WRAP++: Web Discovery Amplified Pretraining

Anonymous ACL submission

Abstract

Synthetic data rephrasing has emerged as a powerful technique for enhancing knowledge acquisition during large language model (LLM) pretraining. However, much of web-scale synthetic rephrasing operates at the *single-document* level, rewriting individual web pages in isolation. This confines synthesized examples to intra-document knowledge, missing cross-document relationships and leaving facts with limited associative context. We propose **WRAP++** (**W**eb discove**R**y **A**mplified **P**retraining), which *amplifies* the associative context of factual knowledge by *discovering* cross-document relationships from web hyperlinks and synthesizing joint QA over each discovered document pair. Concretely, **WRAP++** discovers high-confidence relational motifs including dual-links ($A \leftrightarrow B$) and co-mentions ($A \rightarrow E \leftarrow B$ with $A \rightarrow B$), and synthesizes QA that requires reasoning across both documents. This produces relational knowledge absent from either source document alone, creating diverse entry points to the same facts. Because the number of valid entity pairs grows combinatorially, this discovery-driven synthesis also amplifies data scale far beyond single-document rewriting. Instantiating **WRAP++** on Wikipedia, we amplify ∼8.4B tokens of raw text into 80B tokens of cross-document QA data. On SimpleQA, OLMo-based models at both 7B and 32B scales trained with **WRAP++** substantially outperform single-document approaches and exhibit sustained scaling gains, underscoring the advantage of cross-document knowledge discovery and amplification.

1 Introduction

Synthetic data has become an increasingly important component of large language model (LLM) pretraining. WRAP (Maini et al., 2024) showed that rephrasing noisy web text into QA format can improve pretraining, and later systems scaled this recipe substantially: Nemotron-CC (Su et al., 2025) produced ∼2 trillion synthetic tokens from Common Crawl, Phi-4 (Abdin et al., 2024) used 40% synthetic data in pretraining, and Qwen3 (Yang et al., 2025a) incorporated synthetic data into its training pipeline.

However, much of the web-scale synthetic pretraining literature has focused on intra-document factors such as rephrasing strategy, generator model, and source quality (Nguyen et al., 2025; Niklaus et al., 2026). These methods vary *how* a single document is rewritten while largely preserving *what* evidence is synthesized. Because many factual associations are distributed across multiple documents, this single-document paradigm can leave relational context underrepresented, limiting the ways in which a model can recover and connect related facts.

This limitation motivates a shift toward *cross-document synthesis*: bringing multiple facts into a shared context to learn relational knowledge jointly. However, moving to cross-document synthesis is non-trivial. A naïve approach of randomly pairing documents yields little improvement over single-document baselines (see §4.1), as forcing an LLM to synthesize joint QA from unrelated texts produces fabricated connections and low-quality data. Thus, the *document selection* mechanism is critical: cross-document synthesis only succeeds when the paired documents contain genuinely related facts.

Web hyperlinks provide a broad relevance signal by encoding human-curated judgments of importance (Zhou et al., 2022). For example, the Wikipedia pages of composers Hans Zimmer and Ludwig Göransson are topologically linked through shared collaborations with director Christopher Nolan. While single-document synthesis might only extract isolated facts (e.g., "Göransson won an Oscar for Oppenheimer"), **WRAP++** pairs these connected documents to synthesize multi-hop relational QA. As illustrated in Figure 1, the model is forced to explicitly reason across both

texts—deducing that Zimmer left Tenet to score Dune (his second Oscar), leading to Göransson's hiring and subsequent second Oscar for Oppenheimer. This explicit relational supervision provides vital disambiguation cues and diverse retrieval paths absent from single-document synthesis, saving the LLM from having to implicitly infer these complex connections from massive unstructured text.

Based on this insight, we propose **WRAP++** (**W**eb discov**R**y **A**mplified **P**retraining), which effectively extends the synthetic data paradigm from single-document *rewriting* to cross-document *discovery and amplification*.

