When Do Large Language Models Diverge? Late-Stage Trajectory Divergence in Multiple-Choice QA

Anonymous ACL submission

Abstract

Answering multiple-choice questions (MCQA) with large language models requires reducing complex reasoning to a single-token prediction. Existing evaluation metrics, such as output perplexity (PPL), summarize final predictions but obscure how answers take shape inside the network. To investigate these internal dynamics, we track Layer-wise Choice Probability (LCP) via the Logit Lens for 54,060 items across six instruction-tuned models. We find that items naturally stratify into three calibration groups—Correct, Mixed, and Incorrect—with PPL often failing to distinguish their distinct internal computations. Specifically, we show that (i) each group exhibits a reproducible LCP profile; (ii) Mixed-group trajectories diverge only in the final 12–32% of layers (*late-stage trajectory divergence*); and (iii) reasoning benchmarks disproportionately fall into these late-diverging regimes. Our results demonstrate that depth-resolved LCP probing provides qualitative distinctions between types of uncertainty. [1]

1 Introduction

Answering multiple-choice questions (MCQA) with large language models (LLMs; Ouyang et al. (2022); Dubey et al. (2024); Team et al. (2025); Yang et al. (2024); Chen et al. (2021)) requires the model to reduce complex reasoning into a *single-token prediction* (Gao et al., 2021). The goal is to accurately assign the highest probability to the correct option token based on the input context.

Existing evaluation methods typically rely on output-level confidence metrics, such as perplexity (PPL). While effective for summarizing final predictions, these methods fall short in explaining how answers actually take shape inside the network. Recent interpretability work highlights the need to understand these internal dynamics:

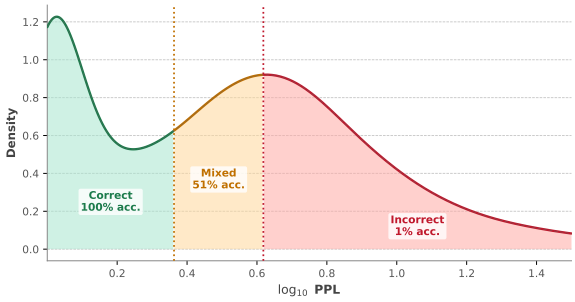

Figure 1: $\log_{10}$ PPL on Llama-3.1-8B-Instruct (example; six models in Appendix Fig. 7).

Geva et al. (2022) demonstrated that transformer layers build predictions iteratively by promoting vocabulary concepts, and Kim et al. (2025) showed that layer-wise trajectories for certain and uncertain outputs often remain surprisingly aligned until the terminal layers. This highlights the need to investigate whether models exhibit distinct *internal processing regimes* when navigating different levels of difficulty.

To address this, we adopt a depth-resolved approach to analyze LLM inference in MCQA. We evaluate 54,060 items from six benchmarks across six instruction-tuned LLMs. By applying the Logit Lens (nostalgebraist, 2020), we track the Layer-wise Choice Probability (LCP) for the gold answer, moving beyond scalar output metrics to capture internal computation.

Our analysis reveals that items naturally stratify into three calibration groups—Correct, Mixed, and Incorrect (Figure 1)—with output PPL often failing to distinguish the distinct internal computation of Mixed and Incorrect items. In summary, our main findings are: (i) each group exhibits a distinct, reproducible LCP profile; (ii) Mixed-group trajectories diverge only in the final 12–32% of layers (*late-stage trajectory divergence*); and (iii) reasoning benchmarks disproportionately fall into these late-diverging regimes. Thus, depth-resolved LCP

[1] anonymous github link https://anonymous.4open.science/r/Late-StageTrajectoryCommitment-5003/README.md

probing provides a qualitative distinction between types of uncertainty.

2 Related Work

Layer-wise probing. The Logit Lens (nostalgebraist, 2020) and Tuned Lens (Belrose et al., 2023) decode intermediate representations across depth to reveal how predictions form. Recent work uses these tools to uncover systematic layer-wise variation in uncertainty and representational geometry (Godey et al., 2024; Razzhigaev et al., 2024). For instance, Geva et al. (2022) show that transformer layers iteratively build predictions by promoting vocabulary concepts. Furthermore, Kim et al. (2025) find that layer-wise trajectories for certain and uncertain outputs often remain aligned throughout most of the network, suggesting that uncertainty does not fundamentally alter early-stage inference dynamics.

Example difficulty and data quality. Prior work estimates example difficulty through structured data design (Xu et al., 2025), multi-dimensional metrics (Zhang et al., 2026), external ratings (Wettig et al., 2024), output-level signals (Swayamdipta et al., 2020), and loss- or gradient-based statistics (Sow et al., 2025; Dai et al., 2025; Thakkar et al., 2023). These approaches summarize examples through final predictions rather than internal layer-wise computation.

3 Methodology

3.1 Experimental Setup

We analyze six instruction-tuned LLMs—Gemma-3 (4B, 12B; Team et al. 2025), Llama-3 (3B, 8B; Dubey et al. 2024), and Qwen3 (4B, 8B; Yang et al. 2025)—on a pooled MCQA corpus of 54,060 items from the train splits of six benchmarks: MathQA (Amini et al., 2019), CommonsenseQA (Talmor et al., 2019), ARC-Easy and ARC-Challenge (Clark et al., 2018), OpenBookQA (Mihaylov et al., 2018), and LogiQA (Liu et al., 2020). Each item is formatted as a four-option MCQ; the model receives the question and options and predicts a single answer token.

All experiments are inference-only: we perform forward passes and record, for each model–sample pair, the predicted answer, correctness, output perplexity, and layer-wise hidden states. No fine-tuning or parameter updates are performed. Perplexity is computed on the gold-answer token given the prompt. Analysis is conducted *per model*: group assignments are computed independently within each model's 54,060 outputs.

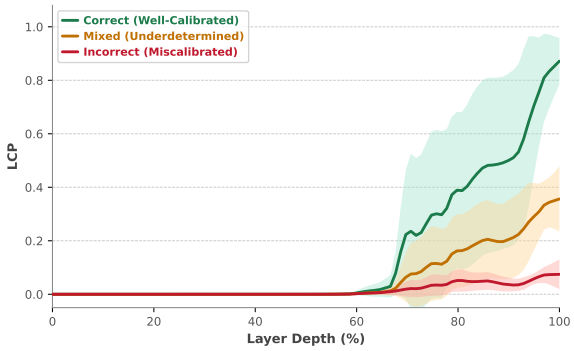

Figure 2: Mean LCP by calibration group (six models).

3.2 Three Calibration Groups

Each forward pass is assigned to one of three groups (per model, on $\log_{10}$ PPL): **Correct**—top-1 correct and PPL in the low tail; **Incorrect**—PPL in the high tail; **Mixed**—all remaining items with intermediate PPL and mixed correctness. Cut points follow the 37th/58th percentile of each model's empirical $\log_{10}$ PPL distribution; see Appendix A for details and a full group-definition table.

