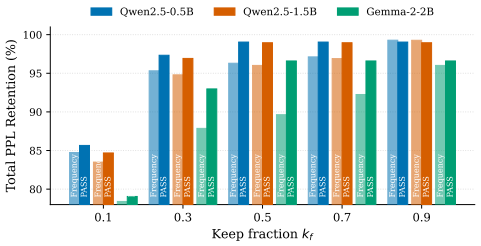

Figure 6: Held-out log-likelihood retained by frequency (light) and PASS (dark) as % of the full head ($|\mathcal{H}_{\text{test}}|$=20M).

➤ **Downstream Benchmarks and Quality Metric:** Table 1 shows how does the likelihood retention carries over to generation on real downstream tasks[1]. At k_f=0.5, PASS is nearly lossless on math and knowledge QA (GSM8K within 0.08 points of the full head on all three models). We observe degradation under heavier compression: at k_f=0.3 where frequency and VT collapse to 0 on GSM8K for both Qwen models while PASS keeps 33.66 and 60.50 (98–99% of baseline), and even at k_f=0.1 it keeps 26.54 and 54.44. On HumanEval PASS is best in every setting (e.g., 20.73 vs. 15.24 for

[1]MMLU is scored by log-likelihood over fixed options (A/B/C/D) which are preserved and hence identical for every method and budget.

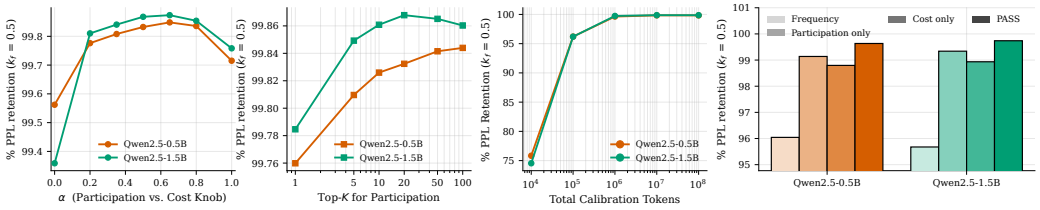

Figure 7: **PASS Ablation at k_f=0.5:** (a) Influence of Stage-1 allocation with α; (b) Influence of Top-K depth within participation cost; (c) Influence of total calibration tokens used for estimation of Equation 5 and 8; (d) Influence of Signal 1 and 2 in isolation relative to Frequency and PASS.

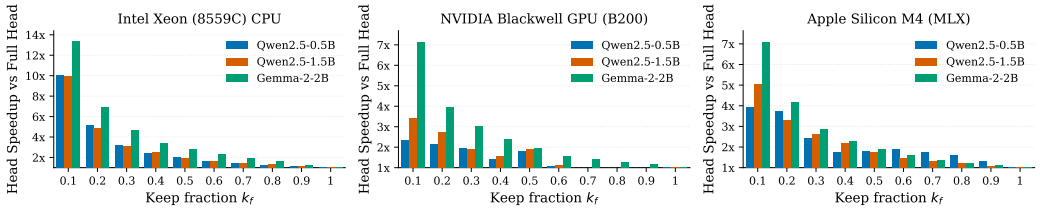

Figure 8: `LM-Head` decode-time Speedup at Batch size 1 with FP16 measured with varying K_f.

VocabTrim and 2.44 for frequency on Qwen2.5-1.5B at k_f=0.3). MBPP is the clear limitation: PASS is still best at k_f=0.5 but keeps only 67-89% of baseline, and every method collapses at $k_f \leq 0.3$, suggesting that code-specific whitespace and indentation pieces are under-sampled by web-text calibration and motivating domain-matched calibration. MMLU is unchanged for every method and keep fraction, consistent with likelihood scoring over option letters that every selector can retain.