Amplifying Associative Context via Relation Discovery. We discover high-confidence relational motifs from web hyperlinks (Zhou et al., 2022)—specifically *dual-links* ($A \leftrightarrow B$) and *co-mentions* ($A \rightarrow E \leftarrow B$ with $A \rightarrow B$). Rather than simply concatenating these documents, we feed the discovered pairs to an instruction-tuned LLM generator subjected to three strict synthesis constraints: enforcing **Cross-Document Dependency** to mandate joint reasoning, requiring **Explicit Factual Chaining** to decode multi-hop logical paths, and ensuring **Context-Free Internalization** by presenting the synthesized QA as standalone knowledge rather than passage-grounded reading comprehension. This process produces genuinely new relational knowledge (comparisons, contrasts, bridging facts) that creates diverse retrieval paths to the same facts. Furthermore, because the number of valid entity pairs grows combinatorially, this discovery-driven synthesis achieves a $\sim$10$\times$ data amplification—scaling a fixed 8.4B-token source corpus into 80B tokens of cross-document QA data—consistently improving the knowledge recoverability of the model.

We instantiate **WRAP++** on Wikipedia and amplify $\sim$8.4B tokens of raw text into 80B tokens of cross-document QA data—compared to only $\sim$5.4B tokens from single-document WRAP. On the SimpleQA benchmark (Wei et al., 2024), OLMo-based models at both 7B and 32B scales trained with **WRAP++** data substantially outperform all single-document baselines, and **WRAP++** demonstrates a more favorable scaling trajectory than single-document approaches.

Our contributions are three-fold:

1. We propose **WRAP++**, a scalable framework for cross-document synthetic pretraining that uses naturally web hyperlink topology as a model-free signal for relation discovery.
2. We instantiate **WRAP++** on Wikipedia and synthesize 80B+ tokens of cross-document QA data from an 8.4B-token source corpus, showing that topology-guided combinatorial relation discovery enables data amplification far beyond single-document synthesis.
3. We show that **WRAP++** improves factual knowledge acquisition at scale: on SimpleQA, OLMo-based 7B and 32B models trained with **WRAP++** substantially outperform single-document baselines with sustained scaling gains, and **WRAP++** can be integrated into a realistic mid-training mixture without degrading general capabilities.

2 Method: WRAP++

WRAP++ is a framework that transitions synthetic pretraining data from single-document rewriting to topology-guided cross-document discovery and amplification. The framework consists of two core stages: Topological Relation Discovery (§2.2) and Joint QA Synthesis (§2.3). Figure 1 illustrates the overall pipeline.

2.1 Problem Formulation and Graph Abstraction

Let $\mathcal{D} = \{d_1, d_2, \ldots, d_N\}$ denote a large-scale web corpus consisting of N documents. The inter-document references (e.g., hyperlinks) naturally induce a directed knowledge graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where each vertex $v_i \in \mathcal{V}$ corresponds to a document d_i, and a directed edge $e_{i,j} \in \mathcal{E}$ exists if d_i explicitly references d_j. Since each document in our corpus describes a single entity, we use "entity pair" and "document pair" interchangeably throughout this paper.

Conventional single-document synthesis operates solely on the local context of v_i, limiting the model's exposure to isolated facts. In contrast, **WRAP++** leverages the topological structure of $\mathcal{G}$ to discover genuine semantic dependencies across documents, bringing related knowledge into a shared synthesis context to amplify the associative context of factual knowledge.

2.2 Topological Relation Discovery

A naïve approach of pairing random documents from $\mathcal{D}$ forces the synthesis model to hallucinate