3.3 Layer-wise Choice Probability (LCP)

Perplexity summarizes model confidence at the *final* layer; to examine *internal* computation, we apply the Logit Lens (nostalgebraist, 2020). Let $\mathbf{h}_l(\mathbf{x}) \in \mathbb{R}^d$ denote the last-token hidden state at layer l for input $\mathbf{x}$, and let L be the total number of layers. We decode each intermediate state through the model's final layer normalization and unembedding matrix $\mathbf{W}_U$:

$$\mathrm{LCP}_l(\mathbf{x}) = \left[\,\mathrm{softmax}\big(\mathbf{W}_U\,\mathrm{LN}(\mathbf{h}_l(\mathbf{x}))\big)\,\right]_{y^*} \quad (1)$$

where y^* is the gold-answer token. $\mathrm{LCP}_l(\mathbf{x})$ measures the model's layer-l support for the correct answer, yielding a depth-resolved trajectory $\{\mathrm{LCP}_l\}_{l=1}^{L}$ per sample. Figure 2 shows mean LCP trajectories for the three calibration groups, aggregated across all six models (shaded: cross-model standard deviation).

4 Results

4.1 Group-Specific LCP Profiles

Table 1 reports per-model counts, mean final-layer LCP_L, and group-wise accuracy on the pooled train corpus. Across all six models, each calibration

Model	Correct			Mixed			Incorrect		
	N	LCP_L	Acc.	N	LCP_L	Acc.	N	LCP_L	Acc.
Gemma-3-12B-it	27,189	0.951	95.1%	365	0.490	50.1%	26,506	0.048	4.8%
Gemma-3-4B-it	22,581	0.894	89.9%	1,018	0.416	43.3%	30,461	0.075	7.5%
Llama-3.1-8B-Inst.	19,883	0.805	93.4%	11,726	0.324	47.7%	22,451	0.136	5.3%
Llama-3.2-3B-Inst.	20,430	0.718	90.5%	8,934	0.313	47.2%	24,696	0.158	6.8%
Qwen3-8B	28,957	0.888	99.2%	3,901	0.126	37.5%	21,202	0.004	3.6%
Qwen3-4B	28,429	0.974	99.4%	765	0.469	57.8%	24,866	0.027	1.8%

Table 1: Per-model group size, LCP_L, and within-group accuracy (benchmark top-1: Appendix Table 3).

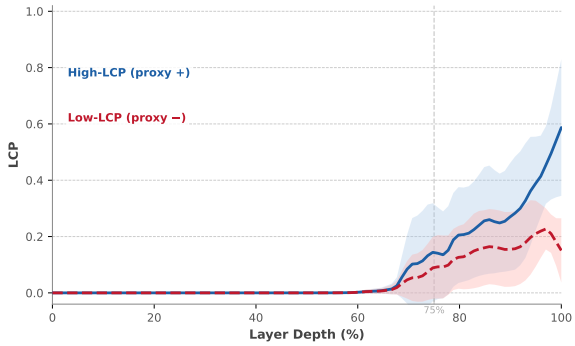

Figure 3: Mixed group: high vs. low final-LCP-rank proxy subsets (Appendix D.2), averaged across six models.

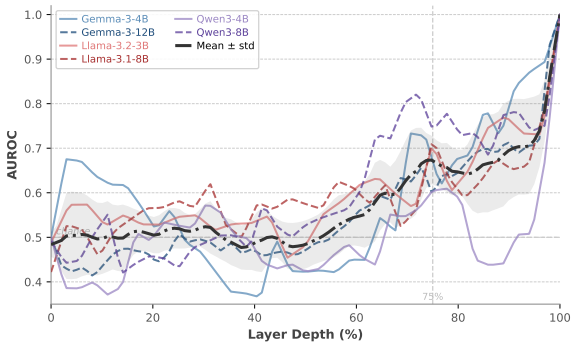

Figure 4: Layer-wise LCP AUROC for predicting final-layer LCP rank in the Mixed group (proxy label; six models).

group exhibits a reproducible internal LCP profile (Figure 2; Table 1; per-model gallery in Appendix B). Correct samples (89–99% group accuracy) show high LCP_L (0.72–0.97), rising monotonically toward the output. Incorrect samples (1.8–7.5% accuracy) maintain $\mathrm{LCP}_L \approx 0$–0.16 throughout depth, with little change toward the gold token. The Mixed group (37–58% group accuracy) occupies an intermediate regime, with LCP_L of 0.13–0.49 depending on the model. The three profiles are similar across all six models: Correct / Mixed / Incorrect groups differ in depth-resolved LCP, not only in scalar PPL or accuracy.

In particular, Mixed and Incorrect samples—both often associated with elevated perplexity—exhibit measurably distinct internal LCP profiles. At the final layer, mean LCP on the gold token differs between the two groups across all six models (Mixed LCP_L: 0.13–0.49 vs. Incorrect LCP_L: 0.00–0.16; Table 1). This dissociation is invisible to output PPL alone and supports treating Mixed and Incorrect as separate processing regimes rather than a single "hard" category.

4.2 Late-Stage Trajectory Divergence in the Mixed Group

Within the Mixed group, output PPL poorly separates correct from incorrect predictions (Table 1). Figure 3 plots mean LCP trajectories for Mixed-group items split by high vs. low final-layer LCP rank (Appendix D.2), aggregated across six models. The trajectories overlap through most of depth, diverging only in the terminal layers. We use this final-rank proxy (rather than ground-truth labels) to isolate internal convergence dynamics from ultimate correctness, measuring *when* decisions form. We leverage large train splits to guarantee sufficient sample sizes for this fine-grained categorization, anticipating future applications in dynamic data filtering. Per-model trajectories and internal token competition dynamics are detailed in Appendices D and F.

To quantify the onset of separation, we apply layer-wise Mann–Whitney U tests with Benjamini–Hochberg (BH) correction to the two proxy subsets of each model's Mixed-group trajectories (Appendix D). Across all six models, the first FDR-significant separation (BH-adjusted $p < 0.01$, Wasserstein distance > 0.01) occurs at 68–88% of depth (mean: 74%), i.e., in the last 12–32% of

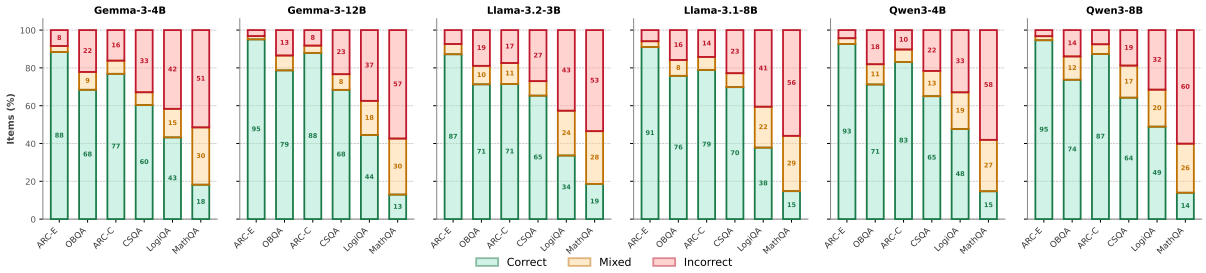

Figure 5: Calibration group distribution by benchmark for each of the six models (stacked bars; segment labels show %, omitted when <8%). The left-to-right shift from knowledge to reasoning benchmarks is reproduced across all model families and sizes.

layers. We term this pattern *late-stage trajectory divergence*: in ambiguous items, layerwise gold-token mass (LCP) for the high-rank proxy subset separates from the low-rank subset only near the output (Appendix D.2).

