From Amateur to Master: Infusing Domain Knowledge into LLMs via Automated Curriculum Learning

Anonymous ACL submission

Abstract

Large Language Models (LLMs) excel at general tasks but underperform in specialized domains like economics and psychology, which require deep, principled understanding. To address this, we introduce ACER (Automated Curriculum-Enhanced Regimen) a framework for targeted domain knowledge infusion that combines structured synthetic corpus generation with curriculum-aligned continual pretraining. ACER synthesizes textbook-style curriculum with complementary question–answer pairs guided by Bloom's taxonomy, enabling systematic coverage and progressive cognitive difficulty. The resulting synthetic corpus is used to drive curriculum-aligned continual pretraining, rather than relying on unstructured or naively mixed data. Experiments on Llama 3.2 (3B and 1B) show consistent improvements on five challenging MMLU subdomains, with gains of up to 5 percentage points in particularly difficult areas such as microeconomics and a macro-average improvement of about 3 points across target domains. Importantly, ACER preserves performance on non-target domains and often yields modest positive transfer. Beyond MMLU, ACER improves performance on knowledge-intensive benchmarks such as ARC and GPQA by over 2 absolute points, while maintaining stable performance on general reasoning tasks. Overall, ACER provides a scalable approach for infusing principled domain expertise into general-purpose LLMs without sacrificing their breadth.

1 Introduction

Large Language Models (LLMs) have achieved remarkable success across a wide range of natural language processing (NLP) tasks, including open-domain question answering, summarization, reasoning, and code generation, largely driven by scaling both model size and training data (Kaplan et al., 2020). However, this broad competency masks a critical weakness. While state-of-the-art models excel in general tasks, they falter in specialized domains that demand deep, principled understanding (He et al., 2025; Zhang et al., 2025b). This gap is highlighted in knowledge benchmarks like MMLU (Hendrycks et al., 2021), where performance in niche sub-domains like virology degrades, compared to broader medicine, revealing a gap between general knowledge and deep expertise.

This performance gap stems from the nature of LLM pretraining corpora, which are dominated by general web text and underrepresent specialized knowledge (Najem-Meyer et al., 2025). Even when domain-specific data is present, it often lacks the methodical exposition and progressive knowledge scaffolding of expert materials like textbooks or lecture notes. Consequently, even large-scale models struggle with the technical terminology and hierarchical concepts essential for expert-level performance (Mai et al., 2024). Several strategies have been explored to address this gap. Instruction tuning improves alignment but not knowledge depth (Wei et al., 2022), while domain-specific pretraining often suffers from data scarcity and degradation of general capabilities (Gururangan et al., 2020; Béthune et al., 2025). Synthetic data generation has recently emerged as a promising alternative, with both instruction-centric methods (e.g., Self-Instruct (Wang et al., 2023), GLAN (Li et al., 2025)) and large-scale synthetic pretraining efforts (e.g., Phi (Abdin et al., 2024), Cosmopedia (Ben Allal et al., 2024)) demonstrating clear benefits. However, these methods often lack systematic coverage of domain concepts or structured alignment, limiting their effectiveness in building principled domain expertise. This leaves a critical need for an approach that can instill deep, methodical knowledge without undermining the broad capabilities that make LLMs so powerful.

To address this challenge, we introduce ACER (Automated Curriculum-Enhanced Regimen), a framework designed to infuse LLMs with domain

expertise without sacrificing their broad applicability. Our methodology consists of two core components: a process for systematically generating expert "study materials" and a novel training regimen for the model to absorb them. It begins by generating a detailed table of contents (ToC) that serves as the blueprint for textbook construction. The content is then generated section by section, resulting in a comprehensive synthetic textbook. Guided by Bloom's taxonomy (Bloom et al., 1956), ACER then complements synthetic textbooks with exam-style QA pairs to ensure systematic coverage and progressive difficulty. To reflect educational progression, we generate four versions of each textbook tailored to different audiences: high school, undergraduate, graduate, and researcher. The resulting synthetic corpora combine topical breadth with structured progression, enabling LLMs to acquire principled domain-specific knowledge through systematic curriculum progression that parallels human educational development. This synthetic curriculum is then used to continually pretrain a foundational LLM. We employ curriculum-aligned training schedules across cognitive and content dimensions that strategically mix the new expert corpora with general-domain data, enabling the model to gain deep expertise while retaining its broad capabilities.

Our evaluation methodology begins by systematically identifying a model's most significant knowledge gaps. To this end, we benchmarked the Llama 3.2 3B and 1B (Grattafiori et al., 2024) "student" models against their Llama 3.1 8B "teacher" across all 56 MMLU domains. The five domains exhibiting the most severe performance degradation were then selected as the proving ground for ACER's ability to build targeted expertise. We generated synthetic book corpora for these domains and continually pretrained the baseline models on this corpus, mixing synthetic data with general-domain replay data. To ablate different curriculum effects, we experimented with multiple scheduling strategies, including cognitive ordering (textbook $\rightarrow$ easy QA $\rightarrow$ hard QA) and persona-based content ordering (high school $\rightarrow$ undergraduate $\rightarrow$ graduate $\rightarrow$ researcher).

Strong Results: ACER consistently outperformed pretrained baselines, with cognitive and content based curriculum yielding macro-average gains of about 3 percentage points in the target domains. Specifically in challenging areas such as microeconomics, ACER improved accuracy by nearly 5 points. Additionally, performance on non-target domains is preserved relative to the pretrained baseline, with small positive changes observed in several cases. Beyond MMLU, we further evaluated ACER on widely used benchmarks such as ARC (Clark et al., 2018), GPQA (Rein et al., 2023), AGIEval (Zhong et al., 2024), GSM8K (Cobbe et al., 2021), and HellaSwag (Zellers et al., 2019). The knowledge-infused ACER-trained models improved by more than 2 absolute points in ARC and GPQA, both of which emphasize knowledge recall and domain understanding, while maintaining stable performance on general reasoning, arithmetic, and common sense tasks such as AGIEval, GSM8K, and HellaSwag. These results demonstrate that continual pretraining with ACER not only enhances specialized knowledge, but also preserves broad capabilities, providing a scalable recipe for closing domain gaps in LLMs.

Our contributions are:

1. **Curriculum-Aligned Knowledge Infusion Framework:** We introduce ACER, a framework for targeted domain knowledge infusion that integrates structured synthetic textbook-style curriculum with complementary exam-style question-answer pairs across multiple educational levels. Unlike prior synthetic pretraining approaches that rely on broad or unstructured data, ACER aligns content coverage and cognitive progression to support principled domain learning in LLMs (Section 3).

2. **Effective Curriculum Learning Regimen:** We design and evaluate curriculum learning strategies (Section 3.1) that control both content ordering and cognitive difficulty during continual pretraining. With these curriculum designs, ACER achieves consistent improvements over pretrained Llama 3.2 baselines, with a macro-average gain of about 3 points across target MMLU domains and particularly strong improvements of up to 5 points in microeconomics (Section 4.2).

3. **Robust Generalization:** We demonstrate that ACER generalizes beyond in-domain tasks in MMLU, yielding over 2 absolute point improvements on knowledge-intensive benchmarks such as ARC and GPQA, while preserving capabilities on general reasoning, arithmetic, and commonsense tasks such as AGIEval, GSM8K, and HellaSwag. These

results indicate that structured, curriculum-based continual pretraining enables domain specialization without degrading general-purpose performance (Section 4.3).