➤ **Ablation Studies on PASS Hyperparameters:** Figure 7 ablates PASS at k_f=0.5. ❶ The Stage-1 allocation α achieves an optimum near $\alpha \approx 0.65$ on both models. Allocating the whole token budget by participation (α=1) or ranking the whole middle band by cost (α=0) illustrate worse performance which confirms the two signals are complementary. ❷ Performance improves as we increase Top-K depth in participation signal and saturates or slightly degrades around K=20 which we use as default across our work. ❸ We found calibration budget significantly improves retention from ∼75% at 10^4 tokens to 99.7% at 10^6 and saturates by 10^7. It aligns with the certificate shrinkage in Figure 3(a). ❹ We found that both of our proposed signal alone can beat frequency by >2.5 points (participation 99.1-99.3%, cost 98.8-98.9%, frequency 95.7-96.0%), and their combination is best (99.6-99.7%).

➤ **Measured Wall-clock Decode Latency:** Figure 8 show that smaller head translates directly into per-token speedups (Figure 8). On Apple M4 (MLX GPU) at $B = 1$ in FP16, pruning to $k_f = 0.5$ reduces the head's per-token decode time from $\sim 1.9\times$ gain on Gemma-2-2B, $\sim 1.75\times$ on Qwen2.5-1.5B, and $\sim 1.8\times$ on Qwen2.5-0.5B, with near-identical gains on B200 (1.8-1.9×). On an Intel Xeon 8559C CPU at 16 threads, the head-side reduction is significantly large acheving $\sim 2.8\times$ on Gemma-2-2B and 1.9-2.0× on the Qwen models. The speedup grows sharply as k_f decreases, reaching 13.3× on CPU and 7.1× on B200 and M4 for Gemma-2-2B at $k_f = 0.1$. For B200 at $k_f \geq 0.6$, where fixed kernel overheads dominate the smaller Qwen heads and gains vanish. The end-to-end benefit scales with the head's share of decode latency, which reaches 6–20% on these checkpoints (Figure 1c), so a 2× head speedup can reduce 3-10% of per-token head decode time.

6 CONCLUSION

In this work, we propose PASS, a calibration-guided, training-free method for static vocabulary-head pruning that addresses key limitations of frequency-based token selection such as substitution blindness, critical token deletion. Instead of counting how often a token appears in text, PASS measures how often the model competes to generate it through top-K participation, and how well the surviving head can absorb its role through a mass-weighted substitution cost. We further introduce a two-stage selector that uses participation to resolve the confident ends of the vocabulary and substitution cost to rank the ambiguous middle band. Since the survival set is fixed once offline, the pruned head remains a smaller dense matrix that runs on existing kernels with no per-token overhead. We empirically establish the benefits of PASS in reducing the overhead of `LM-Head` in large-vocabulary small models while closely matching full-head performance at half the head cost.

AI USE STATEMENT

In this work, we did not use generative AI tools for any task that requires disclosure. We have not used generative AI tools to generate synthetic datasets, develop theoretical models, or propose and refine hypotheses, or interpret results. We used generative AI tools only to edit the paper for readability and grammatical correctness. We have reviewed all AI-assisted edits and verified that they do not alter the technical content, claims, or reported results. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Amine Abdaoui, Camille Pradel, and Grégoire Sigel. Load what you need: Smaller versions of mutililingual BERT. In Nafise Sadat Moosavi, Angela Fan, Vered Shwartz, Goran Glavaš, Shafiq Joty, Alex Wang, and Thomas Wolf, editors, *Proceedings of SustaiNLP: Workshop on Simple and Efficient Natural Language Processing*, pages 119–123, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.sustainlp-1.16. URL https://aclanthology.org/2020.sustainlp-1.16/.

Chameleon. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Patrick Chen, Si Si, Sanjiv Kumar, Yang Li, and Cho-Jui Hsieh. Learning to screen for fast softmax inference on large vocabulary neural networks. In *ICLR*, 2019.

Gemma Team. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Raghavv Goel, Sudhanshu Agrawal, Mukul Gagrani, Junyoung Park, Yifan Zao, He Zhang, Tian Liu, Yiping Yang, Xin Yuan, Jiuyan Lu, et al. Vocabtrim: Vocabulary pruning for efficient speculative decoding in llms. *arXiv preprint arXiv:2506.22694*, 2025.

Edouard Grave, Armand Joulin, Moustapha Cissé, David Grangier, and Hervé Jégou. Efficient softmax approximation for GPUs. In *ICML*, 2017.

