Fine-Tuning Small Language Models for Cybersecurity: Data Ordering, Knowledge Distillation, and the Educator Effect

Ozkan Kilic, Raja Soundaramourty, Ramu Chenchaiah
Cisco Systems, Inc., San Jose, CA, USA
{okilic,rajasoun,rchencha}@cisco.com

Abstract

Data-sovereignty rules forbid cloud-hosted AI in many high-security environments, leaving compact on-premise models as the only path to AI-assisted cybersecurity. We fine-tune three small open-source models, Gemma 2 2B, Phi-3.5 3.8B, Llama 3.1 8B on ~147,600 synthetic cybersecurity QA pairs using QLoRA on V100 GPUs. Under strict MCQ evaluation Gemma 2 gains +9.3 pp, Phi +4.0 pp, and Llama drops −24.0 pp. We term this the educator effect: models trained on pedagogical data internalize explanatory behavior at the expense of format compliance. Severity appears to scale with capacity, though capacity is confounded with architecture and learning rate. A controlled ablation shows randomized ordering outperforms curriculum, without significance testing on the 75-question exam.

1 Introduction

The cybersecurity workforce shortage has driven interest in AI for security operations and training. Large models like GPT-4 (OpenAI, 2023) show strong capability but require cloud inference. Financial institutions (PCI-DSS), energy facilities (NERC CIP), healthcare (HIPAA), and GDPR-bound enterprises all face the same constraint: cloud LLM APIs are forbidden. In these air-gapped environments the choice is between a local model and no model at all.

A substantial portion of the organizations most in need of cybersecurity AI assistance operate under regulatory and policy frameworks that categorically prohibit transmitting operational data to external servers. The development of compact, on-premise language models fine-tuned for cybersecurity thus represents both a practical necessity and a research imperative.

This work addresses three questions: (1) Can small language models (≤8B parameters) be effectively fine-tuned for cybersecurity using curriculum-based training on synthetic data? (2) Does curriculum learning in easy-to-hard progression confer advantages over random presentation during LoRA fine-tuning? (3) How do different model architectures and pre-training methods respond to domain-specific fine-tuning?

We construct a large-scale synthetic cybersecurity dataset from 21 Cisco University courses organized into a 6-tier difficulty curriculum, fine-tune three models using QLoRA on V100 GPUs (reflecting the hardware constraints typical of institutional and governmental computing environments), and evaluate on both certification-style MCQ and open-ended cybersecurity questions. Our contributions include:

- Documenting the educator effect: A systematic behavioral shift in fine-tuned models toward pedagogical explanation at the expense of constrained answer selection. In our three-model sample severity appears to scale with model capacity, but this remains confounded with architecture, tokenizer, and learning rate and is offered as a hypothesis for future controlled study.
- The first controlled curriculum-vs-random ablation study for LoRA fine-tuning of instruction-tuned LLMs, extending Wu et al. (2021) and Campos (2021) to parameter-efficient fine-tuning.
- Preliminary, confounded evidence that knowledge-distillation (KD) pre-training may confer robustness to both behavioral disruption and data ordering effects during fine-tuning; we frame this as a hypothesis requiring a controlled KD-vs-non-KD experiment.

Proceedings of the Second International Conference on Natural Language Processing and Artificial Intelligence for Cyber Security (NLPAICS'2026)
11–12 June, 2026. ©2026 The Authors

- Practical findings on hardware compatibility for resource-constrained deployments (V100 GPU limitations with recent architectures).

2 Related Work

2.1 Curriculum Learning

Curriculum learning (Bengio et al., 2009) orders training examples easy-to-hard, mirroring human education. Surveys (Wang et al., 2021; Soviany et al., 2022) categorize approaches and note that anti-curriculum orderings generally underperform. Xu et al. (2023) show curriculum-based DPO benefits alignment.

The application of curriculum learning to domain-specific fine-tuning remains underexplored. In the medical domain, curricula mirroring the progression of medical education have shown promise, providing theoretical motivation for cybersecurity, which exhibits comparable prerequisite structures (e.g., networking fundamentals must precede network security concepts).

However, the most comprehensive empirical study (Wu et al., 2021) tests thousands of orderings and finds randomly ordered samples perform as well or better, attributing any benefit to dynamic dataset sizing rather than ordering. Campos (2021) corroborates this for LM pre-training. Neither addresses LoRA fine-tuning of instruction-tuned LLMs.

