Knowledge Collapse in LLMs: When Fluency Survives but Facts Fail under Recursive Synthetic Training

Anonymous Author(s)
Affiliation
Address
email

Abstract

Large language models increasingly rely on synthetic data due to human-written content scarcity, yet recursive training on model-generated outputs leads to model collapse, a degenerative process threatening factual reliability. We define knowledge collapse as a distinct three-stage phenomenon where factual accuracy deteriorates while surface fluency persists, creating "confidently wrong" outputs that pose critical risks in accuracy-dependent domains. Through controlled experiments with recursive synthetic training, we demonstrate that collapse trajectory and timing depend critically on instruction format, distinguishing instruction-following collapse from traditional model collapse through its conditional, prompt-dependent nature. We propose domain-specific synthetic training as a targeted mitigation strategy that achieves substantial improvements in collapse resistance while maintaining computational efficiency. Our evaluation framework combines model-centric indicators with task-centric metrics to detect distinct degradation phases, enabling reproducible assessment of epistemic deterioration across different language models. These findings provide both theoretical insights into collapse dynamics and practical guidance for sustainable AI training in knowledge-intensive applications where accuracy is paramount.[1]

1 Introduction and Related Work

Large language models (LLMs) are increasingly trained on **synthetic data** due to cost-effectiveness, accessibility, and the growing contamination of internet sources with AI-generated content [9, 17]. Synthetic data avoids expensive human annotation, provides unlimited scalability, and is often indistinguishable from human-written text [18]. However, recursive training on model outputs can cause **model collapse**, where models progressively **lose the tails of the true data distribution** and converge to repetitive outputs [16, 15, 2]. This degradation arises from statistical sampling error (rare events vanish), functional expressivity error (capacity limits distort distributions), and optimization error (training favors easy patterns), with formal accounts in Appendix A.1.

While collapse signals and evaluation metrics are increasingly documented [3], the specific impact on factual question answering remains underexplored. Critically, models may remain fluent while factual reliability declines, creating "confidently wrong" outputs [21, 19]. This divergence distinguishes **knowledge collapse** from catastrophic forgetting, which concerns cross-task transfer [8]. Knowledge collapse instead occurs within a domain: factual accuracy erodes while surface competence persists. Such epistemic degradation threatens high-stakes applications, with healthcare systems reporting 40% error rates [12, 14], and synthetic feedback loops accelerating information decay across ecosystems [17, 18, 19]. Synthetic data exacerbates these risks: uniform Q&A formats promote pattern

[1] Codes will be publicly available upon acceptance.

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

overfitting and distributional shift, weakening instruction-following [10]. Mitigation has focused on *accumulate* (mixing real and synthetic) versus *replace* (substitution) workflows [7, 5]; while unlearning methods can recover some instruction fidelity [10], domain-specific synthetic training shows stronger preservation of accuracy in specialized areas [22].

This paper makes four contributions: (1) **Defining knowledge collapse** as a three-stage process: *Stage A* (Knowledge Preservation), *Stage B* (Knowledge Collapse, i.e., the "confidently wrong" transition), and *Stage C* (Instruction-following Collapse). (2) **Demonstrating conditional degradation**, showing that collapse trajectory depends on prompt format, unlike traditional prompt-agnostic collapse. (3) **Proposing mitigation via domain-specific training**, delaying accuracy decay by 15× compared to general synthetic training. (4) **Providing an open-source framework** for reproducible evaluation of epistemic degradation across models.

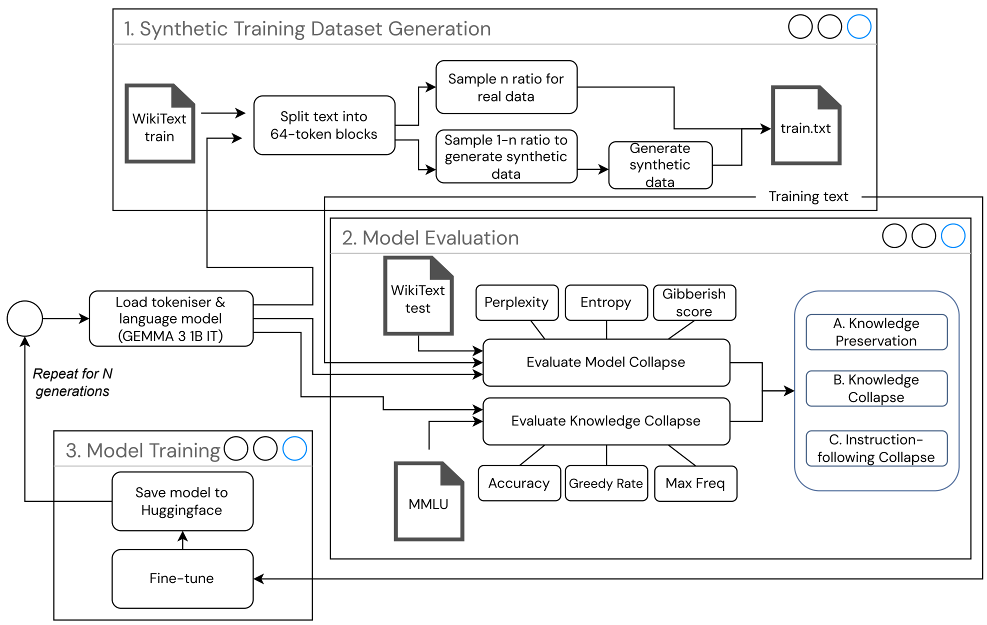

Figure 1: Cyclical workflow for recursive synthetic training: dataset generation, dual evaluation, and fine-tuning repeated across generations. Full step-by-step description is in Appendix A.2.

2 Methodology

We design a three-stage cyclical framework to study knowledge collapse under recursive synthetic training (Figure 1), using GEMMA 3 1B IT for controlled fine-tuning (details in Appendix A.2).

Datasets and training. Training uses WikiText-2 (8,000 64-token prompts) and evaluation covers five MMLU subjects (100 Q&A each, factual recall focus; Appendix A.3). Models are fine-tuned across generations with synthetic fractions $\alpha \in \{0.25, 0.50, 1.0\}$, combining $(1 - \alpha)$ real prompts with α synthetic continuations. Light-touch updates (0.5 epochs) enable gradual drift observation (Appendix A.4).

Evaluation. We combine model-centric signals (perplexity, entropy, gibberish score) with task-centric signals (accuracy, greedy rate, maximum frequency; Appendix A.5). MMLU items are reformatted into short-answer style to isolate factual retention (Appendix A.6).

Mitigation setup. To test whether collapse can be delayed through *distributional anchoring*, we construct a World Religions–focused subset aligned with one evaluation subject, and repeat recursive training under identical corpus sizes, synthetic ratios, and generation schedules. Corpus construction, semantic filtering, and validation appear in Appendix A.7 and B.3.