4.3 When Does LCP Predict Terminal Rank in the Mixed Group?

Having established that proxy-split LCP trajectories commit only at late depth, we ask whether LCP_l can predict *final-layer LCP rank* within the Mixed group (not ground-truth top-1 correctness; Appendix D.3). Figure 4 shows the AUROC of LCP_l as a predictor of the final-layer LCP rank proxy, plotted against relative layer depth for all six models.

AUROC stays near chance through the first 68–88% of depth and increases in the terminal layers, reaching values above chance mainly at or near the final layer. We do not observe a stable early-layer LCP threshold that separates the high- from the low-rank proxy subset in the Mixed group. The onset of separation is model-dependent (68–88% depth; Appendix D), but early depth is uninformative in all six models. Final-layer LCP carries more discriminative signal for terminal rank than intermediate layers, consistent with late-stage trajectory divergence.

4.4 Task-Type Group Distribution

Figure 5 shows, for each of the six models, the proportion of samples assigned to each calibration group per benchmark. Knowledge-oriented datasets (CommonsenseQA, ARC, OpenBookQA) concentrate in the Correct group, whereas reasoning-intensive datasets (MathQA, LogiQA) shift mass toward Mixed and Incorrect groups. This gradient is consistent across all six models and both model families (Gemma-3, Llama-3, Qwen3), ruling out a model-specific explanation. A χ^2 test on task type $\times$ group assignment confirms this association is significant ($\chi^2 = 78{,}818$, $p < 10^{-100}$, Cramér's $V = 0.35$, over all six models and 324,360 items; Appendix D).

5 Discussion and Future Work

While output PPL summarizes confidence, LCP reveals *when* internal trajectories diverge. This late-stage divergence involves active competition between the gold token and top alternatives (Appendix F). Future work should explore causal interventions (e.g., activation steering), apply late-layer LCP to filter training data, and use the Tuned Lens (Belrose et al., 2023) to mitigate early-layer projection artifacts. Ultimately, extending these analyses to free-form generation and chain-of-thought settings will be critical for understanding how models dynamically allocate internal computation to resolve ambiguity.

6 Conclusion

We address the challenge of interpreting model uncertainty in multiple-choice QA by tracking Layer-wise Choice Probability (LCP) across six large language models. By analyzing intermediate representations, we reveal that scalar output metrics obscure distinct internal inference dynamics. Specifically, we identify *late-stage trajectory divergence*, wherein internal trajectories for ambiguous items only separate in the final 12–32% of layers. Through this depth-resolved probing, we demonstrate that reasoning-intensive tasks disproportionately trigger this late-diverging regime, capturing crucial qualitative differences in model uncertainty.

Limitations

First, our methodology relies primarily on associational, inference-only probing. While we establish robust observational links between output calibration and depth-resolved trajectory profiles, this approach cannot fully uncover the underlying causal mechanisms. Future work must incorporate causal interventions, such as activation steering or layer patching, to definitively establish how altering intermediate representations shifts the onset of late-stage divergence and ultimate model predictions.

Second, our experiments are conducted on the train splits of the evaluated benchmarks. We deliberately chose this setting to secure the substantial sample sizes necessary for stable distribution fitting at the statistical tails (the Correct and Incorrect regimes), and to provide groundwork for future confidence-weighted training curricula. Consequently, because instruction-tuned models may have memorized portions of these training sets, our current findings cannot strictly disentangle multi-hop reasoning from sophisticated retrieval. Extending this framework to zero-shot, out-of-distribution evaluation remains a critical next step to validate the generalizability of these internal dynamics.

Finally, the use of the standard Logit Lens at early layers may introduce decoding artifacts, as the model's intermediate hidden states are not yet aligned with the output vocabulary space. Although the persistence of the divergence pattern across disparate model families suggests a genuine architectural phenomenon, applying techniques like the Tuned Lens (Belrose et al., 2023) could help mitigate projection noise. Furthermore, our analysis of the Mixed group depends on a final-layer LCP proxy split to capture internal convergence. Bridging the theoretical gap between this internal consistency metric and objective, ground-truth correctness is an important direction for future alignment research.

Ethical Considerations

This study analyzes publicly available MCQA benchmarks using inference-only forward passes on pretrained instruction-tuned models. It does not involve human subjects, personal data, or model fine-tuning. Our Layer-wise Choice Probability (LCP) trajectories are associational probes and should not be interpreted as causal evidence of reasoning or as reliable confidence estimates for high-stakes use. Analyses are limited to English MCQA benchmarks and required substantial GPU compute; we release code to support reproducibility. LCP should be treated as a descriptive interpretability tool, not a guarantee of model reliability, fairness, or safety.

We used AI-assisted tools solely for grammar checking and linguistic refinement of the manuscript.

References

Aida Amini, Saadia Gabriel, Shanchuan Lin, Rik Koncel-Kedziorski, Yejin Choi, and Hannaneh Hajishirzi. 2019. MathQA: Towards interpretable math word problem solving with operation-based formalisms. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2357–2367, Minneapolis, Minnesota. Association for Computational Linguistics.

Nora Belrose, Igor Ostrovsky, Lev McKinney, Zach Furman, Logan Smith, Danny Halawi, Stella Biderman, and Jacob Steinhardt. 2023. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mohammad Bavarian, Clemens Winter, Philippe Tillet, Felipe Petroski Such, Dave Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William Hebgen Guss, Alex Nichol, Alex Paino, Nikolas Tezak, Jie Tang, Igor Babuschkin, Suchir Balaji, Shantanu Jain, William Saunders, Christopher Hesse, Andrew N. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, Matthew Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, Ilya Sutskever, and Wojciech Zaremba. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *Preprint*, arXiv:1803.05457.

Yalun Dai, Yangyu Huang, Xin Zhang, Wenshan Wu, Chong Li, Wenhui Lu, Shijie Cao, Li Dong, and Scarlett Li. 2025. Data efficacy for language model training. *Preprint*, arXiv:2506.21545.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Leo Gao, Jonathan Tow, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Kyle McDonell, Niklas Muennighoff, Jason Phang, Laria Reynolds, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2021. A framework for few-shot language model evaluation.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 conference on empirical methods in natural language processing*, pages 30–45.

Nathan Godey, Éric Clergerie, and Benoît Sagot. 2024. Anisotropy is inherent to self-attention in transformers. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 35–48, St. Julian's, Malta. Association for Computational Linguistics.

Sunwoo Kim, Haneul Yoo, and Alice Oh. 2025. On the effect of uncertainty on layer-wise inference dynamics. *arXiv preprint arXiv:2507.06722*.