2.2 Domain-Specific LLM Adaptation: Cybersecurity

Parameter-efficient fine-tuning dominates domain adaptation. LoRA (Hu et al., 2021) injects low-rank matrices while freezing pre-trained weights; QLoRA (Dettmers et al., 2023) adds 4-bit NormalFloat quantization, enabling 65B fine-tuning on a single 48 GB GPU at 16-bit quality. These methods reduce but do not eliminate the tension between specialization and general-capability preservation (Ouyang et al., 2022).

Biderman et al. (2024) show LoRA "learns less and forgets less" than full fine-tuning, with learning rate the most critical hyperparameter. Kalajdzievski (2024) establishes that forgetting follows a shifted power law in trainable parameters and update steps. In medicine, MEDITRON (Chen et al., 2023) and Med-PaLM 2 (Singhal et al., 2023) achieve strong results through continued pretraining at 70B+ scale.$5 \times 10^{-5} 5 \times 10^{-4}$

2.3 Language Models for Cybersecurity

CyberMetric (Tihanyi et al., 2024) benchmarks LLM cybersecurity knowledge and finds substantial gaps versus experts. SecureFalcon (Ferrag et al., 2023) shows domain-specific models can outperform general ones (94% vulnerability classification). Air-gapped deployment creates demand for compact on-premise models, and synthetic data from authoritative sources is a standard strategy for domain training (Taori et al., 2023).

3 Methodology

3.1 Data Set Construction

Our training data comes from 21 cybersecurity courses on Cisco University (u.cisco.com), spanning networking, security operations, ethical hacking, cloud security, and incident response. A three-stage prompted pipeline using GPT-4 (Azure OpenAI, temperature 0.7, max 3,000 tokens) produced 147,598 QA pairs: (1) concept extraction returning strict JSON, (2) chain-of-thought generation under a teaching-expert persona, and (3) Cisco-styled variant/reinforcement. Each item has a question, an explanatory answer, concept tags, and a curriculum tier (1–6). We did not apply MinHash/Jaccard deduplication or human re-labelling. Generation used a tenant-isolated Azure OpenAI endpoint; training and inference ran entirely on the on-premise GPU server. Cisco University licensing prevents release of the raw corpus, but we release prompts, training/evaluation code, LoRA adapter checkpoints, and per-question outputs. A substring decontamination check (50-char prefix) between the 213 evaluation prompts and 118,549 unique training prompts found 0 exact and 0 confirmed partial matches. Topical adjacency between courses and exam objectives remains.

Tier	Focus Area	Courses	Samples
1	Foundational Knowledge (OS, Networking, Crypto)	4	42,579
2	SOC Fundamentals	6	29,559
3	Threats & Attacks	4	40,912
4	Detection & Monitoring	2	13,573

5	Incident Response	3	13,721
6	Cloud Security	2	7,254
	Total	**21**	**147,598**

Table 1: Curriculum Structure.

Data is arranged in a difficulty-based curriculum (Bengio et al., 2009): the model first encounters all 21 courses at easy level, then difficult, then hard. Within each difficulty level, courses are presented in tier order (1→6).

3.2 Models

We fine-tune three instruction-tuned models selected for resource-constrained deployment:

Model	Params.	Pre-training	Source
Gemma 2 2B IT	2.0B	Knowledge Distillation	Gemma Team (2024)
Phi-3.5 Mini Instruct	3.8B	Next-Token Prediction	Abdin et al. (2024)
Llama 3.1 8B Instruct	8.0B	Next-Token Prediction	Dubey et al. (2024)

Table 2: List of the base models fine-tuned.

A fourth model, Gemma 3 4B IT, was attempted but failed catastrophically on V100 hardware due to BFloat16 incompatibility.

3.3 Fine-Tuning Configuration

All models use QLoRA (Dettmers et al., 2023): 4-bit NF, LoRA rank 16, α=32, dropout 0.05, all linear layers. Optimization is Paged AdamW 8-bit (Loshchilov & Hutter, 2019) with cosine annealing. We train 1 epoch with effective batch 16 (per-device 4, gradient accumulation 4) on a single V100-32GB, gradient checkpointing enabled, max sequence length 2048:

Hyper-param.	Phi-3.5 / Llama	Gemma 2
Learning rate	2×10^{-4}	5×10^{-5}
Warmup ratio	0.03	0.05
Trainable params	1.24% / 0.92%	1.28%

Table 3: Fine-tuning Hyper-parameters.