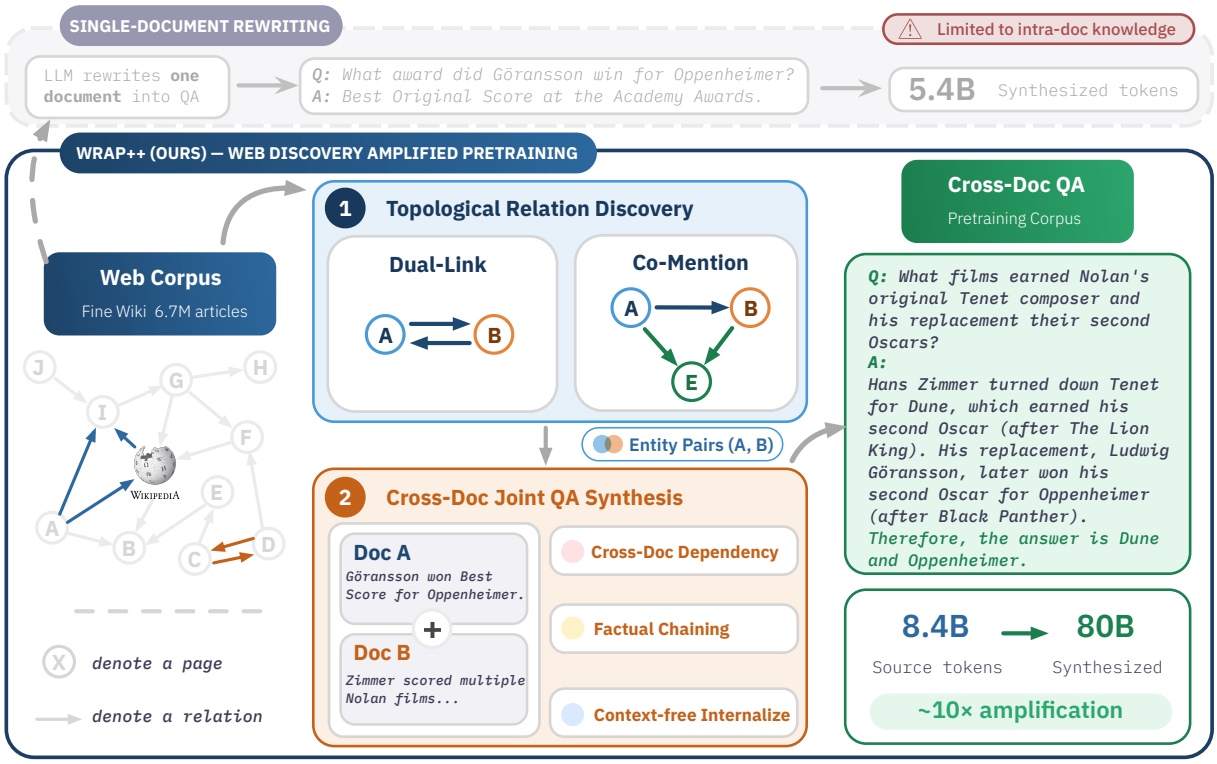

Figure 1: Overview of the **WRAP++** pipeline. Unlike single-document WRAP, which rewrites individual documents, **WRAP++** *discovers* cross-document entity relationships from web topology and *amplifies* them into pretraining data through joint QA synthesis.

spurious connections. To ensure the semantic validity of cross-document synthesis, we discover high-confidence relational motifs directly from $\mathcal{G}$. We focus on two topological structures that provide strong inductive biases for relational reasoning:

Dual-link Motif. Two documents u and v form a dual-link relationship if they mutually reference each other. Formally, a dual-link pair (u, v) satisfies the bidirectional constraint:

$$e_{u,v} \in \mathcal{E} \wedge e_{v,u} \in \mathcal{E} \tag{1}$$

This mutual dependency typically indicates a strong, foundational semantic correlation (e.g., a notable director and their magnum opus, or a scientist and their core discovery). Discovering this motif ensures the underlying entity pair is highly coupled.

Co-mention Motif. Documents u and v share a co-mention relationship if they both reference a common structural hub E, while maintaining a direct link between themselves. Formally, the triplet (u, v, E) satisfies:

$$e_{u,E} \in \mathcal{E} \wedge e_{v,E} \in \mathcal{E} \wedge e_{u,v} \in \mathcal{E} \tag{2}$$

The shared structural context E imposes implicit analogical, hierarchical, or comparative relationships (e.g., two competing theories cited in the same survey article). This motif explicitly encourages the subsequent synthesis model to generate relational knowledge that contrasts and compares the related entities, thereby amplifying their associative context.

2.3 Cross-Document Joint QA Synthesis

Given a discovered document pair (d_u, d_v) connected by a valid topological motif, we employ an instruction-tuned LLM generator $\mathcal{M}_\theta$ to synthesize a set of composite QA instances $\mathcal{S}_{u,v} = \{(q_i, c_i, a_i)\}_{i=1}^{K}$, where q_i is the question, c_i is the intermediate factual chain, and a_i is the final answer.

The generation process is conditioned on a structured prompt $\mathcal{P}$ and the concatenated document context:

$$\mathcal{S}_{u,v} \sim \mathcal{M}_\theta(d_u \oplus d_v, \mathcal{P}) \tag{3}$$

To amplify the associative context of the synthesized data and prevent the generator from degrading to shallow single-document summarization, $\mathcal{P}$

enforces three functional constraints on the output space:

- **Strict Cross-Document Dependency:** The generated question q_i should not be answerable from either document alone. Deriving the correct answer a_i must require complementary premises from both d_u and d_v, ensuring that the synthesis captures cross-document associations rather than merely rephrasing isolated single-document facts.

