

Surfing the Activation Wave: Analyzing the Information and Geometry of LLM Hidden States for Speculative Decoding

Raghav Goel*, Sudhanshu Agrawal*^ψ, Mukul Gagrani, Risheek Garrepalli, Mingu Lee

Qualcomm AI Research and ^ψ Stanford University

{raghgoel,rgarrepa,mgagrani,mingul}@qti.qualcomm.com, sudhagra@stanford.edu

Abstract

Decoder hidden states from different layers of large language models (LLMs) often encode complementary information. However, existing speculative decoding methods typically rely on a small set of target-model layers chosen using fixed, position-based heuristics. Selecting an optimal subset of layers is inherently a combinatorial problem, making exhaustive evaluation infeasible even for moderately deep models. In this work, we present a principled framework for multi-layer hidden-state selection for speculative decoding based on two complementary criteria: informativeness, measured through matrix-entropy analysis, and diversity, quantified using centered kernel alignment (CKA). We investigate layer-wise Shannon entropy, aggregate entropy under hidden-state concatenation, and pairwise CKA dynamics across depth, revealing distinct geometric and informational compartments within decoder representations. Motivated by these findings, we propose entropy and CKA-based layer selection algorithms that favor hidden-state sets that are both highly informative and minimally redundant. As part of this framework, we introduce token-entropy filtering, which removes hidden states associated with low predictive-entropy tokens, thereby improving the quality of layer-selection statistics. In EAGLE3-style speculative decoding, our selected sets outperform fixed baseline choices, including the default $[2, N//2, N-3]$ rule, while providing a clearer explanation of why some defaults can remain competitive. Our results provide a practical recipe for multi-layer hidden-state selection for spec-

ulative decoding, insights into the geometry and information in LLM decoder states, and a foundation for extending similar analysis to related applications.

1 Introduction

Hidden states in decoder-only LLMs capture the progressive transformation of token representations through depth. While prior studies analyze individual layers through rank, sparsity, and entropy, modern inference pipelines increasingly use *multiple* target-model hidden states jointly. In speculative decoding methods such as EAGLE3, selected target-model layers are fused and provided to the draft model. However, in practice these layers are usually selected by simple fixed rules (e.g., positional heuristics) rather than an explicit objective that balances informativeness and redundancy.

This creates both a practical and scientific gap. Practically, multi-layer selection is combinatorial. For a 36-layer model, choosing only three layers yields thousands of combinations, and the search space grows rapidly for larger tuples. Scientifically, it remains unclear which layer combinations are jointly informative versus geometrically redundant. As a result, default position-based heuristics (uniform selection with the initial and final layers offset) may perform well in some regimes without providing a principled explanation of when or why they work.

In this work, we view multi-layer hidden-state selection through the lenses of information maximization and

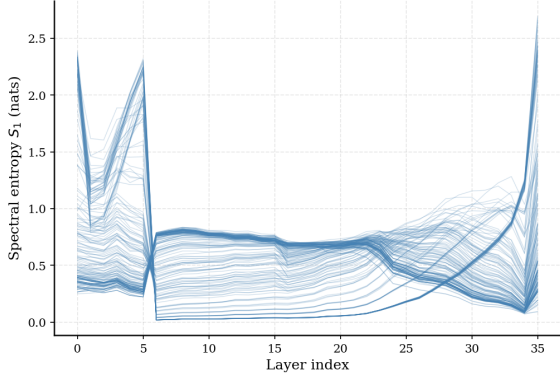


Figure 1: Prompt-wise per-layer Shannon entropy curves for Qwen3-4B-Instruct. The curves show a depth-wise non-uniform profile (an “activation wave”), with distinct early (0–6), middle (7–28), and late-layer (29–35) regimes.

redundancy minimization. We use matrix entropy to quantify informativeness of layer sets and linear CKA to quantify representational redundancy. We first introduce **Entropy-based Layer Set Iteration (ELSI)**, a practical entropy-based procedure that scales to realistic model depths. ELSI uses beam search to efficiently identify high-entropy layer sets by evaluating entropy on concatenated hidden states across layers. We then introduce **ELSI-C**, an entropy-initialized CKA-transition extension that promotes diversity-aware layer expansion. Alongside ELSI/ELSI-C, we introduce token-entropy filtering during calibration: hidden states of low predictive-entropy tokens are removed so that selection statistics focus on significant and informative positions.

We empirically study Qwen3-4B-Instruct-2507 and Llama-3.1-8B-Instruct. Our analyses reveal consistent trends across both models: (i) a non-uniform per-layer Shannon entropy profile (an *activation wave* across depth, Fig. 1), (ii) non-uniform entropy gain under layer concatenation, and (iii) CKA transition boundaries that indicate depth-wise representational compartments. These observations motivate ELSI, an entropy-based selector, and ELSI-C, a diversity-aware extension. Moreover, they clarify why fixed baselines can still be competitive when they overlap with high-scoring entropy-beam regions.

Contributions.

- We propose a principled framework for studying and selecting *multiple* target-model hidden layers using matrix entropy of concatenated representations and CKA transitions between depths.
- We introduce **ELSI**, a practical entropy-based selec-

tor that utilizes beam search to create informative layer sets.

- We introduce **ELSI-C**, an entropy-initialized CKA-transition extension for diversity-aware expansion.
- We introduce calibration token-entropy filtering, where hidden states of low predictive-entropy tokens are removed to sharpen layer-discriminative statistics.
- We demonstrate practical gains for EAGLE3-style conditioning using ELSI/ELSI-C-selected sets relative to the default $[N-1]$ and $[2, N//2, N-3]$ rules, while improving interpretability.

2 Related Work

Layer specialization and redundancy in transformers. Prior work shows that transformer layers are not uniformly useful and that representational redundancy is often concentrated in middle layers. Transformer layer-function studies report that many middle layers are partially interchangeable, while early and late layers remain more functionally distinct [1]. Related pruning/acceleration work similarly finds that substantial middle-layer computation can be removed with limited quality loss [2]. Representation-geometry analyses also suggest structured evolution of hidden states across depth [3, 4]. These studies motivate transition-aware reasoning: the depth axis can be divided into coherent compartments separated by sharp representational changes. To algorithmically identify these compartments, we introduce CKA-transition diagnostics.