3 Results and discussion

Our first experiment establishes knowledge collapse as a distinct phenomenon with three identifiable stages under recursive synthetic training. **Stage A (Knowledge Preservation)** represents reliable factual accuracy with high instruction adherence. **Stage B (Knowledge Collapse)** demonstrates the critical transition where factual accuracy deteriorates while task format adherence persists, the "confidently wrong" phenomenon where models produce well-formatted but factually incorrect responses. **Stage C (Instruction-following Collapse)** indicates complete breakdown where accuracy approaches random baselines ($\leq$0.28) and outputs become incoherent.

This three-stage framework distinguishes knowledge collapse from general distributional degradation by focusing on epistemic rather than linguistic competence. Stage B represents the critical "valley of dangerous competence" most threatening to downstream applications, where traditional quality metrics fail to detect underlying knowledge erosion while factual reliability degrades. Figure 2 demonstrates how different synthetic ratios drive distinct stage transitions: 25% synthetic ratio reveals prolonged Stage A stability with transition to Stage B occurring only at later generations, 50% synthetic exhibits quicker transition from Stage A to Stage B at mid-generation, while 100% synthetic training shows rapid transition from Stage A to Stage B in early generations before continuing into Stage C collapse. Detailed distributional analysis (in Appendix B.1) reveals that knowledge collapse follows distinct patterns across synthetic ratios, with 100% synthetic training causing rapid entropy decline and vocabulary narrowing while 25-50% ratios preserve discriminative capability despite accuracy degradation. In addition, another qualitative trajectory analysis is shown in Appendix C.1.

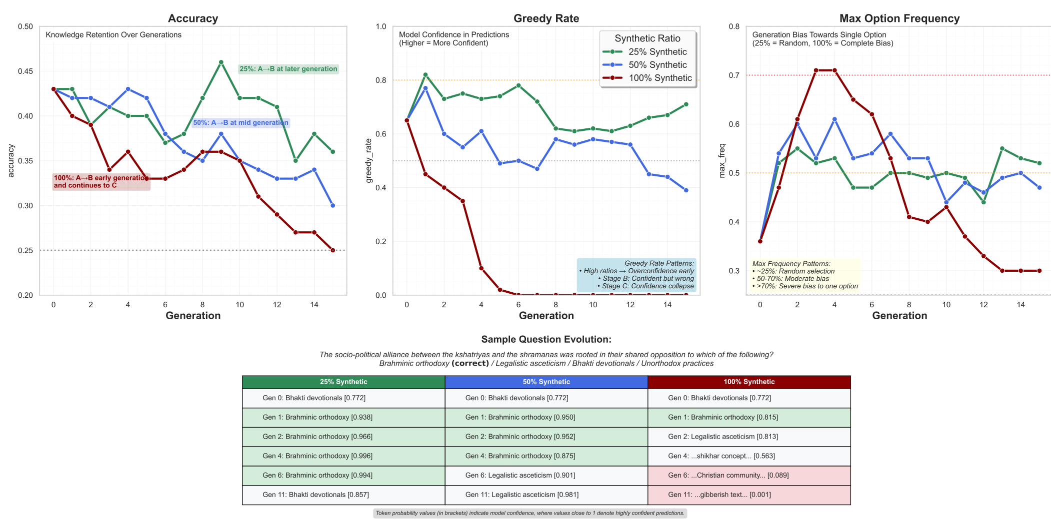

Figure 2: Knowledge collapse stages across synthetic ratios showing three critical metrics and sample evolution. The top tracks accuracy decline, model confidence (greedy rate), and option bias across generations, while the bottom demonstrates response degradation from accurate answers through confidently wrong responses to a complete breakdown.

Instruction-following collapse as conditional degradation. Stage C resembles traditional model collapse (loss of coherence, near-random outputs) but differs in being prompt-dependent. As shown in Figure 3, short-answer prompts remain above the random baseline until about Generation 8, few-shot prompts collapse rapidly by Generation 6, and zero-shot prompts degrade more gradually. Thus, collapse trajectory and timing are mediated by instruction format rather than occurring uniformly. Whereas model collapse is often described as a global, prompt-agnostic drift, instruction-following collapse shows that prompt structure can accelerate or delay degeneration. This prompt-dependent collapse reflects differences in how instruction complexity interacts with synthetic training artifacts. Few-shot prompts introduce structural dependencies through exemplars that become corrupted under recursive training, leading to rapid instruction-following breakdown [10]. The exemplar format exposes more surface patterns for overfitting and repetition, accelerating distributional shift toward template-driven outputs that disregard task requirements. In contrast, short-answer prompts restrict the response space without heavy structural demands, reducing vulnerability to pattern overfitting [20]. These results align with evidence that complex prompt structures amplify systematic biases, with greater formatting complexity correlating with faster degradation under recursive training [18]. Instruction format therefore mediates whether models retain competence during knowledge erosion (Stage B) or transition rapidly to instruction-following failure (Stage C). Additional distributional analysis and statistical validation are in Appendix B.2.

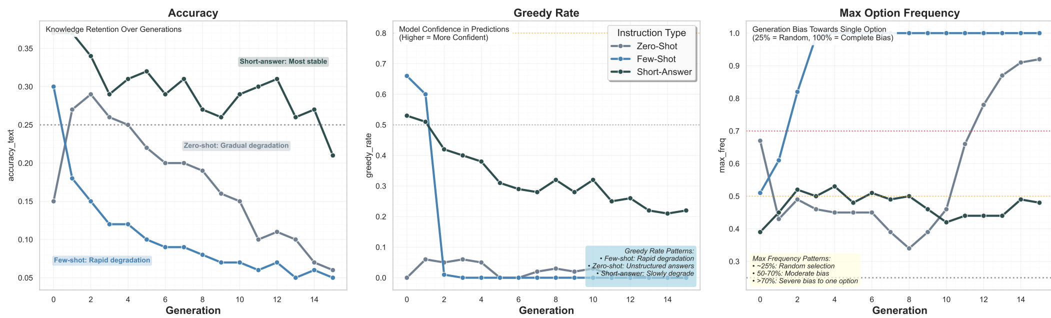

Figure 3: Instruction format sensitivity in knowledge collapse showing conditional degradation patterns. Experiment conducted on High School US History at 50% synthetic ratio. Short-answer prompts maintain stability the longest, while few-shot formats exhibit rapid collapse by Generation 6. Zero-shot prompts show intermediate degradation, demonstrating that collapse trajectory and timing depend on instructional format rather than representing uniform, prompt-agnostic drift.

Domain-specific Mitigation. Our third experiment tests whether restricting synthetic training to a subject-aligned corpus can delay knowledge collapse through distributional anchoring. We construct a World Religions–focused corpus (Appendix A.7) and compare it with original WikiText recursive training on equivalent-sized datasets. Figure 4 shows that domain-specific training yields greater stability, with a decay rate of -0.00054 accuracy per generation versus -0.00837 for general training, a 15× improvement with significant interaction effects ($p < 0.001$). It also prevents large perplexity growth (35 vs. 170) and maintains stable confidence, whereas general training drifts early and entrenches in incorrect outputs. Additional distributional analysis and qualitative trajectories are provided in Appendix B.3 and C.2.

