# ProbeRoute: Probes as Routing Priors for Frozen-Backbone Multi-Token Prediction

Ali Uyar
Independent Researcher

**Abstract**

Frozen autoregressive language models are trained for next-token prediction, but their hidden states can still encode information about several future tokens. ProbeRoute tests whether that latent structure can be converted into a better explicit multi-token adapter without unfreezing the backbone. The method first runs a one-time offline probe stage across depth, then uses probe-derived top-5 scores to initialize a sparse top-$m$ router over frozen hidden states. Under a stage-gated protocol with mandatory probes, screening baselines, a finalist-selection step, and a final 1B rerun, the resulting sparse adapter beats the strongest selected dense frozen-backbone baseline on the paper's two headline held-out metrics. On the final 1B comparison, ProbeRoute improves `mean_top1_h2_h4` from 0.1162 to 0.1172 and the speculative draft acceptance proxy `mean_accept_len` from 1.1110 to 1.1188, while changing `mean_nll_h1_h4` only from 5.3769 to 5.3780. The practical point is therefore stronger than a small raw delta: the gain comes from a more selective routing interface, not from a heavier backbone adaptation recipe. The mechanistic story is equally important. Probe heatmaps reveal horizon-dependent depth structure at both 410M and 1B, the learned sparse router concentrates mass on the same depth bands, and the random-initialization ablation is weaker than probe-initialized sparse routing. By contrast, loss warmup and deeper far-horizon heads do not improve over the base sparse configuration at the tested budget. The evidence therefore supports a focused conclusion: in this frozen-backbone setting, future-token probes are not merely descriptive diagnostics; they are a useful routing prior for explicit multi-token prediction.

## 1 Introduction

Explicit multi-token prediction is attractive because it asks a language model for more than the next token at each decoding position. In principle, that creates room for richer future supervision during adaptation and for block-level acceptance proxies at evaluation time. In practice, however, a pretrained autoregressive backbone is optimized for one-step prediction, so a naive multi-token adapter can easily read from the wrong part of the network. The central question in this paper is whether a frozen backbone already contains enough horizon-specific future structure to support a better adapter than standard last-layer or dense layer-mixing baselines.

ProbeRoute answers that question with a deliberately constrained recipe. We keep the pretrained backbone frozen, probe every layer for future-token predictability, and use the probe scores to initialize a sparse per-horizon layer router. The key idea is simple: if future-token information is distributed non-uniformly across depth, then probes should identify which layer bands matter for each horizon, and that information should be useful as an adaptation prior rather than remaining a diagnostic artifact. The paper's practical pitch follows directly from that design. Instead of adding a denser adaptation interface, ProbeRoute uses a cheap offline diagnostic to build a more selective router.

The resulting study is intentionally focused. It does not ask whether sparse routing universally dominates dense routing, whether frozen adapters beat full finetuning, or whether the acceptance proxy is a direct throughput benchmark. Instead, it asks a sharper question that the artifact packet can answer cleanly: *does probe-initialized sparse routing improve frozen-backbone multi-token prediction relative to the strongest dense baseline selected under the same screening and final-rerun protocol?*

The completed evidence supports a positive answer. Mandatory probe runs at 410M and 1B reveal non-flat, horizon-dependent depth structure. Screening then identifies a strong dense weighted-hidden-state baseline and reruns it at the final budget as the finalist comparator. Against that selected dense finalist, the final 1B sparse run improves the future-token top-1 aggregate and the speculative draft acceptance proxy while changing aggregate NLL only minimally. The ablation stack narrows the interpretation further: random initialization weakens the sparse model, while extra warmup and deeper far-horizon heads do not explain the gain.

This yields three contributions that are fully supported by the packet bundled with this manuscript:

- We show that future-token probe heatmaps reveal horizon-dependent depth structure at both 410M and 1B, with shorter horizons peaking later and farther horizons shifting earlier in depth.
- We turn those probe measurements into a concrete frozen-backbone adapter—a sparse top-$m$ layer router initialized from probe top-5 scores—and evaluate it against screening-selected last-layer and dense weighted-hidden-state baselines.
- We find that the final 1B sparse run beats the strongest selected dense baseline on held-out future-token top-1 and speculative draft acceptance metrics, while relying

on a more selective routing interface rather than a denser adaptation recipe; the random-init ablation is the one that meaningfully weakens that result.