Gemma's lower learning rate reflects empirical instability at higher rates, consistent with QLoRA guidance. Training data uses each model's native chat template with the system prompt: "You are an expert cybersecurity instructor. Answer the following question clearly, accurately, and in detail. Use examples where helpful and always aim to teach the concept behind the answer." This encodes the pedagogical objective whose implications we analyze in §5.1.

3.4 Evaluation

Evaluation uses 222 questions in two disjoint sets: 75 MCQ from the Cisco CyberOps Associate certification exam (Cisco University) and 147 open-ended questions from public sources (Cisco Learning Network, Edureka, GeeksforGeeks, community resources). Exam items follow the standard format (stem + four options A–D, one correct).

We employ three evaluation protocols developed iteratively to address confounds discovered during analysis:

- **Protocol v1 (Educator prompt for all):** Educator system prompt for both exam and general questions, 512-token max. *Issue:* Encourages explanatory responses even for MCQ.
- **Protocol v2 (No prompt for exam):** No system prompt for exam, educator prompt for general, 256-token max for exam. *Rationale:* Isolates the educator persona's effect on exam performance.
- **Protocol v3 (Strict MCQ):** Explicit MCQ instruction for exam questions ("You are taking a multiple-choice certification exam. Reply with ONLY the letter of the correct answer. Do not explain your reasoning.") with 64-token max; educator prompt for general questions with 2048-token max. *Rationale:* Strongest possible format constraint.

Protocol v3 is our primary evaluation. For exam questions we use lenient substring containment: a response is correct if the correct option letter (A–D) appears anywhere in the output. This can over-credit verbose responses, so the negative Llama result (−24.0 pp) is conservative w.r.t. this bias while the positive Gemma 2 result (+9.3 pp, mean 8.4-word responses) is essentially unaffected. For open-ended questions we report coverage and keyword coverage as lexical proxies only, not measures of correctness, utility, or safety, and do not use them to support any accuracy claim. A 0-3

rubric and LLM-as-judge protocol are left to future work (§6).

3.5 Ablation: Curriculum vs. Random Ordering

To isolate the effect of curriculum-based data ordering, we conduct controlled ablation studies using both Phi-3.5 and Gemma 2. Each ablation holds all hyperparameters constant and varies only the data presentation order:

Curriculum condition: difficulty-first order from §3.1, progressing through all courses at each tier before advancing. Randomized condition: the same 147,600 samples fully shuffled (different seed). Both conditions share hardware, learning rate, batch size, and training duration. Running on two architecturally distinct models with different pre-training (KD vs. NTP) tests robustness of ordering effects.

3.6 Infrastructure

Experiments run on a server with 6× NVIDIA Tesla V100-PCIE-32GB GPUs (CC 7.0, CUDA 535.247.01), PyTorch 2.1.2 (CUDA 11.8), Transformers 4.50.3, PEFT 0.14.0, TRL 0.13.0; each model on its own GPU. V100s represent hardware typical of institutional, government, and research environments. Data generation, training, inference, and evaluation all ran on the isolated server, demonstrating the air-gapped workflow and respecting content-licensing constraints.

4 Results

4.1 Training Outcomes

Model	Condition	Loss	Time (h)
Gemma 2 2B	Curriculum	0.611	7.7
Gemma 2 2B	Randomized	1.128	7.7
Phi-3.5	Curriculum	0.879	25.9
Phi-3.5	Randomized	0.859	25.9
Llama 3.1 8B	Curriculum	1.007	44.6

Table 4: Training outcomes for all configurations.

All three successful models show monotonically decreasing loss. Gemma 2 reaches the lowest loss (0.611) in the shortest time (7.7 h), reflecting smaller size and the lower learning rate. The inverse size–loss relationship (2B < 3.8B < 8B) tracks the trainable/total parameter ratio (Gemma 2: 1.28%, Phi: 1.24%, Llama: 0.92%). Gemma 2's curriculum loss (0.611) is nearly half its randomized loss (1.128), while Phi's are comparable (0.879 vs. 0.859); we analyze this divergence in §5.3.

