MEASURING MAXIMUM ACTIVATIONS IN OPEN LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

The dynamic range of activations is a first-order constraint for low-bit quantization, activation scaling, and stable LLM inference. Prior work characterized outlier features and massive activations on pre-2024 LLaMA-style models, and the downstream activation-quantization stack inherits that picture without revisiting it for the post-LLaMA open-model boom. We ask the deployment-oriented question: how *large* can activations get in modern open LLMs, and how does this magnitude vary across families, generations, and training stages? Under a unified pipeline (5,000-sample multi-domain corpus, family-specific tokenization, identical hooks across embeddings, hidden states, attention, MLP/MoE, SwiGLU gates, and final norm), we measure global and layerwise maxima on 27 checkpoints from 8 open families spanning dense, MoE, vision-language, intermediate-training, and instruction-tuned variants. We find that (i) global maxima span over nearly four orders of magnitude at comparable parameter counts, with Qwen3.5 and MoE checkpoints in the 10^2–10^3 range and Gemma3-27B-it reaching $\sim 7 \times 10^5$; (ii) cross-family and cross-generation comparisons break simple monotonic scaling; and (iii) in two matched MoE–dense pairs, MoE checkpoints exhibit 14.0–$23.4\times$ lower peaks, while the residual stream carries the global maximum in 22/24 main-analysis checkpoints. A lightweight INT-8 sanity check shows that measured maxima co-vary with low-bit reconstruction error via activation-scale selection. We conclude that maximum activation magnitude is a protocol-conditioned model statistic associated with family, architecture, and training stage—not a simple byproduct of size—and should be measured and reported alongside any open-weight release before low-bit deployment. An anonymous code repository is linked at `https://anonymous.4open.science/r/Max_act_llm-CB63/`.

1 INTRODUCTION

The activation dynamic range of a large language model (LLM) is not merely a descriptive statistic: it determines the numerical range that inference systems, activation quantizers, and scaling rules must accommodate (DeepSeek-AI et al., 2025). In low-bit inference, for example, a per-tensor activation scale is often chosen to cover the largest magnitude observed on a calibration set. A small number of extremely large activations can therefore dominate the scale, waste most quantization levels on rarely used values, and amplify reconstruction error for ordinary activations. This paper studies a simple but deployment-critical quantity: the *maximum activation magnitude*, defined as the largest absolute activation observed across layers and key components under a fixed evaluation protocol.

Extreme activations, massive activations, and outlier features have been studied from several perspectives, including existence tests, token- or feature-level localization, and functional interventions. However, the deployment question remains less systematically mapped: how large can activations become in recent open LLMs, where do the largest values appear, and how do they change with model family, architecture, model generation, and training stage? This question is increasingly important because modern open models no longer differ only in parameter count. They vary in normalization and training recipes, dense versus MoE computation, vision-language adaptation, instruction tuning, and released intermediate training stages. As a result, parameter scale alone may be an unreliable proxy for activation range.

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.

Prior work on extreme activations falls along two largely separate lineages. The *interpretability* line begins with Dettmers et al. (2022), who defined *emergent outlier features* via a 6σ rule on OPT/BLOOM, and Sun et al. (2024), who introduced *massive activations* as coordinates that are simultaneously large ($|x_i| > 100$) and locally sparse ($\geq 1000\times$ the per-token median); Bondarenko et al. (2023) attributed them to attention heads needing a "no-op" route through the residual stream, and very recent work refines this picture—Gu et al. (2025) trace when attention sinks emerge during pretraining, and Sun et al. (2026) decouple massive-activation "spikes" from sinks and localize them to early-layer step-up blocks under pre-norm transformers. All of these studies treat the phenomenon categorically and analyze a handful of LLaMA-family or single-architecture checkpoints. The *quantization* line treats the same activations as a deployment obstacle: SmoothQuant (Xiao et al., 2023), AWQ (Lin et al., 2024b), GPTQ (Frantar et al., 2023), and Outlier Suppression+ (Wei et al., 2023) migrate or rescale outlier mass; rotation methods QuaRot (Ashkboos et al., 2024), SpinQuant (Liu et al., 2024a), and DuQuant (Lin et al., 2024a) remove the outlier basis; FlatQuant (Sun et al., 2025) learns affine flattening transforms; PrefixQuant (Chen et al., 2024) and KIVI (Liu et al., 2024b) target the KV cache; and FP8 pretraining pipelines (DeepSeek-AI et al., 2025; DeepSeek-AI, 2025) fold analogous mitigations into low-precision training, with Tian et al. (2025) arguing that high-sparsity MoE routing further changes the activation regime. All of these mitigations *transform away* the upper bound rather than measuring how it varies across modern releases.

It remains unclear whether either discovery still holds for the recent wave of post-LLaMA open releases—Qwen2.5/3/3.5 (Qwen Team, 2024; 2025), Qwen2.5-VL (Bai et al., 2025), Gemma 2 and Gemma 3 (Gemma Team, 2024; 2025), the Ling-mini series (Ling Team, 2025), and gpt-oss (OpenAI, 2025)—which diverge from earlier LLaMA-style models along multiple axes simultaneously: normalization stack, gated MLP variants, MoE routing, multimodal adaptation, intermediate-training releases, and instruction tuning. We are not aware of a prior study that reports activation magnitudes across these families under a unified protocol. Our paper is complementary along both axes: we provide a unified-protocol measurement of the global maximum $M = \max|a|$ across 27 post-LLaMA open checkpoints from 8 families, treat M as a continuous protocol-conditioned statistic rather than a binary outlier flag, and connect M to per-tensor INT-8 reconstruction error—inputs that the mechanistic and quantization-mitigation lines do not directly provide.

We address this gap with a unified empirical survey of maximum activations in modern open LLMs. Our main analysis covers 24 checkpoints from 8 model families: Qwen2.5, Qwen2.5-VL, Qwen3, Qwen3.5, Gemma2, Gemma3, Ling, and GPT-OSS. We additionally analyze 3 Qwen2.5-Instruct checkpoints to isolate the effect of supervised fine-tuning. All checkpoints are evaluated on the same 5,000-sample multi-domain corpus, with the text corpus re-tokenized for each model family. During forward inference, we use PyTorch hooks to stream activation statistics from embeddings, layerwise hidden states, attention outputs, MLP or MoE outputs, SwiGLU gate pre-activations, and final normalization outputs. This protocol lets us compare global maxima, layerwise peak trajectories, carrier components, family and generation effects, and matched architectural or training contrasts under the same measurement pipeline.

Contributions. Our work makes the following contributions:

- **Broad cross-family activation survey.** We measure global and layerwise maximum activations on 24 checkpoints from 8 modern open families (Qwen2.5/2.5-VL/3/3.5, Gemma2/3, Ling, GPT-OSS)—spanning dense, MoE, vision-language, intermediate-training, and instruction-tuned variants—under a single unified pipeline (5,000-sample multi-domain corpus, family-specific tokenization, identical hooks across embeddings, hidden states, attention/MLP/MoE outputs, SwiGLU gates, and final norm), moving beyond the LLaMA-derivative monoculture of Dettmers et al. (2022); Sun et al. (2024).
- **Continuous magnitude reformulation of "massive activation."** We complement the binary same-token criterion of Sun et al. (2024) with the deployment-relevant statistic $M = \max|a|$, and show that the two views can disagree across the tested checkpoints.
- **Five matched-design comparisons.** We examine (i) within-family scaling, (ii) same-scale MoE-vs-dense, (iii) same-family vision-language-vs-text-only, (iv) same-backbone Base-vs-Instruct, and (v) same-family training-stage effects on the same measurement substrate, providing an observational comparison of scale, family, generation, architecture, modality, and training progress for activation peaks.

- **Empirical findings with deployment implications.** (a) Global maxima vary by orders of magnitude across families at comparable parameter counts and break simple monotonic scaling; (b) the residual stream carries the global maximum in 22/24 main checkpoints; (c) the two matched MoE–dense pairs show lower peaks for the MoE checkpoints by 14.0–23.4×; (d) SFT is accompanied by compression of late-layer peaks in the examined Qwen2.5 pairs; (e) the examined Ling training-stage sequence shows a monotonic increase in the global maximum; and (f) a lightweight per-tensor INT-8 probe shows higher M is associated with lower SQNR under the tested recipe.
- **Open pipeline and per-checkpoint statistics.** We describe a hook-based measurement pipeline and report per-checkpoint activation statistics to support reproducibility and future quantization, scaling, and architecture research. We do *not* claim a causal mechanism for the observed differences (see Appendix D).

2 Measurement Protocol

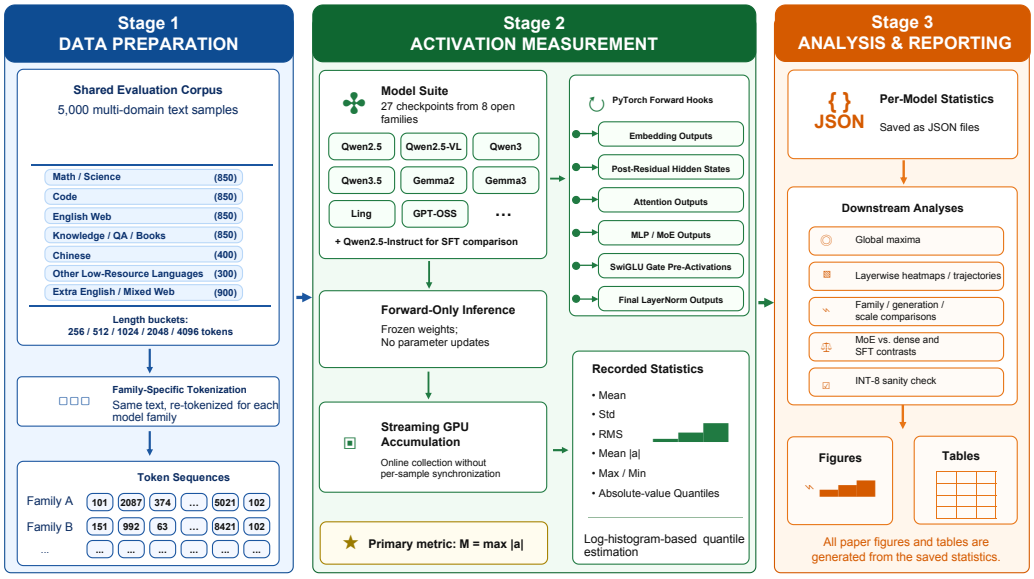

Figure 1: Overview of the pipeline.

The overall pipeline is shown in Figure 1, which consists of three steps: Data Preparation, Activation Measurement, and Analysis. We use a unified offline evaluation protocol. We first construct a multi-domain text corpus, tokenize the same text with each model family's tokenizer, run forward inference on each checkpoint, and record layerwise activation statistics. All figures and tables are generated from the resulting per-model statistics.