Sébastien Jean, Kyunghyun Cho, Roland Memisevic, and Yoshua Bengio. On using very large target vocabulary for neural machine translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2015.

Haitao Mi, Zhiguo Wang, and Abe Ittycheriah. Vocabulary manipulation for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 124–129, 2016.

Andriy Mnih and Geoffrey E Hinton. A scalable hierarchical distributed language model. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2008. URL https://proceedings.neurips.cc/paper_files/paper/2008/file/1e056d2b0ebd5c878c550da6ac5d3724-Paper.pdf.

Frederic Morin and Yoshua Bengio. Hierarchical probabilistic neural network language model. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics*, volume R5 of *Proceedings of Machine Learning Research*, pages 246–252. PMLR, 06–08 Jan 2005. URL https://proceedings.mlr.press/r5/morin05a.html. Reissued by PMLR on 30 March 2021.

Kyuhong Shim, Minjae Lee, Iksoo Choi, Yoonho Boo, and Wonyong Sung. SVD-softmax: Fast softmax approximation on large vocabulary neural networks. In *NeurIPS*, 2017.

Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Emu: Generative pretraining in multimodality. In *International Conference on Learning Representations (ICLR)*, 2024.

Chaofan Tao, Qian Liu, Longxu Dou, Niklas Muennighoff, Zhongwei Wan, Ping Luo, Min Lin, and Ngai Wong. Scaling laws with vocabulary: Larger models deserve larger vocabularies. *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Asahi Ushio, Yi Zhou, and Jose Camacho-Collados. Efficient multilingual language model compression through vocabulary trimming. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14725–14739, 2023a.

Asahi Ushio et al. Efficient multilingual language model compression through vocabulary trimming. In *EMNLP Findings*, 2023b.

Chengyue Wu, Xiaokang Chen, Zhiyu Wu, Yiyang Ma, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, Chong Ruan, et al. Janus: Decoupling visual encoding for unified multimodal understanding and generation. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12966–12977. IEEE, 2025a.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, et al. Vila-u: a unified foundation model integrating visual understanding and generation. In *International Conference on Learning Representations*, volume 2025, pages 93620–93638, 2025b.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Minjia Zhang, Xiaodong Liu, Wenhan Wang, Jianfeng Gao, and Yuxiong He. Navigating with graph representations for fast and scalable decoding of neural language models. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 6311–6322, 2018.

Weilin Zhao, Tengyu Pan, Xu Han, Yudi Zhang, Ao Sun, Yuxiang Huang, Kaihuo Zhang, Weilun Zhao, Yuxuan Li, Jie Zhou, Hao Zhou, Jianyong Wang, Zhiyuan Liu, and Maosong Sun. FR-spec: Accelerating large-vocabulary language models via frequency-ranked speculative sampling. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3909–3921, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.198. URL https://aclanthology.org/2025.acl-long.198/.

A Related Works

The cost of a language model's output head projection grows linearly with the vocabulary size V. The problem has become more extreme as LLM vocabularies have grown to 150K–300K tokens, driven by multilingual coverage, code, and evidence that larger models benefit from larger vocabularies (Tao et al., 2024). When it comes to multimodal models, new modality further presses the cost by adding discrete visual codebook entries to the same output head (Chameleon, 2024; Wu et al., 2025b; Sun et al., 2024; Wu et al., 2025a). Existing approaches to optimize this cost differ in *when* and *how* the vocabulary is reduced.

A.1 Efficient Softmax for Large Vocabularies

One category of work focuses on the cost of training with a large softmax derived by the output head projection. Jean et al. (2015) approximates the softmax normalization with importance sampling over a small subset of the target vocabulary, which keeps the per-update training cost nearly constant in $\mathcal{V}$. Since the sampling approach applies only to the training objective, the full output projection is still required at inference unless it is paired with task-specific heuristics such as the alignment based candidate lists.

Another category of work changes the structure of the output layer itself, so the model must be trained with the modified output head. Hierarchical softmax (Morin and Bengio, 2005; Mnih and Hinton, 2008) factorizes each token's probability into a sequence of decisions along a tree over the vocabulary. Adaptive softmax (Grave et al., 2017) partitions the vocabulary by frequency into a small head clsuter evaluated for every token and tail clusters evaluated only for tokens that fall in them, with cluster sizes chosen to minimize a GPU computation-time model. Moreover, these savings apply when scoring a given token. Computing the full next-token distribution still requires the entire vocabulary as required for sampling.