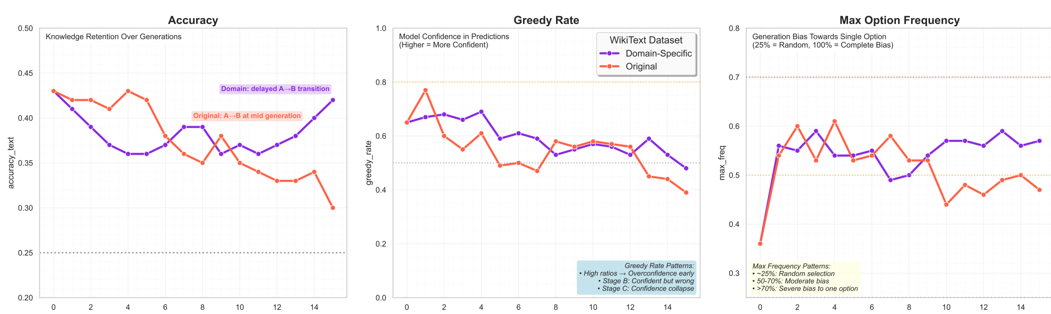

Figure 4: Domain-specific mitigation through WikiText filtering to focus on World Religions content demonstrates superior collapse resistance. By training on domain-aligned synthetic data, the approach stabilizes accuracy within MMLU World Religions while maintaining similar confidence and option bias patterns to the original baseline, contrasting with general training's rapid accuracy decline.

Limitations and Future Work. We identified knowledge collapse as a three-stage phenomenon in large language models, with domain sensitivity and mitigation through domain-specific synthetic training (15× slower decay, $p < 10^{-3}$). Our experiments were limited to GEMMA 3 1B IT and five MMLU subjects. Future work should evaluate collapse across scales and domains and develop predictive frameworks for "collapse-aware" training.

References

[1] Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F., & Vaughan, J. W. (2010). A theory of learning from different domains. *Machine Learning*, 79(1-2), 151-175.

[2] Dohmatob, E., Feng, Y., & Kempe, J. (2024). Strong model collapse. *arXiv preprint arXiv:2410.04840*.

[3] Dohmatob, E., Feng, Y., Yang, P., Charton, F., & Kempe, J. (2024). A tale of tails: Model collapse as a change of scaling laws. *International Conference on Machine Learning 2024*.

[4] Feng, Y., Dohmatob, E., Yang, P., Charton, F., & Kempe, J. (2024). Beyond model collapse: Scaling up with synthesized data requires reinforcement. *arXiv preprint arXiv:2407.10603*.

[5] Gerstgrasser, M., Schaeffer, R., Dey, A., Rafailov, R., Korbak, T., Sleight, H., et al. (2024). Is model collapse inevitable? Breaking the curse of recursion by accumulating real and synthetic data. *Advances in Neural Information Processing Systems 37*.

[6] Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2021). Measuring massive multitask language understanding. *Proceedings of ICLR*.

[7] Kazdan, J., Schaeffer, R., Dey, A., Gerstgrasser, M., Rafailov, R., Donoho, D. L., & Koyejo, S. (2025). Collapse or thrive? Perils and promises of synthetic data in a self-generating world. *arXiv preprint arXiv:2410.16713*.

[8] Kirkpatrick, J., Pascanu, R., Rabinowitz, N., Veness, J., Desjardins, G., Rusu, A. A., ... & Hadsell, R. (2017). Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13), 3521-3526.

[9] Langley, H., & Chowdhury, H. (2024). The AI world's most valuable resource is running out, and it's scrambling to find an alternative: 'fake' data. *Business Insider*.

[10] Liu, R., Jia, R., Zhang, W., Wang, D., Zhao, Y., Ping, Q., & Chen, D. (2024). Unveiling the flaws: Exploring imperfections in synthetic data and mitigation strategies for large language models. *Findings of the Association for Computational Linguistics: EMNLP 2024*, 14855–14865.

[11] Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*.

[12] IBM Think. (2024). What is model collapse? Retrieved from https://www.ibm.com/think/topics/model-collapse

[13] Peterson, A. J. (2025). AI and the problem of knowledge collapse. *AI & Society*.

[14] Ramachandran, H. (2025). Mitigating AI hallucinations in generative models with HITL. *Indium Blog*.

[15] Seddik, M. E. A., Chen, S.-W., Hayou, S., Youssef, P., & Debbah, M. (2024). How bad is training on synthetic data? A statistical analysis of language model collapse. *arXiv preprint arXiv:2404.05090*.

[16] Shumailov, I., Shumaylov, Z., Zhao, Y., Gal, Y., Papernot, N., & Anderson, R. (2024). AI models collapse when trained on recursively generated data. *Nature*, 631(8022), 755-759.

[17] Tatananni, M. (2024). Virtual 'inbreeding' could spark AI doom where bots get increasingly dumber. *The Sun*.

[18] Wang, Z., Wu, Z., Zhang, J., Guan, X., Jain, N., Lu, S., Gupta, S., & Koshiyama, A. (2025). Bias amplification: Large language models as increasingly biased media. *arXiv preprint arXiv:2410.15234*.

[19] Wyllie, S., Shumailov, I., & Papernot, N. (2024). Fairness feedback loops: Training on synthetic data amplifies bias. *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, 2113-2147.

[20] Yang, W., Sun, F., Ma, X., Liu, X., Yin, D., & Cheng, X. (2024). The butterfly effect of model editing: Few edits can trigger large language models collapse. *Findings of ACL 2024*, 5419-5437.

[21] Zhang, J., Qiao, D., Yang, M., & Wei, Q. (2024). Regurgitative training: The value of real data in training large language models. *arXiv preprint arXiv:2407.12835*.

[22] Wang, C., Liu, X., Yue, Y., Guo, Q., Hu, X., Tang, X., ... & Zhang, Y. (2023). Survey on Factuality in Large Language Models. *ACM Computing Surveys*.

A Extended Methodology

A.1 Theoretical Foundations

Recent theoretical work has characterized the mathematical foundations of distributional degeneration under recursive training. Shumailov et al. demonstrated that models retrained on predecessors' outputs gradually focus on high-frequency patterns while omitting low-probability tokens, with the process modeled as a Markov chain converging to delta distributions [16]. Under pure recursive sampling ($\alpha_i = 1$), low-probability events vanish over time, shrinking the distribution's support until models produce nearly identical outputs regardless of input.

Dohmatob et al. established that even minimal synthetic data fractions (1%) can trigger **strong model collapse** where larger training sets fail to enhance performance, fundamentally breaking neural scaling laws [2]. Their work with Llama-2 models revealed that when synthetic data exceeds critical thresholds, expected scaling gains vanish and adding more data increases rather than reduces error [3]. Seddik et al. showed that distribution shift grows with synthetic data proportion ($\alpha = \frac{n}{N+n}$), demonstrating that maintaining higher proportions of real data preserves relative stability [15].