2.1 Corpus construction

The target evaluation corpus contains 5,000 samples. The data are sampled from RedPajama (Shen et al., 2024) sources and bucketed by content type. The target category counts are 850 mathematical or scientific samples, 850 code samples, 850 English web samples, 850 knowledge-oriented samples such as encyclopedic, book, or Q&A text, 400 Chinese samples, 300 samples in other low-resource languages, and 900 additional English or mixed web samples. This design reduces the risk that maximum-activation statistics are dominated by a single domain and ensures that the corpus covers formal text, natural web text, knowledge-intensive text, code, and multilingual content.

The corpus also controls sequence-length diversity. Samples are randomly truncated to 256, 512, 1024, 2048, or 4096 tokens with target proportions of 1%, 1%, 2%, 3%, and 93%, respectively. The corpus is therefore dominated by long-context inputs while retaining a small number of short and

medium-length sequences. The resulting corpus has an average length of approximately 3899 tokens, corresponding to roughly 19.5M tokens in total.

To avoid tokenizer mismatch artifacts, the text corpus is held fixed while tokenization is performed separately for each model family. Thus, models receive semantically identical text but token sequences aligned with their own tokenizer, reducing activation-statistics bias caused by tokenizer incompatibility.

2.2 Model suite and instrumentation

We select models according to three principles. First, we cover recent mainstream open LLM families rather than restricting the study to earlier LLaMA-style models. Second, we include multiple parameter scales and architectural forms, allowing us to separate the effects of scale, family, and architecture. Third, we include special variants such as MoE models, vision-language models, intermediate training checkpoints, and instruction-tuned models, so that we can examine whether maximum activations change with routing, modality adaptation, training progress, or supervised fine-tuning (SFT).

The main experiment contains 24 checkpoints from 8 families: Qwen2.5, Qwen2.5-VL, Qwen3, Qwen3.5, Gemma2, Gemma3, Ling, and GPT-OSS. Except for the publicly released Gemma3 checkpoints, which are instruction-tuned models, we treat the main-analysis checkpoints as base or intermediate-training checkpoints, as shown in Table 2. Therefore, the Gemma2/Gemma3 comparison should be interpreted as a public-checkpoint family-level contrast rather than a strict base-to-base generational ablation.

2.3 Recorded statistics and peak stability

The statistics pipeline has three stages. First, the shared text corpus is converted into token sequences with each model family's tokenizer. Second, each checkpoint is loaded with its full weights and evaluated with forward inference only; no parameters are modified. During inference, PyTorch forward hooks collect six classes of activation tensors: embedding outputs, layerwise hidden states after residual updates, layerwise attention outputs, layerwise MLP outputs or MoE block outputs, MLP gate pre-activations in SwiGLU-style architectures, and final LayerNorm outputs. Third, per-model JSON statistics are used to generate all figures and tables. For each captured component, we record the mean, standard deviation, RMS, mean absolute value, maximum value, minimum value, and streaming estimates of absolute-value quantiles.

Because the global maximum activation is an extreme statistic, we first verify that it is not triggered by a small number of accidental samples. For four representative models, we construct category-proportional subsamples of 1,000 and 2,000 examples from the original 5,000-example corpus. Each subsample size is repeated 5 times, and each repeat runs the full activation scan. The resulting peaks consistently reproduce the order of magnitude of the 5k reference run. The largest coefficient of variation across 1k repeats is 10.1% for Qwen3-30B-A3B, and the largest coefficient of variation across 2k repeats is 8.2%. These results indicate that the reported maximum activations are not accidental single-sample artifacts and that the measurements are statistically robust at the scale studied here.

2.4 Domain-stratified robustness

As a complementary check, we scanned seven input domains using 100 sequences per domain, three repeats, and 2,048-token inputs. The Qwen3 dense–MoE peak gap appears across all seven tested domains, while domain-to-mixed-corpus median ratios remain close to one for the Qwen3 checkpoints. This check concerns medians rather than the global maximum, so it supports robustness of the observed ordering within the tested domains but not stability of M on unseen deployment inputs; detailed results are given in Appendix E.

3 FROM BINARY MASSIVE ACTIVATIONS TO CONTINUOUS PEAKS

The empirical story proceeds from definition to mechanism to comparison. We first connect our deployment-oriented maximum M to the binary massive-activation criterion used in prior work, then ask where the largest values are carried, and finally compare families, generations, architectures, and training stages under matched designs.

3.1 RELATIONSHIP TO THE SUN CRITERION

Our main metric is the global maximum activation, $M = \max |a|$, taken across all six hooked component classes (embeddings, layerwise hidden states, attention outputs, MLP/MoE outputs, SwiGLU gate pre-activations, final LayerNorm) and all layers; this is the value plotted in every bar chart and used in every matched-pair ratio in Section 5 and Appendix C. The Top-k values reported in Table 1 are drawn from a single representative layer chosen for the local-sparsity diagnostic and may therefore differ from M when a different layer carries the global peak. Before turning to this macro-level magnitude analysis, we first drill down into whether the global extrema also satisfy a commonly used local sparsity definition from prior work. We adopt the same-token criterion of Sun et al. (2024): given a hidden state vector $x \in \mathbb{R}^d$ for one token, a coordinate x_i is counted as a massive activation if it simultaneously satisfies $|x_i| > 100$ and $\frac{|x_i|}{\text{median}_{j=1}^{d} |x_j|} > 1000$. At the model level, a checkpoint passes the criterion if any hidden layer contains at least one token-feature coordinate satisfying both thresholds.

3.2 OVERALL EXISTENCE AND FAILURE MECHANISMS

Table 1 summarizes representative activation locations for each checkpoint based on the full layerwise scan. Overall, 20 of the 24 main-analysis checkpoints pass the Sun criterion, indicating that massive activations remain widespread in recent open LLMs.

The four checkpoints marked as failing reveal two distinct failure mechanisms. Qwen2.5-1.5B reaches an absolute peak of 7,968, but the median absolute value within the representative peak token is 13.9, giving a local ratio of roughly 574 and falling below the $1000\times$ threshold. This model therefore exhibits large but relatively dense activations rather than locally sparse massive activations. In contrast, Qwen3.5-0.8B, Qwen3.5-9B, and Qwen3.5-35B-A3B fail because their overall activation scale is systematically suppressed. Figure 2 shows that all failing points avoid the upper-right passing region. This diagnosis confirms that local ratio alone does not fully characterize activation-range risk, motivating our use of the global absolute maximum as the primary metric for quantization and deployment analysis.

We retain the absolute peak, the local-ratio scatter (Figure 2), and the dual reporting in Table 1 as the diagnostic for these failures rather than introducing a normalization-stack-specific re-analysis. The diagnostic is consistent with the architectural account of Sun et al. (2026), in which residual-stream spikes are generated by a small number of early-layer step-up blocks and shaped by the pre-norm normalization stack; for the deployment-oriented question of this paper—"how large can the activation magnitude become"—the absolute peak M is the directly relevant quantity. We therefore treat the binary criterion mainly as a descriptive bridge to prior work and report M as the primary metric throughout the paper.

4 WHERE MAXIMUM ACTIVATIONS FORM

4.1 LAYERWISE INTENSITY DISTRIBUTION

Figure 3 shows the normalized-depth heatmap of hidden-state peak magnitudes for all main-analysis checkpoints. Peak depth has no universal location across architectures: even within the same family, maxima can occur in shallow, middle, or deep layers. Therefore, reporting only the peak layer index is less informative than characterizing the full layerwise trajectory, namely how peak magnitudes accumulate, jump, plateau, or decay with network depth.

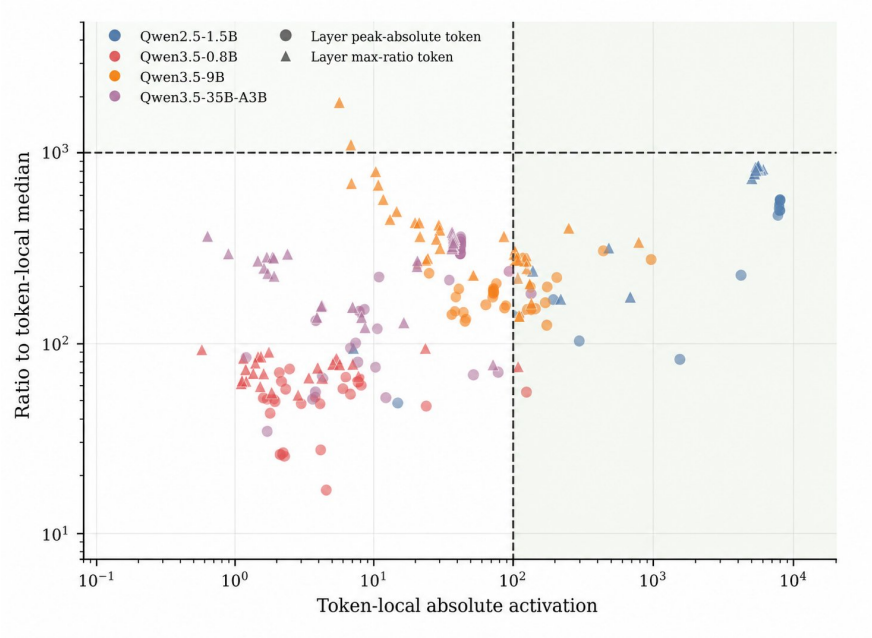

Figure 2: Failure modes for the four checkpoints that do not satisfy the Sun criterion. Colors indicate models; circles denote each layer's peak token, and triangles denote each layer's highest local-ratio token. Dashed lines mark the absolute-magnitude and local-ratio thresholds; the upper-right region is the passing region. No point overlaps this region.

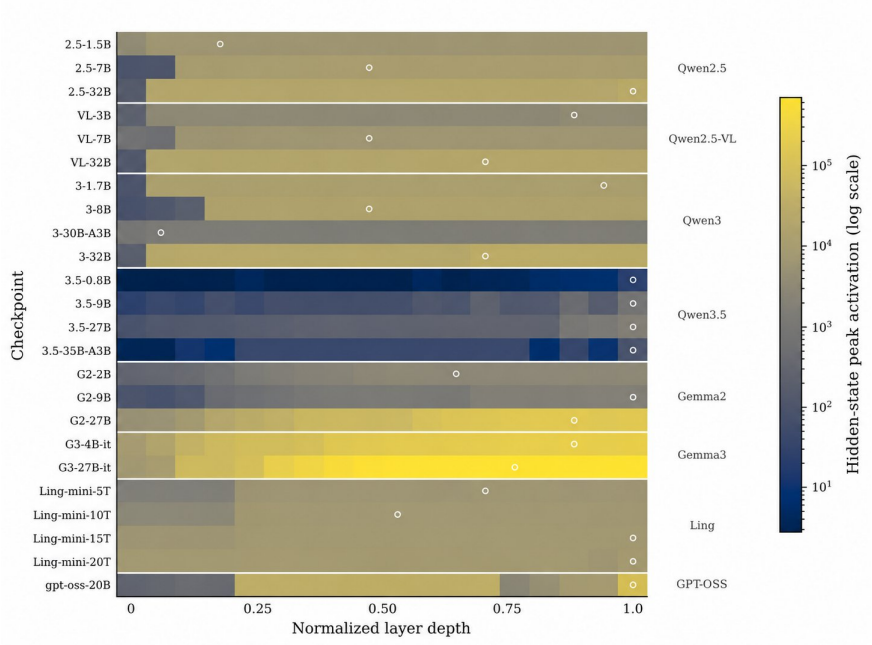

Figure 3: Layerwise heatmap of hidden-state peak magnitudes. The horizontal axis is normalized depth, and color indicates layerwise absolute peak magnitude on a log scale. White hollow markers indicate the peak depth bin for each checkpoint.