Jian Liu, Leyang Cui, Hanmeng Liu, Dandan Huang, Yile Wang, and Yue Zhang. 2020. Logiqa: A challenge dataset for machine reading comprehension with logical reasoning. *Preprint*, arXiv:2007.08124.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

nostalgebraist. 2020. interpreting gpt: the logit lens. `https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens`. Accessed: 2026-05-12.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc.

Anton Razzhigaev, Matvey Mikhalchuk, Elizaveta Goncharova, Ivan Oseledets, Denis Dimitrov, and Andrey Kuznetsov. 2024. The shape of learning: Anisotropy and intrinsic dimensions in transformer-based models. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 868–874, St. Julian's, Malta. Association for Computational Linguistics.

Daouda Sow, Herbert Woisetschläger, Saikiran Bulusu, Shiqiang Wang, Hans-Arno Jacobsen, and Yingbin Liang. 2025. Dynamic loss-based sample reweighting for improved large language model pretraining. *Preprint*, arXiv:2502.06733.

Swabha Swayamdipta, Roy Schwartz, Nicholas Lourie, Yizhong Wang, Hannaneh Hajishirzi, Noah A. Smith, and Yejin Choi. 2020. Dataset cartography: Mapping and diagnosing datasets with training dynamics. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9275–9293, Online. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4149–4158, Minneapolis, Minnesota. Association for Computational Linguistics.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Megh Thakkar, Tolga Bolukbasi, Sriram Ganapathy, Shikhar Vashishth, Sarath Chandar, and Partha Talukdar. 2023. Self-influence guided data reweighting for language model pre-training. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2033–2045, Singapore. Association for Computational Linguistics.

Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. *Preprint*, arXiv:2402.09739.

Hu Xu, Zeyan Li, Rui Wang, and Jianfeng Xu. 2025. Structure trumps size: Rethinking data quality for LLM reasoning. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11489–11513, Suzhou, China. Association for Computational Linguistics.

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

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. 2024. Qwen2. 5 technical report. *arXiv preprint arXiv:2412.15115*.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. 2026. Beyond random sampling: Efficient language model pretraining via curriculum learning. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5776–5794, Rabat, Morocco. Association for Computational Linguistics.

This appendix provides (A) the group assignment protocol and cutoff details (Section A), including benchmark-level train accuracy (Table 3), (B) per-model LCP profile gallery and PPL distributions (Section B), (C) benchmark-level group shares and cross-model consistency (Section C), (D) the statistical protocols and per-model trajectory-commitment gallery (Section D), and (E) a supplementary analysis of internal token competition in the Mixed group (Section F). Together these sections let readers verify that the main-text findings hold individually across all six models, not only in aggregate. All counts refer to train splits; analyses are inference-only.

A Calibration Group Assignment

Table 2 gives the full group-definition rules referenced in Section 3.2.

Group	Assignment rule
Correct (Well-Calibrated)	Top-1 correct *and* $\log_{10}$PPL $\leq$ 37th pct. LCP rises to high values. Acc. 89–99%.
Mixed (Underdetermined)	Neither Correct nor Incorrect: 37th–58th pct. PPL, mixed correctness. Late-stage trajectory divergence. Acc. 37–58%.
Incorrect (Miscalibrated)	$\log_{10}$PPL $\geq$ 58th pct. (predominantly wrong). LCP near zero throughout. Acc. 1.8–7.5%.

Table 2: Calibration group definitions (per model, N=54,060).

Cutoff definition. Low and high ranges are defined from the empirical $\log_{10}(\mathrm{PPL})$ distribution of that model's 54,060 train-split forward passes; cutoffs are not shared across models. Concretely, the low tail is defined by the *bottom 37th percentile* of each model's $\log_{10}(\mathrm{PPL})$ distribution, and the high tail by the *top 42%* (above the 58th percentile); items in the 37th–58th percentile band form the Mixed group. These boundaries are chosen so that Mixed-group accuracy falls near the random-chance range (37–58% for four-option items). Aliases Well-Calibrated, Underdetermined, and Miscalibrated denote the same three groups.

Mixed-group size imbalance. Mixed-group size varies substantially across models (365–11,726 items; Table 1). Trajectory-commitment and AUROC analyses are therefore weighted toward models with larger Mixed cohorts (Llama-3.1-8B, Llama-3.2-3B). For models with small Mixed groups (Gemma-3-12B N=365, Qwen3-4B N=765) the reported divergence-onset statistics should be interpreted with caution.

A.1 Benchmark-Level Train Accuracy

Table 3 reports top-1 accuracy (%) on each benchmark's train split, per model. All items use the same four-option MCQ prompt and single-token answer format as in Section 3.1. Counts per benchmark (per model) are: MathQA N=29,837; CommonsenseQA N=8,520; ARC-Easy N=2,251; ARC-Challenge N=1,119; OpenBookQA N=4,957; LogiQA N=7,376; overall N=54,060. Values are reproduced in `tables/table_benchmark_accuracy.csv`.

B Per-Model LCP Profiles and PPL Distributions

A key question is whether the three-group LCP pattern—visible in the aggregated Figure 2—holds for *each model individually*, or is an artifact of averaging across architectures with different layer counts and token distributions. Figure 6 answers this directly: every model shows the same qualitative ordering (Correct > Mixed > Incorrect in LCP_L), and the trajectory shapes are recognizably similar across Gemma-3, Llama-3, and Qwen3 families. Quantitative differences in absolute LCP values reflect known differences in model accuracy (Table 1).

Figure 7 shows $\log_{10}(\mathrm{PPL})$ density plots for all six models, with group boundaries shaded. Figure 1 (main text) shows an illustrative single-model panel (Llama-3.1-8B-Instruct).

Cross-model PPL structure. Several patterns are consistent across all six models. First, each model's $\log_{10}(\mathrm{PPL})$ distribution is approximately bimodal: a low-PPL peak (items the model assigns high probability to the gold token) and a high-PPL shoulder or secondary peak (items where the model places most mass elsewhere). The Mixed group occupies the valley between these two modes. Second, the absolute PPL range differs substantially across models—smaller models (3B/4B) generally show higher baseline perplexity on the same items—

Model	MathQA	CSQA	ARC-E	ARC-C	OBQA	LogiQA	Overall
Gemma-3-4B-it	25.8	62.5	89.6	78.0	71.3	46.7	42.4
Gemma-3-12B-it	31.7	74.3	96.1	90.2	84.2	57.1	50.6
Llama-3.2-3B-Inst.	32.3	66.7	88.3	74.4	73.8	42.2	46.1
Llama-3.1-8B-Inst.	32.5	71.3	91.5	80.3	77.4	45.3	47.9
Qwen3-4B	35.6	75.9	95.2	88.9	80.3	63.1	53.4
Qwen3-8B	35.6	78.3	96.6	91.4	84.2	65.1	54.5

Table 3: Top-1 accuracy (%) by model and benchmark on train splits. CSQA = CommonsenseQA; ARC-E/C = ARC-Easy / ARC-Challenge; OBQA = OpenBookQA.