Spectral and rank-based analyses. A complementary line of research studies rank and spectral structure for finetuning and compression. Adaptive rank-allocation methods for LoRA finetuning [5] demonstrate that rank requirements vary by layer/module and should not be treated uniformly [6, 7]. Other methods use low-rank approximations of transformer components for efficiency [8]. Random-matrix and singular-value analyses further show that spectral tails and non-dominant directions can materially affect behavior [9]. However, these works primarily deal with over-parameterization and compression - they do not address the problem of *selection* of sets of activations.

Speculative decoding and hidden-state conditioning. Recent speculative decoding frameworks such as EAGLE [10, 11] and DFlash [12] condition draft models on selected target-model hidden states. In current practice, those target layers are usually chosen by fixed, position-based heuristics (e.g., $[N-1]$ or $[2, N//2, N-3]$) which

may lead to redundant representations. Our work addresses this gap through ELSI, which maximizes information and diversity through matrix entropy and CKA-transition structure signals for principled multi-layer selection. We further introduce beam-overlap diagnostics to explain when and why position-based baseline rules still remain competitive.

Positioning of this work. Relative to prior layer-analysis and rank-based compression literature, our contribution is to *jointly* optimize hidden-layer sets for downstream draft conditioning: (i) quantify set informativeness via entropy of fused representations with ELSI, (ii) control redundancy via CKA-transition-guided expansion with ELSI-C, and (iii) evaluate practical gains in EAGLE3-style speculative decoding with interpretable diagnostics.

3 Preliminaries

Let the target model in the speculative decoding pipeline have N transformer layers indexed by $\{0, \dots, N-1\}$ and hidden dimension d . For a sequence of length T , let the hidden states output by layer- l be

$$h_l \in \mathbb{R}^{T \times d}. \quad (1)$$

For any selected layer set $\mathcal{L} \subseteq \{0, \dots, N-1\}$, define the concatenated representation

$$\bar{h}_{\mathcal{L}} = \text{concat}(\{h_l\}_{l \in \mathcal{L}}, \text{feature dim}) \in \mathbb{R}^{T \times |\mathcal{L}|d}. \quad (2)$$

Following EAGLE3-style conditioning, fused states are down-projected as

$$U_{\mathcal{L}} = \bar{h}_{\mathcal{L}} W_{\text{down}}, \quad W_{\text{down}} \in \mathbb{R}^{|\mathcal{L}|d \times d}. \quad (3)$$

Thus, the set of states $\mathcal{L}$ acts as a bottleneck, determining the information seen by the draft model.

3.1 Matrix Entropy

Following [13, 4, 14] for $X \in \mathbb{R}^{T \times D}$, we define $K_X = XX^\top$ and the normalized spectrum

$$p_i(X) = \frac{\lambda_i(K_X)}{\text{tr}(K_X)} = \frac{\sigma_i(X)^2}{\|X\|_F^2}, \quad (4)$$

where $\sigma_i(X)$ are the singular values. The Rényi- α matrix entropy is

$$S_\alpha(X) = \frac{1}{1-\alpha} \log \sum_i p_i(X)^\alpha, \quad \alpha > 0, \alpha \neq 1. \quad (5)$$

The limit as $\alpha \rightarrow 1$ is

$$S_1(X) = \lim_{\alpha \rightarrow 1} S_\alpha(X) = - \sum_i p_i(X) \log p_i(X). \quad (6)$$

For a set of hidden states $\mathcal{L}$, we consider the concatenated representation $\bar{h}_{\mathcal{L}}$ defined in Eq. (2) and compute the matrix entropy as $S_\alpha(\bar{h}_{\mathcal{L}})$. Notably, as compared to matrix-based *joint entropy* [13], this formulation of the entropy of a set of hidden states better represents the linear combination of features in the EAGLE-3 style conditioning that we see in Equation (3). In our experiments we use S_1 to study per-layer trends and select high entropy layers.

3.2 Centered-Kernel Alignment (CKA)

For representations $X \in \mathbb{R}^{T \times D_x}$ and $Y \in \mathbb{R}^{T \times D_y}$, we have Gram matrices $K = XX^\top$ and $L = YY^\top$, with the CKA defined as

$$\text{CKA}(X, Y) = \frac{\langle HKH, HLH \rangle_F}{\|H\|_F^2}, \quad H = I - \frac{1}{T} \mathbf{1}\mathbf{1}^\top. \quad (7)$$

Let $M_{ij} = \text{CKA}(h_i, h_j)$. We define CKA transition score at boundary between layers $\ell \rightarrow \ell + 1$ as

$$\delta_\ell = \|M_{\ell+1,:} - M_{\ell,:}\|_2, \quad \ell = 0, \dots, N-2. \quad (8)$$

Large peaks indicate compartment boundaries. The above equation can be extended for concatenated layers as well.

4 Method

We propose a three-stage layer-selection pipeline for speculative decoding: (i) token-entropy filtering within hidden states, (ii) hidden-state aggregation to build layer representatives, (iii.a) entropy-driven beam search for informative multi-layer sets (ELSI), and (iii.b) transition-aware CKA expansion for diversity-oriented refinement (ELSI-C). This section formalizes each stage and the evaluation protocol used in our experiments.

4.1 Token-Entropy Filtering

For token position t , predictive entropy is

$$\mathcal{H}_t = - \sum_{v \in \mathcal{V}} p_t(v) \log p_t(v). \quad (9)$$

We remove low-entropy hidden states within the sequence by top-fraction filtering

$$\Omega = \text{TopFrac}(\{\mathcal{H}_t\}_{t=1}^T, \rho), \quad \rho \in (0, 1]. \quad (10)$$

Hidden states restricted to retained positions are denoted $h_l^{(\Omega)}$. In our experiments, we perform filtering with $\rho = 0.3$.

4.2 Hidden State Collection and Block Averaging

For each model and calibration prompt, we collect hidden states and apply token filtering ($h_l^{(\Omega)}$). We then partition the token axis into non-overlapping blocks of size $B = 32$ and average across blocks/samples to obtain layer representatives

$$Z_l \in \mathbb{R}^{B \times d}, \quad l = 0, \dots, N-1. \quad (11)$$

These $\{Z_l\}$ are used for entropy/CKA analyses and layer-set scoring. In our experiments, we use 200 calibration samples from ShareGPT [15].

4.3 Matrix Entropy-based Expansion (ELSI)

ELSI performs beam search with matrix entropy of concatenated hidden states as the expansion criterion. We use Z_l to denote the layer representatives from Eq. (11) and $\bar{Z}_{\mathcal{L}} \triangleq \text{concat}(\{Z_l\}_{l \in \mathcal{L}}, \text{feature dim})$.