## 2 Related Work and Positioning

This paper touches three nearby lines of work, but it occupies a narrower space than any of them.

**Multi-token prediction and multi-head decoding.** Recent work has revisited the standard next-token objective by asking language models to predict multiple future tokens in parallel. Gloeckle et al. [4] study multi-token prediction during large-scale language-model training and show gains in sample efficiency and downstream generation quality. A separate line of work uses multiple decoding heads primarily for faster inference. Speculative decoding introduces a draft-and-verify framework for exact acceleration without changing the output distribution [8]. Medusa replaces a separate draft model with lightweight decoding heads on top of the base model [3], while Hydra improves those heads by making them sequentially dependent [2]. EAGLE pushes the same acceleration theme further with feature-level extrapolation for speculative sampling [10]. Our setting is different from all of these in two ways: the backbone remains frozen throughout, and the main question is not raw decoding speed but whether horizon-specific future structure can be exploited to build a stronger frozen-backbone multi-token adapter.

**Probes as analysis tools.** The probe stage draws on a well-established interpretability tradition. Linear classifier probes were introduced as a way to inspect intermediate representations without perturbing the original training dynamics [1]. Later work emphasized that probe accuracy alone can be misleading unless probe capacity and selectivity are controlled carefully [5]. We adopt the core probing intuition from that literature, but use probes differently. In this paper, probes are not only descriptive diagnostics for analyzing hidden states after the fact. They are used operationally to define a routing prior for the downstream adapter.

**Frozen-backbone adaptation.** The broader motivation also overlaps with parameter-efficient adaptation methods, which keep most or all pretrained parameters fixed while learning a lightweight task-specific interface. Adapters [6], prefix tuning [9], and LoRA [7] all pursue that general objective. ProbeRoute differs in what is being adapted. Instead of inserting low-rank weight updates or learned prompts, the method learns how to access existing hidden-state depth structure in a frozen model, with probes supplying the initial guess for which layers matter at each horizon.

Taken together, these neighboring literatures help clarify the paper's contribution. We are not proposing a general speculative decoding system, a new pretraining recipe for large language models, or a generic parameter-efficient fine-tuning method. The contribution supported by the packet is narrower and more mechanistic: future-token probes reveal usable depth structure, and that structure can be converted into a sparse routing prior that improves frozen-backbone multi-token prediction relative to the strongest dense baseline selected under the same protocol.

## 3 Method

### 3.1 Problem setup

Let a frozen autoregressive backbone produce hidden states $h_t^{(\ell)}$ at token position $t$ and layer $\ell \in \{0, \dots, L-1\}$. The adaptation problem is to predict future tokens $x_{t+k}$ for horizons $h \in \{1, 2, 3, 4\}$ while changing only a lightweight router and horizon heads. The study compares three adapter families:

- **Last-layer baselines**: decode each horizon only from the final hidden state.
- **Dense weighted-hidden-state (WHS) baselines**: learn a dense mixture over all layers for each horizon.
- **Sparse top-$m$ routing**: learn a mixture over only a small horizon-specific subset of layers.

The proposed method differs from the sparse random-init ablation only in how that subset and its initial weights are chosen.

### 3.2 Probe stage

The first stage trains horizon-specific future-token probes over every layer of the frozen backbone. The packet's canonical probe defaults use rank-64 probes, a frozen backbone, and validation top-5 as the initialization metric. For each layer $\ell$ and horizon $k$, the probe stage produces a validation score $q_{\ell,k}$, which we treat as a measurement of how useful that layer is for predicting $x_{t+k}$.

The probes are not an end in themselves. Their role is to summarize a layer × horizon landscape before the main adapter is trained. Operationally, this stage is a one-time offline diagnostic run once per backbone family to extract architecture priors before adapter training, not a recurring component of test-time inference. In the final paper narrative, this is the key methodological shift: probe heatmaps are used as an *input* to the adapter, not merely as a post hoc visualization.

### 3.3 From probe scores to sparse routers

For dense WHS, each horizon learns a full mixture over all layers. For ProbeRoute, we instead form a sparse support set $S_k$ by taking the top-$m$ layers under probe-derived top-5 scores, with $m = 4$ in all main sparse runs. The router is then initialized from z-scored probe top-5 values on that support and trained jointly with the horizon heads while the backbone remains frozen.

