
Towards Characterizing Question Format Bottlenecks for Fine-Tuned Small Language Models

Paolo Cudrano, Matteo Matteucci

Department of Electronics Information and Bioengineering (DEIB)
Politecnico di Milano
Milano, Italy
`{name.surname}@polimi.it`

Abstract

Small language models (SLMs) are commonly specialized to a task or domain before deployment, yet the success of that specialization depends partly on how the learned knowledge is queried. We study the performance disparity between binary and multiple-choice (MC) formats after format-specific supervised fine-tuning (SFT), evaluating 15 models from 58M to 3B parameters on binary and multiple-choice QA. On WikiDoc-BMCQA, whose binary and MC questions are generated from the same medical source corpus, closed-book binary verification remains close to chance across much of the SLM range, while four-way MC is learned substantially more strongly relative to its chance baseline. The same qualitative asymmetry appears on public benchmarks, where MC is learned strongly on SciQ and MedMCQA while StrategyQA and the Mintaka yes/no subset show smaller gains over their respective baselines. Supplying the relevant source passage raises binary accuracy to 0.79–0.89 on WikiDoc, showing that verification itself is not intrinsically too hard at this scale when evidence is available. Preliminary diagnostics find that a global yes/no threshold correction yields negligible improvement, while a last-token linear probe exposes no more answer signal than the model’s own output margin. Whether the remaining closed-book bottleneck lies in acquisition, retrieval, or the decision learned by SFT remains open.

1 Introduction

Small language models are attractive when memory, latency, privacy, or deployment constraints make larger models impractical. Their limited capacity, however, changes how they are commonly used, making adaptation to narrow tasks or domains more natural than expecting them to be broadly knowledgeable out of the box. Their evaluation presents a related difficulty. Language models are commonly assessed with closed-response benchmarks because they are simple and reliable to score, despite being susceptible to shortcuts and other artifacts introduced by the response format (Chandak et al., 2025). Open-ended generation can provide a more faithful assessment by requiring the model to produce the answer without supplied candidates, but for models at the scale we study, free-form text generation can be too unreliable for such evaluation to be informative. Closed-response benchmarks are therefore particularly important in this regime, making it crucial to understand how their format affects the capabilities they appear to measure.

Language-model evaluation is well known to be sensitive to prompt and option representation (Sclar et al., 2024; Zheng et al., 2024). Controlled reformulations across MC, true/false, and other answer formats likewise show substantial format effects (Oh et al., 2024; Zhou et al., 2024b; Song et al., 2025), though without a uniform ordering across formats. Additionally, MC formulations expose the candidate answers, enabling elimination and option-side shortcuts (Balepur et al., 2024).

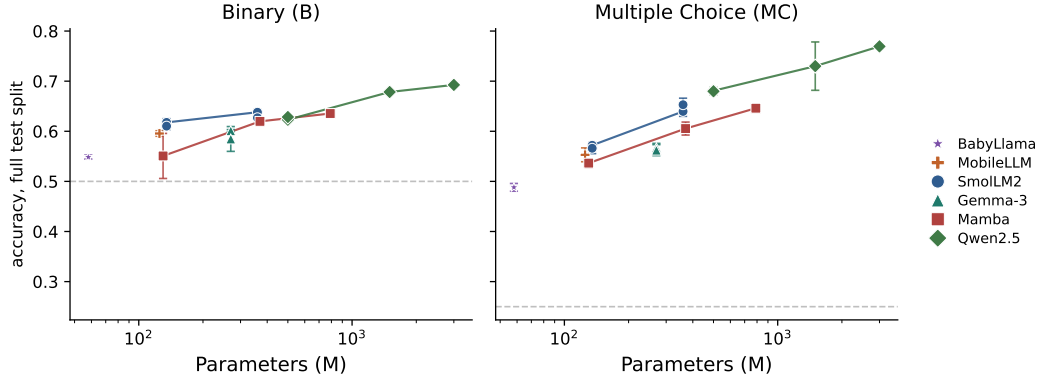


Figure 1: Full WikiDoc-BMCQA test split accuracy after format-specific SFT, binary (left) and four-way multiple choice (right). Dashed lines mark the chance rate of each format. Markers and bars are model means $\pm$ standard deviation over five SFT seeds. Models of the same family share marker and color. Base variants are joined by a line; instruction-tuned variants are overlapped. Although the two panels sit at similar accuracies, the improvement over chance is substantially larger in MC.

Relatedly, Chandak et al. (2025) formalize the distinction between discrimination and per-candidate verification, showing that verification is strictly harder in an oracle formulation. Recent work also shows that multi-format SFT can improve cross-format robustness (Liu et al., 2026). Most of this evidence, however, concerns large models or inference-time prompting, leaving unclear whether format dependence persists after target-format SFT in small models, exactly where fixed-response evaluation is particularly difficult to avoid.