ELSI is detailed in Algorithm 1 which begins with entropy-based singleton seeding ([S1]) and then performs beam expansion over set sizes ([S2]). At each depth, candidates are deduplicated ([S3]) and pruned by entropy score ([S4]) before selecting the best size- m set ([S5]).

Algorithm 1: ELSI: Entropy-based Layer Set Iteration

Input: Layer representatives $\{Z_l\}_{l=0}^{N-1}$, target set size m

Input: Beam hyperparameters ($K_{\text{expand}}, K_{\text{beam}}$)

Output: Selected layer set $\mathcal{L}^*$ with $|\mathcal{L}^*| = m$

[S1] initial beam: top-entropy singletons

$\mathcal{B}_1 \leftarrow \arg \text{top-}K_{\text{beam}} \text{ } l \in \{0, \dots, N-1\} S_1(Z_l)$

[S2] entropy-driven expansion across depths

for $d \in \{2, \dots, m\}$ **do**

$d = 2$ forms pairs; $d = 3$ forms triples, etc.

$\mathcal{C}_d \leftarrow []$

for each $\mathcal{L} \in \mathcal{B}_{d-1}$ **do**

top extension candidates

$\mathcal{E}(\mathcal{L}) \leftarrow \arg \text{top-}K_{\text{expand}} \text{ } l \notin \mathcal{L} S_1(\bar{Z}_{\mathcal{L} \cup \{l\}})$

$\mathcal{C}_d \leftarrow \mathcal{C}_d \cup \{\mathcal{L} \cup \{l\} : l \in \mathcal{E}(\mathcal{L})\}$

end for

[S3] deduplicate identical sets from different parents

$\mathcal{C}_d \leftarrow \text{Dedup}(\mathcal{C}_d)$

[S4] beam prune by set entropy

$\mathcal{B}_d \leftarrow \arg \text{top-}K_{\text{beam}} \text{ } \mathcal{L} \in \mathcal{C}_d S_1(\bar{Z}_{\mathcal{L}})$

end for

[S5] final selection at target size m

$\mathcal{L}^* \leftarrow \arg \max_{\mathcal{L} \in \mathcal{B}_m} S_1(\bar{Z}_{\mathcal{L}})$

return $\mathcal{L}^*$

A visualization of the beams produced by ELSI using $K_{\text{beam}} = 10$ is shown in Supplementary Fig. S3.

4.4 ELSI-C: Entropy-Initialized CKA Expansion

ELSI-C reuses entropy initialization and entropy scoring, but restricts each expansion to a transition-aware candidate pool from conditional CKA. Algorithm 2 first initializes the beam using entropy ([C1]). For each beam member, it computes a transition score for each candidate using CKA ([C2]). Candidate layers with the highest transition scores are then kept for each beam ([C3]). Next, candidate expansions are aggregated with deduplication and the beam is pruned based on entropy ([C4]). The final size- m set is selected as the top beam element ([C5]).

Algorithm 2: ELSI-C: Entropy-based Layer Set Iteration with CKA Filtering

Input: Layer representatives $\{Z_l\}_{l=0}^{N-1}$, target set size m

Input: Hyperparameters ($K_{\text{beam}}, K_{\text{pool}}$)

Output: Selected set $\mathcal{L}^*$ with $|\mathcal{L}^*| = m$

[C1] initial beam: top-entropy singletons

$\mathcal{B}_1 \leftarrow \arg \text{top-}K_{\text{beam}} \text{ } l S_1(Z_l)$

for $d \in \{2, \dots, m\}$ **do**

$d = 2$ forms pairs; $d = 3$ forms triples, etc.

$\mathcal{C}_d \leftarrow []$

for each $\mathcal{L} \in \mathcal{B}_{d-1}$ **do**

conditional CKA matrix

$M_{\mathcal{L}}(i, j) \leftarrow \text{CKA}(\bar{Z}_{\mathcal{L} \cup \{i\}}, \bar{Z}_{\mathcal{L} \cup \{j\}})$

[C2] calculate transition scores for each candidate

$\delta_{\mathcal{L}}(r) \leftarrow \|M_{\mathcal{L}}(r, :) - M_{\mathcal{L}}(r-1, :)\|_2$

[C3] candidate layers with highest transition

scores

$\mathcal{Q}_{\mathcal{L}} \leftarrow \arg \text{top-}K_{\text{pool}} \text{ } l \notin \mathcal{L} \delta_{\mathcal{L}}(l)$

$\mathcal{C}_d \leftarrow \mathcal{C}_d \cup \{\mathcal{L} \cup \{l\} : l \in \mathcal{Q}_{\mathcal{L}}\}$

end for

[C4] deduplicate + beam prune with entropy

$\mathcal{C}_d \leftarrow \text{Dedup}(\mathcal{C}_d)$

$\mathcal{B}_d \leftarrow \arg \text{top-}K_{\text{beam}} \text{ } \mathcal{L} \in \mathcal{C}_d S_1(\bar{Z}_{\mathcal{L}})$

end for

[C5] final selection at target size m

$\mathcal{L}^* \leftarrow \arg \max_{\mathcal{L} \in \mathcal{B}_m} S_1(\bar{Z}_{\mathcal{L}})$

return $\mathcal{L}^*$

5 Experiments

We use open-source models from two different families: Qwen3-4B-Instruct-2507 [16] and Llama3.1-8B-Instruct [17] to analyze the per-layer entropy (section 5.1), concatenated entropy (section 5.2), and CKA transitions across layers (section 5.4). We additionally include the layer candidates selected using ELSI and ELSI-C for these models (section 5.5).

5.1 Per-layer matrix entropy

We first analyze per-layer Shannon matrix entropy S_1 . In Figure 2, we see that across models, entropy varies substantially with depth, indicating non-uniform layer informativeness. Llama3.1-8B-IT shows high entropy for the first layer which then drops sharply and then rises, peaking around layer 27. Based on this we add another condition when selecting a single layer: in addition to having high-entropy, the layer index should occur after the middle layer index. Qwen3-4B-IT shows high-low-high entropy peaking at layer 28. While per-layer entropy can inform the selection of a single layer, it does not immediately reveal favorable *sets* of layers, motivating deeper analysis.