Neither categories reduce the head cost of a trained LLM. Sampling-based method leave the inference-time of output head unchanged and structured output layers require the model to be re-trained with the new architecture.

A.2 Dynamic Candidate Selection

Dynamic selection methods exploit the fact that, under decoding, each step is driven by a few highest-scoring tokens rather than the full distribution. Instead of computing all $\mathcal{V}$ logits, they cheaply identify a small set of candidate rows for each context and compute only those exactly. SVD-softmax (Shim et al., 2017) factorizes the output head and uses a low-rank preview of all $\mathcal{V}$ logits to choose a small set of rows that are then computed exactly. FGD (Zhang et al., 2018) builds a graph index over the rows of the head and searches it to approximately retrieve the highest-scoring tokens for each context. L2S (Chen et al., 2019) learns a lightweight screening model that assigns each context to one of many clusters, each with its own candidate set learned to preserve the full model's top-k predictions.

Since the candidate set adapts to each context, these methods can reduce computation aggressively. On device, however, this flexibility comes at a cost. The full head must be kept in memory and each step adds work over auxiliary parameters or indices. The selected rows also change from context to context which makes memory access irregular. SVD-softmax additionally reads every row of the head at reduced width. These overheads show up in measured speedups. SVD-softmax achieves only a 1.3× speedup out of a theoretical 5.5× because the selected rows are scattered in memory (Zhang et al., 2018). On a small language model, SVD-softmax is even slower (0.8×) than the full softmax, while FGD is only 1.3× faster (Chen et al., 2019). PASS instead makes a single offline decision. The pruned head is a smaller dense matrix. As a result, it reduces the bytes read per token, adds no per-token overhead, and runs on existing matrix-vector kernels.

A.3 Static Vocabulary Selection

Unlike the dynamic methods, static methods including PASS, fix a reduced vocabulary before decoding begins and keep it unchanged across decoding steps. Since the same vocabulary must serve every step, the central design question becomes which tokens to keep. Prior work restricted decoding to a candidate list built once per source sentence. These lists combine a fixed set of frequent target

words with translation candidates for the source words, drawn from an alignment-based dictionary or a phrase table (Jean et al., 2015; Mi et al., 2016). For pre-trained multilingual models, vocabulary trimming keeps only tokens that occur in a target language corpus, optionally only the most frequent ones. This shrinks the embedding matrix of encoders (Abdaoui et al., 2020) and, for generative models, the output head as well (Ushio et al., 2023a). In speculative decoding, vocabulary reduction has been applied to the drafter. FR-Spec (Zhao et al., 2025) restricts the draft LM head to tokens that are frequent in a pre-training corpus. VocabTrim (Goel et al., 2025) limits the draft head to tokens that the target model most often generates on a calibration set. Apart from the source-conditioned candidate lists, these methods rank tokens by a single per-token count. PASS instead ranks tokens by top-K participation and substitution cost, which addresses the failure modes of corpus frequency identified in Section 3.

B Details of Frequency Bins for Figure 4(c)

The Nemotron-CC-v2 High-Quality subset is streamed through Qwen2.5-1.5B with the Qwen2.5 BPE tokenizer ($|\mathcal{V}| = 151{,}936$, sequence length $T = 1{,}024$), producing a calibration stream of $\sim$20M tokens over $\sim$19,500 sequences. Corpus token counts are computed on this stream where tokens with zero observed frequency remain in the analysis because static head pruning is a *per-row* decision on the output matrix $W \in \mathbb{R}^{V \times d}$ and every row either observed or not remains a candidate for removal.

Table 2: Frequency decile bin reference for Figure 4(c). Nemotron-CC-v2 corpus, Qwen2.5 tokenizer ($|\mathcal{V}| = 151{,}936$), $\sim$20M calibration tokens. Each bin contains $\approx$15,194 token *types* assigned by ascending corpus frequency (rank-based, ties broken arbitrarily). Frequency ranges shown are typical for this setup; exact boundaries depend on the specific stream sampled.