A preliminary non-archival study introduced WikiDoc-BMCQA and observed a large difference between binary and MC QA performance after SFT around the 100M-parameter scale (Cudrano et al., 2026). WikiDoc-BMCQA contains balanced yes/no questions (WikiDoc-B) and four-way MC questions (WikiDoc-MC) generated from the same medical source corpus. The two formats were generated independently and do not in general ask the same atomic facts. We substantially extend that study to 15 models from six model families (BabyLlama, MobileLLM, SmolLM2, Gemma-3, Qwen2.5, Mamba), spanning Transformer and state-space architectures, four instruction-tuned variants, six benchmarks, five SFT seeds, and several controls.

Our results show that the contrast persists well beyond the original 100M-scale observation. Binary SFT produces only modest gains over a 0.5 chance baseline for many models, whereas MC is learned much more strongly relative to its chance level. This pattern is not simply removed by instruction tuning or by moving to larger models within the range we study. We finally ask what could explain this behavior, and report preliminary diagnostics testing two simple explanations. Identifying the full mechanism is left to future work.

Our main **contributions** are:

1. **A wider characterization of the binary–MC contrast after format-specific SFT.** Extending Cudrano et al. (2026), we evaluate 15 models from 58M to 3B on WikiDoc-BMCQA, spanning six pretraining model families, Transformer and state-space architectures, and four instruction-tuned variants. Four public benchmarks provide complementary evidence on an eight-model subset spanning 58M–1.5B (Section 3 and Section B).
2. **Preliminary analysis and diagnostics** towards a characterization of the phenomenon. A validation-fitted global decision threshold recovers negligible accuracy, and a linear probe on hidden states never exceeds the model’s own output margin (Section 5).

2 Experimental setup

Models. We evaluate 15 models from 58M to 3B parameters: BabyLlama 58M (Timiryasov & Tastet, 2023); MobileLLM 125M (Liu et al., 2024); SmolLM2 135M and 360M (Allal et al., 2025), with their instruction-tuned variants; Gemma-3 270M and its instruction-tuned variant (Team et al.,

2025); Qwen2.5 0.5B (base and instruct), 1.5B, and 3B (Yang et al., 2024); and Mamba 130M, 370M, and 790M (Gu & Dao, 2023). The models span transformer and selective state-space architectures. As their pretraining data differ by orders of magnitude, we avoid direct comparisons between model families.

Benchmarks. WikiDoc-BMCQA (Cudrano et al., 2026) is derived from WikiDoc medical text (Han et al., 2023). WikiDoc-B contains balanced yes/no questions, while WikiDoc-MC contains four-way MC questions generated from the same source passages. Following and extending Cudrano et al. (2026), we also evaluate two public binary benchmarks, the yes/no subset of Mintaka (Sen et al., 2022) and StrategyQA (Geva et al., 2021), and two public MC benchmarks, SciQ (Welbl et al., 2017) and MedMCQA (Pal et al., 2022). WikiDoc and SciQ additionally provide supporting text for open-book evaluation. We report performance on the full test split of each evaluation setting.

Fine-tuning and evaluation. Each model is fine-tuned separately for each benchmark and format. For comparability, instruction-tuned variants use the same benchmark- and format-specific prompts as the base checkpoints during both SFT and evaluation; we do not apply their native chat templates. Training optimizes the answer-token loss only. We select one learning rate per model on a validation split of WikiDoc-B and freeze it across the remaining datasets and conditions, and train with early stopping on validation accuracy. Unless otherwise stated, results are mean $\pm$ standard deviation over five independent SFT seeds. We evaluate following Gao et al. (2024), ending the prompt with `Answer:` and comparing the logits of the allowed answer tokens (yes/no or A–D).

3 Binary verification improves more slowly than multiple choice

Figure 1 shows the main WikiDoc-BMCQA result on its test split. Binary performance generally improves with model size, but it remains close to its 0.5 chance baseline over much of the small-model range. The corresponding four-way MC trainings instead produce a much larger improvement above their 0.25 baselines. For example, MobileLLM-125M reaches 0.597 ± 0.007 and Mamba-130M 0.551 ± 0.045 on WikiDoc-B, while they reach 0.553 ± 0.008 and 0.536 ± 0.009 respectively in 4-way MC. This separation remains visible throughout the studied range, even though overall performance improves with model size in both formats. Full results are reported in Section A.

Unpaired public benchmarks show the same broad asymmetry in learnability. Such benchmarks are not constructed from the same sources and should not be interpreted as a controlled replication, but they show that binary SFT remains comparatively brittle while MC SFT stays readily learnable in other domains as well. Full public-benchmark results are reported in Section B.

We confirm that the different chance levels (0.5 vs 0.25) do not explain this separation between B and MC by experimenting with a two-choice version of WikiDoc-MC (Section E).

4 Relevant context largely removes the binary difficulty

We investigate whether binary verification is inherently too hard for small models. We provide the models with explicit relevant context for making these yes/no decisions, thanks to the source passage present in WikiDoc-BMCQA and SciQ.