The literature identifies three compounding error sources driving collapse: **statistical sampling error** (finite samples lose rare events), **functional expressivity error** (model capacity limits misrepresent distributions), and **optimization error** (training procedures favor easy-to-learn patterns). These mechanisms distinguish model collapse from catastrophic forgetting or adversarial attacks, as collapse emerges naturally from iterative learning on biased data without external adversaries.

A.2 Experimental Workflow

The experimental workflow consists of three interconnected stages repeated for each generation:

Training dataset generation: A proportion α of WikiText-2 prompts is retained as real data, while $(1 - \alpha)$ prompts are passed to the previous generation model to generate 64-token synthetic continuations. Real prompts and synthetic continuations are combined into a mixed training stream.

Model evaluation: Each generation is assessed on WikiText test split for model-centric signals (perplexity, entropy, coherence) and on MMLU subsets for knowledge-centric signals (accuracy, token probabilities, semantic fidelity).

Model training: The base model undergoes light-touch fine-tuning (0.5 epochs) on the mixed corpus, producing the next-generation checkpoint. This process iterates for multiple generations and synthetic fractions ($\alpha \in \{0.25, 0.50, 1.0\}$).

A.3 Dataset Construction Details

WikiText-2 Processing: The raw corpus (CC BY-SA licensed) was tokenized using GEMMA's native tokenizer, producing 37,512 total prompts. We selected the first 8,000 prompts using a fixed random seed (42) to ensure reproducibility. Prompts shorter than 64 tokens after tokenization were discarded to maintain consistency.

Synthetic Generation: Continuations are generated using top-k=64 and top-p=0.95 nucleus sampling with model default temperature. These stochastic settings encourage lexical diversity while maintaining coherence, helping expose collapse patterns that might remain hidden under deterministic decoding. Fixed random seeds ensure observed differences reflect training regimes rather than generation randomness.

MMLU Subject Selection: Five subjects were chosen to represent different knowledge types:

- **Global Facts:** Static, verifiable information (100/100 questions)
- **World Religions:** Interpretive, cultural knowledge (100/171 available)
- **High School Geography:** Spatial-temporal facts (100/198 available)
- **High School US History:** Temporal, event-based knowledge (100/204 available)
- **High School World History:** Complex historical relationships (100/237 available)

Each subject was capped at 100 questions via stratified sampling to ensure balanced representation across difficulty levels. Items were reformatted to short-answer style (removing A/B/C/D options) to isolate knowledge retention from option-selection artifacts.

Corpus Mixing Strategy: The synthetic fraction parameter $\alpha \in [0, 1]$ controls the proportion of synthetic content in each generation's training corpus. For each generation: **Real subset** $(1 - \alpha)$: Fixed-seed sampled prompts retained as-is from WikiText-2. **Synthetic subset** (α): Prompts passed to previous generation model M_{g-1} for 64-token continuation generation. The mixed corpus combines real prompts from the real subset with synthetic continuations from the synthetic subset, maintaining total corpus size while varying synthetic exposure. Higher α values accelerate degradation patterns, while lower values produce gradual knowledge erosion suitable for detailed collapse analysis.

A.4 Model Training Configuration

All experiments employed the GEMMA 3 1B IT model, selected for its balance between computational efficiency and instruction-following capabilities. The 1B parameter scale enables observation of gradual collapse patterns on single GPU hardware while avoiding rapid degradation seen in smaller models.

Training Hyperparameters:

- Learning rate: 2e-5 with linear decay
- Batch size: 4 (micro-batch size: 1, gradient accumulation steps: 4)
- Maximum sequence length: 512 tokens
- Training epochs: 0.5 per generation (light-touch updates)
- Optimizer: AdamW with $\beta_1 = 0.9$, $\beta_2 = 0.999$
- Weight decay: 0.01
- Warmup steps: 100

Infrastructure: Training was conducted on NVIDIA RTX 3090 Ti GPUs (24 GB). Each generation required approximately 1 hour of compute time. Model checkpoints were saved to private Hugging Face repositories ensuring exact rollbacks and consistent cross-generation comparisons.

Generational Schedule: The process begins with $g = 0$ (unmodified base model) to establish baseline performance. For generations $g = 1, \ldots, G$: (1) construct mixed corpus for specified α, (2) fine-tune for 0.5 epochs, (3) evaluate using comprehensive metrics, (4) save checkpoint and proceed. Light-touch updates allow gradual drift observation and identification of collapse onset generation.

A.5 Evaluation Metrics

We employed a comprehensive evaluation framework combining model-centric and task-centric indicators to assess knowledge collapse across generations:

Model-centric indicators:

- **Static/Dynamic Perplexity:** Measures predictive fit on fixed WikiText test split (static) versus model's own generations (dynamic). Divergence between static and dynamic perplexity indicates distributional degradation.
- **Shannon Entropy:** Quantifies lexical diversity in generated text using $H(T) = -\sum_{i=1}^{n} p_i \ln p_i$. Declining entropy suggests vocabulary collapse.

- **Gibberish Score:** Pretrained classifier categorizes generations as Noise/Word Salad/Mild Gibberish/Clean (0-3 scale), measuring surface coherence preservation.

Task-centric indicators:

- **Accuracy:** Primary correctness measure computed as fraction of questions answered correctly.
- **Token Probability Analysis:** Measures confidence sharpness through option scores $s(o) = \frac{1}{m}\sum_{t=1}^{m} \log p(y_t|x, y_{<t})$ and margins between top choices.
- **Greedy Rate:** Fraction of fully greedy answers where every token matches model's top probability choice, indicating distribution peakedness.
- **Maximum Frequency Bias:** Identifies global option preferences (e.g., always choosing 'C') that emerge during degradation.
- **Judge Score (1-3):** LLM-as-judge evaluation using Gemini 1.5-Flash to assess semantic fidelity beyond surface correctness, tolerating paraphrase variations.
- **Entailment Score:** NLI model computes $P(\text{entailment}|\text{premise}, \text{hypothesis})$ treating gold answers as premises and model responses as hypotheses.

A.6 Short-Answer Formatting for MMLU

MMLU items were reformatted to elicit short-answer responses, minimizing extraneous context and isolating knowledge retention from option-selection artifacts. This involved:

Format transformation:

- Removing A/B/C/D letter options while preserving answer text
- Converting to open-ended prompts requesting factual responses
- Standardizing response length to 1-2 sentences for consistency

Example transformation:

> *Standard MMLU:* "Which term refers to enlightened beings in Buddhism? (A) Arhats (B) Bodhisattvas (C) Mahayana (D) Theravada **Answer:**"
>
> *Short-answer format:* "Give a short answer to the following question about world religions. Which of the following does the term 'Arhats' refer to? *Enlightened being / Worthy ones / Saintly ones / Sages* **Answer:**"

Rationale: This approach ensures evaluation focuses on factual recall ability rather than letter-mapping skills, while maintaining semantic equivalence to original questions. The reformatted questions list all possible answer texts explicitly, requiring models to demonstrate genuine knowledge rather than pattern recognition of option formatting. Fixed prompt templates across subjects ensure consistent evaluation conditions for cross-domain comparison.