Gemma 3 Failure. Gemma 3 4B IT failed on V100 under two frameworks (QLoRA and Unsloth), with loss plateauing at ~61.0. The cause is hardware-level: Gemma 3's mixed-precision operations and attention soft-capping require BF16, available only on Ampere (A100) and newer GPUs; V100 (CC 7.0) supports only FP16, whose narrower dynamic range produces NaN or extreme gradients. The same model converges smoothly (loss 1.18) on A10G with identical hyperparameters, confirming a precision incompatibility rather than misconfiguration. Gemma 2 2B IT from the same family trains successfully on V100, so the boundary tracks model generation, not family.

4.2 Exam Performance

Model	Base Accuracy	FT Accuracy	Δ (pp)
Gemma 2 2B	44.0% (33/75)	53.3% (40/75)	**+9.3**
Phi-3.5	52.0% (39/75)	56.0% (42/75)	**+4.0**
Llama 3.1 8B	58.7% (44/75)	34.7% (26/75)	**−24.0**

Table 5: Exam accuracy (Q1-75, Protocol v3, strict MCQ).

Fine-tuning effects are strikingly model-dependent. The smallest model (Gemma 2) improves most (13 questions gained vs. 6 lost); the largest (Llama) degrades catastrophically. Under v2 (no system prompt) both Phi and Llama drop −16.0 pp; under v3 Phi recovers to +4.0 pp while Llama worsens to −24.0 pp. The Phi v2→v3 swing (−16.0→+4.0 pp) shows that fine-tuning's sign can flip purely with the evaluation protocol, cautioning against premature conclusions.

Model	Educator (v1)	No Prompt (v2)	Strict MCQ (v3)
Base Phi-3.5	49.3%	68.0%	52.0%
FT Phi-3.5	41.3%	52.0%	56.0%
Base Llama 3.1	56.0%	62.7%	58.7%
FT Llama 3.1	45.3%	46.7%	34.7%
Base Gemma 2	36.0%	61.3%	44.0%
FT Gemma 2	34.7%	54.7%	53.3%

Table 6: Cross-protocol exam accuracy comparison.

The v2 protocol gives the highest base scores (Phi 68.0%, Llama 62.7%, Gemma 2 61.3%), suggesting any system prompt lowers baselines. Fine-tuned Phi improves under v3 (+4.0 pp) despite degrading under v2 (−16.0 pp). Llama FT degrades under all protocols and worsens with explicit format constraints (v1 45.3%, v2 46.7%, v3 34.7%), confirming deep behavioral entrenchment. Gemma 2 follows Phi's pattern: −1.3 pp (v1), −6.7 pp (v2), +9.3 pp (v3), confirming the strict MCQ prompt is essential to reveal knowledge gains.

4.3 Response Verbosity

Model	Base Words	FT Words
Gemma 2	7.9	8.4
Phi-3.5	11.6	9.6
Llama 3.1	5.9	**49.7**

Table 7: Exam response verbosity (Protocol v3, 64-token max).

Gemma 2 shows virtually no verbosity change after fine-tuning (7.9→8.4 words). Phi becomes slightly more concise (11.6→9.6). Llama saturates the 64-token limit despite explicit "ONLY the letter" instructions with an order of magnitude more verbose than its base (5.9→49.7), a striking override of prompt-level instructions by fine-tuning-induced behavior.

On general questions (Q76–222), Gemma 2 and Phi become 25–40% *more concise* while maintaining keyword coverage (Gemma 2: 528→398 words; Phi: 518→312 words). Llama becomes 33% more verbose (386→512 words), consistent with its exam behavior.

A training-data probe (10 questions, seed=42) confirmed genuine behavioral change rather than degradation: all 20 base-vs-fine-tuned comparisons differed. On familiar training-distribution questions fine-tuned Llama is more concise than base (−63% TLS handshake, −27% WMI exploitation), inverting its verbose MCQ behavior (5.9→49.7 words). The asymmetry shows the educator effect is format-dependent: on familiar formats the model terminates early; on unfamiliar formats (MCQ) it defaults to its trained explanatory mode.

4.4 Ablation: Curriculum vs. Randomized Data Ordering

Model	Base	FT Acc.	Loss
Phi-3.5 (Curr.)	52.0% (39/75)	56.0% (42/75)	0.879
Phi-3.5 (Rand.)	50.7% (38/75)	65.3% (49/75)	0.859
Gemma 2 (Curr.)	44.0% (33/75)	53.3% (40/75)	0.611
Gemma 2 (Rand.)	44.0% (33/75)	54.7% (41/75)	1.128

Table 8: Curriculum vs. randomized: Exam accuracy (Q1–75, Protocol v3).