- **Explicit Factual Chaining:** Before outputting a_i, the generator must explicitly decode the traversal path c_i. By articulating the necessary facts extracted from both documents and linking them step-by-step, the pretraining model internalizes multi-hop knowledge structures, effectively creating diverse associative entry points to the underlying facts.

- **Context-Free Internalization:** The generator should not attribute facts to the provided passages (e.g., "According to Passage A") or rely on local passage labels. Since the source documents are assumed to have already appeared in pretraining, the synthesized QA presents cross-document relations as standalone knowledge, reinforcing associations among known facts rather than repeating the original context as a reading-comprehension exercise.

3 Experiments

3.1 Experimental Setup

Synthesis Data. In principle, **WRAP++** is a general framework applicable to any text corpus containing hyperlinks. In this work, we instantiate it on Wikipedia because of its rich link structure and its widespread use in prior work on synthetic rewriting methods (Maini et al., 2024; Su et al., 2025). Specifically, we use the English subset of FineWiki (Penedo, 2025) as our base corpus ($\mathcal{D}$), which contains approximately 8.4B tokens. We parse the hyperlinks in FineWiki to construct the directed inter-document graph $\mathcal{G}$ used for topological relation discovery.

Synthesis Model. We use Qwen3-30B-A3B-Instruct-FP8 as our instruction-tuned generator $\mathcal{M}_\theta$. The prompt is designed to enforce the strict cross-document dependency and explicit factual chaining constraints described in §1, thereby encouraging high-quality relational QA generation. The full prompt template is provided in Appendix C.

Data Scale. Topological relation discovery substantially amplifies the data scale beyond individual documents. The dual-link motif yields highly coupled entity pairs that produce ∼3B tokens of cross-document QA data. Incorporating the co-mention motif broadens coverage, bringing the combined **WRAP++** dataset to ∼82.7B tokens.

Training Models. To assess the effect of cross-document synthetic data on parametric knowledge, we continue pretraining from the OLMo-3 stage-1 last checkpoint at both the 7B and 32B scales for one epoch. We choose OLMo because it is fully open-source and provides publicly released checkpoints throughout training, making it a suitable platform for controlled continued-pretraining experiments.

3.2 Evaluation Setup

Benchmark. We evaluate on SimpleQA (Wei et al., 2024), a knowledge-intensive benchmark designed to measure short-form factual accuracy while minimizing sensitivity to formatting heuristics. Most SimpleQA questions can be answered directly from Wikipedia-derived knowledge, making it a natural testbed for studying factual knowledge acquisition under our setup.

Metric. We use pass@128 as our primary metric, defined as the empirical probability that at least one of 128 sampled responses contains the correct fact. Our goal is to measure *knowledge recoverability* rather than only top-1 answer accuracy. In this setting, pass@128 is useful because it probes whether the correct factual association can be elicited from the model under sampling, even when it is not the single most likely surface form. We therefore treat it as a more sensitive indicator of parametric knowledge recoverability during continued pretraining than pass@1.

Baselines. To isolate the benefit of cross-document synthesis, we compare **WRAP++** against two closely related single-document baselines derived from the same FineWiki corpus:

1. **WRAP (∼5.4B tokens)**: Standard single-document QA synthesis following the original WRAP recipe (Maini et al., 2024), represent-

Table 1: SimpleQA results after 1-epoch continued training on OLMo-3-7B and OLMo-3-32B using different data recipes. The metric reported is the empirical pass@128 rate (%).

Data Recipe	OLMo-3-7B	OLMo-3-32B
Pretrained Base	34.76	42.35
+ WRAP	39.55	44.43
+ Extended WRAP	43.69	47.91
+ WRAP++	**49.13**	**53.97**

ing the typical yield of intra-document fact extraction.

2. **Extended WRAP (∼17.4B tokens)**: An expanded single-document synthesis utilizing additional prompting strategies (e.g., exhaustive extraction) to push the limits of single-document scaling.

These baselines also illustrate the data-scaling constraint of single-document synthesis: on the same FineWiki corpus, standard WRAP yields only ∼5.4B tokens and Extended WRAP reaches ∼17.4B tokens, both well below the ∼82.7B-token scale of **WRAP++**.