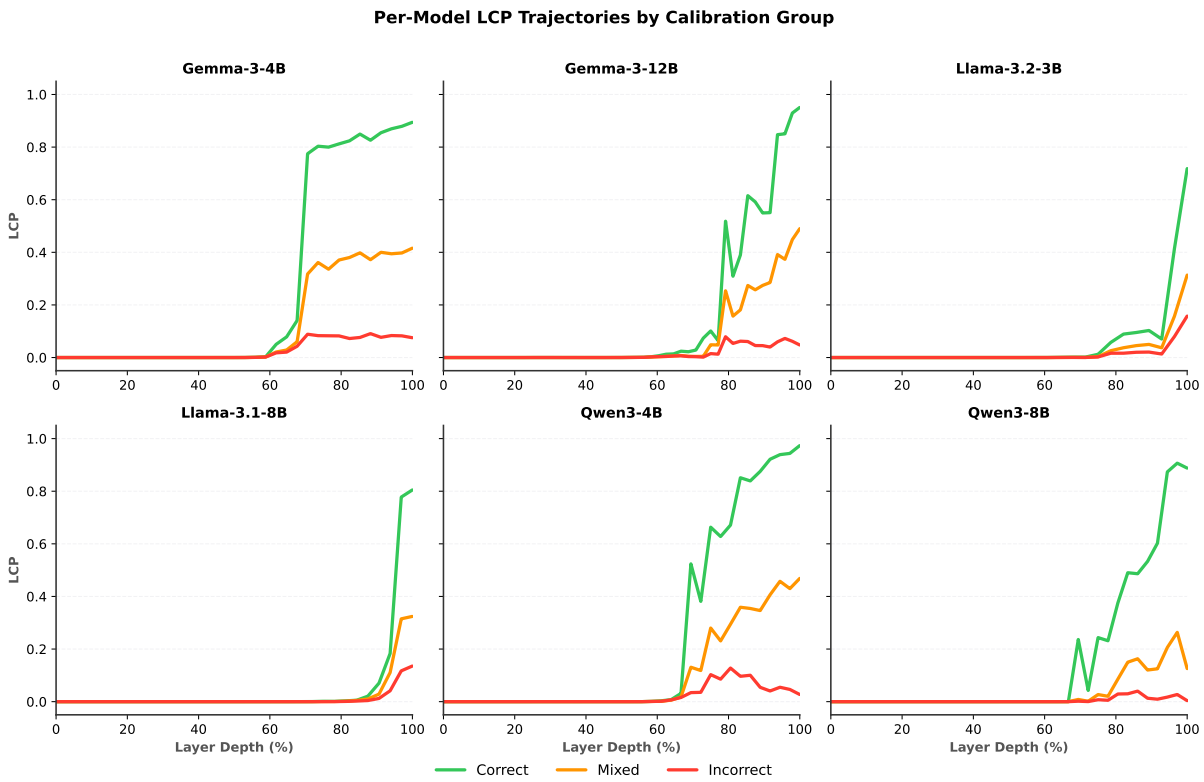

Figure 6: Per-model LCP trajectories for the three calibration groups (2×3 panel). Each panel is independent; the three-group ordering is reproduced across all six models.

but the *shape* of the tripartite division is reproduced in all six cases, justifying per-model percentile cutoffs rather than a shared global threshold. Third, accuracy tracks the PPL partition as expected: within the low-PPL region (Correct group) top-1 accuracy reaches 89–99% across models, whereas within the high-PPL region (Incorrect group) it drops to 1.8–7.5%, and the middle region (Mixed) spans 37–58%, close to the 25% random-chance baseline for four-option items.

What the per-model LCP gallery reveals. Figure 6 complements the PPL analysis: the qualitative LCP ordering (Correct > Mixed > Incorrect throughout depth) is reproduced in every model without exception, ruling out the possibility that the aggregate picture is driven by any single architecture or parameter scale. The within-group trajectory shape is also remarkably consistent: Correct items show a monotone rise; Incorrect items stay near zero; and Mixed items show the characteristic flat-then-rise profile that motivates the late-stage trajectory commitment analysis. Absolute LCP levels scale roughly with model accuracy (Table 1), so the Correct group in a more accurate model has a higher mean LCP_L than the same group in a less accurate model, but the *relative* gap between groups is preserved.

Group-organized trajectory view. Figure 6 above shows the per-model view (one panel per model, all three groups overlaid). Figure 8 shows the complementary *group-organized* view: one panel per calibration group, with individual model

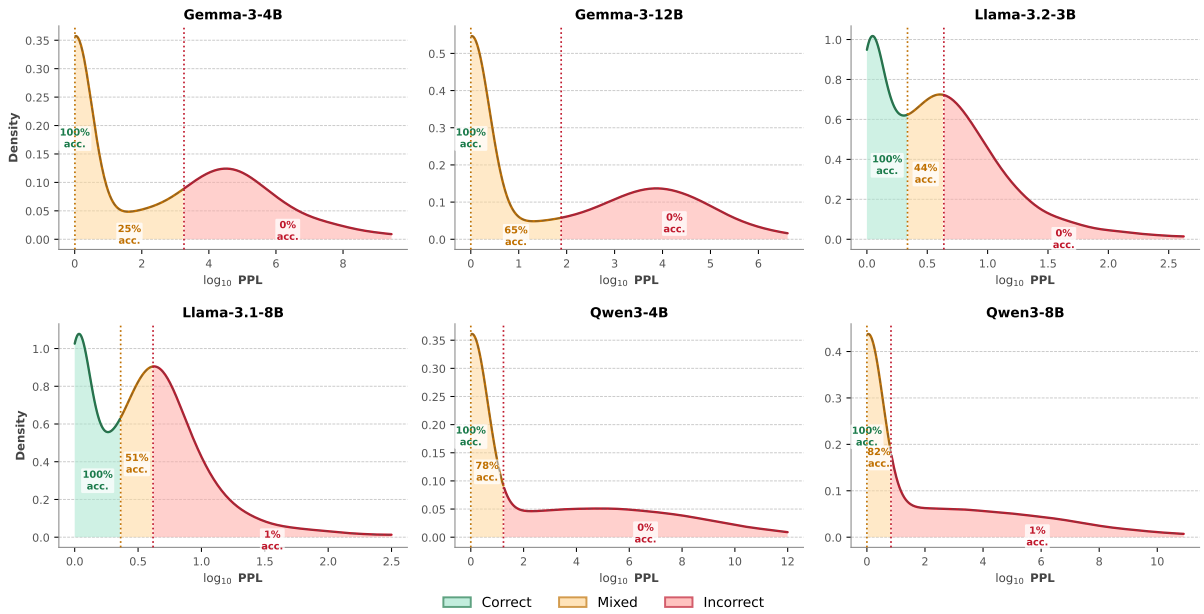

Figure 7: $\log_{10}$ PPL distributions for six models

lines and a cross-model mean band. This makes within-group model variation visible—the spread of individual lines around the mean band reflects model-to-model differences in absolute LCP level rather than profile shape. Figure 9 summarises the same information as exact final-layer LCP values, serving as a visual companion to Table 1.