Bin	Freq. range	Types	Description	Role
1	0	15,194	Zero-frequency & unused BPE pieces	**Rare**
2	0–2	15,194	Byte fallbacks, rare multilingual fragments	**Rare**
3	2–8	15,193	Rare technical vocabulary, foreign compounds	**Rare**
4	8–25	15,194	Uncommon content words	**Mid**
5	25–70	15,194	Infrequent content words	**Mid**
6	70–180	15,193	Moderate-frequency content words	**Mid**
7	180–450	15,194	Fairly common content words	**Mid**
8	450–1,200	15,194	Common words & mid function words	**Common**
9	1,200–5,000	15,193	High-frequency & function words	**Common**
10	5,000–350,000	15,193	Ubiquitous function, punctuation, subwords	**Common**
Rare examples (Bins 1–3)			`<|endoftext|>`, byte fallbacks, rare CJK/Cyrillic pieces	
Mid examples (Bins 4–7)			`_volcano`, `_diocese`, `_battalion`, `_sculptor`	
Common examples (Bins 8–10)			`_the`, `_of`, `_and`, `_to`, `_in`, `.`, `,`	

Decile assignment of token types. All $V = 151{,}936$ token types (no structural filtering; see rationale above) are sorted by ascending corpus frequency, breaking ties by token id, and partitioned into $B = 10$ equal-cardinality bins of $\approx$15,194 types each. The bin index of token v is assigned by

$$b(v) \;=\; \min\left(\left\lfloor \frac{\mathrm{rank}(v)}{|\mathcal{V}|} \cdot B \right\rfloor,\; B-1\right) \;\in\; \{0, 1, \ldots, B-1\}, \tag{11}$$

and displayed on Figure 4(c) as $D_{b(v)+1}$. Assignment is determined by rank alone; the absolute frequency values in Table 2 establish the ordering but do not enter the plot. Every bin contains the same number of token *types*, but the *token stream* shares are highly non-uniform under Zipf's law: bin B=10 alone accounts for the overwhelming majority of the $\sim$20M calibration tokens, while bins 1–3 collectively contribute a negligible fraction.

Throughout Section 4.2 and the discussion of Figure 4(c), bins $\{1, 2, 3\}$ (equivalently, D_1–D_3) are referred to as **rare**; bins $\{4, 5, 6, 7\}$ (equivalently, D_4–D_7) as **mid-frequency**; and bins $\{8, 9, 10\}$ (equivalently, D_8–D_{10}) as **common**. The within-decile $\log_{10}$-spread of $\mathrm{cost}(i)$ reported on the panel is computed *per decile* (all ten values) before summarizing as median (headline) and maximum (caption), so the rare/mid/common binning governs the narrative discussion but not the numerical values on the plot itself.

Why cost's within-decile spread is largest in the rare tier. A token v contributes to bins 1–3 whenever it is rare in text, but its substitution cost is set by two independent quantities: the mass $p^\star(\boldsymbol{h})_v$ the model puts on v at decode time (which can be substantial for structurally special pieces even when frequency is zero, e.g. end-of-text and format markers), and the ℓ_2 head-row distance to its nearest survivor $\|\boldsymbol{w}_v - \boldsymbol{w}_{\mathrm{nn}(v)}\|_2$ (which can be very small for interchangeable byte-fallback pieces and very large for directionally isolated special tokens). The rare bins therefore mix (i) genuinely discardable tokens with tiny mass and close substitutes alongside (ii) rare-but-load-bearing tokens with meaningful mass and no close substitute. This mixture is exactly what produces the 10^{15} within-decile spread seen at D_1–D_3 in Figure 4(c) and what the frequency signal is structurally unable to distinguish (Failure Mode ❷).