Writing the learned router weights as $\alpha_{\ell,k}$, the mixed representation for horizon $k$ is

$$\tilde{h}_t^{(k)} = \sum_{\ell \in S_k} \alpha_{\ell,k} \, \mathrm{LN}\left(h_t^{(\ell)}\right), \tag{1}$$

where $\mathrm{LN}(\cdot)$ is the stateless layer normalization used by the packet's canonical MTP adapter. A residual MLP head then maps $\tilde{h}_t^{(k)}$ to logits for the token at horizon $k$.

The dense WHS baselines use the same frozen-backbone setting and the same family of horizon heads. The comparison therefore isolates the routing strategy rather than head capacity or backbone updates.

## 4 Experimental Protocol

### 4.1 Data, models, and budgets

The packet's experiment plan uses deterministic processed FineWeb-Edu subsets throughout. Probes are run at 410M and 1B with sequence length 1024 and budgets of 5M train / 1M validation / 1M test tokens. Screening and ablation runs use Pythia-1B at sequence length 2048 with 20M / 2M / 2M token budgets. Final 1B reruns use 50M / 5M / 5M token budgets. The low-cost confirmatory path reruns the ProbeRoute recipe at 410M with a second seed and a 20M / 2M / 2M budget.

The canonical MTP defaults in the plan are fixed across the relevant runs: frozen backbones, horizons $\{1, 2, 3, 4\}$, residual-MLP horizon heads, stateless layer normalization, router temperature 0.5, entropy penalty $\beta = 10^{-3}$, and top-$m = 4$ for sparse routing.

### 4.2 Screening and finalist selection

The five mandatory screening runs are:

(i) `BASE_LAST_LINEAR_1B`,
(ii) `BASE_LAST_MLP_1B`,
(iii) `BASE_DENSE_WHS_RANDOM_1B`,
(iv) `BASE_DENSE_WHS_PROBE_1B`, and
(v) `MAIN_SPARSE_PROBE_1B`.

Finalist selection is deliberately conservative: only completed *non-sparse* screening baselines are eligible, and the primary score is the validation future-token top-1 aggregate at the selected best checkpoint. Under that rule, Dense WHS Probe-Init is the finalist baseline and is promoted to the 50M-token final comparison against the final sparse rerun.

### 4.3 Metrics

The study tracks future-token quality at horizons 1–4 using top-1, top-5, and negative log-likelihood (NLL). The paper's main aggregate metric is the mean top-1 over horizons 2–4, because it emphasizes the explicitly multi-token part of the task rather than allowing the horizon-1 term to dominate. The second headline metric is mean accepted-prefix length, measured with $K_{\max} = 4$ on deterministic prefixes. We treat it as a speculative draft acceptance proxy: following the experiment plan, the evaluation rolls the frozen base model greedily for up to four tokens, compares the proposed MTP block against that rollout, records the longest matching prefix, appends exactly one base-greedy token, and repeats. This is a practical proxy, not an end-to-end throughput benchmark.

All non-probe scientific results are evaluated from best checkpoints chosen on validation data. That contract is part of the evidence policy bundled with the packet and is important to the paper's interpretation.

## 5 Probes Reveal Depth Structure Across Horizons

The probe results establish the paper's mechanism story before the main adapter is trained. At 1B, the best layer for horizon 1 is layer 13, while the best layers shift earlier for horizons 2–4 (layers 10, 9, and 9 respectively). At 410M, the shift is even larger: the best layer moves from 22 at horizon 1 to 15, 13, and 12 for horizons 2–4. The heatmaps are therefore not flat; they show that the depth band with the strongest future-token signal depends on how far into the future the model is asked to predict.

The spread of probe top-5 accuracy across layers reinforces the same point. At 1B, the range from worst to best layer is 5.90pp for horizon 1, 3.11pp for horizon 2, 1.59pp for horizon 3, and 1.12pp for horizon 4. At 410M, the corresponding spreads are 7.07pp, 3.35pp, 1.18pp, and 0.78pp. The far-horizon signal is weaker, but it is still structured rather than uniform.