Randomized ordering outperforms curriculum for both models, but the magnitudes differ. Phi gains nearly 4× more from randomization (+14.7 pp vs. +4.0 pp); Gemma 2's gap is within noise (+10.7 pp vs. +9.3 pp, one question). Per-question, Phi has 9 randomized-only correct vs. 2 in reverse; Gemma 2 is near-symmetric (9 vs. 8). Without McNemar or bootstrap tests on the small set the Phi gap is suggestive while the Gemma 2 gap is effectively null.

Base model scores are nearly identical across conditions (Phi 38 vs. 39; Gemma 2 33 vs. 33), confirming differences lie in fine-tuning. Phi

converges to similar losses across conditions (0.879 vs. 0.859), while Gemma 2's curriculum loss (0.611) is nearly half its randomized loss (1.128) for only +1.4 pp accuracy difference as analyzed in §5.3.

Model	Condition	Avg Words	Keyword Cover.
Phi-3.5	Curr. FT	312	55.8%
Phi-3.5	Random FT	421	59.3%
Phi-3.5	Base	518	58.0%
Gemma 2	Curr. FT	398	54.1%
Gemma 2	Random FT	446	55.5%
Gemma 2	Base	528	53.3%
Llama 3.1	Curr. FT	512	57.8%
Llama 3.1	Base	386	55.8%

Table 9: General question metrics (Q76–222, Protocol v3).

Both fine-tuned conditions produce more concise responses than base while maintaining or improving keyword coverage, except for Llama, which becomes 33% more verbose (386→512 words). Randomized Phi reaches the highest absolute keyword coverage of any fine-tuned model (59.3%), exceeding even its base (58.0%); curriculum Phi drops to 55.8%. For Gemma 2 both conditions edge above base, randomized slightly ahead (55.5% vs. 54.1%). Curriculum induces stronger length compression, potentially discarding useful content.

5 Discussion

5.1 The Educator Effect

The most salient finding is the educator effect: fine-tuned models adopt pedagogical response styles, which includes detailed explanations, conceptual context, instructional scaffolding, inappropriate for constrained formats such as MCQ. This is consistent with the training system prompt instructing the model to "teach the concept behind the answer": the model becomes a better teacher but a worse test-taker.

In our three-model sample severity appears to scale with model capacity, but the three models differ simultaneously in size, architecture, tokenizer, instruction-tuning procedure, and learning rate; we therefore frame the ordering below as a hypothesis, not a demonstrated capacity–entrenchment law:

- **Llama 3.1 8B** (0.92% trainable) exhibits deep behavioral entrenchment: verbose explanations override format instructions under all protocols, averaging 49.7 words per exam response under strict MCQ constraints. Fine-tuned Llama's evaluation time is more than double base (51.1 vs. 22.1 minutes under v3), consistent with maximum-length response generation.
- **Phi-3.5** (1.24% trainable) internalizes the pedagogical style while retaining format modulation, producing concise exam responses (9.6 words) and improving accuracy by +4.0 pp. Fine-tuned Phi is also slightly faster than base (25.3 vs. 28.9 minutes), reflecting more concise outputs.
- **Gemma 2 2B** (1.28% trainable) shows virtually no educator effect on exam responses (8.4 words, matching base), while improving accuracy by +9.3 pp.

This gradient is consistent with larger models absorbing both knowledge and behavioral patterns from fine-tuning data while smaller models preferentially absorb knowledge, but three points across three families cannot distinguish a true capacity effect from family- or learning-rate-specific behavior. A controlled within-family scan is required.

5.2 Understanding Divergent Accuracy

We ground the divergence in three factors:

Capacity–entrenchment tradeoff. Biderman et al. (2024) show LoRA acts as implicit regularization against forgetting; our result inverts the naive expectation: The smallest model benefits most, the largest degrades most. Larger capacity appears to absorb not only facts but also the training data's pedagogical patterns, which dominate under format-constrained evaluation. Gemma 2's smaller representational space may filter stylistic patterns while retaining factual content.