3.3 Main Results

Table 1 presents the main results of 1-epoch continued training across different synthesis recipes. We highlight two principal findings.

WRAP++ substantially outperforms single-document baselines. Across both model scales, continued pretraining with **WRAP++** yields substantially higher pass@128 on SimpleQA compared to all single-document approaches (*+9.5 pp on 7B, +9.8 pp on 32B over WRAP; +5.4 pp on 7B, +6.1 pp on 32B over Extended WRAP*). This advantage reflects two complementary factors. First, cross-document synthesis produces higher-quality relational knowledge *per token*: at a matched budget of ∼8B tokens, **WRAP++** already outperforms Extended WRAP by +2.48 pp (detailed in §4.5), confirming a genuine quality advantage independent of data scale. Second, the combinatorial nature of relation discovery amplifies this quality advantage to a far larger data space (∼80B tokens) that single-document methods cannot access, yielding further gains as training progresses (Figure 2).

Surpassing the single-document scaling bottleneck. Single-document methods face an inherent data bottleneck: the finite number of extractable facts within an individual page. While Extended WRAP attempts to push this limit through exhaustive extraction (reaching ∼17.4B tokens), it ultimately depletes the source material. The resulting diminishing returns (+4.1 pp on 7B, +3.7 pp on 32B over standard WRAP) suggest information saturation under the single-document paradigm. In contrast, because the number of valid cross-document entity pairs grows combinatorially, **WRAP++** amplifies the same FineWiki source corpus into ∼80B tokens of relational knowledge—a data space fundamentally inaccessible to single-document methods. Single-document methods *cannot* close this gap simply by training longer, since their source material is already exhausted. We analyze the resulting scaling dynamics in detail next.

3.4 Scaling and Training Dynamics

Figure 2 plots pass@128 as a function of training tokens consumed. The trajectories confirm the scaling bottleneck discussed above: single-document recipes plateau early, whereas **WRAP++** maintains a steady upward trend all the way to 80B tokens without obvious saturation, demonstrating that the combinatorial data space opened by relation discovery translates into sustained knowledge gains.

To further dissect how this scaling translates into knowledge recoverability, we track the evolution of the pass@k curves throughout training. Figure 3 visualizes the unbiased SimpleQA pass@k for both models. As training progresses (indicated by the light-to-dark blue gradient representing the 80B token influx), the curves exhibit a monotonic upward shift across all values of k.

Crucially, this improvement spans the entire logarithmic k-spectrum. The persistent lift at small k (the leftmost regions of the curves) shows that the model's top-ranked answers increasingly contain the correct fact, reflecting higher precision. Simultaneously, the parallel gains at larger k indicate a broader and more robust set of associative retrieval paths to the same knowledge.

4 Ablations and Analysis

We conduct extensive ablations to validate each component of **WRAP++**. Due to experimental costs, unless otherwise noted, all ablation experiments use OLMo-3-7B continued pretraining with ∼8B tokens and report the results in terms of SimpleQA pass@128.