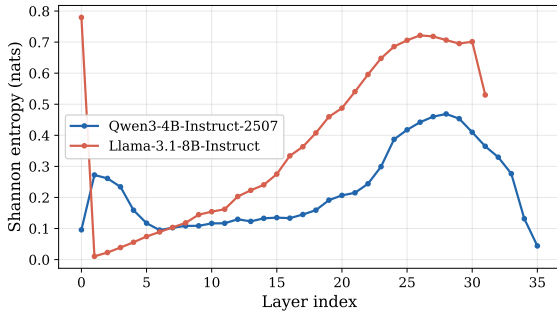


Figure 2: Per-layer Shannon entropy (S_1) across depth for Qwen3-4B-Instruct-2507 and Llama-3.1-8B-Instruct.

5.2 Entropy of layer pairs

We next evaluate the entropy of concatenations of layer pairs. Figure 3 visualizes top- k seed layers (by per-layer entropy) against all companion layer concatenations. We see non-uniform gain patterns: some layer pairs provide strong complementary directions while others are largely redundant. Moreover, different seed layers benefit from different companion layers. These results justify the need to optimize sets of layers jointly and entropy of concatenated representations as a metric to achieve this.

5.3 Token-Entropy Filtering

Across models, entropy varies substantially with depth and also across prompts as seen in Fig. 1 where the entropy variance across prompts is very high for initial and final layers. This can result in a prompt dependent layer ordering which is infeasible as the set of layers selected remains fixed through training and inference. We therefore use token filtering which reduces the cross-prompt entropy variance by 48.4% (early layers 0–6), 45.8% (middle layers 7–28), and 35.6% (late layers

Method	Qwen3-4B-Instruct		Llama-3.1-8B-Instruct	
	Layers selected	Entropy	Layers selected	Entropy
m=1				
EAGLE2	{L35}	0.044	{L31}	0.530
ELSI	{L28}	0.469	{L26}	0.718
ELSI-C	{L28}	0.469	{L26}	0.718
m=2				
ELSI	{L23, L35}	0.625	{L22, L31}	0.824
ELSI-C	{L25, L35}	0.588	{L25, L31}	0.797
m=3				
EAGLE3	{L2, L18, L33}	0.441	{L2, L16, L29}	0.687
ELSI	{L5, L23, L35}	0.627	{L1, L22, L31}	0.910
ELSI-C	{L6, L28, L35}	0.593	{L1, L25, L31}	0.889

Table 1: Layer selections and entropy for Qwen3-4B-Instruct and Llama-3.1-8B-Instruct, grouped by selection size m .

29–35), while preserving the same depth-wise activation-wave profile (Fig. 4).

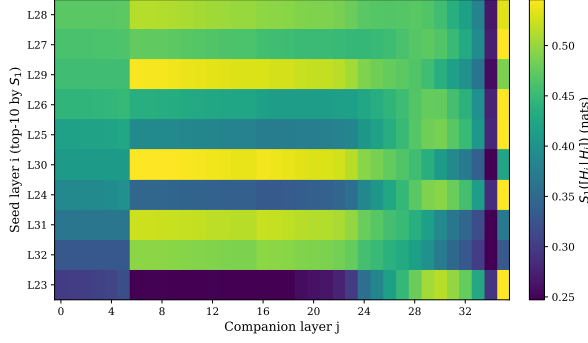
5.4 CKA compartments and transitions

Figure 5 shows pairwise CKA with transition boundaries derived from Eq. 8 overlaid. The transition analysis reveals clear depth-wise representational compartments. Qwen3-4B-IT has more transitions than Llama3.1-8B-IT, showing more variation as we go deeper into the network, with two large compartments in the middle. Llama3.1-8B-IT on the other hand has similar CKA throughout except for the first and last layers.

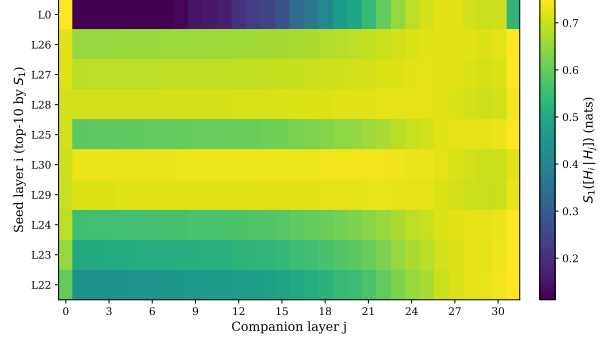
From this perspective, EAGLE2-like methods which only used last layer hidden-states might miss out on other high information layers. EAGLE3 which selects layers following $[2, N//2, N - 3]$ selects layers in different compartments, thus resulting in improved performance over EAGLE2, though it might still be sub-optimal by having overlapping information.

5.5 Selection tables and baseline intuition

We show in Table 1 the layers selected by ELSI and ELSI-C when choosing singletons, pairs, and triples of layers. We also show how the entropy of the layer sets selected by ELSI compares to the entropy of the layer choices of EAGLE2 and EAGLE3. We present this for both Qwen3-4B-IT and Llama3.1-8B-IT models and observe that layers selected using ELSI and ELSI-C always have higher entropy than those selected naively in EAGLE2 and EAGLE3.



(a) Qwen3-4B-Instruct-2507



(b) Llama-3.1-8B-Instruct

Figure 3: Entropy of layer pairs. Top- k layers by entropy are chosen as seeds and paired with the remaining layers $S_1(\{H_i, H_j\})$.

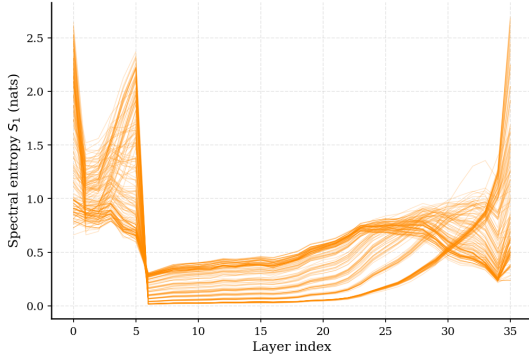
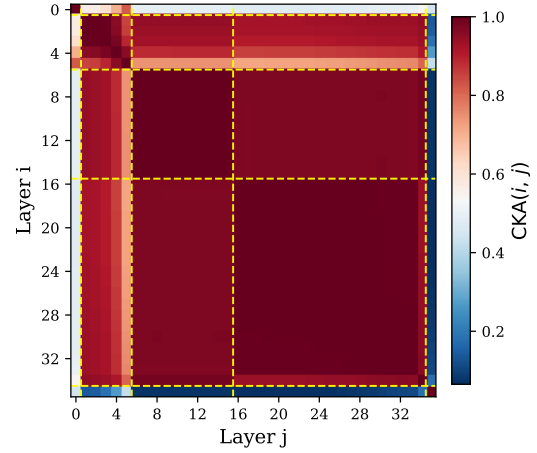
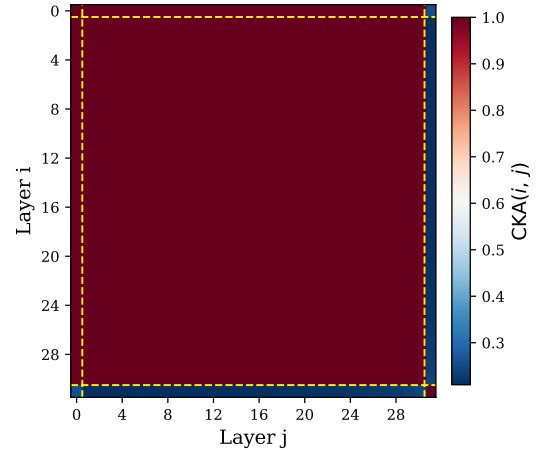