C Benchmark-Level Group Shares and Cross-Model Consistency

Table 4 summarises group proportions averaged across models. Figure 5 (main text) shows the per-model stacked-bar breakdown; the heatmap (Figure 10) provides exact numerical read-off per Benchmark $\times$ Model cell.

Knowledge versus reasoning gradient. The six benchmarks form a near-monotone gradient from knowledge-dominated to reasoning-dominated group occupancy: ARC-Easy (91% Correct) $\rightarrow$ OpenBookQA (76%) $\rightarrow$ ARC-Challenge / CommonsenseQA ($\approx$70–79%) $\rightarrow$ LogiQA (38% Correct, 22% Mixed, 41% Incorrect) $\rightarrow$ MathQA (15% Correct, 29% Mixed, 56% Incorrect). This gradient aligns with the intuitive notion of *retrieval* versus *multi-step inference*: items the model resolves confidently (high accuracy, low PPL) tend to be factual recall tasks, while items that concentrate in the Mixed and Incorrect groups tend to require chained reasoning or symbolic manipulation.

Role of the Mixed group. The Mixed group is not merely a residual category. For LogiQA and MathQA, Mixed-group proportions reach 22% and 29% respectively—large enough to be analytically important. These are precisely the benchmarks where output-level PPL fails to separate gold-label correct from incorrect answers before the terminal layers, and where late-stage LCP divergence is most consequential. For ARC-Easy, by contrast, the Mixed group is only 3.1%: most items are resolvable at the output level, and the LCP profile is correspondingly cleaner.

Model-level consistency. Figure 10 shows that the benchmark ordering is not an artifact of model choice: for every benchmark, the three models from three families (Gemma-3, Llama-3, Qwen3) agree on the direction of the Correct / Mixed / Incorrect split, even when absolute proportions vary. The cross-model standard deviation on the Correct-group share is largest for CommonsenseQA ($\pm$9 pp) and smallest for ARC-Easy ($\pm$2 pp), consistent with the view that commonsense reasoning is more model-dependent than straightforward factual recall.

D Statistical Tests

D.1 Task type $\times$ group (χ^2)

For each model–item pair we assign a group (Correct / Mixed / Incorrect) and a task type: Knowledge (ARC-Easy, OpenBookQA, Com-

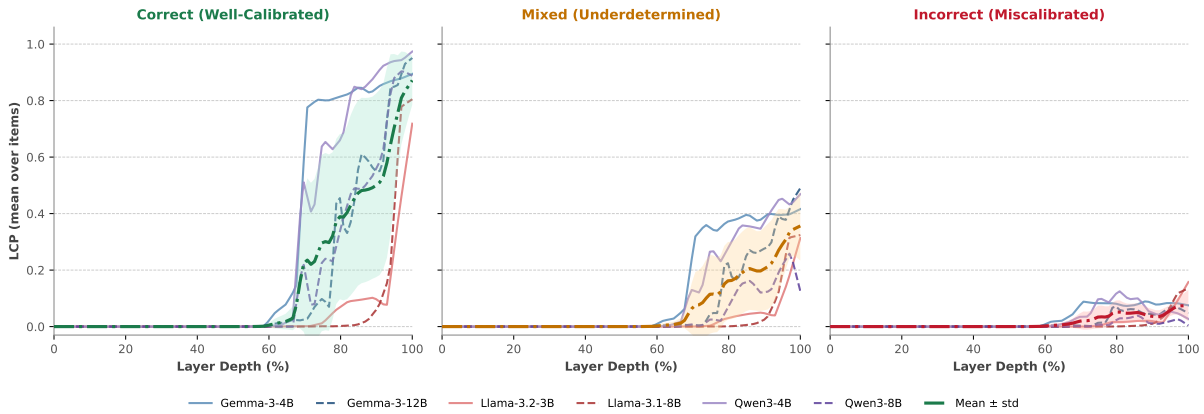

Figure 8: Per-model LCP trajectories organized by calibration group (1×3 panels). Solid lines = smaller models (3B/4B); dashed = larger (8B/12B). Dash-dotted line and shaded band = cross-model mean ± std. All six models follow the same qualitative shape within each group.

Figure 9: Final-layer mean LCP for each Model × Group combination. Visual companion to Table 1; colours follow a Red–Yellow–Green scale (0–1).

Benchmark	Correct	Mixed	Incorrect
ARC-Easy	91.0	3.1	5.9
OpenBookQA	75.8	8.4	15.8
CommonsenseQA	69.8	7.4	22.8
ARC-Challenge	78.9	6.9	14.2
LogiQA	37.8	21.6	40.6
MathQA	14.8	29.3	55.9

Table 4: Group share (%) by benchmark (six-model mean).

monsenseQA, ARC-Challenge) or Reasoning (MathQA, LogiQA). Pooling all six models yields $N = 324{,}360$ counts. Table 5 gives the contingency table; Pearson's $\chi^2 = 78{,}818$ ($p < 10^{-100}$). Given the large N, p-value alone is uninformative; the effect size is Cramér's $V = \sqrt{\chi^2/(N \cdot (k-1))} = \sqrt{78{,}818/(324{,}360 \cdot 2)} \approx 0.35$.

	Correct	Mixed	Incorrect
Knowledge	73,047	8,923	19,112
Reasoning	46,690	59,470	117,118

Table 5: Task type × group counts (N=324,360).

D.2 Late-stage trajectory commitment (Mixed group)

Trajectory split. For each model, we use LCP trajectories of items in the Mixed group. Items are split into two subsets for layer-wise comparison using a *final-layer LCP rank proxy*: sort all Mixed-group items by LCP_L, then mark the top k as "positive," where $k = \lfloor \text{Mixed-group accuracy} \times N \rceil$. This proxy is *not* the ground-truth top-1 correctness label; it aligns subset sizes with the observed accuracy, but defines membership by the final-layer LCP score itself. As a consequence, a substantial portion of the late-commitment signal is structurally built into the split: items ranked high by LCP_L will necessarily show higher LCP_l in the terminal layers. Sensitivity analyses splitting by ground-truth top-1 correctness are left for future work; results may differ, particularly for models with small Mixed cohorts (e.g., Gemma-3-12B N=365, Qwen3-4B N=765). This split is used consistently for Figure 3 and Figure 4.