The router analysis in figure 2 shows that the learned sparse adapter respects this measured structure. In the final 1B sparse run, the selected layer sets are tightly concentrated around the best probe layers for each horizon. Horizon 1 remains late, horizons 3 and 4 move earlier, and horizon 2 falls between them. This alignment matters because it links the one-time offline diagnostic to the downstream adapter rather than treating the two as unrelated analyses.

Together, the mandatory probes support the paper's first claim: future-token information in a frozen autoregressive backbone is depth-structured and horizon-dependent, and that structure is strong enough to guide an explicit routing prior.

## 6 Final 1B Comparison

### 6.1 Screening selects a strong dense baseline

Before making any headline claim, the study runs a full screening sweep at the 20M-token budget. This step is important because it prevents the final comparison from using a weak straw baseline. Among the non-sparse screening baselines, dense weighted-hidden-state mixing with probe initialization achieves the highest validation `mean_top1_h2_h4` and is therefore promoted to the final-budget rerun.

The screening results already suggest a useful distinction between adaptation goals. The last-layer residual-MLP baseline attains the best screening NLL, but it lags badly on `mean_top1_h2_h4`. By contrast, the dense probe-initialized WHS baseline and the sparse screening run are much closer on the explicitly multi-token metrics that the paper cares about. This is why the final comparison should be read as a comparison between two serious frozen-backbone MTP adapters, not between a proposed method and a deliberately weak baseline.

### 6.2 ProbeRoute wins the final held-out comparison

The final 1B comparison is the paper's central quantitative result. Relative to the selected dense finalist, ProbeRoute improves the future-token top-1 aggregate from 0.1162 to 0.1172 and mean accepted-prefix length from 1.1110 to 1.1188.

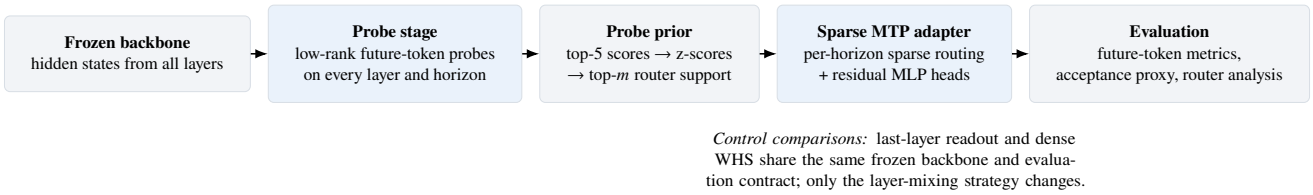


**Figure 1: Study design.** ProbeRoute converts probe measurements into a routing prior. Every stage keeps the pretrained backbone frozen; the comparison is therefore about how future information is accessed, not about unfreezing or large-scale retraining.