Table 1: Representative absolute activations for 24 checkpoints: a hidden layer containing a same-token sparse extreme for passing models, or the full-model peak location for failing models. Median is the median absolute value across the d dimensions of the token vector carrying the reported peak. $\times$ marks failure of the Sun criterion due to insufficient local ratio or absolute magnitude. Additional ranks appear in Table 7.

Model	Top 1	Top-100	Top 1%	Top 10%	Median
Qwen2.5-1.5B$^{\times}$	7,968	7,776	3.72	1.37	13.9
Qwen2.5-7B	13,248	12,224	3.94	1.37	5.00
Qwen2.5-32B	22,144	21,120	12.9	3.94	8.62
Qwen2.5-VL-3B	3,248	2,944	6.05	1.97	1.84
Qwen2.5-VL-7B	8,256	7,552	5.57	1.91	3.23
Qwen2.5-VL-32B	22,144	21,248	14.6	4.37	9.31
Qwen3-1.7B	14,208	13,952	103.9	52.5	4.19
Qwen3-8B	17,664	16,896	4.05	1.32	3.14
Qwen3-30B-A3B	1,512	1,392	0.348	0.099	0.578
Qwen3-32B	35,328	34,304	11.7	3.61	12.4
Qwen3.5-0.8B$^{\times}$	122.0	3.23	12.2	6.28	2.20
Qwen3.5-9B$^{\times}$	956.0	87.5	12.1	5.48	3.42
Qwen3.5-27B	167.0	127.0	0.965	0.445	0.225
Qwen3.5-35B-A3B$^{\times}$	132.0	40.8	1.37	0.595	0.719
gemma-2-2b	2,992	884.0	6.71	2.34	50.5
gemma-2-9b	1,656	294.0	16.4	6.11	1.50
gemma-2-27b	157,696	83,456	564.7	298.5	122.0
gemma-3-4b-it	245,760	165,888	828.7	461.2	86.5
gemma-3-27b-it	696,320	667,648	602.4	316.6	31.2
Ling-mini-5T	7,648	1,976	16.8	7.03	3.59
Ling-mini-10T	9,024	3,296	9.59	3.14	3.39
Ling-mini-15T	9,600	8,192	10.4	3.31	4.06
Ling-mini-20T	10,240	8,512	3.52	1.05	1.32
gpt-oss-20b	43,008	42,240	48.3	12.0	5.94

4.2 Two layerwise patterns

The layerwise trajectories broadly fall into two patterns, illustrated in Figure 4. The first is a jump-and-plateau pattern: activation magnitude rises sharply in early or middle layers and then remains high over a long layer interval, as in Qwen2.5 and GPT-OSS. The second is a gradual-accumulation pattern: activation magnitude increases more smoothly with depth and often reaches its maximum in later layers, as in Qwen3.5 and Gemma. This distinction indicates that maximum activations are governed not only by the physical depth of the peak layer, but also by the dynamics through which the peak forms. The pattern is strongly associated with model family and architecture rather than being a monotonic function of parameter scale.

We treat this dichotomy as a qualitative description of the depth-normalized layerwise heatmap in Figure 3 and the representative trajectories in Figure 4, not as a quantitative classifier. The two patterns are consistent with the architectural account of Sun et al. (2026), in which spike formation is concentrated in a small number of early-layer step-up blocks under pre-norm transformers, and the cross-family heatmap already shows the relevant separation; introducing a numerical classifier on $n = 24$ trajectories would not change the matched-design contrasts in Section 5 and Appendix C, which use M rather than the trajectory shape.

4.3 Carrier components

Across the 24 main-analysis checkpoints, 22 global maxima occur in layerwise hidden states. GPT-OSS-20B is a component-level exception whose global maximum comes from the MLP output, and the failing Qwen3.5-0.8B checkpoint peaks at the final LayerNorm output. If we restrict attention to the 20 checkpoints that satisfy the local sparsity criterion, all qualifying coordinates occur in

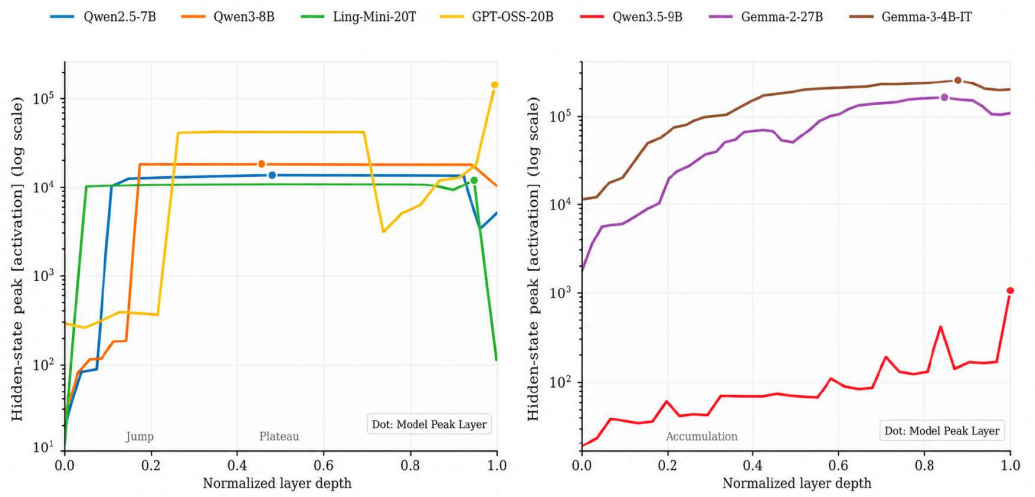

Figure 4: Representative layerwise trajectories for the two main emergence patterns. Left: jump-and-plateau models rise sharply in early or middle layers and remain high afterward. Right: gradual-accumulation models grow more smoothly with depth and often peak in later layers. Each curve corresponds to one representative checkpoint; markers indicate the peak layer.

layerwise hidden states. The residual stream is therefore the dominant carrier through which extreme activation magnitudes are propagated and preserved.

4.4 Layerwise prevalence beyond the global carrier

The expanded 27-checkpoint analysis shows that a large hidden-state peak is often not confined to the single layer carrying the global maximum: 21/27 checkpoints have $\mathrm{Coverage@50} \geq 0.5$, while the global maximum is carried by a post-residual hidden state in 25/27 cases. The correlation between $\log_{10} M_{\mathrm{hidden}}$ and $\mathrm{Coverage@50}$ is modest ($\rho = 0.165$, permutation $p = 0.4138$), so magnitude and prevalence should be reported as complementary statistics rather than treated as interchangeable; the full comparison appears in Appendix E.

5 What Controls Peak Magnitude?

5.1 Within-family scaling

Figure 12 compares checkpoints of different sizes within the same family and model form. Most families, including Qwen2.5, Qwen3.5, and Gemma3, show a stable within-family scale effect: the global maximum activation increases with parameter count. Gemma2 is the main local non-monotonic exception, with the 9B checkpoint peaking below the 2B checkpoint before the 27B checkpoint rises again. These results suggest that, when model family and training form are fixed, model size often amplifies activation extremes, although individual checkpoints can still deviate due to training or recipe differences.

5.2 Cross-family magnitude differences

Figure 5 summarizes the global maximum activation magnitudes of all 24 main-analysis checkpoints. Cross-family variation is much larger than within-family scaling variation, spanning several orders of magnitude. For example, Qwen3.5 is concentrated in a low-magnitude regime around hundreds to low thousands, whereas Gemma3-27B-it reaches a global maximum of 696,320. This directly shows that the severity of maximum activations is strongly reshaped by family-level architecture and training choices.

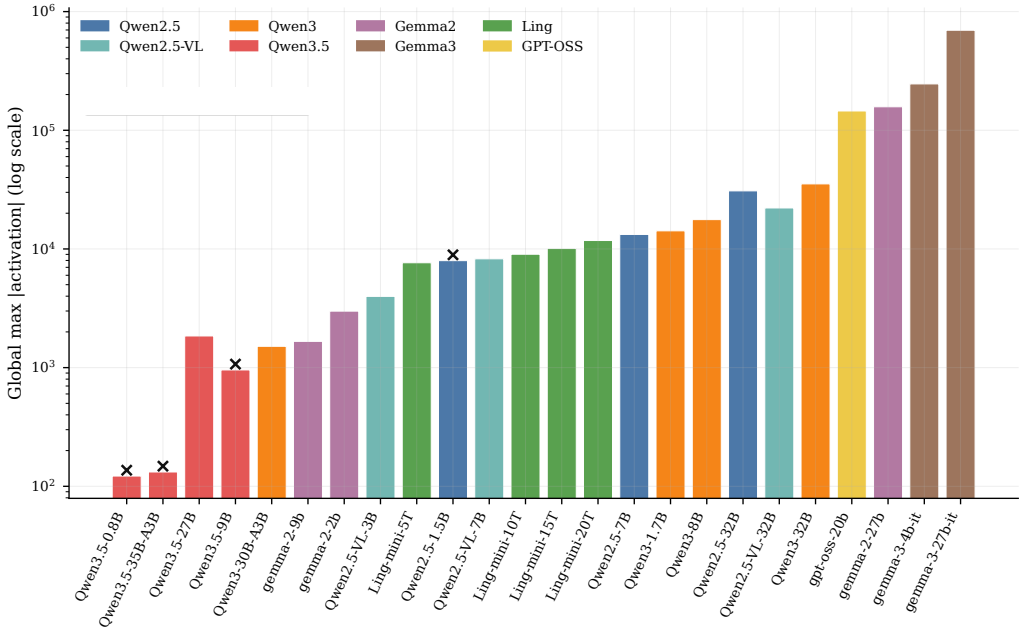

Figure 5: Global maximum activation magnitudes for the 24 main-analysis checkpoints. The vertical axis is on a log scale. A $\times$ marker indicates a model that does not satisfy the massive-activation existence criterion. The main analysis includes publicly released instruction-tuned Gemma3 checkpoints; this comparison is not a strict base-to-base generational comparison.

5.3 Non-monotonic generational evolution

Appendix Figure 6 compares generational trends at similar model sizes. Maximum activation magnitude does not monotonically shrink or grow with release time; instead, it is highly family-dependent. Across three size groups, Qwen exhibits an inverted-V trajectory: maximum activations increase from Qwen2.5 to Qwen3 and are then strongly suppressed in Qwen3.5. In contrast, Gemma shows a sharp increase from Gemma2 to Gemma3 across both matched size groups.

5.4 Empirical rule of thumb

Taken together, these measurements suggest a practical rule of thumb. Within a fixed family, increasing parameter count often increases maximum activation magnitude. Once family or generation changes, however, design and training differences can override this monotonicity. For inference and quantization boundaries, family identity and generation are therefore at least as important as parameter count. Matched MoE-vs-dense, vision-language-vs-text, Base-vs-Instruct, and training-stage contrasts provide additional evidence for this interpretation.