As shown in Figure 2, models fine-tuned in the open-book condition on WikiDoc-B reach 0.792–0.891 when they can access the related source passage from context at test-time. Providing an unrelated passage from the same corpus or no passage removes most of the gain, indicating that the improvement depends on relevant test-time context rather than merely on the open-book fine-tuning setup. Additional open-book results are given in Section C.

We conclude that small models can read relevant evidence and map it to the binary answer when the evidence is supplied, excluding that binary verification is intrinsically too hard for models of this scale.

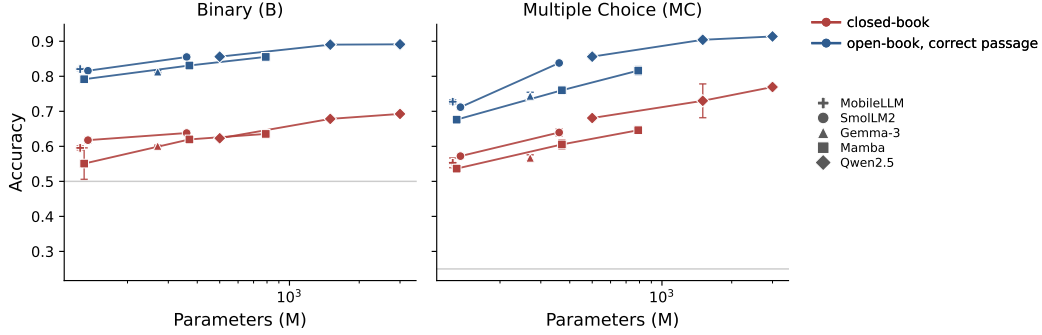


Figure 2: Open-book performance on binary (left) and multiple choice (right): supplying the passage (blue) lifts binary accuracy far above the closed-book level (red). Markers and bars represent model means $\pm$ standard deviation over five SFT seeds. Families are joined as in Figure 1.

5 Simple readout explanations do not resolve the gap

Global decision threshold. Closed-response language-model evaluation can exhibit systematic preferences for particular answer labels, and several calibration methods correct these preferences through global label-score adjustments (Zhao et al., 2021; Zhou et al., 2024a; Bhatt & Ivanova, 2026). For a WikiDoc-B question x , define the output margin as the logits difference $m(x) = \ell_{\text{yes}}(x) - \ell_{\text{no}}(x)$. In the binary case, any fixed additive correction to the two logit scores is equivalent to replacing the default decision rule $m(x) > 0$ with $m(x) > \tau$. We therefore choose τ separately for each checkpoint to maximize its validation accuracy and apply it unchanged on the test data. The resulting change in accuracy is essentially null across models (maximum per-model mean gain $+0.011$), indicating that a single global yes/no score offset does not account for the closed-book binary deficit. Full results are in Section D.

Preliminary representation probe. We test whether more answer information is linearly available in the hidden states than is already expressed by the model’s own output margin. For four representative models (58M, 270M, 0.5B, and 3B; one seed), we train logistic-regression probes on final-token hidden states, selecting layer and regularization on validation. The probe becomes predictive in the larger models, but never exceeds the model’s own output-margin AUROC; the probe-minus-margin differences range from -0.024 to -0.003 . Fine-tuning does increase linear decodability relative to the pretrained base model, but this experiment does not reveal a large additional pool of answer information hidden behind the final output. Given the small number of models and the single seed, we treat this as a preliminary result worth further investigation; per-model values are in Section G.

6 Discussion and Limitations

Our results place the known verification–discrimination distinction in the operational regime of small models. After task-specific SFT, we observe a persistent separation between closed-book binary verification and MC QA across the models studied. Across 58M–3B models, closed-book binary verification improves only gradually above its balanced baseline, whereas MC QA is learned much more strongly relative to its baseline. Yet the binary task itself is not intrinsically out of reach, as supplying the relevant source passage raises accuracy to 0.79–0.89 and removes much of the small-model bottleneck. A global answer threshold and a preliminary linear readout do not explain the remaining closed-book difficulty.

These results leave at least three possible bottlenecks: the relevant information may never have been acquired; it may be stored but difficult to retrieve in response to a standalone statement; or binary SFT may learn a poorer decision function than candidate-ranking supervision. The present experiments cannot cleanly separate these explanations.

We cannot exclude data biases in WikiDoc-BMCQA or in the public benchmarks as confounders. Multiple choice inherently supplies the model with additional information; its items can carry option-level artifacts, which we quantify in Section F; and none of the binary and MC benchmarks we use are matched fact by fact.

The useful empirical observation is instead that, across the benchmarks and model families studied, candidate-selection QA is learned substantially more strongly under format-specific SFT than standalone binary verification, while supplying the source evidence largely removes the binary difficulty.