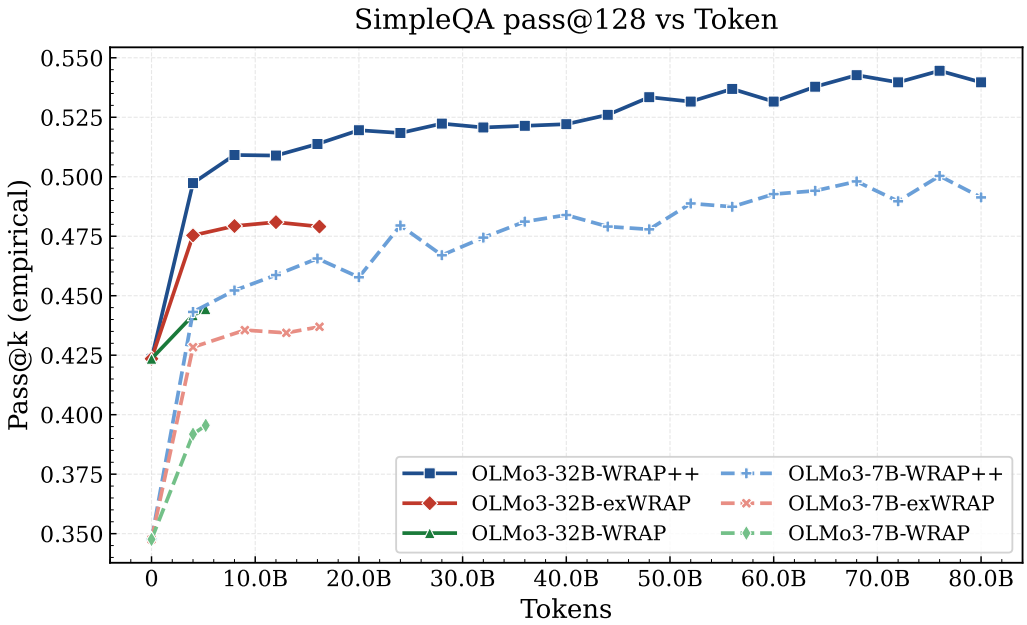

Figure 2: **SimpleQA pass@128 vs. training tokens.** Single-document recipes (WRAP and Extended WRAP) reach a data bottleneck early, limiting further knowledge acquisition. In contrast, the combinatorial nature of **WRAP++** allows it to scale effectively up to 80B tokens, improving performance without obviously plateauing.

Table 2: Ablation results across different design choices (OLMo-3-7B, ∼8B tokens). We report SimpleQA pass@128. **WRAP++** (default) uses topological relation discovery, combined topologies, joint QA synthesis, and the Qwen3-30B-A3B synthesis model.

Ablation Axis	Variant	pass@128
Pairing strategy	Random entity pairing	43.46
	Topological discovery	**45.11**
Relation topology	Dual-link only	44.24
	Co-mention only	44.36
	Combined	**45.11**
Synthesis Method	Raw document concatenation	35.43
	QA with source docs prepended	38.93
	Cross-document joint QA	**45.11**
Synthesis model	Qwen3-30B-A3B	45.11
	Qwen3-235B-A22B	**47.70**

4.1 Necessity of Topological Relation Discovery

We explore whether the topological relation discovery is essential, or whether randomly pairing Wikipedia pages would suffice. Table 2 (Pairing Strategy rows) shows a clear performance drop (from 45.11 to 43.46) when entities are paired randomly rather than via dual-link or co-mention relations. Qualitatively, random pairing forces the synthesis model to fabricate relationships between unrelated entities, producing factually incorrect comparisons and superficial connections. This confirms that principled relation discovery—specifically, dual-link and co-mention motif discovery—is important for high-quality cross-document synthesis.

4.2 Topology Comparison: Dual-Link vs. Co-Mention

We explore the contribution of each relation type at a matched token budget of ∼8B. As shown in Table 2 (Relation Topology rows), both topologies provide strong relational signal at this budget. Co-mention retains a slight edge over dual-link (44.36 vs. 44.24), while their combination yields the best overall performance (45.11). This suggests that bidirectional links and shared structural context capture yet complementary aspects of cross-document knowledge.

4.3 Necessity and Format of QA Synthesis

We explore the optimal data format for learning cross-document relationships by comparing our joint QA synthesis against two alternatives: (1) raw concatenation of related documents (no QA), and (2) prepending source documents to the synthesized QA pairs. Table 2 shows raw concatenation performs only marginally above the pretrained base (35.43 vs. 34.8), indicating that explicit *synthesis* is essential to convert document proximity into learnable relational knowledge. Moreover, prepending source documents to QA pairs causes a notable performance drop (38.93), likely by allowing the model to superficially copy answers rather than