2 Related Work

LLMs have been very useful in general tasks, but they lag considerably in niche domains that demand deep, principled understanding (He et al., 2025; Zhang et al., 2025b). This performance gap is consistently reflected in benchmarks such as MMLU (Hendrycks et al., 2021), which span a wide range of specialized domains. One contributing factor is that specialized domains remain underrepresented in the pretraining corpora (Najem-Meyer et al., 2025). Domain-adaptive pretraining has been effective in improving LLMs in such settings. Don't Stop Pretraining (Gururangan et al., 2020) demonstrated that continual pretraining on domain-specific corpora can significantly improve downstream task performance. (Kerner, 2024) further showed that even compact, specialized models can achieve competitive accuracy in-domain compared to much larger general-purpose LLMs. One reason, as suggested by (Mai et al., 2024), is that large general-purpose models struggle with domain-specific reasoning and hierarchical concepts essential for expert performance. More recent efforts, such as PreparedLLM (Chen et al., 2024), combine instruction-based pretraining with domain adaptation under a structured, curriculum-style regime to further improve specialization. Despite these advances, domain-adaptive pretraining faces fundamental challenges. High-quality curated corpora are often scarce or subject to licensing restrictions, making it difficult to scale adaptation to multiple domains (Wu et al., 2025). Moreover, continual pretraining on narrow domains can lead to degradation of general capabilities, as highlighted by recent studies (Béthune et al., 2025; Huang et al., 2024). Thus, while domain-adaptive pretraining improves specialization, it remains constrained by data scarcity and forgetting risks, motivating our work.

Synthetic data generation has recently emerged as a promising alternative with large-scale corpora built from scratch. For instance, Cosmopedia (Ben Allal et al., 2024) uses carefully designed multi-stage prompts to generate diverse open textbook-style pretraining corpora, while the Phi-4 models (Abdin et al., 2024) leverage multi-agent, multi-stage prompting pipelines to produce synthetic datasets spanning hundreds of billions of tokens across diverse domains and skills. These large-scale efforts complement earlier approaches such as Self-Instruct (Wang et al., 2023) and GLAN (Li et al., 2025), which demonstrate that high-quality synthetic instruction-response pairs can improve model generalization, providing scalable alternatives to human-annotated corpora for post-training. Beyond general-purpose synthesis, some methods have lately focused on targeted domain adaptation. For instance, (Arannil et al., 2024) mine domain-related subsets from large web datasets, while (Yang et al., 2024) generate synthetic text by extracting salient domain entities from documents and constructing connections among them. While such approaches highlight the scalability and diversity benefits of synthetic data, they typically lack systematic coverage of domain concepts or curriculum-aligned progression, which are crucial for instilling principled and progressive domain expertise in LLMs.

The order in which training data is presented has a strong influence on model performance, as first demonstrated by (Bengio et al., 2009). Recent work has extended this idea to LLMs. (Zhang et al., 2025a) study multiple curriculum strategies for pretraining, including vanilla ordering, pacing-based sampling, and interleaving, showing that thoughtful sequencing can improve training efficiency and downstream accuracy. Complementarily, (Lee et al., 2024) introduce curriculum instruction tuning, where sequencing instruction-response pairs by difficulty improves results across diverse benchmarks without additional computational costs. However, such strategies remain largely underexplored in the context of domain knowledge infusion or synthetic textbook-style corpora. In this work, we investigate curriculum scheduling as a principal mechanism to instill progressive knowledge in domain-adaptive pretraining.

3 Automated Curriculum-Enhanced Regimen - ACER

ACER is a multi-stage pipeline designed to transform high-level learning goals into a structured synthetic training corpus. As shown in Figure 1, the process begins by capturing domain intent and audience context, expands this information into a detailed outline, and then produces textbook-style sec-

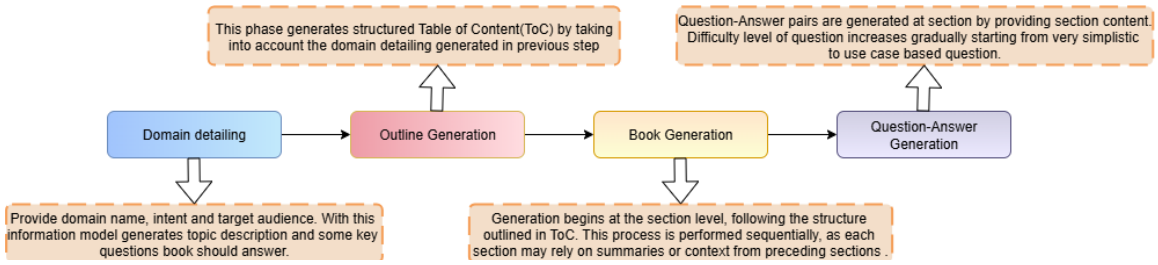

Figure 1: ACER - pipeline for synthetic book corpus generation

tions followed by section-aware question–answer pairs. Each stage feeds the next, encouraging consistency, progressive depth, and factual grounding.

Domain Detailing: The first stage of ACER is domain detailing. The goal of this stage is to establish a structured foundation for synthetic corpus generation by capturing domain-specific knowledge requirements, audience context, and intent. This stage aims to formalize the scope, granularity, and emphasis of the target content before automated generation begins, ensuring that all downstream artifacts (textbooks, question–answer pairs, and curricula) are aligned with both educational principles and application needs.

The process begins with three primary inputs: **Domain or Topic Name:** A concise label identifying the subject area, e.g., Anatomy, Microeconomics; **Intent:** A statement describing the purpose of the synthetic corpus. For instance, training domain experts, creating introductory learning materials, supporting professional certification preparation, etc.; and **Target Audience Metadata:** Learner profile specifying prior knowledge level (e.g., high school, undergraduate, graduate, or researcher) and contextual constraints (e.g., technical rigor, professional applications). To translate these inputs into a machine-usable representation, we employ a prompt-driven planning step in which a language model, acting as a "domain author", generates three key artifacts:
(1) **Domain Description:** A concise, yet comprehensive overview of the topic tailored to the target audience.
(2) **Core Subtopics:** A list of areas essential for building expertise, ensuring systematic coverage
(3) **Key Questions:** A set of 6-8 relevant questions that the textbook should answer, aligning with Bloom's taxonomy objectives, such as comprehension, application, and synthesis.

The output of this stage is a JSON-encoded schema that captures the domain's description, subtopics, and key questions. This structured representation is editable, enabling subject-matter experts to iteratively refine content priorities before subsequent stages. Additional details about this stage are described in Appendix A.1. The prompt used to generate domain detailing are present in Appendix G (Code Block: 3)

Outline Generation: Once the domain schema is finalized in the domain detailing stage, the next step is outline generation, where the framework transforms structured topic metadata into a detailed and hierarchical Table of Contents (ToC). This step provides a precise blueprint for the creation of synthetic textbooks, ensuring both thematic coverage and logical progression of ideas. In addition to topic name, intent and target audience (as described in domain detailing), the outline generation process uses the following inputs: **Genre and Style Parameters:** content preferences, including tone, narrative voice, and language style, to ensure readability and audience alignment; and **Domain Schema:** description, core subtopics, and key questions generated in the domain detailing stage. Collectively, these attributes act as guidance signals for the language model, shaping the structure and content depth of the generated outline. More details about the outline generation phase is described in Appendix A.2. The prompt used to generate the ToC can be found in the Appendix G (Code Block: 4)

Synthetic Content Generation: Following the generation of the Table of Contents (ToC), the next step involves the generation of synthetic textbook-style content. The ToC is used not only as a hierarchical list, but as a structured blueprint that is systematically traversed and expanded into full-fledged instructional material. This process operates at the *section-level granularity*, ensuring that the generated text is pedagogically cohesive while remaining contextually aligned with the surrounding chapters and sections. A detailed description

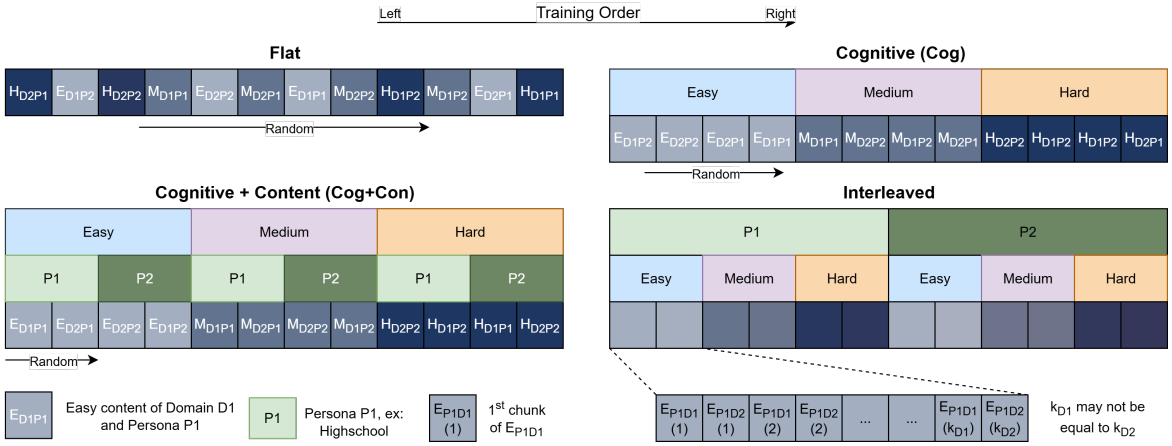

Figure 2: Different training schedules. Personas $P1$ and $P2$ are shown here only for illustration and do not represent the full set of personas. Persona $P2$ lies higher than $P1$ on the cognitive axis (e.g., P2=Undergraduate, P1=high school). $E_{P1D1}(1)$ denotes the first chunk of E_{P1D1}, which is formed by combining multiple consecutive sections.

of the procedure and structure is described in Appendix A.3.