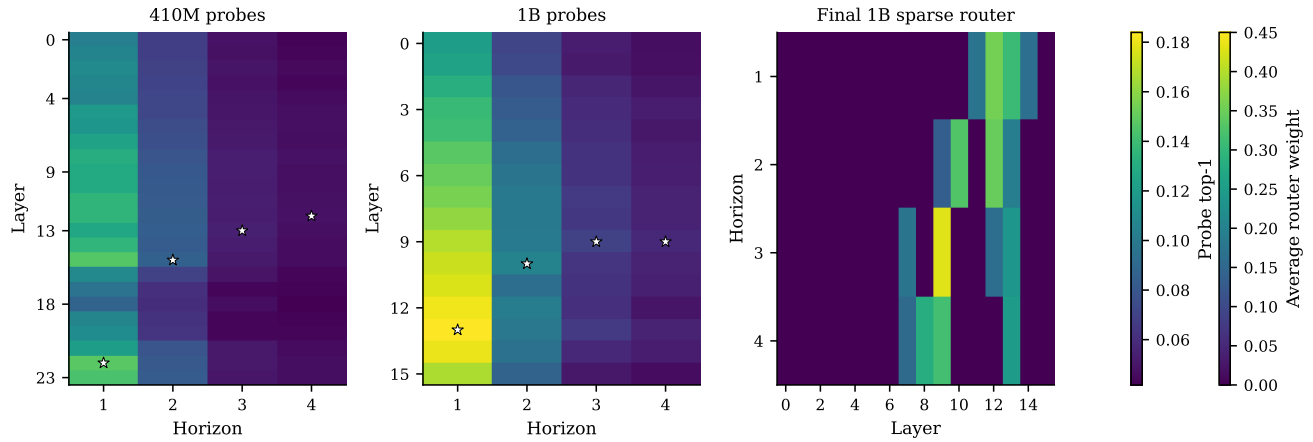


**Figure 2: Mandatory probe heatmaps and final sparse-router support.** Left and center: probe top-1 heatmaps for the 410M and 1B mandatory probe runs. The brightest layers move earlier as the prediction horizon increases. Right: average learned router weights for the final 1B sparse model. The key visual result is that the learned sparse router autonomously recovers the same horizon-dependent depth bands highlighted by the offline probes. In particular, the final sparse router selects layers {12,13,11,14}, {12,10,13,9}, {9,13,7,12}, and {9,8,13,7} for horizons 1–4, closely matching the probe peaks at layers 13, 10, 9, and 9.

The dense finalist keeps a tiny likelihood advantage in the aggregate, but the difference in mean NLL is only 0.00110. In relative terms, ProbeRoute improves the main multi-token accuracy aggregate by roughly 0.85% and the speculative draft acceptance proxy by roughly 0.70%, while the NLL change is only about 0.02%. That combination is the practical hook of the result: a one-time offline probe pass yields a more selective routing interface than dense WHS mixing, yet still improves the task-aligned held-out metrics.

The per-horizon view in figure 3 is especially useful. Top-1 improves at every horizon, not just in the aggregate. The largest gain appears at horizon 3, which is where explicit multi-token structure is already far enough from the next-token objective to make routing matter, but not so far that the signal collapses. The NLL trade-off is concentrated at horizon 4; the sparse model is actually slightly better on NLL at horizons 1–3.

A second, subtler point is that the final 50M-token rerun reverses the narrow screening deficit of the sparse model relative to the dense finalist. We do not turn that into a scaling claim, because the packet contains only one final-budget comparison, but the reversal does make the final win more credible: the result is not inherited from screening, it emerges under the deliberate final-rerun protocol.

# 7 Ablations, Confirmatory Evidence, and Discussion

## 7.1 The probe prior is the decisive ingredient

The sparse ablations are informative precisely because they do *not* all move in different directions. The random-init ablation is the one that clearly weakens the method on the targeted metrics, while warmup and deeper far-horizon heads do not improve on the base sparse configuration at the packet's reported precision.

At the 20M-token budget, the base sparse screening run reaches a future-token top-1 aggregate of 0.1061 and a mean accepted-prefix length of 0.9863. Removing the probe prior drops these to 0.1037 and 0.9754. That is the cleanest mechanism test in the paper: changing the sparse model's initialization while leaving the rest of the frozen-backbone recipe intact weakens the two metrics that define success in this study.

The warmup and deephead ablations tell a different story. Both match the base sparse configuration at the reported aggregate precision. This makes the positive result more convincing, not less: the final gain is not well explained by generic optimization polish, loss-scheduling tricks, or extra head complexity. The paper's most defensible interpretation is therefore also its sharpest one—the probe prior itself is the decisive ingredient.

**Table 1: Screening results at the 20M-token budget.** The finalist baseline is selected from non-sparse baselines by validation `mean_top1_h2_h4`. The sparse screening run is shown for context but is not eligible for baseline selection. Higher is better for `mean_top1_h2_h4` and `mean_accept_len`; lower is better for `mean_nll_h1_h4`.

| Method | Init | Val top-1 h2–4 | Test top-1 h2–4 | Test accept len | Test NLL h1–4 |
|---|---|---|---|---|---|
| Last-layer linear | – | 0.0932 | 0.0941 | 0.6530 | 6.1853 |
| Last-layer residual MLP | – | 0.1004 | 0.1011 | 1.0712 | **5.4016** |
| Dense WHS | random | 0.1031 | 0.1039 | 0.8690 | 5.7361 |
| Dense WHS | probe-init | **0.1056** | **0.1066** | 0.9987 | 5.6230 |
| Sparse top-$m$ (proposed) | probe-init | 0.1053 | 0.1061 | 0.9863 | 5.6412 |

DENSE WHS PROBE-INIT is the strongest non-sparse baseline by validation `mean_top1_h2_h4` and becomes the final comparator. The sparse screening run trails narrowly at screening budget, which makes the later final-budget win more informative.