Figure 4: Token-entropy filtered prompt-token set ($\rho = 0.30$) lowers the entropy variance across prompts while preserving the activation-wave shape.



(a) Qwen3-4B-Instruct-2507



(b) Llama-3.1-8B-Instruct

Figure 5: Pairwise CKA with calculated transition boundary overlaid in yellow.

5.6 Impact on Speculative Decoding

5.6.1 Baselines and Evaluation.

We train EAGLE-style draft models using ELSI and ELSI-C and compare against layer selections based on: (i) EAGLE2 $[N - 1]$, (ii) EAGLE3 $[2, N//2, N - 3]$, and (iii) low-entropy negative sets.

5.6.2 Training and Evaluation Setup.

All the draft models were trained using a distillation dataset generated using ShareGPT [15] and Ultrachat [18] on 8 H100 GPUs. For evaluation we ablate over (a) number of draft tree nodes: 30 and 60, and (b) temperature sampling: $T=0.0$ (greedy) and $T=1.0$. We evaluate on SpecBench [19], which covers a wide range of tasks and report average accepted tokens (τ). Note that average accepted tokens will directly lead to higher token per second in this scenario for same number of hidden-states chosen. For a smaller set of target states, we expect higher

throughput due to the usage of a smaller down-projection layer of size $|\mathcal{L}|d \times d$, though this might not be visible on GPUs which are not memory-bound. We leave it to future work to experiment with ELSI-generated layer sets on edge-devices.

We show our main results for Qwen3-4B-IT in Table 2 and make several observations in this data:

- ELSI ($m=3$), matching the number of EAGLE3 layers, is **the best performing method on average across all settings**, improving over EAGLE3 by 1.6–2.2% ($+0.085\tau$ on average) and over EAGLE2 by 6.1–7.3%.
- Across all four settings, ELSI ($m=1$) matches or exceeds EAGLE3 on average ($+0.01$ to $+0.04\tau$), showing that **a single well-chosen layer matches EAGLE3’s default 3-layer heuristic**.
- **ELSI tends to provide the most benefit in categories where EAGLE3 performs weakest**, for example Translation (2.6–5.6%) and Summarization (2.1–4.2%), but can also improve acceptance rate significantly in performant categories such as STEM (2.4–3.9%).
- **ELSI provides consistent benefits across sampling temperatures and draft tree sizes configurations**, and thus may be applied to diverse deployment scenarios.

We also include a low-entropy control experiment wherein we choose the layer set with the lowest entropy resulting in layers $\{0, 1, 35\}$. We train an EAGLE-3 model with this configuration and see that it consistently results in the lowest acceptance rate. This shows that using low entropy hidden-states (which could have overlapping information) doesn’t provide a strong conditioning to the draft model.

Table 3 reports the acceptance rate for Llama3.1-8B-IT. We show a subset of SpecBench categories (CODING, QA, ROLEPLAY, WRITING, M/R), and report the average (AVG) across all categories. Entropy-guided selection remains highly efficient on this model: **ELSI ($m=1$) exceeds EAGLE2 by 2.8–4.9% while attaining $\sim 99.0\%$ of EAGLE3’s block efficiency using a single hidden state rather than three**. However, increasing the number of selected layers yields no further gain, with AVG varying by only 0.017τ across $m = 1, 2, 3$. On average, ELSI ($m=3$) trails EAGLE3 by 1.0%. However, at $T=1.0$, ELSI ($m=3$) improves over EAGLE3 on Roleplay by 4.7–5.9% and on Coding by 3.0–3.3%, whereas at $T=0.0$ the two are closer. The corresponding losses on Writing at $T=1.0$ leave the average unchanged.

We attribute the reduced headroom overall to the greater interchangeability of Llama’s decoder layers, visible as broad regions of high pairwise CKA in Figure 5. When neighboring layers encode similar information, the choice of layers matters correspondingly less, and a uniformly spaced selection already performs well. Framing layer selection through ELSI thus allows us to identify these differing regimes across models: where representations are diverse, selection results in significant benefits, and where they are redundant, it instead allows for a smaller layer-set selection.

6 Conclusion

We studied multi-layer hidden-state selection for speculative decoding through the lens of information maximization and geometric diversity, moving beyond fixed, positional heuristics. Our analysis of decoder representations across Qwen3-4B-Instruct and Llama-3.1-8B-Instruct revealed consistent structural signals: a non-uniform entropy profile across layers (the activation wave), non-uniform entropy gains with layer concatenation, and CKA transition boundaries that indicate distinct representational compartments. Together, these observations motivate *joint selection of layer sets* to maximize information and minimize redundancy, instead of relying on hand-crafted layer indices. Building on these signals, we introduced ELSI and ELSI-C as practical algorithms to select layer combinations for typical model sizes. We further enhanced these algorithms by introducing token-entropy filtering to reduce cross-prompt entropy variance, improving the stability of selection statistics. Empirically, we demonstrate (i) choice of layers matters more than the number of layers, and (ii) layer selection can result in sets that outperform EAGLE2 and EAGLE3 baselines. We further provide analysis of *why* currently used layer combinations do perform well in practice, thus enhancing the interpretability of current speculative decoding techniques. Overall, this work provides a practical and interpretable framework for multi-layer hidden-state selection in speculative decoding. This framework is designed to be general, with the potential for future work that experiments with a broader variety of target models and draft-model conditioning strategies. Further, the selection of sets of layers may find application in a broader range of tasks beyond speculative decoding including embedding tasks and interpretability, and we encourage these as future directions for exploration.