6 Deployment Takeaways and Conclusion

The preceding sections support three deployment takeaways. First, the observed maximum under a fixed protocol varies with family and architecture; matched-checkpoint differences are associations, not isolated causal effects of routing or training. Second, the residual stream is the dominant carrier, so quantization policies should inspect hidden-state peaks as well as attention and MLP outputs. Third, the INT-8 reconstruction probe and three-checkpoint W8A8 evaluation (Appendices D and E) associate larger peaks with greater error under the tested recipes, without establishing a universal predictor of quantization quality. Layerwise prevalence, domain checks, and routing analyses provide complementary context, not a distribution-wide upper bound.

AI USE STATEMENT

Generative AI tools assisted with manuscript editing, LaTeX formatting, and consistency checks. The authors are responsible for the content, analyses, claims, and reported results.

REPRODUCIBILITY STATEMENT

The main paper describes the shared corpus, family-specific tokenization, activation hooks, streaming statistics, subsampling stability checks, and matched-model comparisons. The appendix provides checkpoint coverage, layerwise trajectories, quantization settings, W8A8 results, domain-stratified stability checks, and MoE routing analyses. The anonymous code link in the abstract and the reported per-checkpoint statistics are intended to support independent reproduction of the measurements.

REFERENCES

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=rLX7Vyyzus`.

Anand, Umberto Cappellazzo, Stavros Petridis, and Maja Pantic. Mitigating attention sinks and massive activations in audio-visual speech recognition with LLMs. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2026.

Saleh Ashkboos, Amirkeivan Mohtashami, Maximilian L. Croci, Bo Li, Pashmina Cameron, Martin Jaggi, Dan Alistarh, Torsten Hoefler, and James Hensman. QuaRot: Outlier-free 4-bit inference in rotated LLMs, 2024.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-VL technical report, 2025.

Yelysei Bondarenko, Markus Nagel, and Tijmen Blankevoort. Quantizable transformers: Removing outliers by helping attention heads do nothing. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Mengzhao Chen, Yi Liu, Jiahao Wang, Yi Bin, Wenqi Shao, and Ping Luo. PrefixQuant: Static quantization beats dynamic through prefixed outliers in LLMs, 2024.

Yinjie Chen, Zipeng Yan, Chong Zhou, Bo Dai, and Andrew F. Luo. Vision transformers with self-distilled registers. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations (ICLR)*, 2024.

DeepSeek-AI. Insights into DeepSeek-V3: Scaling challenges and reflections on hardware for AI architectures, 2025.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, H. Zhang, Han Bao, Hanwei Xu, Haocheng Wang, Haowei Zhang, Honghui Ding, Huajian Xin, Huazuo Gao, Hui Li, Hui Qu, J. L. Cai, Jian Liang, Jianzhong Guo, Jiaqi Ni, Jiashi Li, Jiawei Wang, Jin Chen, Jingchang Chen, Jingyang Yuan, Junjie Qiu, Junlong Li, Junxiao Song, Kai Dong, Kai Hu, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Lei Xu, Leyi Xia, Liang Zhao, Litong Wang, Liyue Zhang, Meng Li, Miaojun Wang, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingming Li, Ning Tian, Panpan Huang, Peiyi Wang, Peng Zhang, Qiancheng Wang, Qihao Zhu, Qinyu Chen, Qiushi Du, R. J. Chen, R. L. Jin, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, Runxin Xu, Ruoyu Zhang, Ruyi Chen, S. S. Li, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shaoqing Wu, Shengfeng Ye, Shengfeng Ye, Shirong Ma, Shiyu Wang, Shuang Zhou, Shuiping Yu, Shunfeng Zhou, Shuting Pan, T. Wang,

Tao Yun, Tian Pei, Tianyu Sun, W. L. Xiao, Wangding Zeng, Wanjia Zhao, Wei An, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, X. Q. Li, Xiangyue Jin, Xianzu Wang, Xiao Bi, Xiaodong Liu, Xiaohan Wang, Xiaojin Shen, Xiaokang Chen, Xiaokang Zhang, Xiaosha Chen, Xiaotao Nie, Xiaowen Sun, Xiaoxiang Wang, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xingkai Yu, Xinnan Song, Xinxia Shan, Xinyi Zhou, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, Y. K. Li, Y. Q. Wang, Y. X. Wei, Y. X. Zhu, Yang Zhang, Yanhong Xu, Yanhong Xu, Yanping Huang, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Li, Yaohui Wang, Yi Yu, Yi Zheng, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Ying Tang, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yu Wu, Yuan Ou, Yuchen Zhu, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yukun Zha, Yunfan Xiong, Yunxian Ma, Yuting Yan, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Z. F. Wu, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhen Huang, Zhen Zhang, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhibin Gou, Zhicheng Ma, Zhigang Yan, Zhihong Shao, Zhipeng Xu, Zhiyu Wu, Zhongyu Zhang, Zhuoshu Li, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Ziyi Gao, and Zizheng Pan. DeepSeek-V3 technical report, 2025. URL `https://arxiv.org/abs/2412.19437`.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. LLM.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Elias Frantar, Saleh Ashkboos, Torsten Hoefler, and Dan Alistarh. GPTQ: Accurate post-training quantization for generative pre-trained transformers. In *ICLR*, 2023.

Gemma Team. Gemma 2: Improving open language models at a practical size, 2024.

Gemma Team. Gemma 3 technical report, 2025.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *International Conference on Learning Representations (ICLR)*, 2025.

Prannay Kaul, Chengcheng Ma, Ismail Elezi, and Jiankang Deng. From attention to activation: Unravelling the enigmas of large language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. DuQuant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024a.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. AWQ: Activation-aware weight quantization for on-device LLM compression and acceleration. In *MLSys*, 2024b.

Ling Team. Every FLOP counts: Scaling a 300B mixture-of-experts LING LLM without premium GPUs, 2025.

Zechun Liu, Changsheng Zhao, Igor Fedorov, Bilge Soran, Dhruv Choudhary, Raghuraman Krishnamoorthi, Vikas Chandra, Yuandong Tian, and Tijmen Blankevoort. SpinQuant: LLM quantization with learned rotations, 2024a.

Zirui Liu, Jiayi Yuan, Hongye Jin, Shaochen Zhong, Zhaozhuo Xu, Vladimir Braverman, Beidi Chen, and Xia Hu. KIVI: A tuning-free asymmetric 2bit quantization for KV cache. In *International Conference on Machine Learning (ICML)*, 2024b.

Iuri Macocco, Nora Graichen, Gemma Boleda, and Marco Baroni. Not a nuisance but a useful heuristic: Outlier dimensions favor frequent tokens in language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, 2025. URL `https://aclanthology.org/2025.blackboxnlp-1.6/`.

OpenAI. gpt-oss-120b & gpt-oss-20b model card, 2025.

Enrique Queipo-de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in LLMs are two sides of the same coin, 2026. URL `https://openreview.net/forum?id=c5TFhCJ6fs`.

Qwen Team. Qwen2.5 technical report, 2024.

Qwen Team. Qwen3 technical report, 2025.

Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, and Eric Xing. SlimPajama-DC: Understanding data combinations for LLM training, 2024. URL https://arxiv.org/abs/2309.10818.

Mingjie Sun, Xinlei Chen, J. Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *Conference on Language Modeling (COLM)*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks, 2026. URL https://arxiv.org/abs/2603.05498.

Yuxuan Sun, Ruikang Liu, Haoli Bai, Han Bao, Kang Zhao, Yuening Li, Jiaxin Hu, Xianzhi Yu, Lu Hou, Chun Yuan, Xin Jiang, Wulong Liu, and Jun Yao. FlatQuant: Flatness matters for LLM quantization. In *International Conference on Machine Learning (ICML)*, 2025.

Changxin Tian, Kunlong Chen, Jia Liu, Ziqi Liu, Zhiqiang Zhang, and Jun Zhou. Towards greater leverage: Scaling laws for efficient mixture-of-experts language models, 2025.

Xiuying Wei, Yunchen Zhang, Yuhang Li, Xiangguo Zhang, Ruihao Gong, Jinyang Guo, and Xianglong Liu. Outlier suppression+: Accurate quantization of large language models by equivalent and effective shifting and scaling. In *EMNLP*, 2023.

Guangxuan Xiao, Ji Lin, Mickael Seznec, Hao Wu, Julien Demouth, and Song Han. SmoothQuant: Accurate and efficient post-training quantization for large language models. In *International Conference on Machine Learning (ICML)*, 2023.

A SUPPLEMENTARY EXPERIMENTS MODEL DETAILS

Table 2: The 24 checkpoints included in the main analysis. Gemma3 uses publicly released instruction-tuned checkpoints; therefore, the Gemma2/Gemma3 comparison is not interpreted as a strict base-to-base ablation.

Model family	Checkpoints	Role in analysis
Qwen2.5	1.5B, 7B, 32B	Dense scaling baseline
Qwen2.5-VL	3B, 7B, 32B	Vision-language adaptation
Qwen3	1.7B, 8B, 30B-A3B, 32B	Dense/MoE comparison
Qwen3.5	0.8B, 9B, 27B, 35B-A3B	Low-outlier counterexample family
Gemma2	2B, 9B, 27B	Non-monotonic scaling case
Gemma3	4B-it, 27B-it	High-magnitude family
Ling	5T, 10T, 15T, 20T	Training-stage evolution
GPT-OSS	20B	Component-level special case

B SUPPLEMENTARY MAIN-TEXT FIGURES

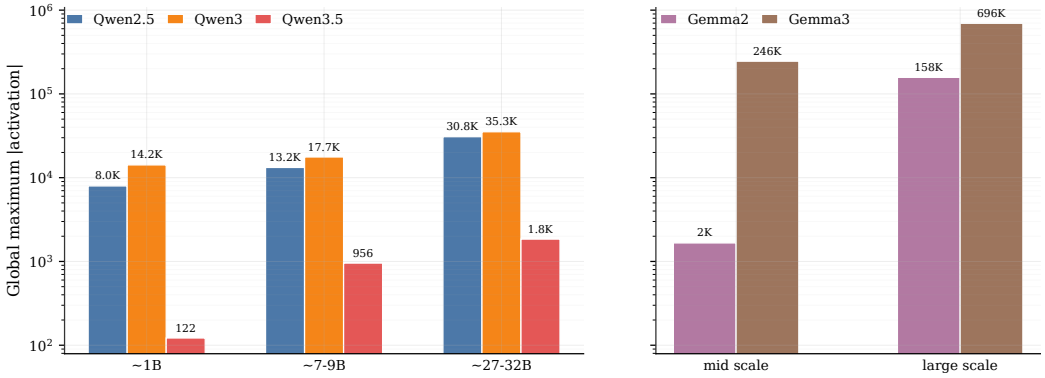

Figure 6: Generational evolution at similar sizes. Left: Qwen shows a Qwen2.5→Qwen3 increase followed by a Qwen3→Qwen3.5 decrease across three size groups. Right: Gemma2→Gemma3 increases in both size groups. The vertical axis is global maximum activation magnitude on a log scale.