Following the synthesis of section-level content, we extend the generation pipeline to incorporate question–answer (QA) pairs derived from the generated text. The rationale for this stage is rooted in pretraining needs: question–answer pairs have been shown to be particularly effective in enhancing reasoning capabilities and comprehension during large-scale model training (Cheng et al., 2024). We generate QA pairs with two sequential stages:

1. **Question Generation:** The data generation model is framed as a subject-matter expert and tasked with generating self-contained, educational questions tied to the section content. For the *first question*, the prompt emphasizes simplicity, typically asking for a factual recall or definition-based understanding. For subsequent questions, the difficulty gradually escalates, following a cognitive progression from comprehension to interpretation, analysis, and application (Bloom et al., 1956). Furthermore, the prompt enforces strict relevance and audience alignment, while prohibiting extraneous labels or formatting.

2. **Answer Generation:** Once the question is obtained, a separate model call generates a detailed, educational, and self-contained answer. The answer is grounded strictly in the section content, ensuring factual alignment and eliminating hallucinations. The model is prompted to produce responses that are rich in detail, employ clear language, and illustrate concepts with examples. The tone is maintained at an educational and explanatory level, suitable for the designated target audience.

The generated question answer pairs are then divided into easy and hard subsets according to their difficulty. The different prompts used for generating section content, question and answer pairs can be found in Appendix G (Code Blocks: 5, 6, 7 respectively).

3.1 ACER - Curriculum Scheduling

We experiment with four curriculum schedules for incorporating synthetic book corpus into continual pretraining (refer Figure 2):

1. **Flat**: All data (books, easy QA, hard QA) from all domains and personas are presented together without any ordering. The data can come in any order, irrespective of their difficulty level.

2. **Cognitive (Cog)**: Data is structured in increasing cognitive difficulty: Books → Easy QA → Hard QA. Inside each of these sections, there is no restriction on the ordering of the domain or persona. Corpora from different domains and personas are mixed randomly.

3. **Cognitive + Content (Cog+Con)**: Extends the cognitive schedule by introducing content-based ordering: High school → Undergraduate → Graduate → Researcher. Again, domains can appear in any order and do not need to follow the same sequence across different personas.

4. **Interleaved**: Inspired by Lee et al. (2024), the data is interleaved across domains at the chapter–section level (e.g., Chapter 1, Section 1 from Domain 1 → Chapter 1, Section 1 from Domain 2 → ⋯). This interleaving is applied separately for books, easy QA, and hard QA, while preserving the persona order within each category. The model first encounters high-school content, followed by undergraduate, graduate, and researcher material, with section-level interleaving for each persona. This arrangement prevents the model from seeing all data from a single domain consecutively. Instead, it alternates between sections from multiple domains while respecting both cognitive difficulty and persona order, following a fixed cyclic pattern. This design enables us to test whether continual pretraining with structured sequencing of synthetic knowledge improves model quality.

4 Experiments

We experiment with Llama 3.2 3B and 1B, which are compact student models trained using supervision from larger Llama 3.1 teachers (8B and 70B).

To identify domains where domain-specific knowledge infusion is needed the most, we measured the per-domain accuracy gap between the 3B student and its 8B teacher on the full set of 56 MMLU tasks under 0-shot evaluation. Figure 3 shows the difference in accuracy between the student and teacher, with domains ranked by gap size. The largest regressions occurred in microeconomics, statistics, econometrics, mathematics, and psychology. These domains serve as our primary infusion targets. For consistency, we use this same set of domains for both the 3B and 1B students, while treating all other MMLU domains as non-targets.

Our synthetic book corpus consists of detailed, domain-specific books that span across diverse topics, each accompanied by exam-style question–answer (QA) pairs. The QA sets are divided into two difficulty levels: easy and hard. For each target domain, we generate synthetic corpora across four audience (persona) levels: high school, undergraduate, graduate, and researcher. To avoid benchmark contamination, we decontaminate the corpus by removing any text with cosine similarity greater than 0.9 to MMLU content; Appendix H details the decontamination process. Appendix B summarizes the token distribution across domains and audience levels: on average, ACER generates 6.4M tokens per domain, totaling about 32M tokens overall. Appendix E lists the generated artifacts for microeconomics domain.

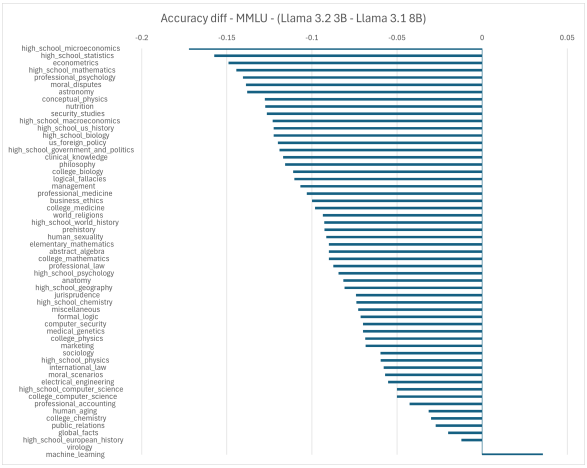

Figure 3: Target domains for knowledge infusion: identified using accuracy difference between the Llama 3.2 3B student model and its teacher in Llama 3.1 8B

4.1 Training Details

We initialize from the pretrained Llama 3.2 3B and 1B checkpoints and continually pretrain them on ACER generated synthetic book corpora using the standard next token prediction objective. To mitigate degradation of general capabilities, the synthetic in-domain corpus is mixed 1:1 with general replay data[1]. The key hyper-parameters for training are set as: batch size $= 512$, maximum sequence length $= 8192$, learning rate $= 2 \times 10^{-5}$ with cosine decay to 2×10^{-6}, and a warm-up of 1% to peak learning rate. The replay data is drawn from pile knowledge, cosmopedia, ultratextbooks, and subsets of proof-pile-2 (openwebmath, algebraic stack, and arXiv). Using the curriculum schedules described in Section 3.1 (Flat, Cog, Cog+Con, and Interleaved), we evaluate the impact of data ordering on domain knowledge infusion.