Layer-wise test. At each layer l, we compare the two subsets' LCP_l distributions with a two-sided Mann–Whitney U test and compute Wasserstein distance W_l. Across layers we apply Benjamini–Hochberg FDR at $\alpha = 0.01$. The divergence-onset layer is the smallest l such that adjusted $p_l < 0.01$ and $W_l > 0.01$. Table 6 and Figure 11 report depth

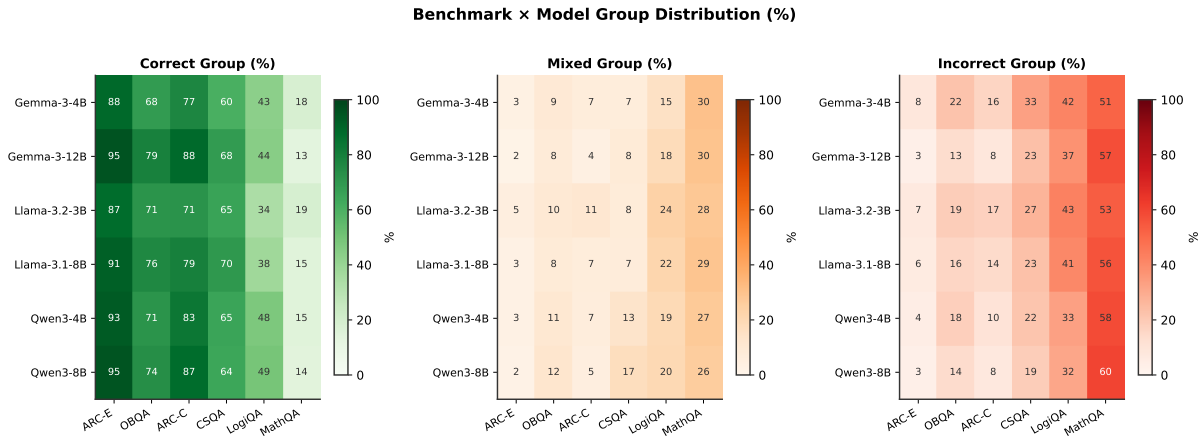

Figure 10: Group proportion (%) for each Benchmark × Model combination (heatmap). Color intensity encodes proportion within each group panel. Complements Figure 5 with exact numerical read-off per cell.

Model	L	Layer	Depth (%)	W_l
Gemma-3-4B-it	35	24	68.6	0.228
Gemma-3-12B-it	49	36	73.5	0.013
Llama-3.2-3B-Inst.	29	22	75.9	0.014
Llama-3.1-8B-Inst.	33	29	87.9	0.013
Qwen3-4B	37	26	70.3	0.053
Qwen3-8B	37	25	67.6	0.019

Table 6: First significant divergence-onset layer in the Mixed group (proxy split).

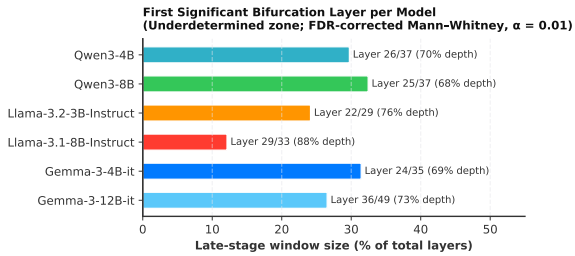

Figure 11: Divergence-onset depth (as % of total layers) for each of the six models, within the Mixed group. Onset is the first layer with BH-adjusted $p<0.01$ and Wasserstein distance >0.01 (Table 6) between high- and low-LCP-rank proxy trajectories. All models exhibit late onset (65–88% depth), confirming that trajectory divergence is a terminal-layer phenomenon rather than a gradual separation.

as a percentage of total layers L.

Per-model trajectory-commitment profiles. Figure 12 shows the proxy-split LCP trajectories for each model individually. A recurring visual pattern emerges: the two curves (high-LCP and low-LCP proxy subsets) run close together through most of depth and separate sharply near the final layers. Models with smaller Mixed cohorts (Gemma-3-12B N=365, Qwen3-4B N=765) show noisier trajectories, so the annotated commitment-onset depth (dashed line) should be treated with more caution than for models with larger Mixed groups. Despite this variability, *no* model shows a separation that begins before 65% depth.

D.3 Layer-wise AUROC (Mixed group)

For each model and layer l, we compute AUROC using LCP_l as a score and the final-layer LCP rank proxy as the binary label (1 = top-k by LCP_L; see Section D.2 for the proxy definition). AUROC is estimated via the Mann–Whitney U statistic: $\mathrm{AUROC} = U/(|P| \cdot |N|)$, where $|P|$ and $|N|$ are positive and negative subset sizes.

What AUROC near 0.5 means. An AUROC of 0.5 at layer l indicates that LCP_l is no better than random chance at ranking items by their final-layer LCP score. In other words, the ordering of items in the intermediate representation—as measured by the logit-lens probability of the gold token—has not yet settled into its terminal form. The practical implication is that no intermediate layer provides a reliable threshold for predicting whether a Mixed-group item will eventually favor the gold answer: the decision is genuinely unresolved until the last 12–32% of depth.

AUROC trajectory and its alignment with divergence onset. Figure 4 shows AUROC rising from near 0.5 to values between 0.75 and 0.95 in the terminal layers, with the inflection point broadly aligned with the divergence-onset depths reported in Table 6. This coherence between two independently defined statistics (Mann–Whitney

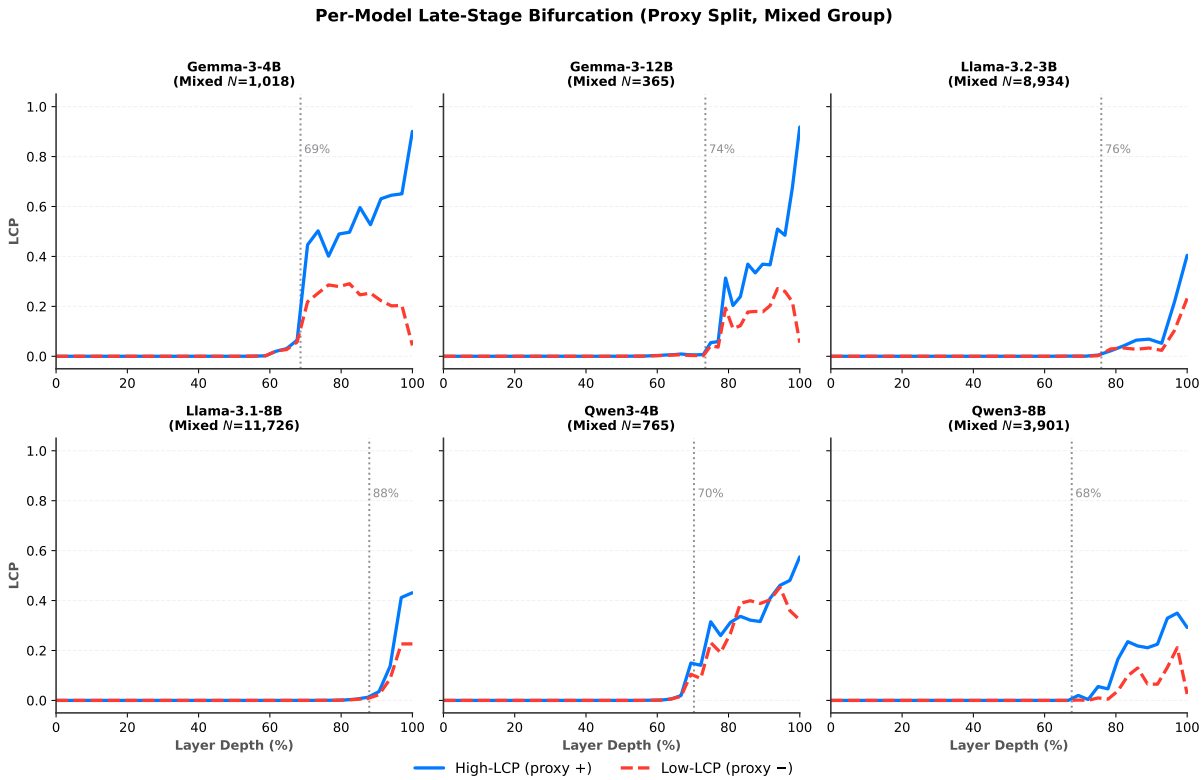

Figure 12: Per-model late-stage trajectory divergence (proxy split, Mixed group). Each panel shows high-LCP-rank (blue, solid) and low-LCP-rank (red, dashed) mean trajectories. Dashed vertical lines mark the first statistically significant divergence-onset layer (Table 6); N in each title is the Mixed-group size.

on raw LCP_l trajectories vs. rank-based AUROC) strengthens the inference that the late-layer transition is a genuine feature of the internal computation, rather than an artifact of any single analysis choice.