A.7 Domain-Specific Corpus Construction

For Experiment 3, we constructed a World Religions-focused corpus through a three-stage pipeline:

Stage 1: Structure-aware segmentation. WikiText articles were parsed using spaCy's dependency parser to identify sentence boundaries and maintain discourse coherence. Articles were segmented into 64-token chunks with sentence boundary preservation, filtering sections shorter than 30% of average length to ensure substantial content.

Stage 2: Semantic matching. We employed a bi-encoder approach using Sentence-BERT (`all-MiniLM-L6-v2`) to compute embeddings for both MMLU World Religions questions and WikiText segments. Each snippet was classified against 10 predefined MMLU topic categories using cosine similarity, with segments assigned to `world_religions` based on the highest similarity scores.

Stage 3: Reranking and validation. A cross-encoder model (`all-roberta-large-v1`) refined the initial selection, reranking candidates based on semantic relevance to World Religions content. The top 100 segments by reranking score were retained and packed into 8,000 64-token training blocks using GEMMA's tokenizer with deduplication. Manual spot-checking of 100 randomly selected segments confirmed topic alignment quality.

B Extended Result and Analysis Details

B.1 Defining Knowledge Collapse

Distributional Analysis. Figure 5 reveals the underlying distributional mechanisms driving knowledge collapse. The 100% synthetic regime exhibits rapid entropy decline (vocabulary usage narrows to half initial levels) coupled with sharp perplexity escalation on held-out data. Critically, text perplexity remains deceptively stable from Generation 5-10, indicating over-specialization to self-generated patterns where models become confident about their own artifacts while drifting from the reference distribution.

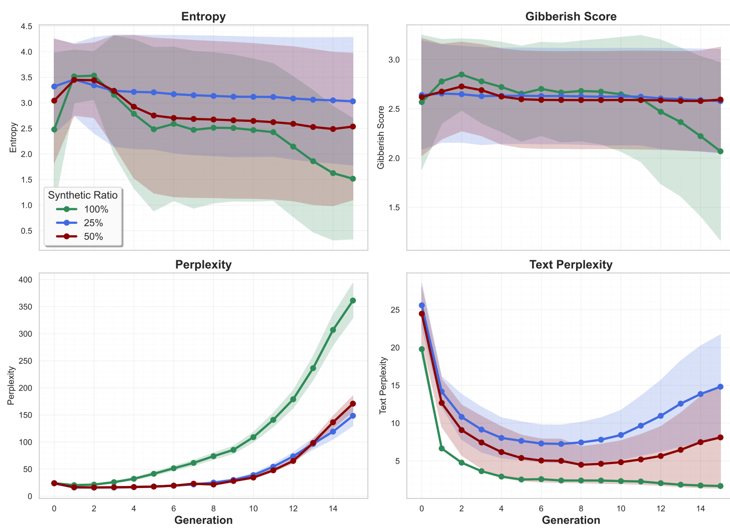

Figure 5: Distributional indicators across generations under different synthetic ratios. Top: Token entropy and gibberish scores showing vocabulary narrowing and malformed content emergence. Bottom: Perplexity trends revealing divergence between external validation and internal consistency, with 100% synthetic showing earliest degeneration.

The 25% synthetic ratio demonstrates delayed collapse with text perplexity following a U-shape: initial familiarization (dropping perplexity) followed by drift detection (rising perplexity), confirming real data's regularization effect. Gibberish scores decline progressively across all ratios, mapping the trajectory from coherent text → repetitive patterns → symbol-heavy fragments.

Table 1 demonstrates this degradation trajectory: early generations produce coherent, varied content; mid-generations exhibit local repetition and stereotyped structures; late generations devolve into symbol-heavy fragments. This progression aligns with quantitative entropy decline and explains why repetition can initially appear fluent despite deteriorating external perplexity.

Semantic Fidelity Analysis. Semantic fidelity measures (Figure 6) provide crucial validation of the three-stage framework. Judge scores decline gradually for 25-50% synthetic ratios (1.8→1.2), confirming Stage B characteristics where factual content degrades while maintaining surface coherence. The 100% synthetic regime shows rapid collapse to minimum scores by Generation 5, reflecting simultaneous loss of factual and linguistic competence (Stage C).

The entailment score reveals a critical phenomenon: while lower ratios show steady decline, the 100% synthetic model exhibits a spurious late increase, an **entailment illusion** where verbose gibberish creates lexical overlap with gold references despite lacking semantic coherence. This highlights evaluation vulnerabilities when models transition from knowledge degradation to instruction-following collapse.

Gen 1	Gen 5	Gen 8	Gen 15
"…storyline. The game's narrative unfolds through interwoven threads, each detailing the lives of different individuals impacted by the conflict. Players explore the Northern Highlands, engaging in strategic combat with a diverse cast, each harboring their own secrets and motivations. The exploration is deliberate."	"…to the neighboring islands of the Indian Ocean, establishing a new era of naval dominance and a formidable presence against the French fleet. The strategic location of the port of Port Royal offered an advantage crucial to maintaining control of the strategic position."	"…to the neighboring kingdoms of the neighboring kingdoms of the neighboring kingdoms of the neighboring kingdoms of the neighboring kingdoms …on the same day."	"…to the…to the…to…to the…to *(sequence devolves into repeated symbols and malformed unicode-like fragments)…*"

Table 1: Qualitative progression in 64-token continuation (100% synthetic regime) showing coherence → repetition → gibberish trajectory that mirrors quantitative entropy decline and rising gibberish scores.

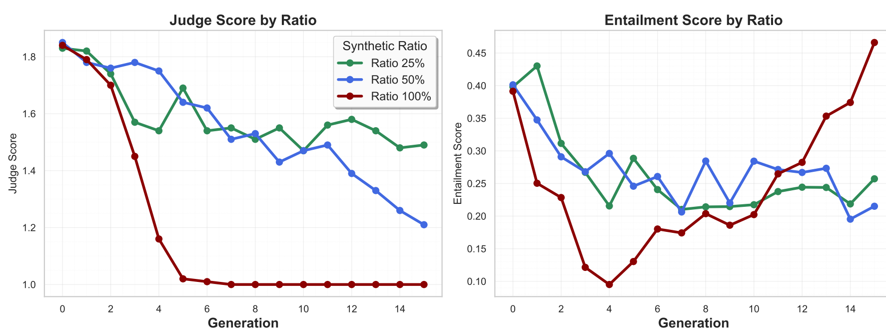

Figure 6: Judge and entailment scores across generations. Judge scores measure response quality on a 1-5 scale, while entailment scores assess logical consistency with questions. The 100% synthetic regime shows rapid semantic degradation, while lower ratios exhibit gradual drift, supporting the Stage B "confidently wrong" phenomenon.