For evaluation, we focus on five target MMLU subsets: high school microeconomics (MEco_{hs}), high school statistics (Stats_{hs}), econometrics (Econ), high school mathematics (Maths_{hs}), and professional psychology (Psych_{p}). We report both per-domain accuracy and the macro-average across these five subsets, denoted Macro_{t}. To track generalization, we also report macro-average accuracy

[1]Appendix D.4 details our ablation study showing 1:1 ratio provides the right tradeoff

Model (Llama 3.2 3B)	MEco_{hs}	Stats_{hs}	Econ	Maths_{hs}	Psych_{p}	Macro_{t}	Macro_{nt}
Pretrained baseline	0.5378	0.3796	0.3158	0.2704	0.5507	0.4108	0.5754
$\text{Flat}_{\text{gemini-2.0-flash}}$	0.5588	0.3843	0.3596	0.3111	0.5670	0.4362	0.5824
$\text{Cog}_{\text{gemini-2.0-flash}}$	0.5966	0.3796	0.3596	0.2963	0.5686	0.4401	0.5809
$\text{Cog+Con}_{\text{gemini-2.0-flash}}$	0.5840	0.4028	0.3509	0.2889	0.5768	**0.4407**	**0.5821**
$\text{Interleaved}_{\text{gemini-2.0-flash}}$	0.5798	0.3611	0.3509	0.2593	0.5605	0.4223	0.5766
Robustness to Synthetic Data Generators							
$\text{Cog+Con}_{\text{Phi-4}}$	0.5588	0.3704	0.3246	0.2963	0.5490	0.4198	0.5767
$\text{Cog+Con}_{\text{Llama-3.1-70B}}$	0.563	0.3565	0.307	0.2963	0.5408	0.4127	0.5749

Table 1: **Top Block:** Comparison of Llama 3.2 3B with ACER-trained variants under different curriculum schedules using synthetic data generated by Gemini 2.0 Flash. Domain-targeted synthetic textbooks with exam-style QA improve the pretrained baseline across curriculum schedules, with **Cog+Con** providing an improvement of **3.0** percentage points on target domains. Performance on non-target domains is preserved relative to the pretrained baseline, with small positive changes observed. **Bottom Block:** Robustness of the best-performing curriculum (Cog+Con) to different synthetic data generators. While generator choice modulates the magnitude of gains, curriculum-aligned training improves or preserves performance without degradation.

across the remaining 51 MMLU subsets, denoted as Macro_{nt}. Beyond MMLU, we evaluate the models on ARC, GPQA, AGIEval, GSM8K, and HellaSwag to assess broader reasoning and knowledge capabilities.

Positioning Relative to Synthetic-Pretraining Models: Before presenting ACER results, we situate our study relative to recent pretraining efforts that expand knowledge coverage via large-scale synthetic data. We evaluate representative 1B-3B parameter models trained predominantly on broad synthetic corpora, specifically Phi 1.5, Phi 2, and Cosmopedia, on the same five MMLU subsets considered in this work. As detailed in Appendix C, these models exhibit low performance in specialized domains such as econometrics and professional psychology, indicating limited domain-specific knowledge despite their scale and diversity. Although these results are not directly comparable to ACER due to differences in architecture and training objectives, they suggest that general-purpose synthetic pretraining alone does not reliably close fine-grained domain gaps, motivating the targeted, curriculum-aligned knowledge infusion explored in ACER.

4.2 Results: with Llama 3.2 baselines

We generate synthetic book corpora for the identified target domains and continually pretrained the baseline Llama 3.2 3B model using different curriculum schedules. Table 1 compares these trained models with the pretrained 3B baseline. Across all schedules, ACER consistently improves performance on the target domains. The **Flat** regime, which uniformly mixes books and QA pairs without ordering, yields a strong gain of +2.5 points in Macro_t over the baseline. Building on this, the **Cognitive (Cog)** schedule, which progresses from books $\rightarrow$ easy QA $\rightarrow$ hard QA, provides further incremental improvements. Extending this with persona-based content progression in **Cog+Con** delivers the best overall performance, indicating that curriculum design amplifies the benefits of synthetic corpora in continual pretraining.

In contrast, the **Interleaved** schedule, which mixes book sections and QA pairs across domains within training stages, underperforms relative to the other curriculum schedules, with particularly large drops in mathematics and statistics. This contrasts with Lee et al. (2024), where interleaving educational content improved performance of the model in the fine-tuning phase. We conjecture that, while interleaving may be effective for instruction-following fine-tuning, its fragmented sequencing dilutes the supervision signal in continual pretraining. For smaller-scale domain infusion tasks, such as ours, frequent task-switching introduced by interleaving likely overwhelms model capacity, leading to degraded performance rather than gains.

As shown in Table 1, domain knowledge infusion through targeted continual pretraining provides the largest benefits in microeconomics and econometrics. These niche areas are relatively underrepresented in the pretraining corpus of the 3B baseline, leaving substantial room for improvement. By contrast, domains such as statistics and mathematics are more prevalent in web-scale text, so

Benchmark	#shots	Setting	Llama 3.2 3B		Llama 3.2 1B	
			Pre-trained	Ours	Pre-trained	Ours
ARC Challenge	25	Acc (Weighted)	0.4701	**0.4837**	0.3652	**0.3831**
GPQA	0	Acc (Mean)	0.2656	**0.2879**	**0.2545**	0.2478
MMLU	5	Acc (Weighted)	0.5408	**0.569**	0.3099	**0.3277**
AGIEval	0	Acc (Weighted)	0.2255	0.2253	0.1867	0.1877
GSM8K	5	Strict (EM)	0.2805	0.2798	0.0667	0.0591
GSM8K	5	Flexible (EM)	0.2767	0.276	0.0644	0.0523
HellaSwag	0	Acc (Mean)	0.5522	0.5489	0.4777	0.4782

Table 2: Evaluation across benchmarks comparing our knowledge-infused Llama 3.2 models (3B and 1B) with their pretrained versions. The knowledge-infused models improve tasks requiring knowledge recall and understanding (ARC, GPQA, and MMLU) by more than 2 absolute percentage points, without regressing on general capabilities such as language understanding, reasoning, and mathematics (AGIEval, GSM8K, and HellaSwag).

the baseline already demonstrates strong competence, resulting in only modest gains. Overall, we observe an improvement of approximately 3 percentage points in Macro_t. Notably, our training regimen preserves performance on non-target domains (Macro_{nt}), with small positive changes observed in several cases. Appendix F shows an example where ACER enabled the model to correctly answer a microeconomics question that the pretrained baseline answered incorrectly.

We also evaluate the best-performing Cog+Con curriculum on Llama 3.2 1B, observing improvements of up to 2.7 points on target domains, with detailed results reported in Appendix D.3.

We further examine the sensitivity of ACER to the choice of synthetic data generator by regenerating the curriculum using different models, while fixing the curriculum schedule to the best-performing variant. As shown in the bottom block of Table 1, generator choice modulates the magnitude of gains, but curriculum-aligned training consistently improves or preserves performance relative to the pretrained baseline. A detailed analysis is provided in Appendix D.2.

4.3 Impact Beyond MMLU: Generalization to Broader Capabilities

Next, we ask how the proposed knowledge-infusion recipe in ACER transfers beyond MMLU. To this end, we evaluate the knowledge-infused Llama 3.2 models against their pretrained baselines on standard language understanding benchmarks. Table 2 reports results on ARC, GPQA, AGIEval, GSM8K, and HellaSwag. Tasks that emphasize domain-specific knowledge and recall, such as ARC, GPQA, and MMLU (5-shot), benefit the most. Both the 3B and 1B variants achieve gains of more than two absolute percentage points, highlighting the effectiveness of targeted continual pretraining in strengthening reasoning over underrepresented knowledge areas.

Equally important, these improvements do not come at the cost of general capabilities. On benchmarks such as AGIEval, GSM8K, and HellaSwag, which measure general reasoning, arithmetic, and commonsense capabilities, the ACER models remain stable relative to their pretrained baselines. For the 3B model, differences are within 0.3 absolute points on HellaSwag and even smaller across other tasks, while the 1B model shows comparable stability. This indicates that domain infusion not only enhances specialized competence but also preserves broad language abilities, showing the scalability of our framework across diverse evaluation settings.

5 Conclusion

We present ACER (Automated Curriculum-Enhanced Regimen), a novel framework that systematically infuses domain-specific knowledge into large language models through progressive, curriculum-aligned learning sequences interspersed with targeted assessment. Our empirical evaluation demonstrates that ACER achieves substantial performance gains in target domains while maintaining model capabilities across existing tasks. This work establishes a principled foundation for structured knowledge integration in LLMs, offering a scalable pathway toward more capable and domain-aware language models.