Relationship to ground-truth correctness. Because both the AUROC score and binary label are derived from the LCP trajectory itself (see Section D.2), the curve strictly measures self-consistency of the trajectory—how well earlier LCP predicts terminal LCP rank—not whether LCP predicts ground-truth top-1 accuracy. This is an important qualification: even high AUROC in the terminal layers does not guarantee that the model's final-layer LCP ranking matches human-labeled correctness. Sensitivity analyses using ground-truth labels are deferred to future work.

D.4 LCP convergence AUC (Knowledge vs. Reasoning)

A natural alternative explanation for the LCP pattern is that Knowledge and Reasoning items simply cause the network to converge to its final hidden state at different depths: earlier convergence would explain higher LCP at intermediate layers, independently of any late-layer commitment effect. We test this alternative by measuring *how quickly* each item's last-token hidden state approaches its terminal representation, rather than *what* that representation is.

Procedure. We sample 700 items per model stratified by benchmark tag (4,200 total per task type across six models), extract the last-token hidden-state sequence $\{h_l\}_{l=1}^{L}$, normalize each layer's vector to unit norm, and compute cosine similarity $s_l = \langle \hat{h}_l, \hat{h}_L \rangle$ to the final layer L. The convergence AUC is the trapezoidal integral of s_l over normalized relative depth $l/L \in [0, 1]$: higher AUC means the hidden state is closer to its final form earlier in the network.

Result. Pooling all six models, median convergence AUC is 0.144 (Knowledge) vs. 0.147 (Reasoning); Mann–Whitney $U = 1.7 \times 10^6$, $p < 10^{-100}$. Although the p-value is vanishingly small due to the large sample size, the absolute difference (0.003 AUC points) is negligible. Both task types show qualitatively identical convergence profiles: low early similarity to the final layer, with most representational change concentrated in later depth.

E Summary of Main-Text Tests

Table 7 condenses the tests cited in Section 4.

Interpretation. The near-identical convergence AUC across task types rules out the alternative explanation that Knowledge items simply "converge faster." Instead, both Knowledge and Reasoning items undergo most of their representational change in the terminal layers—but the terminal representation they converge *to* differs: Correct-group items converge to a state that assigns high probability to the gold token, while Mixed and Incorrect items converge to states that assign it lower probability. This dissociates *when* the model internal representations diverge (depth of convergence, which is task-type-independent) from *what* it converges to (the final LCP level, which is group- and task-type-dependent), and supports the framing of late-stage trajectory divergence as a terminal-layer *content* decision rather than a timing artefact.

F Internal Token Competition in the Mixed Group

Late-stage trajectory divergence could arise from two distinct dynamics: (a) the gold token's probability simply stays flat until the final layers, or (b) the gold token actively competes with a strong alternative option whose probability it must overtake. These interpretations carry different implications for future interventions.

Figure 13 shows both the gold-token LCP and the top-incorrect-option probability layer by layer, aggregated across all six models for Mixed-group items. Panel (A) shows that the two token probabilities track each other closely through most of depth—both low and nearly equal—before diverging in the terminal layers. Panel (B) shows the per-model margin (gold minus top-incorrect), which stays near zero or slightly negative through early depth and rises only in the final 15–30% of layers.

This *parallel tracking* pattern is more consistent with interpretation (b)—internal competition—than with (a): the gold answer is not absent from early processing but is matched by a competing option that the model only resolves in the final layers. Future causal interventions (e.g., activation steering at a mid-layer) could test whether artificially boosting the margin before the natural divergence-onset point changes output correctness.

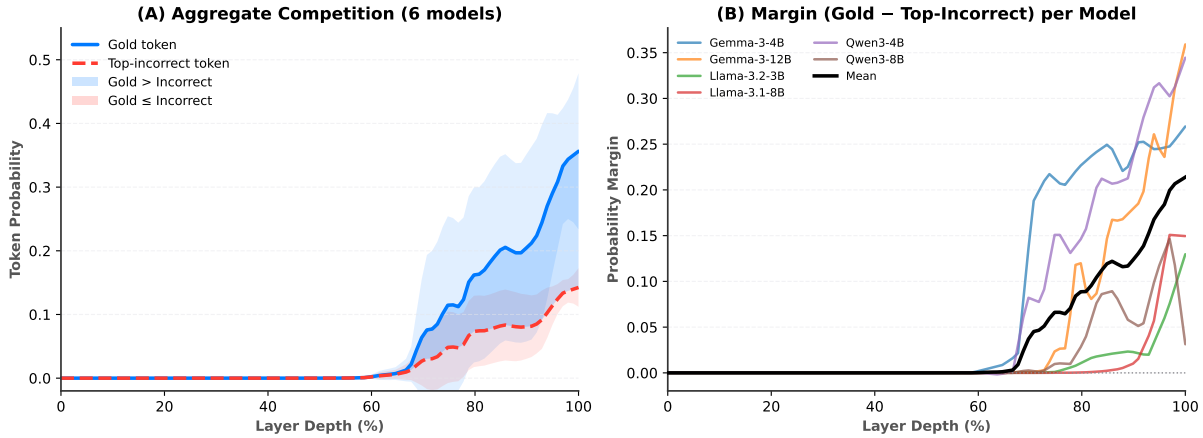

Figure 13: Internal token competition in the Mixed group. (A) Aggregated gold-token LCP vs. top-incorrect-option probability (shaded: cross-model std); filled regions highlight where gold leads (blue) or trails (red). (B) Per-model margin (gold − top-incorrect) across depth; black line is the six-model mean. All six models show the margin rising only in the terminal layers.

Test	Result
Task type × group (χ^2, N=324,360)	78,818, $p < 10^{-100}$
LCP convergence (Mann–Whitney U)	med. AUC 0.144 vs. 0.147, $p < 10^{-100}$
Divergence-onset depth (Mixed, BH FDR)	
Gemma-3-4B / 12B	68.6% / 73.5%
Llama-3.2-3B / 3.1-8B	75.9% / 87.9%
Qwen3-4B / 8B	70.3% / 67.6%

Table 7: Summary of statistical tests (six models).