References

- Loubna Ben Allal, Anton Lozhkov, Elie Bakouch, Gabriel Martín Blázquez, Guilherme Penedo, Lewis Tunstall, Andrés Marafioti, Hynek Kydlíček, Agustín Piqueres Lajarín, Vaibhav Srivastav, et al. SmoLLM2: When smol goes big – data-centric training of a small language model. *arXiv preprint arXiv:2502.02737*, 2025.
- Nishant Balepur, Abhilasha Ravichander, and Rachel Rudinger. Artifacts or abduction: How do llms answer multiple-choice questions without the question? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10308–10330, 2024.
- Om Bhatt and Anna A. Ivanova. RBCorr: Response bias correction in language models. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 540–553. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.gem-main.51.
- Nikhil Chandak, Shashwat Goel, Ameya Prabhu, Moritz Hardt, and Jonas Geiping. Answer matching outperforms multiple choice for language model evaluation. *arXiv preprint arXiv:2507.02856*, 2025. URL <https://arxiv.org/abs/2507.02856>.
- Paolo Cudrano, Greta Corti, Sara Pidò, Valerio Ceccarelli, and Matteo Matteucci. Understanding without knowing: Format and context impact on 100M-parameter domain QA. COLM 2026 Workshop on Methods and Opportunities at Small Scale (MOSS), 2026.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac’h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 2024. URL <https://zenodo.org/records/12608602>.
- Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics (TACL)*, 9, 2021. URL <https://arxiv.org/abs/2101.02235>.
- Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.
- Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel R. Bowman, and Noah A. Smith. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2018.
- Tianyu Han, Lisa C Adams, Jens-Michalis Papaioannou, Paul Grundmann, Tom Oberhauser, Alexei Figuerola, Alexander Löser, Daniel Truhn, and Keno K Bressen. MedAlpaca—an open-source collection of medical conversational ai models and training data. *arXiv preprint arXiv:2304.08247*, 2023.
- June M. Liu, Shaomian Zheng, He Cao, Dingnan Jin, Qing Cui, and Jun Zhou. Improving cross-format robustness in language models with multi-format training. *arXiv preprint arXiv:2606.11643*, 2026.
- Zechun Liu, Changsheng Zhao, Forrest Iandola, Chen Lai, Yuandong Tian, Igor Fedorov, Yunyang Xiong, Ernie Chang, Yangyang Shi, Raghuraman Krishnamoorthi, Liangzhen Lai, and Vikas Chandra. MobileLLM: Optimizing sub-billion parameter language models for on-device use cases. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32431–32454. PMLR, 21–27 Jul 2024.

- Jio Oh, Soyeon Kim, Junseok Seo, Jindong Wang, Ruochen Xu, Xing Xie, and Steven Euijong Whang. ERBench: An entity-relationship based automatically verifiable hallucination benchmark for large language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024.
- Ankit Pal, Logesh Kumar Umapathi, and Malaikannan Sankarasubbu. MedMCQA: A large-scale multi-subject multi-choice dataset for medical domain question answering. In *Conference on health, inference, and learning*, pp. 248–260. PMLR, 2022.
- Adam Poliak, Jason Naradowsky, Aparajita Haldar, Rachel Rudinger, and Benjamin Van Durme. Hypothesis only baselines in natural language inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pp. 180–191, New Orleans, Louisiana, June 2018. Association for Computational Linguistics. doi: 10.18653/v1/S18-2023. URL <https://aclanthology.org/S18-2023/>.
- Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models’ sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2310.11324>.
- Priyanka Sen, Alham Fikri Aji, and Amir Saffari. Mintaka: A complex, natural, and multilingual dataset for end-to-end question answering. In *Proceedings of the 29th International Conference on Computational Linguistics*, pp. 1604–1619, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics. URL <https://aclanthology.org/2022.coling-1.138/>.
- Seok Hwan Song, Mohna Chakraborty, Qi Li, and Wallapak Tavanapong. Is large language model performance on reasoning tasks impacted by different ways questions are asked? In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22066–22081, 2025. doi: 10.18653/v1/2025.findings-acl.1138.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivi re, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.
- Inar Timiryasov and Jean-Loup Tastet. Baby Llama: Knowledge distillation from an ensemble of teachers trained on a small dataset with no performance penalty. In *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning (CoNLL)*. Association for Computational Linguistics, 2023. URL <https://aclanthology.org/2023.conll-baby1m.24/>.
- Johannes Welbl, Nelson F Liu, and Matt Gardner. Crowdsourcing multiple choice science questions. In *Proceedings of the 3rd Workshop on Noisy User-generated Text*, pp. 94–106, 2017.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Tang, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, and Zihan Qiu. Qwen2.5 technical report, 2024. URL <https://arxiv.org/abs/2412.15115>.
- Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 12697–12706. PMLR, 2021.
- Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2309.03882>.
- Han Zhou, Xingchen Wan, Lev Prolee, Diana Mincu, Jilin Chen, Katherine A. Heller, and Subhrajit Roy. Batch calibration: Rethinking calibration for in-context learning and prompt engineering. In *The Twelfth International Conference on Learning Representations*, 2024a.

Wenjie Zhou, Qiang Wang, Mingzhou Xu, Ming Chen, and Xiangyu Duan. Revisiting the self-consistency challenges in multi-choice question formats for large language model evaluation. In *Proceedings of the 2024 joint international conference on computational linguistics, language resources and evaluation (LREC-COLING 2024)*, pp. 14103–14110, 2024b.