B.1 Full specification of the Two-stage selector

Two-Signal Band construction. Given a keep budget $V' = |\mathcal{S}|$ and a Stage-1 allocation $\alpha \in [0, 1]$, let τ_{high} be the $\lceil \alpha V' \rceil$-th largest value of π among tokens with $\pi > 0$. The three bands are

$$\mathcal{A}_{\mathrm{keep}} = \{i : \pi_i \geq \tau_{\mathrm{high}}\}, \quad \mathcal{A}_{\mathrm{prune}} = \{i : \pi_i = 0\}, \quad \mathcal{A}_{\mathrm{mid}} = \mathcal{V} \setminus (\mathcal{A}_{\mathrm{keep}} \cup \mathcal{A}_{\mathrm{prune}}). \tag{12}$$

For each $i \in \mathcal{A}_{\mathrm{mid}}$ we compute $\mathrm{cost}(i \mid \mathcal{R})$ with $\mathcal{R} = \mathcal{A}_{\mathrm{keep}} \cup \mathcal{A}_{\mathrm{mid}}$, and let $\mathcal{A}_{\text{mid-keep}}$ be the top-$(V' - |\mathcal{A}_{\mathrm{keep}}|)$ subset of $\mathcal{A}_{\mathrm{mid}}$ by cost. The survival set is $\mathcal{S} = \mathcal{A}_{\mathrm{keep}} \cup \mathcal{A}_{\text{mid-keep}}$, of size exactly V' barring the coverage undershoot noted below.

Design choices. The candidate set $\mathcal{R}$ excludes $\mathcal{A}_{\mathrm{prune}}$ so that already-doomed tokens cannot "shield" a mid-band token by acting as its substitute; if $\mathrm{nn}_{\mathcal{R}}(i)$ is itself later pruned in Stage 2, the effective substitute is inherited transitively. When calibration coverage is thin, $|\mathcal{A}_{\mathrm{prune}}|$ can exceed $V - V'$; the $\pi = 0$ certificate then licenses more compression than requested, and $|\mathcal{S}| < V'$. We surface this as a diagnostic rather than back-filling certified-dead tokens, since doing so would knowingly restore tokens Proposition 1 has certified as safe to remove.

C Proofs and Deferred Derivations

C.1 Zero-participation greedy invariance

Proof of Proposition 1: With a fixed arbitrary calibration position $\boldsymbol{h}_j \in \mathcal{H}$, we show that the greedy choice under the pruned head at $\boldsymbol{h}_j$ coincides with the greedy choice under the full head at $\boldsymbol{h}_j$. Since $\boldsymbol{h}_j$ was arbitrary, the statement holds at every calibration position. For concreteness we assume any consistent tie-breaking rule (e.g., lowest index) is applied identically to both $\arg\max$ operators.

By the definition of Top-$K(W\boldsymbol{h}_j)$ as the set of K indices with largest logit values at $\boldsymbol{h}_j$, and since $K \geq 1$, the maximum-logit index is necessarily among them. Therefore:

$$i^\star(\boldsymbol{h}_j) \;=\; \arg\max_{i \in \mathcal{V}} \ell_i(\boldsymbol{h}_j) \;\in\; \text{Top-}K(W\boldsymbol{h}_j). \tag{S1}$$

By Eq. 5, $\pi_i(C) = \frac{1}{N}\sum_{j=1}^{N} \mathbb{1}[i \in \text{Top-}K(W\boldsymbol{h}_j)]$ is a non-negative average of $\{0, 1\}$ indicators. Consequently, $\pi_i(C) = 0$ if and only if the indicator is zero for *every* j:

$$\pi_i(C) = 0 \iff i \notin \text{Top-}K(W\boldsymbol{h}_j) \quad \text{for every } j = 1, \ldots, N. \tag{S2}$$

With any $i \in \mathcal{S}_0(C)$, applying (S2) at our fixed $\boldsymbol{h}_j$ gives $i \notin \text{Top-}K(W\boldsymbol{h}_j)$, while (S1) gives $i^\star(\boldsymbol{h}_j) \in \text{Top-}K(W\boldsymbol{h}_j)$. Since one index is in the top-K set and the other is not, they must differ: $i \neq i^\star(\boldsymbol{h}_j)$. As $i \in \mathcal{S}_0(C)$ was arbitrary, we have:

$$i^\star(\boldsymbol{h}_j) \;\notin\; \mathcal{S}_0(C), \quad \text{equivalently,} \quad i^\star(\boldsymbol{h}_j) \;\in\; \mathcal{S}. \tag{S3}$$