Limitations

While ACER demonstrates consistent gains across multiple domains and the tested model scales, several limitations remain. First, although we evaluate sensitivity to different synthetic data generators, the quality and structural properties of the generated curriculum can influence the magnitude of achievable gains (see Appendix D.2 for a qualitative analysis). Designing synthesis pipelines that explicitly optimize for curriculum structure and granularity remains an open direction. Second, our experiments focus on 3B and 1B parameter models. While this allows controlled analysis of domain knowledge infusion and curriculum effects, it remains to be explored how ACER scales to larger models and whether similar curriculum-aligned training yields comparable or amplified benefits at higher capacities.

Ethical considerations

This work focuses on improving domain knowledge acquisition in large language models through synthetic data generation and curriculum-aligned continual pretraining. Our experiments use only publicly available benchmarks and automatically generated data, and do not involve human subjects or personal data. As such, we do not foresee direct ethical risks arising from this work. However, as with all research on language models, downstream use may inherit broader societal risks related to biased or unintended model behavior. We view ACER as a training-time framework, and emphasize that responsible deployment, evaluation, and monitoring of models remain essential in practical settings.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, and 1 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

Vinayak Arannil, Neha Narwal, Sourav Sanjukta Bhabesh, Sai Nikhil Thirandas, Darren Yow-Bang Wang, Graham Horwood, Alex Anto Chirayath, and Gouri Pandeshwar. 2024. Dopamine: Domain-specific pre-training adaptation from seed-guided data mining. *arXiv preprint arXiv:2410.00260*.

Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von Werra. 2024. Cosmopedia.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. 2009. Curriculum learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, page 41–48, New York, NY, USA. Association for Computing Machinery.

Louis Béthune, David Grangier, Dan Busbridge, Eleonora Gualdoni, marco cuturi, and Pierre Ablin. 2025. Scaling laws for forgetting during finetuning with pretraining data injection. In *Forty-second International Conference on Machine Learning*.

B. S. Bloom, M. B. Engelhart, E. J. Furst, W. H. Hill, and D. R. Krathwohl. 1956. *Taxonomy of educational objectives. The classification of educational goals. Handbook 1: Cognitive domain*. Longmans Green, New York.

Zhou Chen, Ming Lin, Zimeng Wang, Mingrun Zang, and Yuqi Bai. 2024. Preparedllm: effective pre-pretraining framework for domain-specific large language models. *Big Earth Data*, 8(4):649–672.

Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. *Preprint*, arXiv:2406.14491.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360, Online. Association for Computational Linguistics.

Yao He, Xuanbing Zhu, Donghan Li, and Hongyu Wang. 2025. Enhancing large language models for specialized domains: A two-stage framework with parameter-sensitive lora fine-tuning and chain-of-thought rag. *Electronics*, 14(10).

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Jianheng Huang, Leyang Cui, Ante Wang, Chengyi Yang, Xinting Liao, Linfeng Song, Junfeng Yao, and Jinsong Su. 2024. Mitigating catastrophic forgetting in large language models with self-synthesized rehearsal. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1416–1428, Bangkok, Thailand. Association for Computational Linguistics.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *CoRR*, abs/2001.08361.

Tobias Kerner. 2024. Domain-specific pretraining of language models: A comparative study in the medical field. *arXiv preprint arXiv:2407.14076*.

Bruce W Lee, Hyunsoo Cho, and Kang Min Yoo. 2024. Instruction tuning with human curriculum. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1281–1309, Mexico City, Mexico. Association for Computational Linguistics.

Haoran Li, Qingxiu Dong, Zhengyang Tang, Chaojun Wang, Xingxing Zhang, Haoyang Huang, Shaohan Huang, Xiaolong Huang, Zeqiang Huang, Dongdong Zhang, Yuxian Gu, Xin Cheng, Xun Wang, Si-Qing Chen, Li Dong, Wei Lu, Zhifang Sui, Benyou Wang, Wai Lam, and Furu Wei. 2025. Synthetic data (almost) from scratch: Generalized instruction tuning for language models. *Transactions on Machine Learning Research*.

Huu Tan Mai, Cuong Xuan Chu, and Heiko Paulheim. 2024. Do llms really adapt to domains? an ontology learning perspective. In *International Workshop on the Semantic Web*.

Sven Najem-Meyer, Frédéric Kaplan, and Matteo Romanello. 2025. Don't stop pretraining! efficiently building specialised language models in resource-constrained settings. In *Proceedings of the 9th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2025)*, pages 252–260, Albuquerque, New Mexico. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. Gpqa: A graduate-level google-proof qa benchmark. *Preprint*, arXiv:2311.12022.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13484–13508, Toronto, Canada. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. *Preprint*, arXiv:2109.01652.

Jian Wu, Hang Yu, Bingchang Liu, Wenjie Yang, Peng Di, Jianguo Li, and Yue Zhang. 2025. Lamdas: Llm as an implicit classifier for domain-specific data selection. *Preprint*, arXiv:2509.06524.

Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. 2024. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.

Yang Zhang, Amr Mohamed, Hadi Abdine, Guokan Shang, and Michalis Vazirgiannis. 2025a. Beyond random sampling: Efficient language model pretraining via curriculum learning. *Preprint*, arXiv:2506.11300.

Yutong Zhang, Lixing Chen, Shenghong Li, Nan Cao, Yang Shi, Jiaxin Ding, Zhe Qu, Pan Zhou, and Yang Bai. 2025b. Way to specialist: Closing loop between specialized llm and evolving domain knowledge graph. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, KDD '25, page 1996–2007, New York, NY, USA. Association for Computing Machinery.

Wanjun Zhong, Ruixiang Cui, Yiduo Guo, Yaobo Liang, Shuai Lu, Yanlin Wang, Amin Saied, Weizhu Chen, and Nan Duan. 2024. AGIEval: A human-centric benchmark for evaluating foundation models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 2299–2314, Mexico City, Mexico. Association for Computational Linguistics.

Appendices

A ACER Synthesis flow

A.1 Domain Detailing

Explicitly encoding domain detailing before corpus synthesis offers two advantages:

1. Provides a transparent knowledge map that facilitates traceability between input requirements and generated content.

Persona	Microeconomics	Statistics	Econometrics	Mathematics	Psychology
Highschool	0.67	0.74	0.55	1.04	0.72
Undergraduate	1.56	1.47	1.67	1.97	1.01
Graduate	2.56	2.10	2.73	1.55	1.27
Researcher	1.66	1.79	2.83	2.21	1.85
Total	6.46	6.10	7.78	6.77	4.85

Table 3: ACER book corpus: #tokens (in Millions) across the five domains and four personas. Total synthetic data #tokens = 31.97 Millions

Model	MEco$_{hs}$	Stats$_{hs}$	Econ	Maths$_{hs}$	Psych$_{p}$	Macro$_{t}$	Macro$_{nt}$
Cosmo-1B	0.2647	0.213	0.2719	0.2556	0.2663	0.2543	0.2626
Phi-1.5	0.4328	0.2685	0.2719	0.37	0.3971	0.3229	0.4266
Phi-2	0.5882	0.4259	0.3158	0.2926	0.5392	0.4323	0.5552
Llama 3.2 1B	0.3361	0.25	0.193	0.2296	0.3448	0.2707	0.4011
Llama 3.2 3B	0.5378	0.3796	0.3158	0.2704	0.5507	0.4108	0.5754

Table 4: Contextual comparison with synthetic-pretrained models: Cosmo 1B (1B parameters), Phi 1.5 (1.3B parameters), and Phi 2 (2.7B parameters), evaluated on the same MMLU subdomains used in this work. Results are provided for contextual reference and are not directly comparable to ACER-trained models due to differences in architecture and training setup.

2. Enables curriculum-aware scaling, where subsequent synthetic data creation (e.g., textbooks and assessments) follows a structured hierarchy rather than ad-hoc sampling.

A.2 Outline Generation

ACER leverages a large language model prompted as a professional author to create a **multilevel outline** in a structured JSON format. The generated outline includes the following.

- **Book Title:** A descriptive title aligned with the topic and purpose.
- **Parts:** Four to six major parts that divide the book into thematic areas, each with a concise summary.
- **Chapters:** Four to six chapters per part, each providing a self-contained exploration of subtopics.
- **Sections and Subsections:** Three to six sections per chapter, with optional subsections for complex concepts, ensuring fine-grained coverage.