A WikiDoc-BMCQA results

Table 1 reports every WikiDoc-BMCQA result behind Figure 1. As chance rates are different, the Δ columns highlight the accuracy gains over the respective chance baselines.

Table 1: Full-test WikiDoc-BMCQA performance after format-specific SFT (mean $\pm$ std over five seeds). Δ is the gain over the task’s chance level, making the different binary (0.50) and four-way MC (0.25) baselines explicit.

Model	Binary		4-way Multiple Choice	
	Accuracy	Δ vs. 0.50	Accuracy	Δ vs. 0.25
BabyLlama-58M	0.548 $\pm$ 0.004	+0.048	0.488 $\pm$ 0.008	+0.238
MobileLLM-125M	0.597 $\pm$ 0.007	+0.097	0.553 $\pm$ 0.008	+0.303
Mamba-130M	0.551 $\pm$ 0.045	+0.051	0.536 $\pm$ 0.009	+0.286
SmolLM2-135M	0.618 $\pm$ 0.007	+0.118	0.572 $\pm$ 0.006	+0.322
SmolLM2-135M-it	0.610 $\pm$ 0.007	+0.110	0.566 $\pm$ 0.011	+0.316
Gemma-3-270M	0.601 $\pm$ 0.002	+0.101	0.567 $\pm$ 0.009	+0.317
Gemma-3-270M-it	0.585 $\pm$ 0.025	+0.085	0.562 $\pm$ 0.011	+0.312
SmolLM2-360M	0.638 $\pm$ 0.005	+0.138	0.640 $\pm$ 0.010	+0.390
SmolLM2-360M-it	0.627 $\pm$ 0.003	+0.127	0.653 $\pm$ 0.013	+0.403
Mamba-370M	0.620 $\pm$ 0.004	+0.120	0.605 $\pm$ 0.013	+0.355
Qwen2.5-0.5B	0.623 $\pm$ 0.004	+0.123	0.681 $\pm$ 0.007	+0.431
Qwen2.5-0.5B-it	0.628 $\pm$ 0.004	+0.128	0.680 $\pm$ 0.007	+0.430
Mamba-790M	0.635 $\pm$ 0.004	+0.135	0.646 $\pm$ 0.005	+0.396
Qwen2.5-1.5B	0.677 $\pm$ 0.002	+0.177	0.751 $\pm$ 0.003	+0.501
Qwen2.5-3B	0.693 $\pm$ 0.004	+0.193	0.769 $\pm$ 0.005	+0.519

B Full benchmark results

We evaluate a subset of eight models spanning 58M–1.5B on four public benchmarks and report in Table 2 the measured closed-book accuracy. On the full test split, WikiDoc-B predictions remain roughly balanced across yes/no, whereas several Mintaka models strongly favor yes and StrategyQA often favors no. For this reason the diagnostics in the main text use WikiDoc-B rather than pooling all binary datasets together.

Table 2: Full-test closed-book accuracy, mean $\pm$ std over five SFT seeds. Baselines are majority accuracy for the binary datasets and random-choice accuracy for the MC datasets. These benchmarks are not paired across formats.

Model	Mintaka yes/no (B)	StrategyQA (B)	SciQ (MC)	MedMCQA (MC)
BabyLlama-58M	0.582 $\pm$ 0.022	0.546 $\pm$ 0.018	0.411 $\pm$ 0.005	0.320 $\pm$ 0.004
MobileLLM-125M	0.568 $\pm$ 0.039	0.592 $\pm$ 0.021	0.661 $\pm$ 0.012	0.375 $\pm$ 0.005
Mamba-130M	0.591 $\pm$ 0.005	0.555 $\pm$ 0.015	0.678 $\pm$ 0.009	0.367 $\pm$ 0.008
Gemma-3-270M	0.588 $\pm$ 0.012	0.615 $\pm$ 0.009	0.767 $\pm$ 0.014	0.399 $\pm$ 0.008
SmolLM2-360M	0.624 $\pm$ 0.015	0.640 $\pm$ 0.024	0.869 $\pm$ 0.003	0.467 $\pm$ 0.005
Qwen2.5-0.5B	0.598 $\pm$ 0.009	0.629 $\pm$ 0.014	0.839 $\pm$ 0.006	0.459 $\pm$ 0.006
Mamba-790M	0.608 $\pm$ 0.010	0.609 $\pm$ 0.009	0.839 $\pm$ 0.007	0.489 $\pm$ 0.007
Qwen2.5-1.5B	0.686 $\pm$ 0.003	0.675 $\pm$ 0.013	0.860 $\pm$ 0.037	0.501 $\pm$ 0.007
Baseline	0.595	0.515	0.25	0.25

C Open-book results

Tables 3 and 4 report the open-book results on WikiDoc-B and SciQ, respectively. For WikiDoc-B, each open-book-trained checkpoint is evaluated three ways on the same test items: with its source passage, an unrelated passage from the same corpus, or no passage. Differences within a row therefore reflect the supplied context rather than a change of model or item set. BabyLlama cannot be evaluated in this condition because its 256-token context cannot hold a WikiDoc passage, and

the instruction-tuned variants were not run open-book. Figure 3 displays the same controls for both WikiDoc-B and WikiDoc-MC. Removing the passage or replacing it with an unrelated passage largely returns performance toward the closed-book level.