Algorithm 1 PASS: Participation And Substitution Selection

Require: Output projection $W \in \mathbb{R}^{V \times d}$; calibration hidden states $\mathcal{H} = \{\boldsymbol{h}_j\}_{j=1}^{N}$; keep budget V'; Stage-1 allocation $\alpha \in [0, 1]$; top-K operating parameter
Ensure: Survival set $\mathcal{S} \subseteq \mathcal{V}$ with per-token provenance labels

// Stage 0: signal accumulation (single calibration pass)
$\pi_i \leftarrow \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}[i \in \text{Top-}K(W\boldsymbol{h}_j)]$ ▷ participation, Eq. 5
$\mathcal{S}_0 \leftarrow \{i : \pi_i = 0\}$ ▷ zero-participation certificate
// Stage 1: partition by participation
$\tau_{\text{high}} \leftarrow \lceil \alpha V' \rceil$-th largest π_i among $\{i : \pi_i > 0\}$
$\mathcal{A}_{\text{keep}} \leftarrow \{i : \pi_i \geq \tau_{\text{high}}\}$ ▷ **auto-keep**
$\mathcal{A}_{\text{prune}} \leftarrow \mathcal{S}_0$ ▷ **auto-prune**
$\mathcal{A}_{\text{mid}} \leftarrow \mathcal{V} \setminus (\mathcal{A}_{\text{keep}} \cup \mathcal{A}_{\text{prune}})$
// Stage 2: rank the middle band by substitution cost
$\mathcal{R} \leftarrow \mathcal{A}_{\text{keep}} \cup \mathcal{A}_{\text{mid}}$ ▷ $\mathcal{A}_{\text{prune}}$ deliberately excluded
for all $i \in \mathcal{A}_{\text{mid}}$ **do**
    $\text{nn}_{\mathcal{R}}(i) \leftarrow \arg\min_{k \in \mathcal{R} \setminus \{i\}} \|\boldsymbol{w}_i - \boldsymbol{w}_k\|_2$
    $\text{cost}(i \mid \mathcal{R}) \leftarrow \frac{1}{N} \sum_{j=1}^{N} p^{\star}(\boldsymbol{h}_j)_i \big(\boldsymbol{w}_i^{\top} \boldsymbol{h}_j - \boldsymbol{w}_{\text{nn}_{\mathcal{R}}(i)}^{\top} \boldsymbol{h}_j\big)$ ▷ Eq. 8
end for
$\mathcal{A}_{\text{mid-keep}} \leftarrow$ top-$(V' - |\mathcal{A}_{\text{keep}}|)$ elements of $\mathcal{A}_{\text{mid}}$ ranked by cost ▷ **mid-keep**
remaining $\mathcal{A}_{\text{mid}} \setminus \mathcal{A}_{\text{mid-keep}} \rightarrow$ **mid-prune**
return $\mathcal{S} = \mathcal{A}_{\text{keep}} \cup \mathcal{A}_{\text{mid-keep}}$
▷ if $|\mathcal{A}_{\text{prune}}| > V - V'$: return $\mathcal{S}$ with $|\mathcal{S}| < V'$ as diagnostic

By the definition of $i^{\star}(\boldsymbol{h}_j)$ as the full-vocabulary argmax, $\ell_{i^{\star}(\boldsymbol{h}_j)}(\boldsymbol{h}_j) \geq \ell_k(\boldsymbol{h}_j)$ for every $k \in \mathcal{V}$, with strict inequality (after tie-breaking) for $k \neq i^{\star}(\boldsymbol{h}_j)$. Restricting to $k \in \mathcal{S}$ preserves this maximum because, by (S3), $i^{\star}(\boldsymbol{h}_j)$ itself lies in $\mathcal{S}$ and therefore remains a candidate:

$$\arg\max_{k \in \mathcal{S}} \ell_k(\boldsymbol{h}_j) \;=\; i^{\star}(\boldsymbol{h}_j). \tag{S4}$$