This structured representation balances clarity, depth, and scalability, allowing the same framework to create outlines of varying complexity based on audience needs.

Explicit outline generation provides three main benefits:

1. **Consistency:** Maintains structural uniformity across different domains and audience levels.
2. **Scalability:** Simplifies automation, as the JSON schema can be directly consumed by subsequent content generation pipelines.
3. **Pedagogical Alignment:** Encourages systematic progression from foundational concepts to advanced material, supporting curriculum-driven pretraining.

A.3 Synthetic Content Creation

The procedure begins with the construction of a tree representation from the ToC. Each node in this tree corresponds to a textual unit—either a *Part*, *Chapter*, *Section*, or *Subsection*. For every node, the system maintains key attributes that include the title, description, content, and node type (an enumerated label designating whether the node belongs to root, part, chapter, section, or subsection). This representation provides a flexible yet structured foundation for downstream content synthesis.

Once the hierarchical structure is established, the system proceeds to parse the tree and generate section-level content. For each node in the section, the input prompt is carefully composed to include

Model (Llama 3.2 1B)	MEco_{hs}	Stats_{hs}	Econ	Maths_{hs}	Psych_{p}	Macro_{t}	Macro_{nt}
Pretrained baseline	0.3361	0.25	0.193	0.2296	0.3448	0.2707	0.4011
$\text{Cog+Con}_{\text{gemini-2.0-flash}}$	0.3445	0.3287	0.2105	0.2481	0.3562	**0.2976**	**0.4051**
$\text{Cog+Con}_{\text{Phi-4}}$	0.3361	0.2917	0.2105	0.2593	0.3317	0.2859	0.4017
$\text{Cog+Con}_{\text{Llama 3.1 70B}}$	0.3529	0.3241	0.2193	0.2407	0.3284	0.2931	0.3962

Table 5: Comparison of pretrained Llama 3.2 1B with ACER-trained variants under the best-performing curriculum (Cog+Con), using different synthetic data generators. Domain-targeted synthetic textbooks with exam-style QA improve performance over the pretrained baseline, with Cog+Con (Gemini 2.0 Flash) achieving a **2.7**-point improvement on target domains. Results across Phi-4 and Llama 3.1 70B generators show that while generator choice modulates the magnitude of gains, curriculum-aligned training improves or preserves performance without degradation.

contextual information such as the title and description of the enclosing part, the title and description of the chapter, the focal section title, and the list of subsections under it. In addition, metadata style guidelines (intent, audience, genre, tone, voice, and language) are explicitly injected into the prompt. This ensures that the generation process is not only content-driven but also stylistically aligned with the pedagogical and editorial goals of the textbook. The LLM is instructed to produce polished, instructional prose while avoiding any meta-commentary, annotations, or explanations of its reasoning.

B Tokens generated using ACER synthesis flow

Table 3 details the number of tokens generated by ACER. It uses Gemini 2.0 Flash apis for synthetic data generation.

C Positioning Relative to Synthetic Pretraining Models

We contextualize our work relative to recent pretraining efforts that expand knowledge coverage via large-scale synthetic data. We evaluate representative 1B-3B parameter models trained predominantly on broad synthetic corpora, specifically Phi 1.5, Phi 2, and Cosmopedia, on the same five MMLU subdomain considered in this work. These models reflect alternative strategies that rely on broad, general-purpose synthetic data rather than targeted curriculum-aligned knowledge infusion.

As reported in Table 4, these models exhibit low performance in specialized domains such as econometrics and professional psychology, indicating limited domain-specific knowledge despite their scale and diversity. We emphasize that these results are not directly comparable to ACER-trained models due to differences in model architecture, scale, and training objectives. Nevertheless, this analysis highlights that broad synthetic pretraining alone does not reliably close fine-grained domain gaps, motivating the need for targeted, curriculum-aware knowledge infusion as explored in ACER.

D Experiments: Additional Details

D.1 Reproducibility Details

The parameters required to reproduce our results, such as token budgets, sequence lengths, batch sizes, optimizer settings, learning rate schedule, training steps, hardware specifications, random seeds, prompts, and de-duplication thresholds—are provided in section 4.1. The prompts used for all types of generations are listed in Appendix G.

D.2 Robustness to Synthetic Data Generators

To assess whether ACER's gains depend on a specific synthesis model, we regenerate the synthetic textbook corpus using Phi-4 and Llama 3.1 70B, while fixing the curriculum to the best-performing variant (Cog+Con) and keeping all other settings unchanged. As shown in Table 1 (bottom block), Cog+Con with Phi-4 generated data continues to improve over the pretrained Llama 3.2 3B baseline on target MMLU domains, whereas Llama 3.1 70B generated data yields smaller gains and performs comparably to the baseline. These results show that ACER is robust to the choice of synthetic data generator: even when using weaker or less structurally detailed generators (e.g. Llama 3.1 70B) instead of Gemini 2.0 Flash, training does not lead to performance degradation, although stronger generators yield larger gains. We find that this behavior correlates with the structural granularity of the synthesized textbooks: Gemini-generated books contain finer-grained outlines (about 100 sections per book on average) compared to Llama 3.1 70B

Ratio	Baseline	1:3	1:1	3:1	9:1
Target Macro	0.4108	0.4257	**0.4407**	0.4284	0.4135
Non-Target Macro	0.5754	0.5759	**0.5821**	0.5805	0.5736

Table 6: Ablation on in-domain to replay data proportion for training: CPT with equal proportion of in-domain and replay data provides the right balance

(about 30 sections). Since ACER's curricula operate at the section level, this reduced granularity weakens the progressive supervision signal during continual pretraining. Corresponding results for Llama 3.2 1B follow similar trends and are reported in Appendix D.3.

D.3 Results: Scaling to Llama 3.2 1B

We apply the best-performing Cog+Con curriculum to the Llama 3.2 1B model to assess whether ACER's benefits extend to smaller model scales. As shown in Table 5, the Cog+Con variant achieves improvements of up to 2.7 points on target MMLU domains relative to the pretrained baseline, while largely preserving performance on non-target domains.

We further evaluate robustness to the choice of synthetic data generator by regenerating the synthetic corpus using Phi-4 and Llama 3.1 70B, while keeping the curriculum schedule fixed. Consistent with observations at the 3B scale, Phi-4 generated data yields comparable improvements, whereas Llama 3.1 70B generated data leads to smaller or neutral gains. These results suggest that, even at smaller scales, ACER does not rely on a specific generator to avoid degradation, although generator choice influences the magnitude of achievable gains.

D.4 Ablations: What is the Right Data Mixture?

Table 6 reports the effect of varying the ratio of in-domain synthetic data to replay data on model performance under the Cog+Con curriculum. The balanced 1:1 setting yields the strongest results, improving Macro$_t$ by about 3 absolute points over the baseline and Macro$_{nt}$ by 0.7 points. Ratios skewed toward replay (1:3) or synthetic data (3:1, 9:1) yield smaller gains and fail to surpass the balanced mixture. The trend is particularly evident for non-target domains: as replay proportion decreases, Macro$_{nt}$ consistently drops, underscoring the role of replay in preserving generalization. Conversely, in-domain synthetic data is the main driver of improvements on target domains. Overall, the balanced 1:1 mixture offers the best trade-off between specialization and generalization, pointing to adaptive mixture schedules as a promising future direction.

E Example domain: Microeconomics

This section shows the generated book content for **Microeconomics** domain (persona: high school), providing snippets for the following:

- Topic detailing
- ToC (shortened version)
- Example section
- Example QA Pair
- Example Easy QA Pair
- Example Hard QA Pair
- A snapshot of the ToC Tree Visualiser

Topic Detailing: Imagine you're running a lemonade stand. Microeconomics is like understanding all the tiny decisions that make your stand successful 2013 how many lemons to buy, what price to charge for your lemonade, and whether to hire a friend to help you out. It's about how individuals, families, and businesses make choices about using limited resources. Instead of looking at the entire economy of a country (that's macroeconomics!), microeconomics zooms in on these smaller, everyday economic activities. It helps us understand why things cost what they do, how markets work, and how people respond to different incentives. Think of it as the science of small-scale economic decisions and their ripple effects.