C ARCHITECTURAL AND TRAINING FACTORS AT MATCHED SCALE

This section provides matched observational comparisons at identical or similar model scales. We compare MoE versus dense models, vision-language versus text-only checkpoints, base versus instruction-tuned checkpoints, and different Ling-mini training stages.

C.1 MOE VS. DENSE MODELS

The MoE-versus-dense comparison is the clearest matched-design contrast we have for an architectural axis. Figure 7 compares two pairs of same-family checkpoints near the 30B scale. Qwen3-30B-A3B has a global peak of 1,512, which is $23.4\times$ lower than Qwen3-32B at 35,328. Qwen3.5-35B-A3B has a global peak of 132, which is $14.0\times$ lower than the Qwen3.5-27B dense counterpart (the value plotted is the global M across all hooked components and layers, which differs from the representative-layer Top-1 reported in Table 1; see Section 3). The compared checkpoints have

similar total parameter counts but also differ in training recipes and other unobserved factors. The observed gap is therefore an association in two matched pairs, not evidence that sparse routing or expert structure alone causes the global peak difference. The follow-up routing analysis supports a narrower statement: routing concentration is associated with MoE-output magnitude in many layers, but it does not explain the residual-stream global peak through a simple positive correlation.

For the representative Qwen3 MoE checkpoint, the routing follow-up finds positive $\mathrm{ARC}_{\mathrm{moe}}$ in 40/48 sparse layers and a median $\mathrm{ARC}_{\mathrm{moe}}$ of 0.3118, while the median $\mathrm{ARC}_{\mathrm{res}}$ is -0.3411. Peak tokens also have a shifted routing distribution relative to all tokens (median $\mathrm{PRS}_{\mathrm{JSD}} = 0.3502$). These results indicate non-uniform routing around peak events and an association with MoE-output magnitude, but the negative residual correlation prevents using routing as an explanation for the residual-stream quantity that defines M; full definitions and additional routing statistics are in Appendix E.

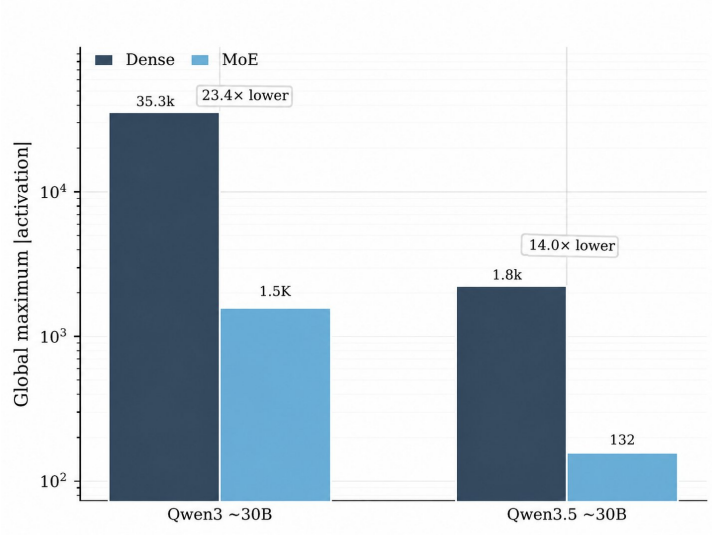

Figure 7: Matched-scale comparison of MoE and dense checkpoints. Each bar group fixes model family and approximate total parameter scale while changing only the dense/MoE form. The vertical axis is global maximum activation magnitude on a log scale.

C.2 Vision-language vs. Text-only models

Vision-language adaptation does not eliminate extreme activations, but matched-scale checkpoints differ measurably in peak magnitude. Figure 8 compares Qwen2.5-VL with text-only Qwen2.5 at 7B and 32B. Qwen2.5-VL-7B reaches a global peak of 8,256, $1.6\times$ lower than Qwen2.5-7B at 13,248. At 32B, Qwen2.5-VL reaches 22,144, $1.4\times$ lower than the text-only Qwen2.5-32B counterpart, whose global M is 30,848 across all hooked components and layers (the corresponding Top-1 entry in Table 1, 22,144, is the value at the representative passing-criterion layer; see the Section 3 convention note). Vision-language checkpoints therefore remain in a high-magnitude regime comparable to their text-only backbones, but modality adaptation and training-recipe changes co-occur with a moderate same-scale shift in the maximum activation. As with the MoE contrast, this is observational evidence based on $n = 2$ matched pairs within a single family.

C.3 Base vs. Instruct

Instruction tuning affects maximum activations differently from model scaling. Figure 9 compares Base and Instruct checkpoints at the same Qwen2.5 backbone and parameter scale. At 1.5B, the global peak remains unchanged at 7,968. At 7B, it increases slightly from 13,248 to 13,312 ($1.005\times$). At 32B, it decreases from 30,848 to 22,144, a $1.4\times$ reduction. The 32B Base value of 30,848 is the global M across all hooked components and layers; the Top-1 entry of 22,144 that appears for

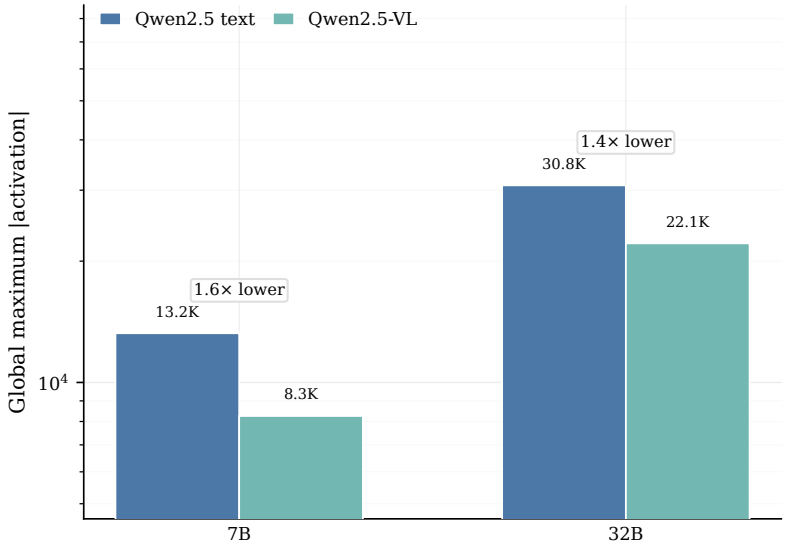

Figure 8: Matched-scale comparison between Qwen2.5-VL and text-only Qwen2.5 checkpoints. The two bar groups correspond to 7B and 32B scales. The vertical axis is global maximum activation magnitude on a log scale.

Qwen2.5-32B in Table 1 is taken at the representative passing-criterion layer, and is by construction $\leq M$ (see the Section 3 convention note). A closer inspection of layerwise hidden-state peaks shows that SFT mainly changes late layers rather than reorganizing the mid-layer structure. For 1.5B, 7B, and 32B, the stable high-peak middle-layer regions remain at $1.000\times$, $1.005\times$, and $1.000\times$ of their base values, respectively. The final-layer peaks decrease from 1,536 to 840 (-45%), from 4,864 to 2,528 (-48%), and from 30,848 to 21,248 (-31%), with only mild compensation in the penultimate layer. The global-peak decrease in Qwen2.5-32B-Instruct is therefore best understood as late-layer compression that exposes an already existing mid-layer high-peak region as the global maximum, rather than the creation of a new activation structure.

C.4 Training-stage evolution

Ling-mini provides a training-stage comparison with fixed family and approximately fixed model scale. Figure 10 shows an increasing activation-peak trajectory from 5T to 20T training tokens. This observed trend motivates recording training stage as a separate observation dimension alongside model size and architecture. It is an association within one training sequence, rather than evidence for a general causal effect of training duration.

Overall, MoE, vision-language adaptation, SFT, and training stage are all associated with measurable differences in maximum activations, but the contrasts differ in magnitude and in the strength of evidence available. The MoE-vs-dense gap is large (an order of magnitude or more) but rests on $n = 2$ matched pairs. The SFT contrast is the strongest matched design ($n = 3$ same-backbone sizes) and reveals a layerwise-localized effect (late-layer compression with preserved middle-layer high-peak regions). The vision-language and training-stage contrasts produce moderate within-family shifts. None of these comparisons constitute causal evidence; each could in principle reflect unobserved training-recipe differences between the paired checkpoints.

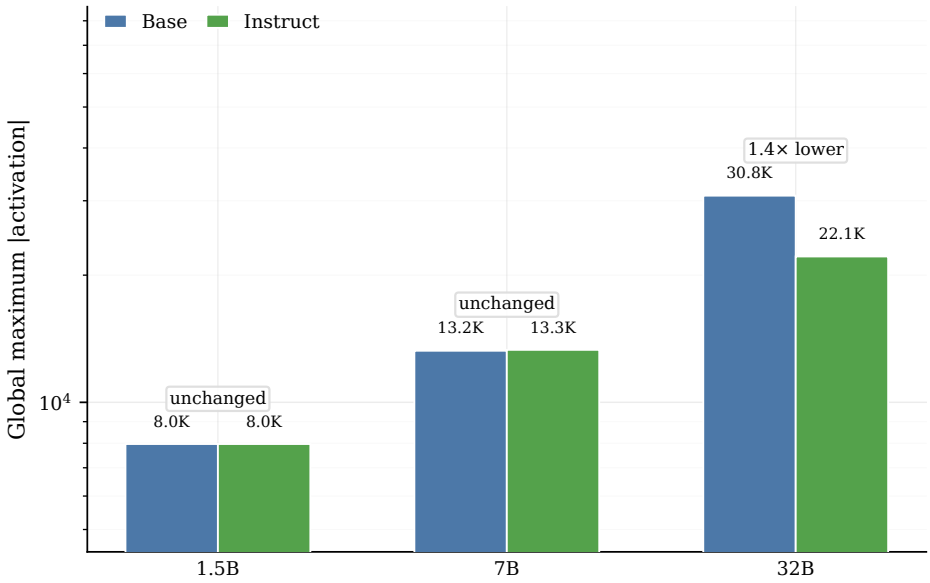

Figure 9: Matched-backbone comparison of Qwen2.5 Base and Instruct checkpoints. Each bar group fixes parameter scale and changes only whether the checkpoint has undergone instruction tuning. The vertical axis is global maximum activation magnitude on a log scale.



Figure 10: Global maximum activation across Ling-mini training stages. The horizontal axis is training tokens, and the vertical axis is global maximum activation magnitude on a log scale.

D DISCUSSION

D.1 DISTINCTION FROM OUTLIER FEATURES

The object of this study is maximum activation magnitude: the largest absolute activation observed across layers and key components under a unified evaluation corpus. This is a protocol-conditioned model-level statistic oriented toward dynamic range and deployment risk. It asks how large an activation value a checkpoint may produce during the tested forward-inference protocol. By contrast, outlier-feature studies often focus on whether certain feature dimensions remain persistently abnormal across many tokens or samples. We therefore do not interpret maximum activation magnitude as a fixed outlier-feature existence test. Instead, we treat it as a useful prior for quantization scales, activation rescaling, and inference-stability analysis, to be interpreted jointly with layerwise and distributional statistics.