Learning rate as confound. Gemma 2 uses a 4× lower learning rate than Phi and Llama, chosen for empirical stability. A gentler rate may update knowledge without disrupting format-following behavior. But rate alone cannot explain the divergence: Phi and Llama share the same rate yet produce opposite outcomes (+4.0 vs. −24.0 pp), so architecture and pre-training also play decisive roles.$5 \times 10^{-5} 2 \times 10^{-4}$

The Knowledge Distillation Hypothesis (speculative). Gemma 2 was pre-trained via knowledge distillation (Hinton et al., 2015; Gemma Team, 2024) rather than standard next-token prediction. We hypothesize, but do not demonstrate, that this could yield representations more robust to behavioral disruption. Because Gemma 2 is the only KD-trained model in our comparison and differs in scale, family, tokenizer, and learning rate, the points below are a hypothesis to be tested:

1. *Distributional awareness.* Soft targets produce calibrated probability mass across completions, making output behavior inherently more stable.
2. *Format robustness.* The Gemma 2 technical report shows σ=2.1 on MMLU across 12 formatting variations, versus 6.9 for Mistral 7B (Gemma Team, 2024, Table 11), inherent format stability surviving fine-tuning.
3. *Double distillation.* Both pre-training and instruction tuning use distillation, doubly reinforcing format-appropriate response generation. When QLoRA subsequently modifies 1.28% of parameters, the distillation-trained majority anchors against behavioral drift.

NTP-trained models (Phi, Llama) optimize for the single most likely next token and may be more susceptible to wholesale mode switching when LoRA introduces new behavior. Gemma 2's observation that IT models are better at "understanding formatted questions" is consistent with KD producing a more modular knowledge/format relationship. A controlled test using KD-trained and from-scratch Gemma 2 variants is left to future work.

An additional factor is objective mismatch. The training objective (maximize likelihood of detailed explanations) is misaligned with the MCQ objective (select a single letter). Impact depends on whether the model can compartmentalize knowledge from behavior: Gemma 2 and Phi do, Llama does not. Its fine-tuned variant produces verbose explanations even under 64-token strict MCQ constraints.

5.3 Why Randomized Ordering Outperforms Curriculum

Our ablation extends Wu et al. (2021) and Campos (2021) to LoRA fine-tuning. When dataset size is fixed (all samples seen in one epoch), curriculum provides no benefit. Our results go beyond "no benefit" to suggest that curriculum ordering can be counterproductive for parameter-efficient fine-tuning.

We propose recency bias interacting with LoRA's capacity bottleneck. Adapters modify only 1.24–1.28% of parameters. Under curriculum ordering the adapter first specializes for easy examples (Tiers 1–2) and must overwrite them as harder examples arrive (Tiers 5–6), biasing it toward the final tier (cloud security). Randomization interleaves difficulty in every batch, forcing the adapter to span the full spectrum.

Training loss corroborates: Gemma 2's curriculum loss (0.611) is strikingly low versus randomized (1.128), likely reflecting overfitting to ordered structure. Predictable curriculum progression yields high-confidence next-token predictions without necessarily better knowledge acquisition; randomization's higher loss reflects the harder learning signal of interleaved difficulty, producing more robust representations.

6 Conclusion

We fine-tune three small language models for cybersecurity using QLoRA on V100 GPUs and present four key findings.

First, fine-tuning effects are model-dependent: the smallest model (Gemma 2 2B) improves most (+9.3 pp), the largest (Llama 8B) degrades most (−24.0 pp), contrary to the assumption that larger models benefit more. We ground this in Biderman et al. (2024) and Kalajdzievski (2024).

Second, we introduce the *educator effect*. It is fine-tuning on pedagogical data induces behavioral entrenchment proportional to model capacity, overriding format instructions in larger models while smaller models retain behavioral flexibility. This has implications for evaluation methodology: measured fine-tuning effectiveness depends on the evaluation protocol, not just the training.

Third, randomized ordering consistently outperforms curriculum for LoRA fine-tuning (Phi: +14.7 pp vs. +4.0 pp; Gemma 2: +10.7 pp vs. +9.3 pp), extending Wu et al. (2021) and Campos (2021) to PEFT. Without significance tests the Gemma 2 gap is within noise; the Phi gap is the main quantitative signal. We propose recency bias interacting with LoRA's capacity bottleneck as a candidate mechanism.