The pruned-head distribution (Eq. 2) at $\boldsymbol{h}_j$ is given as:

$$p^{\mathcal{S}}(\boldsymbol{h}_j)_k \;=\; \frac{\exp \ell_k(\boldsymbol{h}_j)}{Z_{\mathcal{S}}(\boldsymbol{h}_j)}, \qquad Z_{\mathcal{S}}(\boldsymbol{h}_j) \;=\; \sum_{k \in \mathcal{S}} \exp \ell_k(\boldsymbol{h}_j) \;>\; 0, \qquad k \in \mathcal{S}.$$

At fixed $\boldsymbol{h}_j$, every surviving $\exp \ell_k(\boldsymbol{h}_j)$ is divided by the same strictly positive scalar $Z_{\mathcal{S}}(\boldsymbol{h}_j)$, so the ordering of $p^{\mathcal{S}}(\boldsymbol{h}_j)_k$ across $k \in \mathcal{S}$ matches that of $\exp \ell_k(\boldsymbol{h}_j)$; and since exp is strictly monotone, that matches the ordering of $\ell_k(\boldsymbol{h}_j)$ across $k \in \mathcal{S}$. Hence

$$\arg\max_{k \in \mathcal{S}} p^{\mathcal{S}}(\boldsymbol{h}_j)_k \;=\; \arg\max_{k \in \mathcal{S}} \ell_k(\boldsymbol{h}_j) \;\overset{\text{(S4)}}{=}\; i^{\star}(\boldsymbol{h}_j). \tag{S5}$$

Note that (S5) states that the pruned-head greedy decode at $\boldsymbol{h}_j$ equals the full-head greedy decode $i^{\star}(\boldsymbol{h}_j)$. Since $\boldsymbol{h}_j$ was arbitrary, this holds at every calibration position.

C.2 Extended Analysis of Participation Cost on Qwen2.5-0.5B and Gemma-2-2B

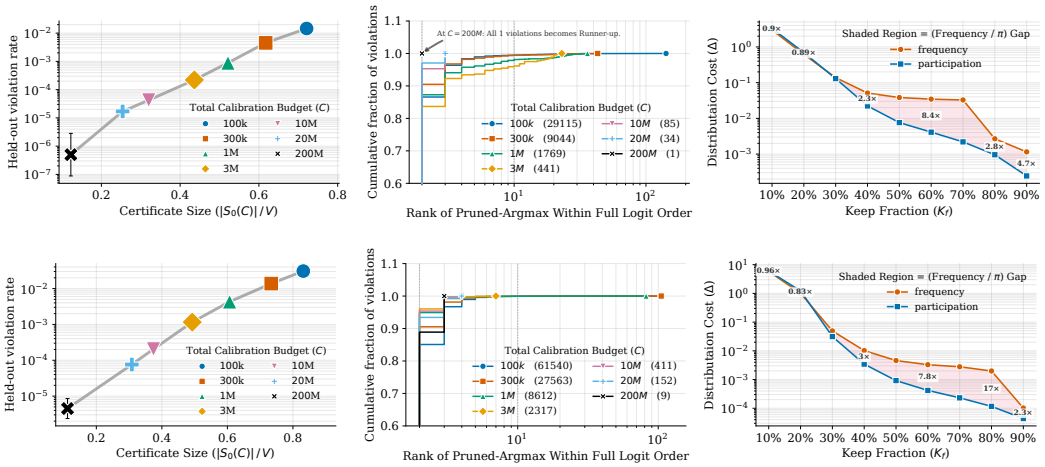

Figure 9: **Investigating Participation Cost on Additional Models:** Top row for Qwen2.5-0.5B (V=151,936) and Bottom row for Gemma-2-2B (V=256,000) (Nemotron-CC-v2 calibration; $|\mathcal{H}_{\text{test}}|$=20M disjoint held-out positions; K=20). **(a)** Certificate size $|\mathcal{S}_0(C)|/V$ and held-out violation rate as the calibration tokens increase C=$10^5 \rightarrow 2 \times 10^8$ tokens. **(b)** Empirical CDF over violations caused by pruned `LM-Head` with K_f=0.5 and corresponding ranking with full logit order. **(c)** Held-out P_{99} of Δ (Eq. 3) for π-only vs. Frequency-only pruning at $f \in \{10, \ldots, 90\}\%$.