1: Shortened ToC (JSON)


```
{
  "title": "Part 1: The Basics of Making
      Choices",
  "description": "This section
     introduces the foundational.... ",
  "chapters": [
```

```
    {
      "title": "Chapter 1: Welcome to
          Microeconomics!",
      "description": "This chapter
          introduces the world of
          microeconomics ...",
      "sections": [
        {
          "title": "What is
              Microeconomics?",
          "subsections": [
            {
              "title": "The 'Micro' in
                  Microeconomics"
            },
            {
              "title": "Lemonade Stands
                  and Economies"
            }
          ]
        },
        {
          "title": "Why Should You Care
              About Microeconomics?",
          "subsections": [
            {
              "title": "Making Smarter
                  Choices"
            },
            {
              "title": "Understanding
                  the World Around You"
            }
          ]
        },
        .
        .
      ]
    }
  ]
}
```


Example Section:
Part title: The Basics of Making Choices
Chapter title: Scarcity: The Reason We Have to Choose
Section title: Dealing With Scarcity: Prioritization and Rationing
Section content: Scarcity, as we've established, means we can't have everything we want. This universal truth forces us to confront some tough questions. If we can't satisfy all our desires, how do we decide which ones to satisfy and which ones to leave unfulfilled? The answer lies in two closely related concepts: prioritization and rationing. ...

Example QA :
Question: What are the three main decision-makers studied in microeconomics?
Answer: Microeconomics focuses on the "small picture" of the economy, studying the decisions made by individual players rather than the entire national or global economy.....

Example Easy QA :
Question: Imagine you're deciding whether to spend your allowance on a new video game or save it for a concert ticket. How does this single decision illustrate the core idea that microeconomics studies choices made by individuals with limited resources?
Answer: Okay, that's a great question! It gets right to the heart of what microeconomics is all about. Let's break it down using the lemonade stand example and the ideas presented in the text.....

Example Hard QA :
Question: Imagine two lemonade stands on the same street. One stand uses only organic lemons and charges a higher price, while the other uses regular lemons and charges a lower price.....
Answer: Okay, that's a great question! It gets right to the heart of how microeconomics works in the real world, even in something as simple as two lemonade stands. Microeconomics helps us understand why both lemonade stands 2013 the one with organic lemons and a higher price and the.....

Tree Visualiser Tool : Figure 4 shows the tree visualizer tool we developed to analyze the generated data.

F Example of Performance Improvement after ACER

We present an example MMLU question from the high-school microeconomics subset to demonstrate the effectiveness of ACER. The base pre-trained model (LLaMA 3.2 3B) could not answer this question correctly, while the ACER trained model was able to find the right answer due to its enhanced domain understanding.

> **Question:** "The elasticity of supply is typically greater when . . . "

Options:

A. Producers have fewer alternative goods to produce.

B. Producers have less time to respond to price changes.

C. Producers are operating near the limits of their production.

D. Producers have more time to respond to price changes.

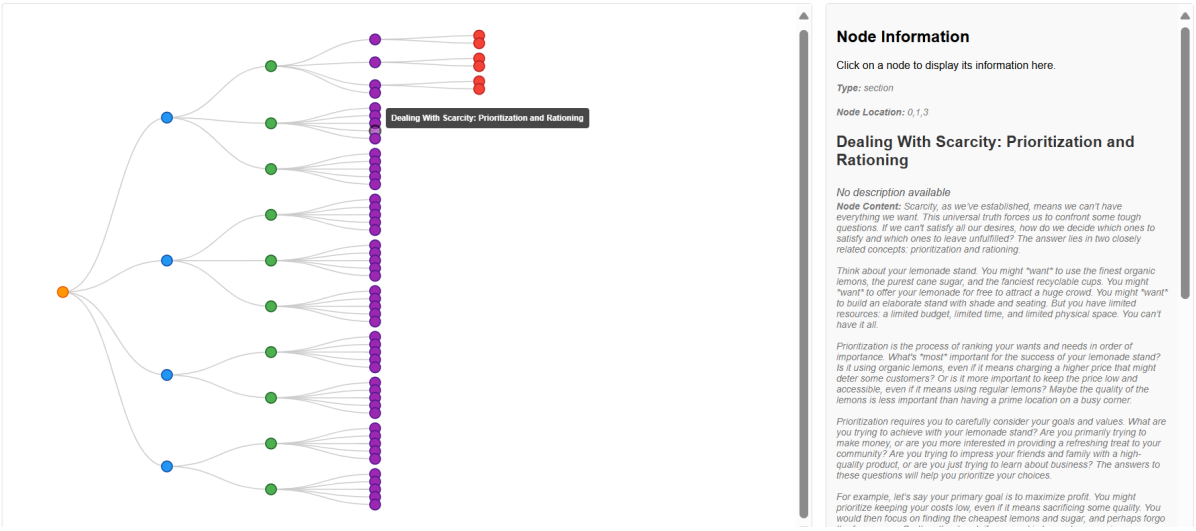

Figure 4: Snapshot of the Tree Visualisation tool we created to analyse the ToC and the associated generated Data

The correct answer is option D. The baseline LLaMA 3.2 3B model predicted option A (incorrect), while our ACER trained model predicted option D (correct).

This improvement can be attributed to the additional knowledge injected during continual pretraining. Our generated microeconomics book included a dedicated section on *"Supply Curve and Elasticity"* An excerpt from the Table of Contents of that book is shown below:

2: Excerpt from the Table of Contents (JSON)

```
{
  "title": "Advanced Producer Theory:
     Cost and Production",
  "description": "This chapter focuses
     on the theory of the firm...",
  "sections": [
    {
      "title": "Profit Maximization and
         Supply",
      "subsections": [
        {
          "title": "Supply Curve and
             Elasticity"
        }
      ]
    },
    {
      "title": "Applications of Producer
          Theory",
      "subsections": [
        {
          "title": "Impact of
             Technological Change"
        }
      ]
    },
    {
      "title": "Summary"
    }
  ]
}
```

G Prompts used in the ACER framework

G.1 Domain Detailing Prompt

The prompt used for domain detailing:

3: Prompt for Domain Detailing

```
prompt = f"""
    You are an expert author preparing
        to write a comprehensive book.
    Topic:\"{topic}\".
    Target audience: \"{audience}\".
    Intent (purpose): \"{intent}\".

    Your task is to generate preparatory
         material that will help you
        structure the book. Based on the
         above, provide the following:
    1. A factual and high-level
        description of the topic,
        suitable for the target audience
        .
    2. A list of 6  8   core themes or
        subtopics that must be covered
        to provide a well-rounded
        understanding of the topic.
    3. A list of 6  8   important
        questions that the book should
        aim to answer from the
        perspective of the target
        audience.

    Present your output in JSON with the
         following keys:
    - "description": A string
    - "subtopics": A list of strings
    - "key_questions": A list of strings
    """
```

G.2 Outline Generation Prompt

The prompt used for outline (Table of Contents) generation:

4: Prompt for Outline generation

```
prompt = f'''
You are a professional book author known
     for clear, structured, and reader-
    friendly writing.
Your task is to design a full book
    outline based on the information
    below.
---
**Topic:** {topic}

**Target Audience:** {audience}

**Intent:** {intent}

**Genre:** {genre}

**Tone:** {tone}
**Voice:** {voice}
**Language Style:** {language}

**Topic Description:**
{description}

**Core Subtopics to Cover:**
{subtopics}

**Key Questions the Book Should Answer
    :**
{key_questions}
---

Based on this, generate a comprehensive
    book structure in JSON with the
    following:
1. A compelling and descriptive "title"
    for the book.
2. Divide the book into **4 to 6 major
    parts**:
   - Each part should have:
     - "title" (clear, theme-based)
     - "description" (3  5  sentence
         summary)
3. Each part should contain **4 to 6
    chapters**:
   - Each chapter should include:
     - "title"
     - "description" (3  5  sentences,
         aligned with purpose)
     - "sections": 3 to 6 entries, each
         with:
         - "title" (section title)
         - "subsections": Optional (to
             describe complex ideas, 2-3
              subsections per section)
             - "title" (subsection title
                 )
   (Include common sections like "
       Introduction" and "Summary" where
        appropriate.)

Return **only** a well-formatted JSON
    object in the following structure:

{{
  "title": "<book_title>",
  "parts": [
    {{
      "title": "<part_title>",
      "description": "<paragraph>",
      "chapters": [
        {{
          "title": "<chapter_title>",
          "description": "<paragraph>",
          "sections": [
            {{
              "title": "<section_name>",
              // Optional subsections
              "subsections": [
                {{
                  "title": "<
                      subsection_name>"
                }},
                ...
              ]
            }},
            ...
          ]
        }},
        ...
      ]
    }},
    ...
  ]
}}
'''
```