D.2 IMPLICATIONS FOR QUANTIZATION AND DEPLOYMENT

Maximum activation magnitude constrains the observed range available to activation-scale selection, and can therefore affect reconstruction error in low-bit activation quantization. Higher peaks make per-tensor quantization more likely to be dominated by a few extreme values; however, quantization error is also shaped by the full distribution at the peak layer, the effective signal magnitude, and the clipping strategy. Maximum activation magnitude should therefore be treated as an important prior for quantization risk, rather than as the sole determinant of end-to-end quantization quality. Appendix Figure 13 provides a deployment-oriented grouping based on global maximum activation magnitude.

To validate this deployment relevance, we conduct a lightweight INT-8 activation quantization sanity check. The experiment covers eight representative models spanning Qwen2.5, Qwen3, Qwen3.5, Gemma3, and MoE/dense contrasts. For each model, we use 128 samples for calibration and 256 samples for evaluation at the peak hidden layer, and compare two per-tensor symmetric quantization strategies: max-abs scaling and 99.9% clipping. Figure 11 shows that Qwen3.5-0.8B maintains the highest SQNR under both strategies, at 29.1 dB and 26.3 dB. Most medium- and high-peak models obtain roughly 10–14 dB under max-abs scaling and further drop to approximately 0.2–0.4 dB under 99.9% clipping. Qwen3.5-9B and Qwen3.5-35B-A3B have relatively low absolute peaks but still show substantially lower SQNR than Qwen3.5-0.8B, indicating that within-layer distributional structure beyond peak magnitude also affects quantization error. Overall, this sanity check associates maximum activation magnitude with observable activation reconstruction error through scale selection, while also showing why richer distributional and task-level measurements are needed.

Representative W8A8 evaluation. The follow-up W8A8 evaluation provides a more end-to-end check on three checkpoints: W8A8 increases PPL by 3.41% for Qwen3-30B-A3B, 282.26% for Qwen3-32B, and approximately $63{,}049\%$ for Gemma-3-27B-it, whereas W8A16 remains close to BF16. This is evidence of an association under one representative fake-quantization recipe, not a general prediction law; the complete settings and table are provided in Appendix E.

We deliberately keep this experiment as a *deployment-relevance sanity check* rather than a calibrated dose–response curve. The eight checkpoints are chosen to span the main regimes identified by the survey—low-peak dense (Qwen3.5-0.8B), low-peak Qwen3.5 variants, medium-baseline Qwen2.5, high-peak dense Qwen3 checkpoints, a MoE checkpoint (Qwen3-30B-A3B), and a high-peak Gemma-family checkpoint (Gemma3-4B-it)—so that any M-dependence of SQNR has a chance to manifest at the same layer hooks used in the rest of the study. The two recipes (max-abs and 99.9% clipping) bracket common per-tensor symmetric scale-selection strategies and are directly affected by the activation upper bound; advanced mitigations such as rotation-based quantization (Ashkboos et al., 2024; Liu et al., 2024a) or KV-cache prefixing (Chen et al., 2024) are designed to neutralize M rather than to expose it, so they are intentionally not part of this probe. We therefore read the M-versus-SQNR relationship as qualitative evidence that maximum activation magnitude translates into observable scale-selection error, not as a generalized scaling law over all 24 checkpoints.

Relation to 2024–2026 quantization advances. A wave of post-2024 quantization techniques—dual-rotation DuQuant (Lin et al., 2024a), learned affine flattening FlatQuant (Sun et al., 2025), and KV-cache-targeted KIVI (Liu et al., 2024b), alongside the earlier rotation-based QuaRot/SpinQuant (Ashk-

boos et al., 2024; Liu et al., 2024a) and prefix-based PrefixQuant (Chen et al., 2024)—has substantially narrowed the SQNR gap that motivates max-abs scaling, and a parallel low-precision-pretraining line (DeepSeek-AI et al., 2025; DeepSeek-AI, 2025) folds analogous mitigations into FP8 training itself. These methods reduce, redistribute, or absorb the numerical effect of the activation upper bound; they are complementary intervention studies rather than direct measurements of the untreated checkpoint range. Our contribution is orthogonal: we measure the upper bound itself across a broad family/architecture/training-stage span, and we show that M varies by nearly four orders of magnitude across modern open releases. The deployment cost of a mitigation may depend on the dynamic range we report, but the relationship is method- and implementation-dependent rather than a universal monotone law.

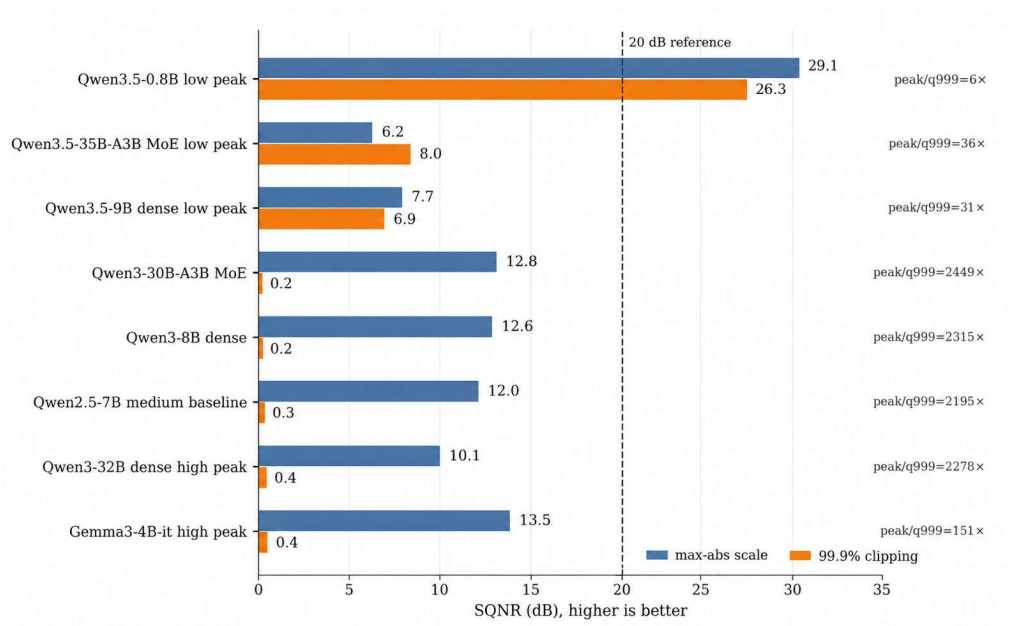

Figure 11: INT-8 activation quantization sanity check for eight representative models. Grouped horizontal bars show SQNR at the peak hidden layer under max-abs scaling and 99.9% clipping. The dashed line marks 20 dB; higher values indicate lower quantization error. Right-side annotations report the ratio between the global peak and the 99.9% clipping threshold.

D.3 Threats to validity and limitations

This is an observational measurement study: family, architecture, normalization, training data, and optimization choices are not independently controlled, so matched comparisons reduce but do not eliminate confounding and should not be interpreted causally. The statistic M is the maximum observed under the stated corpus, tokenizer, context length, and hook set; it is not a guaranteed upper bound for unseen deployment inputs. The domain check supports the reported ordering at the level of tested medians, but does not establish robustness of the global maximum on other distributions.

The quantization evidence is also deliberately scoped. The INT-8 experiment is a reconstruction sanity check, and the W8A8 follow-up covers only three checkpoints under one representative fake-quantization recipe; neither establishes downstream task degradation, serving speed or memory benefits, or a universal prediction law for M. We do not evaluate smoothing-, rotation-based (Ashkboos et al., 2024; Liu et al., 2024a), or prefix-based (Chen et al., 2024) interventions, and the matched MoE–dense and vision-language–text comparisons each contain only two pairs within one family. These limitations define the scope of the empirical claims; full implementation details and checkpoint revisions should accompany the released artifacts.

Table 3: Layerwise prevalence for the parameter-matched Qwen3 MoE–dense pair. Similar coverage but very different M_{hidden} indicates that the MoE–dense contrast is primarily an activation-scale difference rather than a coverage difference.

Model	Architecture	M_{hidden}	Coverage@50
Qwen3-30B-A3B	MoE	1,512	0.958
Qwen3-32B	Dense	35,328	0.891

Table 4: Representative perplexity results for W8A8 activation quantization. BF16, W8A16, and W8A8 report PPL; W8A8 change is relative to BF16. In these three checkpoints, W8A16 remains close to BF16, whereas W8A8 degradation is much larger for the checkpoints with larger M_{hidden}.

Model	M_{hidden}	BF16 PPL	W8A16 PPL	W8A8 PPL	W8A8 change
Qwen3-30B-A3B	1,512	4.0994	4.0998	4.2392	$+3.41\%$
Qwen3-32B	35,328	3.7213	3.7222	14.2252	$+282.26\%$
Gemma-3-27B-it	696,320	8.0013	7.9362	5,052.7429	$\sim +\,63{,}049\%$

Linking each failing checkpoint to specific normalization or attention dynamics—for example via residual-stream RMS profiles or the early-layer step-up blocks identified by Sun et al. (2026)—is left to a mechanistic follow-up.

E Follow-up Experiment Details

This appendix provides the detailed settings and results for four follow-up analyses whose main conclusions are integrated into the corresponding sections on layerwise structure, quantization, input-domain robustness, and MoE routing. The hidden-state carrier counts $22/24$ and $25/27$ refer to the 24 main-analysis checkpoints and the expanded suite including three Qwen2.5-Instruct checkpoints, respectively. Both counts concern the component carrying the global maximum across all six hooked component classes; their denominators differ because the checkpoint sets differ. Separately, M_{hidden} denotes the largest absolute post-residual hidden-state activation, rather than the maximum over all component classes.

E.1 Layerwise prevalence of large hidden activations

For each checkpoint, we define Coverage@50 as the fraction of layers whose hidden-state peak reaches at least 50% of M_{hidden}. Across all 27 checkpoints, the global maximum is carried by a post-residual hidden state in 25 cases. The Spearman correlation between $\log_{10} M_{\text{hidden}}$ and Coverage@50 is 0.165, with a permutation p-value of 0.4138. This test does not establish a monotonic association in this sample, but it also does not establish independence. Large hidden activations are widespread in many checkpoints: $\text{Coverage@50} \geq 0.5$ for 21/27 checkpoints, while only 5/27 have $\text{Coverage@50} < 0.1$.

The dense checkpoint has a $23.37\times$ larger M_{hidden} despite slightly lower coverage. This result supports interpreting magnitude and prevalence as complementary deployment statistics: a model can have high peaks that are broadly distributed, but the breadth of the elevated layerwise range does not by itself determine the peak scale.