Fourth, KD-trained Gemma 2 is insensitive to data ordering despite large training-loss differences (0.611 vs. 1.128). It is suggestive but confounded evidence that distillation-anchored base weights may sandbox adapter-level overfitting. The curriculum model achieves much lower loss yet no accuracy advantage. A controlled KD-vs-from-scratch Gemma 2 test is the most important follow-up (§6).

We additionally argue that certification exams are incomplete measures of cybersecurity domain competence for generative models, and that the observed behavioral shifts may enhance utility in the intended deployment context, as interactive knowledge assistants in air-gapped environments. Our findings suggest that the most promising path for resource-constrained deployment combines compact distillation-trained models with randomized fine-tuning on focused cybersecurity workflows.

7 Future Work and Limitations

Scale dependence. All existing evidence comes from models ≤8B. Whether the educator effect, curriculum ordering effects, and KD interaction persist at 13B–70B scale is unknown. Larger adapters may reduce the recency bias mechanism. A controlled test across model sizes (e.g., Llama 3.1 at 8B and 70B) with identical data would establish whether these effects are scale-invariant.

Controlled KD experiment. Our KD hypothesis is confounded with model scale and learning rate. Fine-tuning both KD-trained and from-scratch Gemma 2 (both publicly available) with identical hyperparameters would isolate the KD variable. This is the most important follow-up experiment.

Evaluation limitations. Our 75-MCQ exam is small and we report no confidence intervals, McNemar tests, or bootstrap analyses; single-question differences should not be read as meaningful. Keyword coverage is a lightweight lexical proxy; a rubric-based or LLM-as-judge evaluation, and multi-turn dialogue, would strengthen claims. CyberOps Associate-style items are topically adjacent to a subset of training courses, so gains may partly reflect source–evaluation overlap; an independent non-Cisco suite is needed.

Training strategy refinements. Mixed-format training (MCQ + explanatory QA) may reduce the educator effect while preserving knowledge gains. Two-stage training, knowledge injection via QA followed by format adaptation via MCQ, is another promising direction.

Workflow-specific fine-tuning. SOC alert triage, incident response playbooks, log analysis, and forensic reporting represent bounded tasks well-suited to small models. They potentially outperform broad domain fine-tuning. These closed-set tasks have constrained output formats and well-defined correctness criteria.

Continued pretraining. Following MEDITRON's approach, continued pretraining on raw cybersecurity text (RFCs, CVEs, MITRE ATT&CK documentation) may improve domain knowledge without disrupting format-following capabilities: Preserving the model's natural response patterns while injecting domain knowledge.

8 References

Marah Abdin et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th International Conference on Machine Learning (ICML)*, pages 41–48.

Dan Biderman et al. 2024. LoRA learns less and forgets less. *Transactions on Machine Learning Research*.

Daniel Campos. 2021. Curriculum learning for language modeling. *arXiv preprint arXiv:2108.02170*.

Zeming Chen et al. 2023. MEDITRON-70B: Scaling medical pretraining for large language models. *arXiv preprint arXiv:2311.16079*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Abhimanyu Dubey et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Mohamed Amine Ferrag et al. 2023. SecureFalcon: Are we there yet in automated software vulnerability detection with LLMs? *IEEE Transactions on Software Engineering*.

Gemma Team, Google DeepMind. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*.

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

Damjan Kalajdzievski. 2024. Scaling laws for forgetting when fine-tuning large language models. *arXiv preprint arXiv:2401.05605*.

Ilya Loshchilov and Frank Hutter. 2017. SGDR: Stochastic gradient descent with warm restarts. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Paulius Micikevicius et al. 2018. Mixed precision training. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Karan Singhal et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *International Journal of Computer Vision*, 130:1526–1565.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford Alpaca: An instruction-following LLaMA model. GitHub repository. https://github.com/tatsu-lab/stanford_alpaca.

Norbert Tihanyi, Mohamed Amine Ferrag, Ridhi Jain, Tamas Bisztray, and Merouane Debbah. 2024. CyberMetric: A benchmark dataset based on retrieval-augmented generation for evaluating LLMs in cybersecurity knowledge. *arXiv preprint arXiv:2402.07688*.

Xin Wang, Yudong Chen, and Wenwu Zhu. 2021. A survey on curriculum learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):4555–4576.

Xiaoxia Wu, Ethan Dyer, and Behnam Neyshabur. 2021. When do curricula work? In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Canwen Xu et al. 2024. Automatic pair construction for contrastive post-training. In *Findings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.