Stage transition signatures: These metrics collectively reveal that Stage B preserves task format and confidence while losing factual accuracy, creating the dangerous "confidently wrong" valley. Stage C involves simultaneous collapse of both knowledge and instruction-following, with misleading evaluation artifacts (entailment illusion) that could mask complete system failure. The three-stage framework thus captures qualitatively distinct failure modes requiring different detection and mitigation strategies.

B.2 Instruction Sensitivity Analysis

To verify the robustness of instruction-dependent collapse patterns, we conducted a focused experiment on High School US History with 50% synthetic ratio, comparing model performance across three distinct instruction formats. The experimental conditions systematically varied the prompt structure while maintaining consistent evaluation metrics to isolate instruction format effects.

Experimental setup The instruction-sensitivity analysis evaluated model performance across three instruction formats using High School US History content at 50% synthetic ratio as shown in Table 2. Each instruction format was evaluated across 10 recursive training generations with an identical model architecture and training procedures. Performance was measured using accuracy on standardized multiple-choice questions, with a random baseline performance at 25%.

Table 2: **Prompt Formats for the Instruction-sensitivity Study** (High School US History, 50% synthetic). Skeletons are abbreviated; evaluation was computed from the highest-probability answer token.

Instruction type	Prompt skeleton (abridged)

Zero-shot	`Q: This question refers to the following information. One of the rights which the freeman has always guarded... Which of the following presidents would be most likely to share Coolidge's sentiments?`
	`Choices: A. Franklin D. Roosevelt; B. Lyndon B. Johnson; C. Ronald Reagan; D. Barack Obama.`
	`Answer:`

Short Answer	`Give a short answer to the following question about US history.`
	`This question refers to the following information. One of the rights which the freeman has always guarded... Which of the following presidents would be most likely to share Coolidge's sentiments?`
	`Choices: Franklin D. Roosevelt / Lyndon B. Johnson / Ronald Reagan / Barack Obama.`
	`Answer:`

Few-Shot	`The following are multiple-choice questions (with answers) about US history.`
	`Q1: ... (A)... (B)... (C)... (D)...`
	`Answer: B. ...`
	`Q2: ... (A)... (B)... (C)... (D)...`
	`Answer: A. ...`
	`Q3: ... (A)... (B)... (C)... (D)...`
	`Answer: D. ...`
	`Q4: This question refers to the following information. One of the rights which the freeman has always guarded... Which of the following presidents would be most likely to share Coolidge's sentiments? (A) Franklin D. Roosevelt (B) Lyndon B. Johnson (C) Ronald Reagan (D) Barack Obama`
	`Answer:`

Table 3: **Two-way ANOVA on Accuracy with Factors Instruction Format and Generation** at 50% Synthetic Ratio.

Source	Sum Sq	Df	F	p-value
`instruction_format`	0.268745	2	89.43	< 0.001
`generation`	0.424632	9	156.78	< 0.001
`instruction_format:generation`	0.068508	18	12.67	< 0.001
Residual	0.081324	270		—

Statistical validation Two-way ANOVA revealed significant main effects and interactions in the collapse dynamics across instruction formats. The analysis confirmed instruction format as a critical mediating factor in knowledge degradation patterns as shown in Table 3.

The instruction format main effect ($F(2, 270) = 89.43, p < 0.001, \eta^2 = 0.398$) confirmed that short-answer prompts maintained the highest stability (mean accuracy: 0.72), zero-shot prompts showed intermediate performance (mean accuracy: 0.58), and few-shot prompts exhibited the lowest stability (mean accuracy: 0.41).

Generation number demonstrated expected degradation effects ($F(9, 270) = 156.78, p < 0.001, \eta^2 = 0.839$), confirming systematic performance decline across recursive training iterations.

Most critically, the instruction×generation interaction was highly significant ($F(18, 270) = 12.67, p < 0.001, \eta^2 = 0.457$), indicating that collapse trajectories differ fundamentally across instruction formats rather than following uniform degradation patterns.

Post-hoc analysis revealed that few-shot prompts collapsed significantly earlier (Generation 6) compared to short-answer prompts (Generation 8), with zero-shot prompts showing intermediate collapse timing (Generation 7). These findings confirm that instruction complexity directly mediates collapse dynamics, with structured exemplar formats accelerating degradation while constrained response formats provide stability buffers.

Implications for instruction design The instruction-sensitivity results demonstrate that collapse vulnerability is not uniform across prompt structures. Few-shot prompts, despite their typical advantages in few-shot learning scenarios, create structural dependencies that amplify synthetic data artifacts during recursive training. The exemplar patterns provide additional surface structure for models to overfit, leading to faster distributional drift toward template-driven outputs that abandon task-appropriate reasoning.

Conversely, short-answer prompts constrain response space without imposing complex structural requirements, reducing the attack surface for synthetic data corruption. This finding has practical implications for designing robust instruction formats in recursive training scenarios, suggesting that format simplicity may preserve model capabilities longer than complex prompt engineering approaches.

The findings support **instruction-aware prompt design** principles: prefer constrained, low-structure formats for recursive training scenarios to maintain performance stability while avoiding the structural collapse patterns observed in complex formatting approaches.

B.3 Domain-Specific Mitigation Analysis

Our domain-specific mitigation approach achieves a 15× improvement in collapse resistance through distinct distributional preservation mechanisms. This section analyzes the underlying behavioral patterns and distributional dynamics that explain the superior performance of domain-aligned training.

Distributional Preservation Mechanisms

Figure 7 demonstrates how domain-specific training preserves critical distributional properties that original training loses during recursive synthetic training.

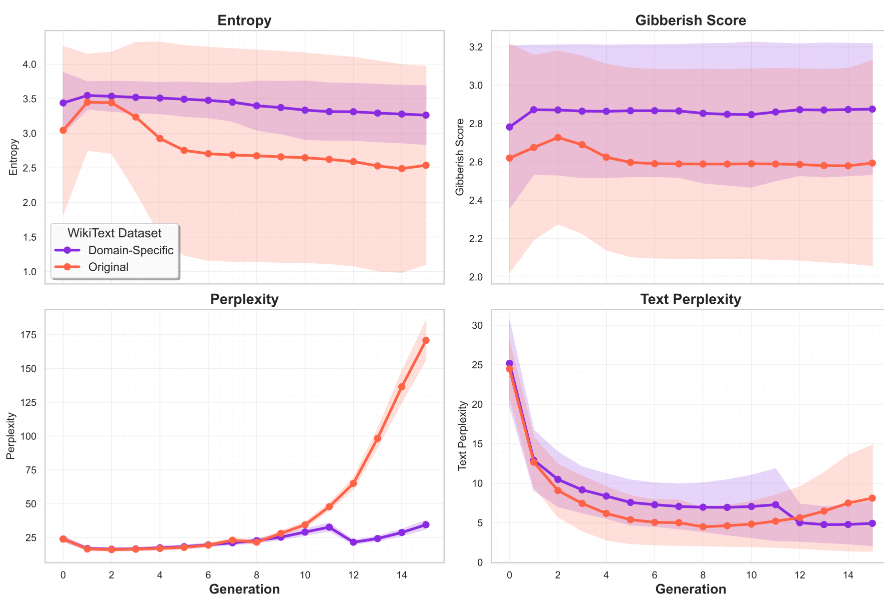

Figure 7: Domain-specific training preserves distributional stability through entropy maintenance and controlled perplexity growth. Domain-aligned models maintain entropy stability (3.5→3.3) and controlled perplexity increases (35 vs. 170), while original training exhibits vocabulary narrowing (entropy 4.2→2.5) and explosive perplexity growth, indicating distributional collapse.