E.2 Representative end-to-end W8A8 evaluation

We next evaluated a representative end-to-end quantization setting using symmetric per-output-channel W8 weights and static per-tensor A8 fake quantization. Calibration and evaluation each used 256 disjoint sequences of 2,048 tokens, or approximately 0.52M tokens per split. Norms, softmax, RoPE, residual connections, nonlinearities, embeddings, the language-model head, and the MoE router remained in BF16. The W8A16 condition isolates weight-only quantization error.

Table 5: DomainRatio ranges across seven domains. The Qwen3 checkpoints remain close to their multi-domain references, while Gemma-3-27B-it varies more, especially on Chinese and minority-language inputs.

Model	DomainRatio range
Qwen3-30B-A3B	0.981–1.012
Qwen3-32B	0.942–1.047
Gemma-3-27B-it	1.000–1.233

Table 6: Routing statistics for Qwen3-30B-A3B. Peak events are non-uniformly routed and routing concentration tracks MoE-output magnitude in most sparse layers.

Metric	Result
Median $\mathrm{ERC}_{\mathrm{entropy}}$	0.1413
Median $\mathrm{ERC}_{\mathrm{margin}}$	0.0221
Median $\mathrm{ARC}_{\mathrm{moe}}$	0.3118
Layers with positive $\mathrm{ARC}_{\mathrm{moe}}$	40/48
Median $\mathrm{ARC}_{\mathrm{res}}$	-0.3411
Median $\mathrm{PRS}_{\mathrm{JSD}}$	0.3502
Median $\mathrm{PRS}_{\mathrm{top5}}$	0.6139
Global peak-routing JSD	0.0876

W8A16 remains essentially unchanged from BF16 for all three checkpoints, while W8A8 shows much larger degradation in the higher-M_{hidden} cases. Within the Qwen3 pair, the dense model has a $23.37\times$ larger M_{hidden} and an approximately $82.76\times$ larger relative perplexity increase. These results strengthen the deployment relevance of the original SQNR sanity check, but they remain a representative three-checkpoint evaluation rather than a universal law relating M to downstream task quality.

E.3 Domain-stratified stability

To test whether the measured ordering is specific to the aggregate corpus, we scanned seven domains using 100 sequences per domain, three repeats, and 2,048-token inputs. For each model, *DomainRatio* is the domain median divided by the median from a same-sized multi-domain reference sample.

The reported Qwen3 dense-to-MoE ratios are $24.000\times$ on math, $24.900\times$ on code, $25.600\times$ on web text, $26.872\times$ on knowledge-oriented text, $27.170\times$ on Chinese, $26.200\times$ on minority-language inputs, and $25.679\times$ on mixed or other inputs. These measurements show a dense–MoE gap across all seven tested domains. However, stability of domain medians does not establish stability of the global maximum M, nor robustness to unseen input distributions.

E.4 MoE routing and activation peaks

Finally, for Qwen3-30B-A3B we aligned routing and activation statistics on 100 sequences across all 48 sparse layers. Peak events are the top 0.1% post-residual token activations within each layer. We compute the routing concentration statistics and activation–routing correlation statistics separately for each sparse layer, then report medians across the 48 layer-level values; the count of positive $\mathrm{ARC}_{\mathrm{moe}}$ values is likewise taken over layers. $\mathrm{ERC}_{\mathrm{entropy}}$ and $\mathrm{ERC}_{\mathrm{margin}}$ are concentration scores derived from the layer-level routing distributions, while $\mathrm{ARC}_{\mathrm{moe}}$ and $\mathrm{ARC}_{\mathrm{res}}$ are correlation statistics between routing concentration and, respectively, MoE-output peak magnitude and residual-stream peak magnitude. The exact correlation coefficient and entropy normalization are implementation details that should be pinned with the released analysis code; we do not use these statistics to make a causal claim.

$\mathrm{PRS}_{\mathrm{JSD}}$ is the Jensen–Shannon divergence between the routing distribution of peak tokens and that of all tokens, computed at the layer level before taking the reported median; its theoretical maximum is $\ln 2$. $\mathrm{PRS}_{\mathrm{top5}}$ is the total routing mass of peak tokens assigned to the five experts with

the largest peak-token routing mass. For an expert e, expert lift is $\mathrm{ExpertLift}(e) = q_{\mathrm{peak}}(e)/q_{\mathrm{all}}(e)$, where both q values are normalized routing masses over the corresponding peak-token or all-token population at the same analysis level. The separately reported global peak-routing JSD compares the routing distribution at the global peak event with the corresponding all-token distribution and is not a median over layers. The median $\mathrm{PRS}_{\mathrm{JSD}}$ is 50.5% of $\ln 2$, showing a routing shift for peak-activation tokens. The largest expert-lift values are $5.855\times$ (expert 74), $5.576\times$ (expert 96), $5.014\times$ (expert 38), and $4.858\times$ (expert 117). Together with positive $\mathrm{ARC}_{\mathrm{moe}}$ in 40/48 layers, these results associate peak events with concentrated and non-uniform expert routing, while the negative median $\mathrm{ARC}_{\mathrm{res}}$ indicates that the same routing concentration does not explain the residual-stream peak through a simple positive correlation.

F Related Work

Outlier features and massive activations. Large activations in transformer language models were first brought into focus by low-bit inference failures. Dettmers et al. (2022) showed that sufficiently large OPT- and BLOOM-style models develop sparse, high-magnitude "emergent features" whose preservation is necessary for lossless INT8 inference. Sun et al. (2024) sharpened this observation into the notion of *massive activations*: token-feature coordinates that are both absolutely large and locally sparse relative to the surrounding hidden-state distribution. This line of work established two facts that are now central to activation-aware deployment: the extreme coordinates occupy a tiny fraction of the representation, yet ablating or quantizing them naively can cause disproportionate degradation. Subsequent analyses extended the taxonomy beyond hidden-state activations. An et al. (2025) relate activation outliers, weight outliers, and attention outliers, arguing that they are mutually structured rather than independent numerical accidents, while Macocco et al. (2025) show that last-layer outlier dimensions can implement useful high-frequency-token prediction heuristics. These studies motivate treating extreme activations as functional model components, not merely as pathological noise.

Attention sinks and mechanistic accounts. A major mechanistic explanation links massive activations to the softmax attention constraint. When an attention head has no useful context to retrieve, the probability simplex still forces it to allocate its mass somewhere; models often learn to route this mass to a beginning-of-sequence token or another low-information token, creating an attention sink. Bondarenko et al. (2023) connect such no-op behavior to quantization difficulty, and Gu et al. (2025) empirically trace when sink behavior emerges during pretraining. More recent work refines this picture. Kaul et al. (2025) attribute first-token attention dominance to softmax geometry and large hidden-state kurtosis partly to coordinate-wise adaptive optimization, proposing softmax-1 and OrthoAdam as architectural and optimizer-level remedies. Sun et al. (2026) decouple residual-stream activation spikes from local attention sinks, showing that both phenomena co-occur in pre-norm transformers but need not be the same object. Queipo-de Llano et al. (2026) further link attention sinks to compression valleys, proving that extreme residual-stream norms induce representational compression and proposing a mix-compress-refine view of depth-wise computation. Together these works explain why massive activations may be useful for information routing and stabilization, while also clarifying why their magnitude can become a deployment bottleneck.

Quantization and numerical mitigation. The practical importance of massive activations is most visible in post-training quantization. Because per-tensor or per-channel scales are often governed by the largest observed magnitude, even a small number of extreme values can waste quantization levels and increase reconstruction error for ordinary activations. LLM.int8() preserves sparse outlier dimensions in higher precision while quantizing the remaining bulk activations (Dettmers et al., 2022). SmoothQuant instead migrates activation difficulty into weights through an algebraically equivalent rescaling (Xiao et al., 2023), while Outlier Suppression+ shifts and rescales activations to reduce quantization sensitivity (Wei et al., 2023). Weight-only and activation-aware methods such as GPTQ and AWQ reduce memory and bandwidth pressure while protecting salient parameters (Frantar et al., 2023; Lin et al., 2024b). More recent methods transform the representation basis or flatten the activation landscape: QuaRot and SpinQuant use rotations to distribute outlier mass (Ashkboos et al., 2024; Liu et al., 2024a), DuQuant applies dual transformations (Lin et al., 2024a), FlatQuant learns affine flattening transforms (Sun et al., 2025), and PrefixQuant or KIVI target outliers in prefixed activations and KV caches (Chen et al., 2024; Liu et al., 2024b). These methods are complementary

to our study: they reduce or absorb the numerical effect of large activations, whereas we measure the upper-bound activation magnitude itself across current open model families.

Architectural and multimodal interventions. A parallel line of work asks whether outlier-like artifacts arise because transformer architectures lack explicit storage locations for global or null information. In vision transformers, high-norm artifact tokens appear in low-information image patches and harm dense prediction; register tokens provide dedicated learned slots that absorb this global computation and remove the artifact pattern (Darcet et al., 2024). Post-hoc register methods show that similar benefits can be obtained without full retraining by distilling artifact-free dense features into a lightly modified model (Chen et al., 2025). In multimodal language models, Anand et al. (2026) show that attention sinks and massive activations can emerge not only at the BOS token but also at intermediate low-semantic audio-visual tokens during fine-tuning, and that decorrelating intermediate tokens from the BOS representation mitigates both the sink and activation effects. These results suggest that maximum activation behavior is sensitive to modality adaptation, token roles, and the availability of architectural "scratch space."

Position of this work. Existing studies have primarily treated massive activations as a binary phenomenon, a mechanistic artifact, or an obstacle to be removed by a particular quantization method. They also concentrate on a limited set of earlier LLaMA-, OPT-, or BLOOM-style checkpoints. Modern open models now vary along many additional axes, including normalization and training recipes, dense versus MoE computation, multimodal adaptation, instruction tuning, and released intermediate training stages (Qwen Team, 2024; 2025; Bai et al., 2025; Gemma Team, 2024; 2025; Ling Team, 2025; OpenAI, 2025). Our work is therefore complementary to both the interpretability and quantization literatures. Rather than asking only whether a model contains a massive activation under a fixed local criterion, we measure the continuous deployment-relevant statistic $M = \max |a|$ under a unified protocol, compare it across recent open families and matched architectural or training contrasts, and connect it to activation-quantization reconstruction error. This framing turns maximum activation magnitude into a protocol-conditioned model statistic that should be reported alongside open-weight releases together with its measurement protocol and layerwise carrier.

G Appendix Summary

We presented a unified measurement study of maximum activations across 24 checkpoints from 8 modern open LLM families. Our central message is empirical and definitional: *the protocol-conditioned maximum activation magnitude M of an open LLM is a continuous, releasable statistic that is not predicted by parameter count alone*. Within a family, M often grows with model size, but cross-family, cross-generation, and cross-architecture comparisons break simple monotonic scaling, and matched-design contrasts associate MoE routing, vision-language adaptation, supervised fine-tuning, and training stage with measurable shifts in M or in its layerwise structure. The residual stream carries the global maximum in 22 of 24 checkpoints, with GPT-OSS-20B as a single MLP-output exception that any quantization scheme should accommodate. A lightweight INT-8 sanity check further indicates that M co-varies with low-bit reconstruction error through its effect on activation-scale selection. We therefore recommend that M, its protocol, and its layerwise carrier be reported alongside any open-weight release as part of a deployment-oriented model card extension. We deliberately stop short of causal claims about why families differ; explaining the mechanism behind these differences—and intervening on it—is left as future work.