References

- [1] Qi Sun, Marc Pickett, Aakash Kumar Nain, and Llion Jones. Transformer layers as painters. In *Pro-*

Method	Nodes	T	CODING	EXTR.	HUMAN.	QA	RAG	ROLE.	STEM	SUMM.	TRANSL.	WRITING	M/R	AVG
EAGLE2			4.73	4.21	3.69	3.90	3.94	3.67	3.98	3.38	3.39	3.95	4.91	4.00
ELSI (m=1)			4.83	4.79	3.80	3.97	4.17	3.63	4.14	3.56	3.85	4.06	5.21	4.24
ELSI (m=2)			4.90	4.61	3.86	4.05	4.20	3.75	4.22	3.58	3.83	4.17	5.26	4.28
ELSI-C (m=2)	30	0.0	4.93	4.77	3.82	4.07	4.18	3.76	4.22	3.59	3.76	4.19	5.20	4.26
Low-Entropy			4.74	4.41	3.71	3.90	4.03	3.63	4.07	3.38	3.42	4.00	4.96	4.04
EAGLE3			4.88	4.75	3.79	4.02	4.13	3.73	4.14	3.52	3.60	4.07	5.18	4.20
ELSI (m=3)			4.90	4.80	3.87	4.08	4.20	3.86	4.30	3.61	3.80	4.22	5.25	4.29
ELSI-C (m=3)			4.92	4.70	3.85	4.05	4.15	3.76	4.19	3.56	3.71	4.16	5.22	4.24
EAGLE2			4.51	4.41	3.37	3.59	3.76	3.44	3.91	3.26	3.35	3.52	4.85	3.86
ELSI (m=1)			4.55	5.03	3.48	3.60	3.98	3.31	3.95	3.41	3.81	3.63	5.06	4.06
ELSI (m=2)			4.68	4.88	3.51	3.72	4.02	3.44	4.04	3.45	3.79	3.75	5.13	4.11
ELSI-C (m=2)	30	1.0	4.68	4.82	3.49	3.74	3.97	3.42	3.94	3.47	3.74	3.70	5.13	4.09
Low-Entropy			4.50	4.56	3.39	3.66	3.86	3.31	3.87	3.26	3.37	3.62	4.84	3.89
EAGLE3			4.52	4.81	3.42	3.65	3.96	3.37	3.93	3.38	3.59	3.62	5.10	4.02
ELSI (m=3)			4.67	4.96	3.50	3.74	4.01	3.34	4.06	3.45	3.74	3.91	5.17	4.11
ELSI-C (m=3)			4.48	4.85	3.50	3.74	4.00	3.50	3.89	3.44	3.71	3.70	5.11	4.08
EAGLE2			5.03	4.54	4.04	4.19	4.22	3.89	4.36	3.61	3.57	4.27	5.23	4.28
ELSI (m=1)			5.17	4.97	4.07	4.26	4.45	4.01	4.40	3.81	4.02	4.37	5.49	4.50
ELSI (m=2)			5.23	4.85	4.17	4.37	4.48	4.21	4.49	3.82	3.97	4.52	5.56	4.55
ELSI-C (m=2)	60	0.0	5.23	4.94	4.18	4.37	4.43	4.04	4.55	3.83	3.96	4.48	5.54	4.54
Low-Entropy			5.09	4.65	3.98	4.25	4.30	4.01	4.39	3.64	3.59	4.25	5.28	4.33
EAGLE3			5.16	4.91	4.13	4.32	4.42	4.10	4.50	3.79	3.83	4.38	5.46	4.48
ELSI (m=3)			5.23	4.91	4.15	4.40	4.49	4.18	4.61	3.87	4.02	4.51	5.54	4.57
ELSI-C (m=3)			5.19	4.91	4.14	4.32	4.46	4.16	4.49	3.83	3.92	4.47	5.52	4.52
EAGLE2			4.71	4.76	3.70	3.87	4.04	3.57	4.17	3.48	3.54	3.73	5.16	4.12
ELSI (m=1)			4.81	5.04	3.74	3.95	4.16	3.72	4.31	3.63	3.98	3.79	5.41	4.31
ELSI (m=2)			4.88	5.03	3.81	3.99	4.23	3.78	4.30	3.69	3.93	3.94	5.49	4.36
ELSI-C (m=2)	60	1.0	4.95	5.07	3.73	4.05	4.24	3.52	4.35	3.69	4.01	3.98	5.43	4.36
Low-Entropy			4.84	4.57	3.66	3.85	4.10	3.62	4.17	3.51	3.54	3.87	5.22	4.14
EAGLE3			5.03	5.01	3.71	3.99	4.25	3.63	4.23	3.58	3.83	3.88	5.38	4.30
ELSI (m=3)			5.00	5.04	3.77	4.06	4.28	3.64	4.37	3.73	3.93	4.03	5.43	4.37
ELSI-C (m=3)			4.85	5.06	3.72	4.02	4.26	3.78	4.25	3.69	3.90	3.90	5.39	4.33

Table 2: Mean accepted tokens per forward pass, τ , on SpecBench for Qwen3-4B-Instruct-2507, across draft-tree size (Nodes) and sampling temperature (T). Evaluation is performed with **EAGLE2**, **EAGLE3**, and **ELSI** and **ELSI-C** with layer-set sizes $m = 1, 2, 3$. **Low-Entropy** is a negative control using layers $\{0, 1, 35\}$. M/R merges the math, reasoning, and math_reasoning categories, and **AVG** is the sample-weighted mean over all categories. Bold marks the best value per block.