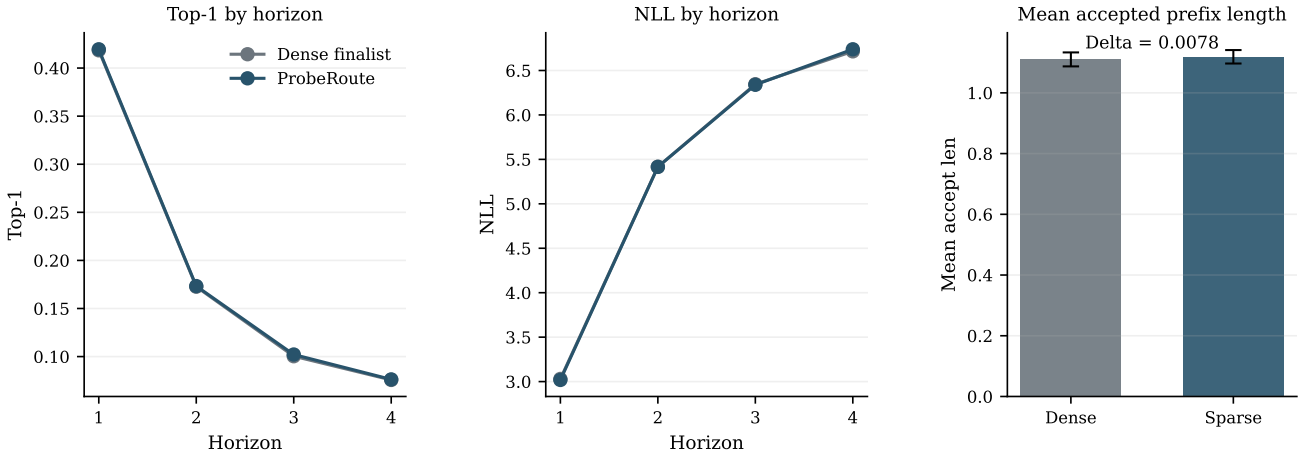


**Figure 3: Final 1B metric profile.** Left: the sparse final run improves top-1 at every horizon. Center: NLL is slightly better for the sparse model at horizons 1–3 and slightly worse at horizon 4, which explains the near-neutral aggregate likelihood change. Right: the sparse model also improves the speculative draft acceptance proxy `mean_accept_len` on the final test set.

**Table 2: Final 1B comparison at the 50M-token budget.** Both finalists use probe-derived initialization. PROBEROUTE wins on the two headline held-out metrics while changing mean NLL only minimally.

| Model | Test top-1 h2–4 | Test accept len | Test NLL h1–4 |
|---|---|---|---|
| Dense finalist | 0.1162 | 1.1110 | **5.3769** |
| PROBEROUTE (proposed) | **0.1172** | **1.1188** | 5.3780 |
| Δ (sparse−dense) | +0.00099 | +0.00781 | +0.00110 |

The NLL change is only +0.00110 in absolute terms, which is about 0.02% relative to the dense finalist.

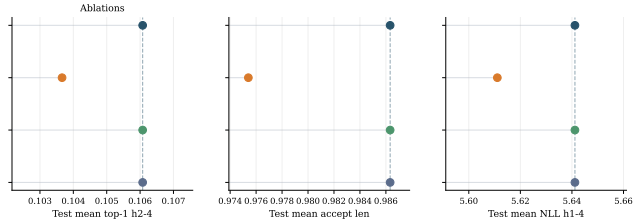


**Figure 4: Screening-budget sparse ablations.** The base sparse configuration uses probe initialization. Random initialization weakens both `mean_top1_h2_h4` and `mean_accept_len`. The warmup and deephead variants coincide with the base sparse metrics at reported precision. Lower NLL does not rescue the random-init ablation because the study's targeted future-token and acceptance metrics both drop.