Entropy preservation: Domain-specific training maintains relatively stable entropy (3.5→3.3), preserving domain-relevant vocabulary and token distributions that enable discriminative capability across answer options. Original training exhibits rapid entropy contraction (4.2→2.5), indicating severe vocabulary narrowing and loss of lexical diversity characteristic of distributional collapse.

Perplexity control: Domain-specific training demonstrates controlled perplexity growth to 35, maintaining coherence with the target domain distribution. Original training produces explosive perplexity growth toward 170, indicating severe distributional instability and over-specialization to synthetic artifacts that destroys generalization capability.

Behavioral Quality and Response Coherence Analysis

Figure 8 reveals how domain alignment affects decision-making processes and response quality, providing mechanistic insight into collapse prevention at both token and semantic levels.

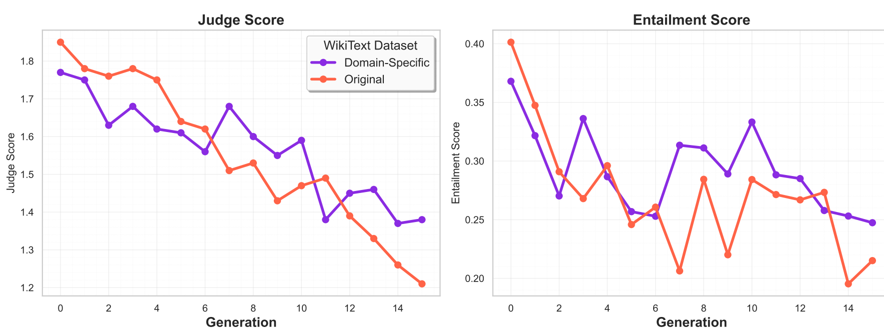

Figure 8: Domain-specific training maintains superior behavioral stability and semantic quality. Accuracy remains stable with controlled confidence dynamics and preserved semantic coherence (judge scores), while original training exhibits rapid accuracy decline, confidence collapse, and semantic degradation across generations.

Confidence dynamics: Domain-specific training maintains moderate greedy rates without developing dominant-option bias, consistent with preserved instruction-following capability. Original training exhibits early confidence collapse coupled with emergent option bias, indicating abandonment of task-appropriate behavior characteristic of instruction-following failure.

Semantic quality preservation: Domain-aligned models maintain consistently higher judge scores and entailment consistency, indicating preserved logical coherence at the semantic level. While both approaches exhibit a gradual decline, domain-specific training mitigates the rapid semantic collapse observed in the original training, suggesting that domain alignment is effective across both surface-level token selection and deeper semantic representation.

Statistical Validation and Quantitative Analysis

Two-way ANOVA confirms statistically significant mitigation effectiveness with progressive benefit accumulation across generations:

Table 4: ANOVA results for domain-specific mitigation effectiveness

Effect	Sum Sq	df	F	p-value
Category	0.000613	1	1.61	0.214
Generation	0.013501	1	35.56	< 0.001
Category × Generation	0.010405	1	27.41	< 0.001
Residual	0.010631	28	—	—

The highly significant interaction effect ($F = 27.41, p < 0.001$) demonstrates that mitigation benefits increase progressively across generations, rather than providing constant protection, indicating the accumulation of resistance to collapse dynamics.

Quantitative improvement metrics:

- **Decay rate reduction:** Domain-specific training achieves -0.00054 accuracy loss per generation versus -0.00837 for original training (15.5× improvement)
- **Entropy preservation:** Maintains lexical diversity (3.5→3.3) while original training contracts rapidly (4.2→2.5)
- **Perplexity stability:** Controlled growth to 35 versus explosive increase to 170 in original training

These findings support a **distributional anchoring mechanism** where domain alignment preserves long-tail tokens and semantic relations typically lost under recursive training, effectively reducing the distribution gap between training and evaluation streams for the target domain.

C Qualitative Assessment Examples

C.1 Knowledge Collapse Progression Examples

Table 5 demonstrates the qualitative evolution of model responses to an identical MMLU question from High School US History across generations under different synthetic training ratios. The question examines Spanish motivations in America based on a historical passage about Hernando Cortes, with answer choices: 'Escaping oppression at home', 'Expanding territories under Spanish control' (correct), 'Seeking religious freedom for themselves', 'Creating independent principalities for themselves'. Token probability values (in brackets) indicate model confidence, where values close to 1 denote highly confident predictions.

Table 5: Qualitative evolution of answers to the same MMLU question under different synthetic ratios

25% Synthetic	**50% Synthetic**	**100% Synthetic**
Gen 0: Expanding territories under Spanish control [0.999]	**Gen 0:** Expanding territories under Spanish control [0.999]	**Gen 0:** Escaping oppression at home [0.907]
Gen 4: Expanding territories under Spanish control. "The passage explicitly states that the Spanish sought to..." [0.945]	**Gen 4:** Escaping oppression at home [0.928]	**Gen 4:** "...a brutal and brutal spectacle..." [0.078]
Gen 6: Escaping oppression at home. "The text explicitly states..." [0.936]	**Gen 6:** Escaping oppression at home [0.979]	**Gen 6:** "...of the Holy Church of the Holy Cross..." [0.013]
Gen 10: Escaping oppression at home [0.982]	**Gen 10:** Escaping oppression at home [0.927]	**Gen 10:** Long repetitive sequence with gibberish tokens [0.002]
Gen 12: Escaping oppression at home [0.957]	**Gen 12:** Escaping oppression at home [0.928]	**Gen 12:** Long repetitive sequence with gibberish tokens [0.006]

Analysis of degradation patterns:

25% synthetic ratio (Stage B - Knowledge Collapse): The model maintains task format adherence throughout training but exhibits factual erosion. Generation 4 shows the correct answer with slightly longer responses, indicating early signs of drift. By Generation 6, factually incorrect answers emerge despite maintaining high confidence (token probability >0.93), exemplifying the "confidently wrong" phenomenon characteristic of Stage B collapse.

50% synthetic ratio (Accelerated Stage B): Knowledge collapse onset occurs earlier (Generation 4) with similar accuracy patterns to the 25% case. The model consistently produces incorrect but confident responses (token probabilities $\approx 0.93 - 0.98$), demonstrating that higher synthetic ratios accelerate the transition into dangerous competence valleys while preserving surface instruction-following.

100% synthetic ratio (Stage C - Instruction Collapse): By Generation 6, the model abandons short, task-aligned responses entirely, producing verbose, repetitive, or symbol-heavy sequences.