Table 3: WikiDoc-B open-book on the full test split. One open-book-trained checkpoint is evaluated with its associated source passage, an unrelated passage, or no passage. Mean accuracy $\pm$ standard deviation over five seeds.

Model	Source passage	Wrong passage	No passage
MobileLLM-125M	0.821 $\pm$ 0.004	0.543 $\pm$ 0.008	0.545 $\pm$ 0.010
Mamba-130M	0.792 $\pm$ 0.006	0.546 $\pm$ 0.007	0.532 $\pm$ 0.011
SmolLM2-135M	0.816 $\pm$ 0.003	0.542 $\pm$ 0.014	0.532 $\pm$ 0.011
Gemma-3-270M	0.813 $\pm$ 0.003	0.545 $\pm$ 0.005	0.525 $\pm$ 0.003
SmolLM2-360M	0.855 $\pm$ 0.006	0.560 $\pm$ 0.016	0.574 $\pm$ 0.010
Mamba-370M	0.831 $\pm$ 0.004	0.550 $\pm$ 0.013	0.556 $\pm$ 0.021
Qwen2.5-0.5B	0.856 $\pm$ 0.004	0.561 $\pm$ 0.008	0.558 $\pm$ 0.002
Mamba-790M	0.855 $\pm$ 0.002	0.570 $\pm$ 0.010	0.599 $\pm$ 0.012
Qwen2.5-1.5B	0.890 $\pm$ 0.003	0.625 $\pm$ 0.010	0.629 $\pm$ 0.005
Qwen2.5-3B	0.891 $\pm$ 0.001	0.645 $\pm$ 0.008	0.649 $\pm$ 0.006

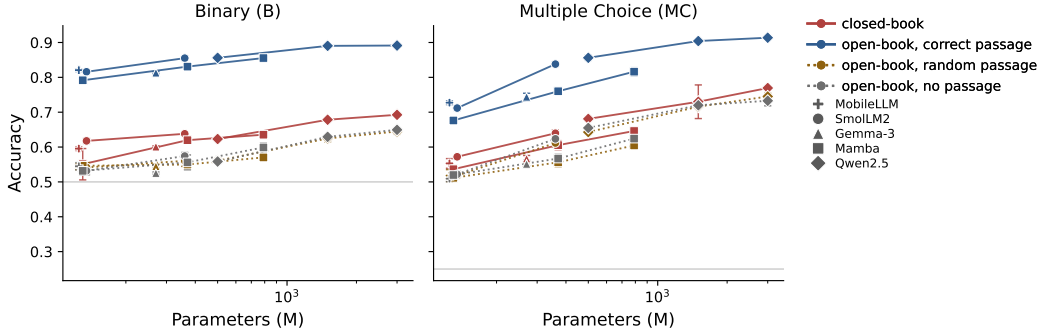


Figure 3: WikiDoc-BMCQA open-book controls. Open-book-trained checkpoints are evaluated with the item’s source passage, an unrelated passage from the same corpus, or no passage; closed-book checkpoints are shown for comparison. Markers and bars denote mean $\pm$ standard deviation over five SFT seeds. The open-book gain largely disappears when relevant context is removed or replaced.

SciQ is the only other benchmark in our set that provides support passages. As shown in Table 4, supplying the passage raises accuracy from 0.411–0.869 closed-book to 0.831–0.965 open-book across the eight models tested. The gain ranges from +0.093 to +0.419.

Table 4: SciQ closed- and open-book accuracy on the full test split. The open-book condition supplies the benchmark’s support passage. Mean accuracy $\pm$ standard deviation over five seeds; Δ is open-book minus closed-book accuracy. BabyLlama is included because SciQ support passages fit within its 256-token context.

Model	Closed-book	Open-book	Δ
BabyLlama-58M	0.411 $\pm$ 0.005	0.831 $\pm$ 0.007	+0.419
MobileLLM-125M	0.661 $\pm$ 0.012	0.913 $\pm$ 0.002	+0.253
Mamba-130M	0.678 $\pm$ 0.009	0.910 $\pm$ 0.007	+0.232
Gemma-3-270M	0.767 $\pm$ 0.014	0.937 $\pm$ 0.004	+0.170
SmolLM2-360M	0.869 $\pm$ 0.003	0.962 $\pm$ 0.004	+0.093
Qwen2.5-0.5B	0.839 $\pm$ 0.006	0.965 $\pm$ 0.001	+0.125
Mamba-790M	0.839 $\pm$ 0.007	0.955 $\pm$ 0.003	+0.116
Qwen2.5-1.5B	0.860 $\pm$ 0.037	0.960 $\pm$ 0.003	+0.100

D Decision-threshold analysis

For each WikiDoc-B checkpoint, we choose a scalar threshold τ^* that maximizes validation accuracy and apply it unchanged to test. This global label-score correction is equivalent to changing the threshold on $m(x) = \ell_{\text{yes}}(x) - \ell_{\text{no}}(x)$. We use this post-hoc decision threshold as a diagnostic, testing whether the observed binary weakness can be explained by a constant offset between the yes and no scores.