Method	Nodes	T	CODE	QA	ROLE	WRIT.	M/R	AVG
EAGLE2	30	0.0	5.36	4.09	4.15	4.41	4.89	4.40
ELSI (m=1)			5.51	4.19	4.25	4.54	5.15	4.58
ELSI (m=2)			5.51	4.22	4.26	4.57	5.14	4.58
ELSI-C (m=2)			5.49	4.19	4.13	4.54	5.08	4.53
EAGLE3			5.54	4.24	4.25	4.62	5.22	4.64
ELSI (m=3)			5.55	4.22	4.30	4.62	5.14	4.59
ELSI-C (m=3)	30	1.0	5.47	4.15	4.24	4.51	5.10	4.52
EAGLE2			4.59	3.24	3.06	3.36	4.01	3.64
ELSI (m=1)			4.53	3.27	3.47	3.76	4.34	3.82
ELSI (m=2)			4.65	3.20	3.38	3.82	4.32	3.81
ELSI-C (m=2)			4.77	3.29	3.22	3.79	4.32	3.81
EAGLE3			4.67	3.32	3.21	3.89	4.44	3.88
ELSI (m=3)	60	0.0	4.81	3.43	3.36	3.72	4.33	3.86
ELSI-C (m=3)			4.68	3.21	3.20	3.79	4.35	3.77
EAGLE2			5.58	4.36	4.48	4.61	5.24	4.66
ELSI (m=1)			5.73	4.42	4.57	4.85	5.47	4.84
ELSI (m=2)			5.70	4.49	4.54	4.84	5.42	4.83
ELSI-C (m=2)			5.67	4.44	4.46	4.81	5.42	4.79
EAGLE3	60	1.0	5.75	4.51	4.60	4.88	5.51	4.89
ELSI (m=3)			5.75	4.47	4.59	4.85	5.45	4.84
ELSI-C (m=3)			5.66	4.39	4.52	4.80	5.37	4.78
EAGLE2			4.72	3.48	3.57	4.06	4.41	3.95
ELSI (m=1)			5.16	3.40	3.64	3.74	4.66	4.06
ELSI (m=2)			4.55	3.64	3.69	3.86	4.60	4.06
ELSI-C (m=2)	60	0.0	5.05	3.43	3.69	4.02	4.62	4.06
EAGLE3			4.91	3.67	3.56	3.87	4.59	4.12
ELSI (m=3)			5.07	3.65	3.77	3.72	4.54	4.06
ELSI-C (m=3)			4.94	3.60	3.48	3.93	4.71	4.05

Table 3: Mean accepted tokens per forward pass, τ , on SpecBench for Llama-3.1-8B-Instruct, across draft-tree size (Nodes) and sampling temperature (T). A subset of categories is shown for space. **AVG** is the sample-weighted mean over all SpecBench categories. M/R merges the math, reasoning, and math_reasoning categories. Bold marks the best value per block.

ceedings of the AAAI Conference on Artificial Intelligence, volume 39, pages 25219–25227, 2025.

- [2] Xin Men, Mingyu Xu, Qingyu Zhang, Bingning Wang, Hongyu Lin, Yaojie Lu, Xianpei Han, and Weipeng Chen. Shortgpt: Layers in large language models are more redundant than you expect. *arXiv preprint arXiv:2403.03853*, 2024.
- [3] Lucrezia Valeriani, Diego Doimo, Francesca Cuturello, Alessandro Laio, Alessio Ansuini, and Alberto Cazzaniga. The geometry of hidden representations of large transformer models. *Advances in Neural Information Processing Systems*, 36:51234–51252, 2023.
- [4] Oscar Skean, Md Rifat Arefin, Dan Zhao, Niket Nikul Patel, Jalal Naghiyev, Yann Lecun, and Ravid Shwartz-Ziv. Layer by layer: Uncovering hidden representations in language models. In *International Conference on Machine Learning*, pages 55854–55875. PMLR, 2025.
- [5] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Liang Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *Iclr*, 1(2):3, 2022.
- [6] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adalora: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.
- [7] Vishnuprasadh Kumaravelu, Sunil Gupta, and PK Srijith. Post-optimization adaptive rank allocation for lora. *arXiv preprint arXiv:2604.27796*, 2026.
- [8] A El Ichi and K Jbilou. Low-rank tensor approximation of weights in large language models via cosine lanczos bidiagonalization. *arXiv preprint arXiv:2601.17112*, 2026.
- [9] Max Staats, Matthias Thamm, and Bernd Rosenow. Small singular values matter: A random matrix analysis of transformer models. *Advances in Neural Information Processing Systems*, 38:153545–153573, 2026.
- [10] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle-2: Faster inference of language models with dynamic draft trees. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pages 7421–7432, 2024.

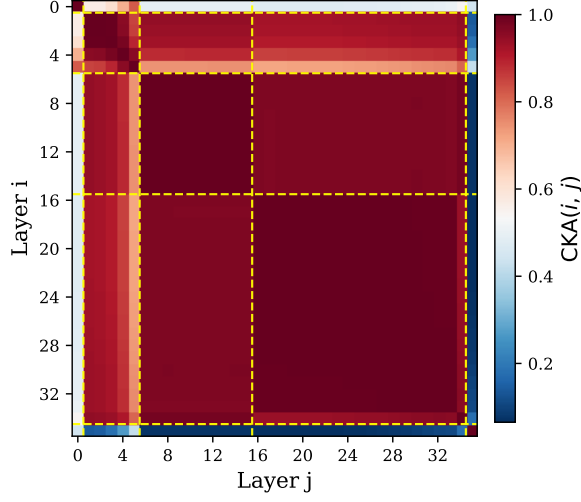
- [11] Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. Eagle-3: Scaling up inference acceleration of large language models via training-time test. *Advances in Neural Information Processing Systems*, 38:136737–136756, 2026.
- [12] Jian Chen, Yesheng Liang, and Zhijian Liu. Dflash: Block diffusion for flash speculative decoding. *arXiv preprint arXiv:2602.06036*, 2026.
- [13] Oscar Skean, Jhoan Keider Hoyos Osorio, Austin J Brockmeier, and Luis Gonzalo Sanchez Giraldo. Dime: Maximizing mutual information by a difference of matrix-based entropies. *arXiv preprint arXiv:2301.08164*, 2023.
- [14] Luis Gonzalo Sanchez Giraldo, Murali Rao, and Jose C Principe. Measures of entropy from data using infinitely divisible kernels. *IEEE Transactions on Information Theory*, 61(1):535–548, 2014.
- [15] ShareGPT Community. Sharegpt: A dataset of conversational data scraped from chatgpt shared links, 2023.
- [16] 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.
- [17] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.
- [18] Ning Ding, Yulin Chen, Bianfang Xu, Yujia Qin, Zonghan Zheng, Shengding Hu, Zhiyuan Liu, Maosong Sun, and Bowen Zhou. Enhancing chat language models by scaling high-quality instructional conversations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3000–3015, 2023.
- [19] Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. Unlocking efficiency in large language model inference: A comprehensive survey of speculative decoding. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics ACL 2024*, pages 7655–7671, Bangkok, Thailand and virtual meeting, August 2024. Association for Computational Linguistics.