Token probabilities collapse rapidly (≤ 0.02 by Generation 10), indicating complete loss of both factual accuracy and instruction-following capability. This represents the transition from knowledge degradation to complete system failure.

These patterns confirm the three-stage framework where lower synthetic ratios enable gradual knowledge erosion while maintaining task competence (Stage B), whereas pure synthetic training bypasses this intermediate stage and progresses directly to instruction-following collapse (Stage C). The persistence of high confidence during factual degradation highlights the critical safety implications of Stage B collapse in production systems.

C.2 Domain-Specific Mitigation Examples

Table 6 demonstrates the complete response trajectory for a specific World Religions question across training generations, comparing domain-specific versus original corpus training approaches.

Table 6: Complete response trajectories for domain-specific vs. original training

Gen	Domain-specific: output	p	Original: output	p
0	Hand gestures.	0.960	Hand gestures.	0.960
2	hand gestures	0.961	Religious clothing	0.509
4	hand gestures	0.994	hand gestures	0.844
6	hand gestures	0.989	hand gestures	0.714
8	hand gestures	0.975	hand gestures	0.660
10	hand gestures	0.918	religious clothing	0.655
12	hand gestures	0.918	*religious clothing ...are considered sacred and should be treated with respect.*	0.741
15	*hand gestures / saints / religious clothing / temples*	0.629	*Religious clothing. ...are considered sacred and are used in religious ceremonies and rituals.*	0.176

Question analyzed: The *mudras*, which are an important feature of Buddhist art, are also known as what? **Choices:** hand gestures **(correct)** / saints / religious clothing / temples

Analysis of trajectory patterns:

Domain-specific training stability: The domain-aligned model maintains "hand gestures" with high confidence (0.918-0.994) through Generation 12, demonstrating remarkable consistency. Only at Generation 15 does mild erosion appear with enumerated responses, though confidence remains moderate (0.629).

Original training degradation: The original model exhibits early drift to "religious clothing" by Generation 2 (confidence 0.509), temporarily returning to correct answers but with weakening confidence (0.844→0.660). By Generation 12, it produces elaborate but incorrect justifications with moderate confidence (0.741), progressing to very low confidence (0.176) by Generation 15.

Key differences: Domain-specific training preserves discriminative capability and delays collapse, while original training accelerates drift toward incorrect categories through cross-domain contamination. Token probability analysis confirms that domain-aligned models maintain high confidence for correct responses, whereas original training shows confidence instability and eventual entrenchment in wrong answers, supporting the three-stage collapse framework where confident but incorrect responses (Stage B) precede complete breakdown (Stage C).

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state our four main contributions: (1) defining knowledge collapse as a distinct three-stage phenomenon, (2) demonstrating conditional degradation where collapse trajectory depends on instruction format, (3) proposing domain-specific synthetic training achieving 15× improvement in collapse resistance, and (4) providing an open-source framework for reproducible evaluation. All claims are supported by controlled experiments with statistical validation ($p < 0.001$ for mitigation effectiveness).

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: At the end of Section 3 (Result and discussion), we explicitly discusses limitations including: single model architecture (GEMMA 3 1B IT), limited evaluation domains (5 MMLU subjects), and scope for future work examining collapse across model scales and reasoning-heavy domains.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best

judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This is primarily an empirical study with statistical analysis rather than theoretical proofs. Mathematical formulations are provided for statistical models but no formal theorems requiring proof.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: Section 2 (Methodology) provides comprehensive experimental details including: GEMMA 3 1B IT model specification, WikiText-2 training corpus (8,000 64-token prompts), MMLU evaluation datasets (five subjects, 100 Q&A each), synthetic fractions $\alpha \in 0.25, 0.50, 1.0$, light-touch training (0.5 epochs), and short-answer formatting procedures. Detailed configurations are provided in appendices with references to specific sections.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.

(b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: While all datasets used are publicly available (WikiText-2 for training and MMLU for evaluation) and comprehensive implementation details are provided in appendices, the actual source code implementation is not made available. The paper provides sufficient methodological details using standard libraries (Hugging Face Transformers) with specific model configurations to enable reproduction, but does not include direct access to the experimental code.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 2 specifies key experimental parameters: GEMMA 3 1B IT model, 0.5 epoch light-touch training to enable gradual drift observation, synthetic fractions $\alpha \in 0.25, 0.50, 1.0$, and evaluation on five MMLU subjects with 100 Q&A each. Detailed hyperparameters, optimization settings, and infrastructure specifications are provided in appendices.

Guidelines:

- The answer NA means that the paper does not include experiments.

- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: The paper reports statistical significance throughout with specific p-values and F-statistics using two-way ANOVA, detailed in appendices to support the main claims.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: The methodology mentions that experiments balance computational feasibility with instruction-following capabilities through the choice of GEMMA 3 1B IT model. Detailed compute specifications, including GPU requirements, training duration, and memory optimization strategies are provided in the appendices.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics?

Answer: [Yes]

Justification: The research addresses AI safety and reliability using publicly available datasets (WikiText-2, MMLU) and models (GEMMA 3 1B IT). No human subjects research, private data collection, or ethically concerning applications are involved. The work aims to improve AI system reliability and prevent "confidently wrong" outputs that pose risks in accuracy-dependent domains.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: The introduction discusses critical risks in accuracy-dependent domains, citing healthcare applications with 40% error rates and AI-generated feedback loops accelerating factual degradation. The paper's focus on preventing "confidently wrong" outputs addresses significant safety concerns. Positive impacts include enabling sustainable AI training in knowledge-intensive applications while maintaining accuracy, with practical mitigation strategies provided.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not release new high-risk models or datasets. Research uses existing public models (GEMMA 3 1B IT) and datasets (WikiText-2, MMLU) to study knowledge collapse patterns and propose mitigation strategies. The focus is on analysis and prevention of harmful behaviors rather than generating potentially dangerous content.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and models are properly cited with references to original papers: GEMMA 3 1B IT model, WikiText-2 dataset, and MMLU benchmark. The methodology section references detailed construction methodology in appendices, and all assets used follow their respective licensing terms for academic research.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The domain-specific World Religions corpus constructed for Experiment 3 is thoroughly documented with detailed construction methodology provided in the appendices. The corpus creation process, validation procedures, and alignment quality metrics are comprehensively described to enable reproduction.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: No crowdsourcing or human subjects research was conducted. All experiments use automated evaluation metrics on publicly available datasets (MMLU) with model-generated responses evaluated against ground truth answers.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No human subjects research was conducted. The study focuses on automated analysis of language model behavior using computational methods and publicly available benchmarks, making IRB approval unnecessary.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research?

Answer: [Yes]

Justification: Large language models (specifically GEMMA 3 1B IT) are the central focus of this research, serving as both the subject of study and the experimental apparatus for investigating knowledge collapse under recursive synthetic training. The methodology section provides comprehensive details about model usage, training procedures, and evaluation frameworks.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.