G.3 Section Content Generation Prompt

The prompt used for section content generation:

5: Prompt for section content generation

```
prompt= f'''
Generate detailed, high-quality, and
    cohesive book content for the
    following section. Expand fully on
    each idea and subsection with in-
    depth explanations, compelling
    examples, and smooth transitions.
    Follow the style guidelines strictly
    . Avoid meta-commentary or
    structural notes. Write as if this
    is the final, polished content ready
     for publication."
'''

metadata_str = '''Style Guidelines:
- Intent: {intent}
- Audience: {audience}
- Genre: {genre}
- Tone: {tone}
- Voice: {voice}
- Language: {language}
- Description: {description}
'''

end_prompt = f'''Instructions
- Compose a flowing narrative or
    exposition that naturally integrates
     the listed subsections.
- Be generous with elaboration, examples
    , and insights.
```

```
- Ensure thematic and stylistic
    continuity with the previous content
    .
- Do not include headers, instructions,
    or artificial markers     only write
     the finished prose.
'''
```

G.4 QA Pair Generation Prompts

The prompt used for question generation is:

6: Prompt for question generation

```
prompt = f"""
You are a professor specializing in the
    subject of {topic_name}. Your task
    is to generate **educational, self-
    contained questions** to help
    students understand a section from a
     textbook titled: "{book_title}".
This book is for {target_audience}, with
     the primary goal to: **{intent}**

### Global Context Start ###
- **Topic**: {topic_name}
- **Topic Description**: {
    topic_description}
- **Questions the book attempts to
    answer**: {guiding_questions}
### Global Context End ###

You will be generating questions for the
     following part of the book:
- Chapter: {chapter_title}
- Section: {section_title}
--- Section Content Start ---
{section_text}
--- Section Content End ---
"""

if previous_question:
    prompt += f"""
### TASK ###

You are generating the next question in
    a sequence.

1. Analyze the difficulty of the
    previous question.
2. Generate a new question that is **
    slightly more challenging** than the
     one before. Increase cognitive
    d e p t h move progressively from
    understanding     interpretation
     analysis     application.
3. Ensure the question is different in
    focus or angle from the previous one
    .
4. Do **not** include the answer.
5. **Strictly** maintain relevance to
    the given section and suitability
    for {target_audience}.
6. **Only output the question** do   not
     include any prefixes like "
    Generated Question", "Q1:", or
    anything else.

Previous Question:
- {previous_question}

### Question ###
Question:"""
else:
    prompt += f"""
### TASK ###

You are generating the **first question
    ** in a learning sequence.

1. Start with a **simple question**
    focused on factual recall or basic
    definitions.
2. Make the question self-contained and
    directly based on the section
    content.
3. Do **not** include the answer.
4. Ensure clarity, relevance, and
    suitability for {target_audience}.
5. **Only output the question** do   not
     include any prefixes like "
    Generated Question", "Q1:", or
    anything else.

### Question ###
Question:"""
```

The prompt used for answer generation is:

7: Prompt for answer generation

```
prompt = f"""
You are a professor specializing in the
    subject of {topic_name}. Your task
    is to generate an **educational,
    self-contained answer** to a
    question based on a section from a
    textbook titled: "{book_title}".
This book is for {target_audience}, with
     the primary goal to: **{intent}**

### Global Context Start ###
- **Topic**: {topic_name}
- **Topic Description**: {
    topic_description}
### Global Context End ###

You are answering a question based on
    the following section:
- Chapter: {chapter_title}
- Section: {section_title}

--- Section Content Start ---
{section_text}
--- Section Content End ---

### TASK ###

Your job is to write a complete,
    thoughtful, and educational answer
    to the student's question below.
    Follow these guidelines:

1. Ensure the answer is **directly
    grounded in the section content**
    provided above.
2. Make the answer **rich in detail**,
    using clear language and examples
    when helpful.
3. The answer should be **self-contained
```

```
    **, meaning the reader should not
    need to refer to the original
    content to understand it.
4. Ensure the tone is educational and
    appropriate for {target_audience}.
5. Do **not invent any information** not
     present in the section.

### Question ###
{question}

### Answer ###
"""
```

H Decontamination- data analysis

We performed semantic deduplication between the synthetic data we generated and the benchmark datasets. The motivation for using semantic deduplication is that it is substantially more thorough than MinHash-based deduplication, as it can detect near-duplicates beyond surface-level lexical overlap. Specifically, we first computed embeddings for both the generated samples and the benchmark data. For each generated sample, we then identified its nearest neighbors in the benchmark set using cosine similarity. Samples with similarity scores above a predefined threshold were manually analyzed. Across all domains, we observed little to no contamination with respect to the benchmark data. For the few cases where the cosine similarity exceeded 0.9, the generated samples were found to discuss canonical concepts that are commonly covered when acquiring foundational knowledge of the domain, rather than reproducing benchmark-specific content. A detailed decontamination report is provided in Table 7, along with representative example fragments in Table 8.

Domain	**Generated Samples**	**Similarity $\geq$ 0.8**
Econometrics	4125	0.0969
Mathematics	4110	0.0243
Psychology	4350	0.2758
Statistics	4200	0.1428
Microeconomics	3900	0.051282

Table 7: Domain statistics for generated data (similarity threshold = 0.8). The final column reports the percentage of benchmark samples with cosine similarity ≥ 0.8

Contents
Domain: Econometrics **Generated Sample:** What is the primary consequence of omitting a relevant variable from a multiple linear regression model? **Closest Benchmark Sample:** If a relevant variable is omitted from a regression equation, the consequences would be that: i) The standard errors would be biased ii) If the excluded variable is uncorrelated with all of the included variables, all of the slope coefficients will be inconsistent. iii) If the excluded variable is uncorrelated with all of the included variables, the intercept coefficient will be inconsistent. iv) If the excluded variable is uncorrelated with all of the included variables, all of the slope and intercept coefficients will be consistent and unbiased but inefficient. **Cosine Similarity:** 0.8456
Domain: Mathematics **Generated Sample:** A university club with 15 members needs to form a committee of 5 to organize a fundraising event. * How many different committees can be formed? * The club decides that the committee must have a president, a vice-president, a treasurer, a secretary, and a public relations officer. The president and vice-president must be chosen from among 8 senior members, while the treasurer, secretary, and public relations officer can be chosen from the remaining members (seniors or otherwise). In how many ways can such a committee be formed, ensuring each member has a distinct role? **Closest Benchmark Sample:** How many different possible committees of 5 people can be chosen from a group of 15 people? **Cosine Similarity:** 0.8338
Domain: Psychology **Generated Sample:** What is the primary difference between classical conditioning and operant conditioning? **Closest Benchmark Sample:** What is the major difference between classical and operant conditioning? **Cosine Similarity:** 0.9871
Domain: Statistics **Generated Sample:** What is a sampling distribution, and how is it constructed? **Closest Benchmark Sample:** What is a sampling distribution? **Cosine Similarity:** 0.929
Domain: Microeconomics **Generated Sample:** What fundamental economic problem does microeconomics primarily address? **Closest Benchmark Sample:** The primary focus of microeconomics is **Cosine Similarity:** 0.8312

Table 8: Examples of generated samples, closest benchmark samples and their cosine similarity