### 7.2 Confirmatory fallback

The project's low-cost confirmatory path reruns PROBEROUTE at 410M with a second seed rather than using the more expensive 2.8B extension. The confirmatory run completes with a future-token top-1 aggregate of 0.0989, a mean accepted-prefix length of 0.9852, and an aggregate NLL of 5.7844. This is supporting evidence only. It does not replace the final 1B headline comparison, and it is not used to claim cross-scale superiority. Its value is narrower: it shows that the same recipe completes cleanly in a second-seed, smaller-model setting and preserves the qualitative shape of a functioning PROBEROUTE adapter.

### 7.3 Discussion and limitations

The strongest supported story in the packet is now clear. Future-token probes reveal useful depth structure in frozen autoregressive backbones. Turning those probes into a sparse routing prior yields a better explicit multi-token adapter than the strongest selected dense frozen-backbone baseline under the same screening and final-rerun contract. The ablations then isolate why that happens: probe initialization matters, while the tested warmup and deeper-head variants do not

explain the gain.

The study remains deliberately narrow. It is confined to the Pythia family, frozen-backbone adaptation, a single final 1B seed, and a low-cost 410M confirmatory fallback rather than a larger confirmatory extension. The effect sizes are real but modest, and `mean_accept_len` remains a block-acceptance proxy rather than a production-throughput measurement. We therefore present the result as a focused contribution paper, not as a universal claim about all multi-token predictors or all sparse adapters.

## 8 Conclusion

This paper asked a targeted question: can future-token probes be turned into a useful routing prior for explicit multi-token prediction on top of a frozen pretrained backbone? Under the completed ProbeRoute protocol, the answer is yes. Mandatory probe runs at 410M and 1B show horizon-dependent depth structure, the final sparse router places mass on the same layer bands, the final 1B sparse model beats the strongest selected dense baseline on held-out `mean_top1_h2_h4` and the speculative draft acceptance proxy `mean_accept_len`, and the random-init ablation is the one that meaningfully weakens the sparse recipe.

That combination of results is what makes the contribution coherent. The paper does not rely on a single positive comparison or on a process-heavy narrative. It makes a specific mechanistic claim and backs it with probes, screening, final reruns, and ablations. The broader takeaway is that ProbeRoute turns a cheap offline diagnostic into an architectural prior: a one-time probe stage yields a more selective router than the dense finalist while improving the task-aligned held-out metrics that matter for explicit multi-token drafting. Within this frozen-backbone setting, probes are not only descriptive. They are useful routing priors.

## References


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

[2] Zachary Ankner, Mansheej Paul Abraham, Yu Bai Chen, Kenji Cho, Jonas Geiping, Tom Goldstein, RJ Gupta, Mark McKenzie, Marco Noci, Jonathan Pilault, et al. Hydra: Sequentially-dependent draft heads for medusa decoding. *arXiv preprint arXiv:2402.05109*, 2024.

[3] Tianle Cai, Yuhong Li, Zhengyang Geng, Hongwu Peng, Jason D. Lee, Deming Chen, and Tri Dao. Medusa: Simple LLM inference acceleration framework with multiple decoding heads. In *Proceedings of the 41st International Conference on Machine Learning*, pages 5209–5235. PMLR, 2024.

[4] Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Roziere, David Lopez-Paz, and Gabriel Synnaeve. Better & faster large language models via multi-token prediction. *arXiv preprint arXiv:2404.19737*, 2024.

[5] John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 2733–2743. Association for Computational Linguistics, 2019. doi: 10.18653/v1/D19-1275.

[6] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[7] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[8] Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19274–19286. PMLR, 2023.

[9] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 4582–4597. Association for Computational Linguistics, 2021.

[10] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. EAGLE: Speculative sampling requires rethinking feature uncertainty. In *Proceedings of the 41st International Conference on Machine Learning*, pages 28935–28948. PMLR, 2024.