The mean gain across models is approximately zero, and the largest per-model mean improvement is +0.011 (Mamba-130M). AUROC, which is invariant to the choice of threshold, ranges from 0.565 to 0.768 and generally increases with scale within model families. Thus the output margin contains increasing ranking signal, but the low binary accuracy is not explained by a misplaced global decision threshold.

Table 5: WikiDoc-B decision margin, closed-book, mean $\pm$ std over five seeds. AUROC is threshold-free, while τ^* is fitted on validation and applied unchanged to test. The threshold correction changes test accuracy negligibly, showing that the closed-book deficit is not explained by a single global yes/no score offset.

Model	AUROC	Acc. at 0	Acc. at τ^*	Gain
BabyLlama-58M	0.565 $\pm$ 0.006	0.548	0.547	−0.001
MobileLLM-125M	0.630 $\pm$ 0.005	0.596	0.592	−0.004
Mamba-130M	0.590 $\pm$ 0.050	0.551	0.562	+0.011
SmolLM2-135M	0.662 $\pm$ 0.001	0.618	0.618	+0.000
SmolLM2-135M-it	0.649 $\pm$ 0.004	0.610	0.610	−0.000
Gemma-3-270M	0.640 $\pm$ 0.004	0.601	0.602	+0.001
Gemma-3-270M-it	0.616 $\pm$ 0.032	0.585	0.587	+0.002
SmolLM2-360M	0.693 $\pm$ 0.010	0.638	0.636	−0.002
SmolLM2-360M-it	0.679 $\pm$ 0.007	0.627	0.627	+0.000
Mamba-370M	0.669 $\pm$ 0.007	0.620	0.617	−0.003
Qwen2.5-0.5B	0.673 $\pm$ 0.007	0.623	0.622	−0.001
Qwen2.5-0.5B-it	0.677 $\pm$ 0.004	0.628	0.625	−0.004
Mamba-790M	0.692 $\pm$ 0.006	0.635	0.639	+0.004
Qwen2.5-1.5B	0.746 $\pm$ 0.004	0.678	0.679	+0.001
Qwen2.5-3B	0.768 $\pm$ 0.003	0.693	0.690	−0.003

E Controlling for MC chance-level

To test whether four-way MC appears easier only because its guessing rate is 0.25, we construct a two-choice WikiDoc-MC task containing the gold answer and one distractor, with the gold position balanced between A and B, and fine-tune the models reported below directly on that format. Results are on the test split over three seeds (Table 6, Figure 4).

The control shows that MC supervision remains learnable when the nominal floor is matched to binary QA. It does not remove the more important limitation that WikiDoc-B and WikiDoc-MC contain different questions.

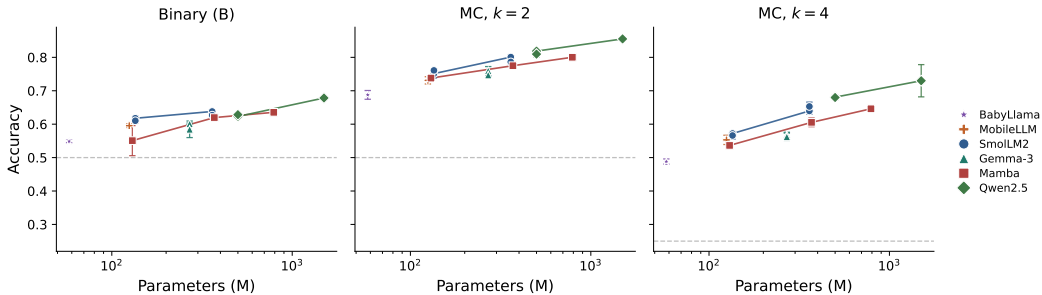


Figure 4: WikiDoc accuracy for binary QA and two- and four-choice MC after format-specific SFT.

Table 6: Two-choice WikiDoc-MC control. Models are fine-tuned directly on a two-option MC task (gold answer plus one distractor, with balanced A/B positions), so chance is 0.50 as in binary QA. Results are test accuracy, mean $\pm$ standard deviation over three seeds.