H Additional Figures

The appendix presents complementary evidence for the main-text conclusions, including within-family scaling, deployment-oriented magnitude tiers, per-family layerwise trajectories, and component-level carrier differences in representative models. These views provide additional verification of the reported activation patterns at the family, layer, and component levels.

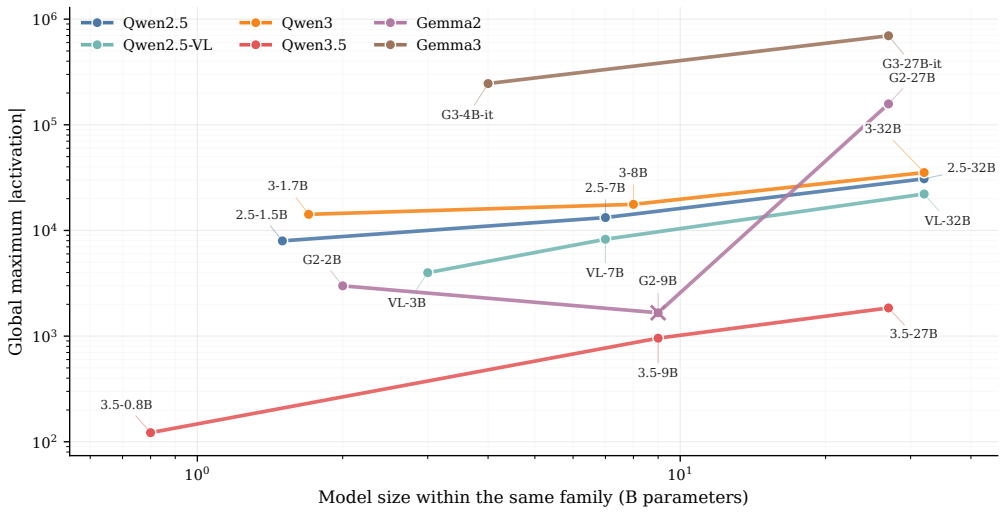

Figure 12: Within-family scaling effects. The figure compares size changes only within the same family and model form; crosses mark local non-monotonic cases. The vertical axis is global maximum activation magnitude on a log scale.

H.1 Deployment-oriented tiers

Figure 13 groups all main-analysis checkpoints into five order-of-magnitude tiers according to global maximum activation magnitude. The figure is not intended to propose a new threshold standard; rather, it provides an intuitive deployment-oriented reference for selecting quantization and activation-scaling strategies.

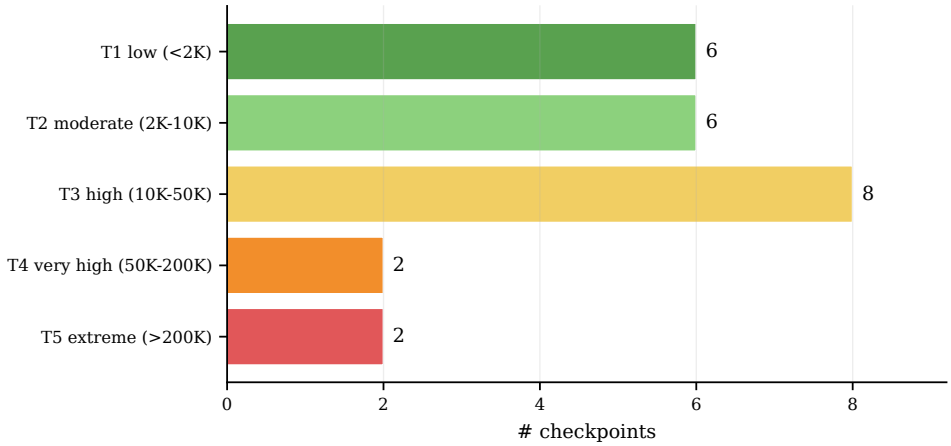

Figure 13: Deployment-oriented tiers based on global maximum activation magnitude. The horizontal bar chart counts checkpoints in each tier. The horizontal layout avoids crowded tier labels. The figure is based only on observed activation magnitudes and is intended to indicate that different models may require different quantization treatment.

H.2 PER-FAMILY LAYERWISE MAXIMUM ACTIVATIONS

Figure 14 shows hidden-state layerwise maximum-activation trajectories within each model family. The horizontal axis is normalized depth, and the vertical axis is hidden-state layerwise maximum absolute activation. Because peak magnitudes span several orders of magnitude across families, the vertical axis uses a log scale.

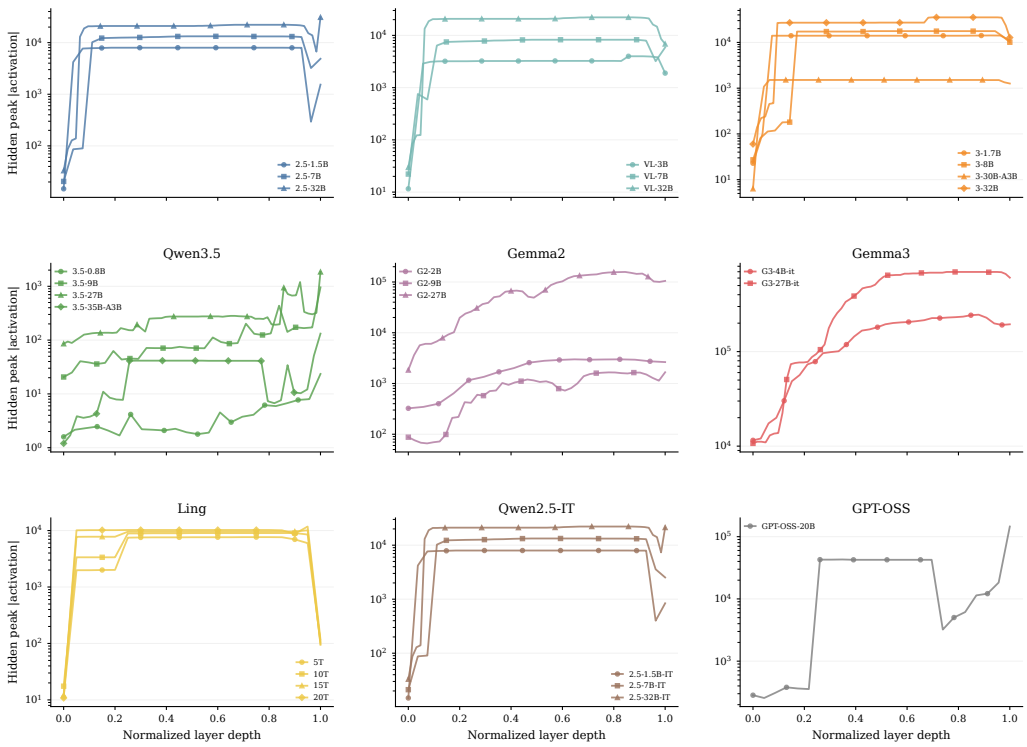

Figure 14: Hidden-state layerwise maximum-activation trajectories within each model family. Each subplot corresponds to a family or training form, and each curve corresponds to one checkpoint in that group. The vertical axis uses a log scale to accommodate cross-family magnitude differences.

H.3 COMPONENT-LEVEL MAXIMUM ACTIVATION DISTRIBUTIONS

Figure 15 presents component-level layerwise trajectories for three representative checkpoints, complementing the main-text discussion of carrier components. The three subplots correspond to a low-peak model, a high-peak model, and the GPT-OSS component-level exception. Compared with plotting all checkpoints, these representative panels more clearly show the magnitude differences among hidden states, attention outputs, and MLP/MoE outputs.

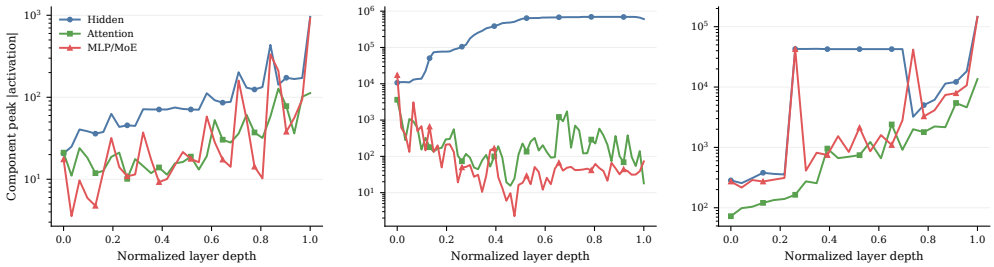

Figure 15: Component-level maximum-activation trajectories for representative models. The three subplots show a low-peak model, a high-peak model, and the GPT-OSS component-level exception. Curves correspond to hidden states, attention outputs, and MLP/MoE outputs; the vertical axis is component-level maximum absolute activation on a log scale.

I SUPPLEMENTARY ACTIVATION RANKS

Table 7: Additional absolute-activation ranks at the same representative locations as Table 1 for all 24 main-analysis checkpoints. Model names and × failure markers follow that table.

Model	Top 2	Top 3	Top 4	Top 5	Top-10
Qwen2.5-1.5B×	7,968	7,968	7,968	7,968	7,968
Qwen2.5-7B	13,248	13,248	13,248	13,248	13,184
Qwen2.5-32B	21,888	21,760	21,760	21,760	21,632
Qwen2.5-VL-3B	3,248	3,248	3,248	3,088	3,056
Qwen2.5-VL-7B	8,256	8,256	8,256	8,256	8,256
Qwen2.5-VL-32B	22,144	22,144	22,144	22,144	22,144
Qwen3-1.7B	14,144	14,144	14,144	14,144	14,080
Qwen3-8B	17,664	17,664	17,664	17,664	17,664
Qwen3-30B-A3B	1,512	1,512	1,512	1,512	1,512
Qwen3-32B	35,328	35,328	35,328	35,328	35,328
Qwen3.5-0.8B×	122.0	114.5	114.5	114.0	22.5
Qwen3.5-9B×	920.0	920.0	912.0	908.0	848.0
Qwen3.5-27B	167.0	167.0	167.0	160.0	152.0
Qwen3.5-35B-A3B×	131.0	129.0	126.5	125.0	121.0
gemma-2-2b	2,992	2,976	2,944	2,944	2,848
gemma-2-9b	1,560	1,160	1,112	1,040	808.0
gemma-2-27b	156,672	155,648	153,600	151,552	149,504
gemma-3-4b-it	243,712	242,688	241,664	241,664	237,568
gemma-3-27b-it	696,320	696,320	696,320	696,320	696,320
Ling-mini-5T	7,648	7,648	7,648	7,648	7,648
Ling-mini-10T	9,024	9,024	9,024	9,024	9,024
Ling-mini-15T	9,600	9,600	9,600	9,600	9,600
Ling-mini-20T	10,048	9,984	9,920	9,856	9,664
gpt-oss-20b	42,496	42,496	42,496	42,496	42,496