# A Additional Tables and Provenance

## A.1 Probe summary

**Table 3:** Best probe layer by horizon and top-5 spread across layers. The best layer shifts earlier as the horizon increases in both mandatory probe runs. Top-5 spread is reported in percentage points.

| Model | Horizon | Best layer | Best top-1 (%) | Top-5 spread (pp) |
|---|---|---|---|---|
| 1B | 1 | 13 | 18.39 | 5.90 |
| 1B | 2 | 10 | 10.54 | 3.11 |
| 1B | 3 | 9 | 6.97 | 1.59 |
| 1B | 4 | 9 | 5.82 | 1.12 |
| 410M | 1 | 22 | 14.76 | 7.07 |
| 410M | 2 | 15 | 8.57 | 3.35 |
| 410M | 3 | 13 | 5.41 | 1.18 |
| 410M | 4 | 12 | 4.86 | 0.78 |

## A.2 Ablation table

**Table 4:** Aggregate test metrics for the sparse screening configuration and its planned ablations.

| Variant | Test top-1 h2–4 | Test accept len | Test NLL h1–4 |
|---|---|---|---|
| ProbeRoute | 0.1061 | 0.9863 | 5.6412 |
| Sparse random-init | 0.1037 | 0.9754 | 5.6111 |
| Sparse + warmup | 0.1061 | 0.9863 | 5.6412 |
| Sparse + deephead | 0.1061 | 0.9863 | 5.6412 |

## A.3 Resource summary

**Table 5:** Execution summary for the authoritative runs included in the paper packet.

| Experiment | Stage | Backbone | Train budget | Best step |
|---|---|---|---|---|
| PROBE 410M | probe | Pythia-410m | 5M | 77 |
| PROBE 1B | probe | Pythia-1b | 5M | 77 |
| BASE DENSE WHS PROBE 1B | screen | Pythia-1b | 20M | 153 |
| MAIN SPARSE PROBE 1B | screen | Pythia-1b | 20M | 153 |
| FINAL BEST BASELINE 1B | final | Pythia-1b | 50M | 382 |
| MAIN SPARSE PROBE 1B FINAL | final | Pythia-1b | 50M | 375 |
| ABL SPARSE RANDOM 1B | ablation | Pythia-1b | 20M | 153 |
| ABL SPARSE WARMUP 1B | ablation | Pythia-1b | 20M | 153 |
| ABL SPARSE DEEPHEAD 1B | ablation | Pythia-1b | 20M | 153 |
| CONFIRM BEST 410M SEED2 | confirm | Pythia-410m | 20M | 153 |

## A.4 Resolved config summary

**Table 6:** Resolved experiment identities from the packet's config summary table.

| Experiment | Stage | Backbone | Layer mix | Router init | Seq. length |
|---|---|---|---|---|---|
| ABL SPARSE DEEPHEAD 1B | ablation | pythia-1b | sparse-topm | probe-zscore-top5 | 2048 |
| ABL SPARSE RANDOM 1B | ablation | pythia-1b | sparse-topm | random | 2048 |
| ABL SPARSE WARMUP 1B | ablation | pythia-1b | sparse-topm | probe-zscore-top5 | 2048 |
| CONFIRM BEST 410M SEED2 | confirm | pythia-410m | sparse-topm | probe-zscore-top5 | 2048 |
| FINAL BEST BASELINE 1B | final | pythia-1b | dense-whs | probe-zscore-top5 | 2048 |
| MAIN SPARSE PROBE 1B FINAL | final | pythia-1b | sparse-topm | probe-zscore-top5 | 2048 |
| PROBE 1B | probe | pythia-1b | sparse-topm | none | 1024 |
| PROBE 410M | probe | pythia-410m | sparse-topm | none | 1024 |
| BASE DENSE WHS PROBE 1B | screen | pythia-1b | dense-whs | probe-zscore-top5 | 2048 |
| BASE DENSE WHS RANDOM 1B | screen | pythia-1b | dense-whs | random | 2048 |
| BASE LAST LINEAR 1B | screen | pythia-1b | last-layer | none | 2048 |
| BASE LAST MLP 1B | screen | pythia-1b | last-layer | none | 2048 |
| MAIN SPARSE PROBE 1B | screen | pythia-1b | sparse-topm | probe-zscore-top5 | 2048 |

## A.5 Claim-evidence map

**Table 7:** Artifact-backed claim map bundled with the paper packet.

| Claim | Evidence source | Status |
|---|---|---|
| C1 | Probe score matrices from `PROBE_1B` and `PROBE_410M` | supported |
| C2 | Final 1B comparison registry (`main_results.csv`) | supported |
| C3 | Final acceptance-proxy metrics for the dense and sparse 1B finalists | supported |
| C4 | Screening-budget sparse ablation registry (`ablation_results.csv`) | supported |