A Transition boundary comparison with and without token filtering

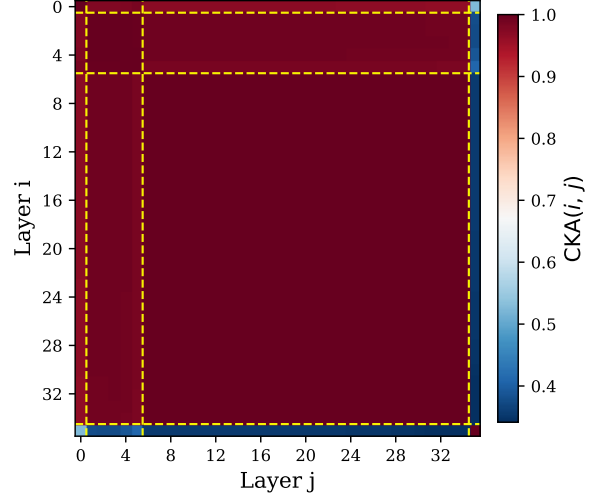
To isolate the effect of token filtering on representational compartment boundaries, we compare default and token-filtered CKA transition plots side by side for both models.

B ELSI Beam Analysis

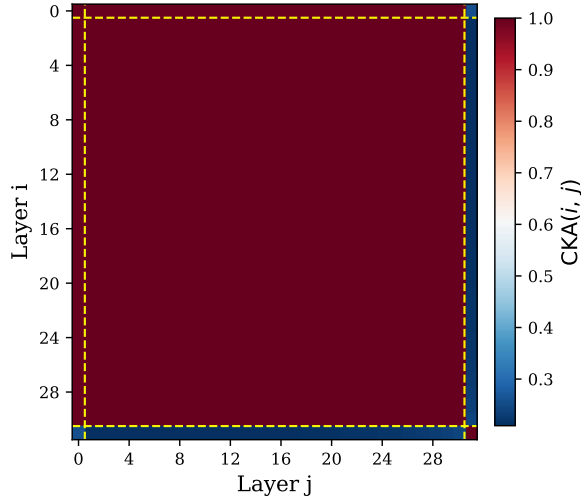
We show the entire beam tree using ELSI, expanding based on high joint entropy of hidden-states post token-filtering in Fig 7. The top-1 at each depth of the beam are layers picked for best single, pair, and triple layers respectively.



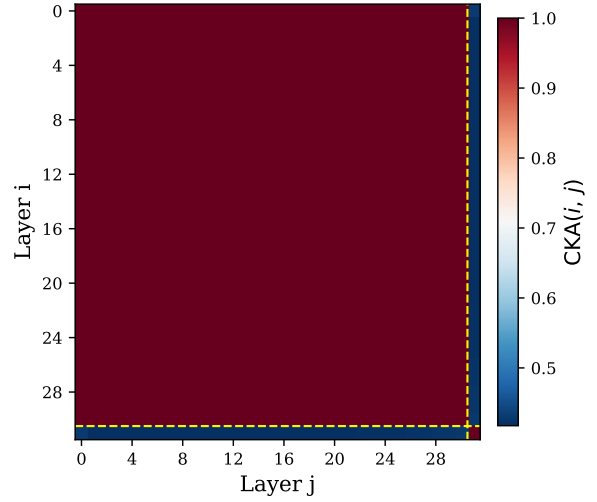
(a) Qwen3-4B, default tokens



(b) Qwen3-4B, token-filter ($\rho = 0.30$)



(c) Llama-3.1-8B, default tokens



(d) Llama-3.1-8B, token-filter ($\rho = 0.30$)

Figure 6: With and without token-filter CKA transition comparison. Yellow dashed lines mark selected transition boundaries under matched detector settings.

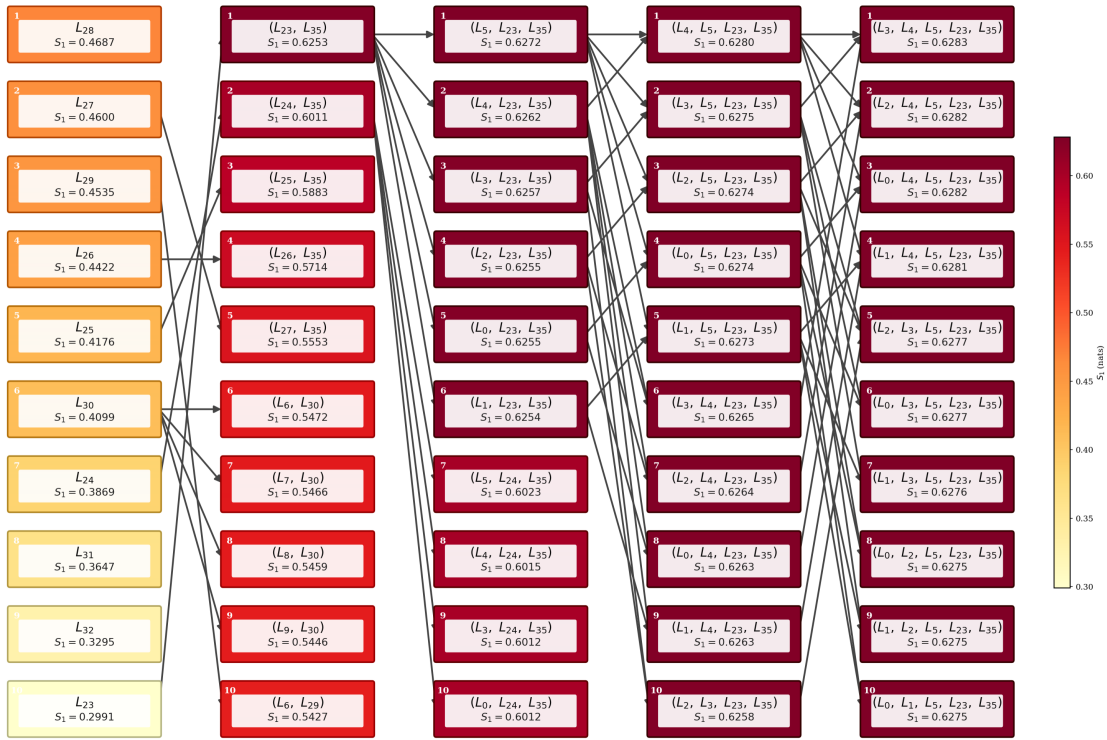


Figure 7: Top-M entropy beam frontier for Qwen3-4B-IT