Model	Accuracy	Δ vs. 0.50
BabyLlama-58M	0.688 $\pm$ 0.013	+0.188
MobileLLM-125M	0.731 $\pm$ 0.011	+0.231
Mamba-130M	0.738 $\pm$ 0.010	+0.238
SmolLM2-135M	0.751 $\pm$ 0.013	+0.251
SmolLM2-135M-it	0.761 $\pm$ 0.001	+0.261
Gemma-3-270M	0.760 $\pm$ 0.013	+0.260
Gemma-3-270M-it	0.747 $\pm$ 0.004	+0.247
SmolLM2-360M	0.801 $\pm$ 0.006	+0.301
SmolLM2-360M-it	0.786 $\pm$ 0.003	+0.286
Mamba-370M	0.775 $\pm$ 0.009	+0.275
Qwen2.5-0.5B	0.819 $\pm$ 0.005	+0.319
Qwen2.5-0.5B-it	0.810 $\pm$ 0.008	+0.310
Mamba-790M	0.800 $\pm$ 0.005	+0.300
Qwen2.5-1.5B	0.855 $\pm$ 0.004	+0.355

F Controlling for MC option artifacts

Multiple-choice items can contain predictive regularities in the answer options independently of their relation to the question. In the spirit of partial-input artifact analyses (Poliak et al., 2018; Gururangan et al., 2018; Balepur et al., 2024), we train a TF-IDF logistic-regression scorer using only option text. Each option is treated independently during training and labeled as gold or non-gold; at test time, the four options of an item are scored independently and the highest-scoring option is selected. The scorer therefore captures global surface regularities that distinguish correct answers from distractors, but cannot use the question, relations among the candidate answers, or answer position.

The option-only scorer reaches 0.448 accuracy on WikiDoc-MC, compared with 0.25 random choice, confirming substantial predictive signal in the option strings alone. Nevertheless, every fine-tuned WikiDoc-MC model exceeds this baseline, from 0.488 at 58M to 0.769 at 3B, with a median advantage of +0.158. The same diagnostic reaches 0.430 on SciQ and 0.344 on MedMCQA.

Option-side artifacts are thus a genuine confounder, but the specific surface regularities captured by this baseline do not account for the full MC performance.

Table 7: Item-level accuracy of the option-only TF-IDF diagnostic. Each candidate is scored independently as gold versus non-gold without access to the question, and the highest-scoring option is selected.

Benchmark	Option-only accuracy	Chance
WikiDoc-MC	0.448	0.25
SciQ	0.430	0.25
MedMCQA	0.344	0.25

G Probe details

The probe analysis uses WikiDoc-B and four representative models: BabyLlama-58M, Gemma-3-270M, Qwen2.5-0.5B, and Qwen2.5-3B. We use one SFT seed. At each layer, a logistic-regression probe predicts the yes/no target from the final-token hidden state; layer and regularization are selected on validation and test is evaluated once. The probe is compared against the model’s own threshold-free output-margin AUROC on the same items. The probe-minus-output differences are -0.024 , -0.012 , -0.017 , and -0.003 , respectively. Fine-tuning increases linear decodability relative to the pretrained base model, but under this particular readout we do not observe more answer information than is already expressed by the output margin. Per-model values are given in Table 8 and visually represented in Figure 5.

Table 8: AUROC for linear probes on WikiDoc-B, one SFT seed. Layer and regularization are selected on validation and test is scored once. *Shuffle* is the same probing procedure trained on randomly permuted labels, and Δ_{shuf} is Probe minus Shuffle, reported as a sanity check. *Margin* is AUROC of the model’s own yes/no output margin on the same items, and Δ_{margin} is Probe minus Margin. $\text{Probe}_{\text{pretr.}}$ is probe AUROC on the corresponding pretrained checkpoint. The fine-tuned probe rises above its shuffled-label control for the larger tested models, but never exceeds the model’s own output margin.

Model	Layer	Probe	Shuffle	Δ_{shuf}	Margin	Δ_{margin}	$\text{Probe}_{\text{pretr.}}$
BabyLlama-58M	16	0.542	0.529	+0.013	0.565	−0.024	0.519
Gemma-3-270M	15	0.628	0.517	+0.111	0.640	−0.012	0.569
Qwen2.5-0.5B	13	0.656	0.505	+0.151	0.673	−0.017	0.572
Qwen2.5-3B	36	0.765	0.492	+0.274	0.768	−0.003	0.696

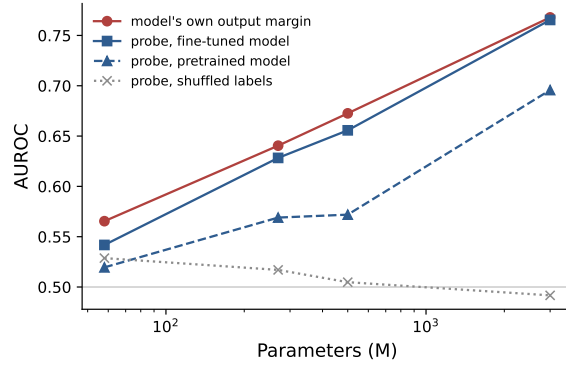


Figure 5: Linear probes on WikiDoc-B final-token hidden states, one SFT seed. The comparison is against the model’s own threshold-free output margin on the same items. The fine-tuned probe rises above its shuffled-label control for the larger tested models and fine-tuning increases decodability over the pretrained model. It nonetheless remains below the model’s own margin for every tested model, so under this probe setup, we recover no additional linearly decodable answer signal beyond the output margin.