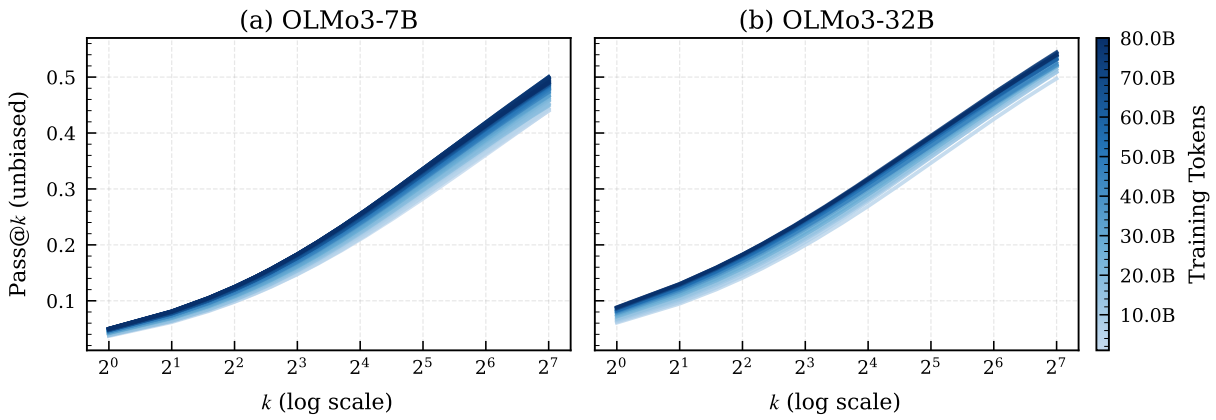

Figure 3: **Evolution of pass@k performance during training.** The curves illustrate the unbiased pass@k of OLMo-3-7B (a) and OLMo-3-32B (b) on SimpleQA ($k \in [1, 128]$ in log scale). The color gradient (from light to dark blue) tracks the accumulation of consumed **WRAP++** tokens (from 10B to 80B). The strictly monotonic upward shift across all values of k indicates robust, unsaturated knowledge internalization.

Table 3: Performance comparison of **WRAP++** mixed with other single-document strategies. Each mixture contains ∼8B tokens total.

Single-Document Component	Proportion of **WRAP++** in Mixture	
	0% (Baseline Only)	50% (1:1 Mix)
Pretrained Model	34.76	
Raw FineWiki	39.23	41.80
Distill	38.12	41.59
Extract Knowledge	38.69	43.35
Knowledge List	38.60	42.03
WRAP++ (100%, Ours)	45.11	

parametrically internalizing them. Thus, joint QA synthesis provides the most effective format for amplifying associative context.

4.4 Effect of Synthesis Model Scale

We explore the effect of synthesis model scale on **WRAP++** quality by comparing two synthesis models of different scales: Qwen3-30B-A3B and Qwen3-235B-A22B (22B active parameters). Table 2 (Synthesis Model Scale rows) shows that the larger model produces higher-quality cross-document QA, leading to better downstream pass@128. In practice, the choice between synthesis models involves a cost–quality tradeoff: the larger model is preferable when generation budget is not the bottleneck, while the smaller model enables broader coverage at lower compute cost.

4.5 Comparison with Other Single-Document Strategies

We contextualize the performance of **WRAP++** against other representative single-document rephrasing strategies Su et al. (2025) applied to the identical FineWiki source corpus, including: (a) **Distill**—rewriting into cleaner, more concise prose while preserving information; (b) **Extract Knowledge**—extracting key factual statements and discarding redundancy; (c) **Knowledge List**—outputting structured knowledge in list format.

Table 3 presents both the isolated performance (0% and 100%) and the mixing dynamics (50% blending) at a strictly restricted budget of ∼8B tokens. When evaluated in isolation at the 8B-token budget, pure **WRAP++** (45.11) substantially outperforms all listed single-document baselines. This margin at a restricted data scale reveals an important dynamic in pretraining efficiency: discovery-driven synthesis already yields stronger knowledge recoverability, even before exploiting its data amplification headroom. When blending **WRAP++** with other strategies in a 1:1 ratio, we observe a clear *uplift effect*: injecting **WRAP++** into any weaker baseline consistently improves upon its 0% counterpart (e.g., Knowledge List rises from 38.60 to 42.03), further highlighting the advantage of discovery-driven synthesis.

4.6 Integration with OLMo-3 Mid-Training Data

We further explore whether **WRAP++** data can be integrated into a realistic mid-training pipeline without harming general capabilities. We augment OLMo-3's 100B-token mid-training mixture with **6B tokens of WRAP++** data and train for the full schedule. As a baseline, we train on the original OLMo-3 mid-training mixture under identical conditions. Following Niklaus et al. (2026), we evalu-

Table 4: Integrating **WRAP++** into OLMo-3-7B 100B-token mid-training. We report SimpleQA pass@128 and the average over 12 general benchmarks (including MMLU Redux, HellaSwag, etc. See Appendix E for the full list of tasks).

Setting	SimpleQA	Gen. Avg
	pass@128	(12 tasks)
Pretrained Base	34.76	57.79
+ Midtrain (100B)	34.74	**68.24**
+ **WRAP++** Mix (100B)	**37.58**	68.16

ate both SimpleQA and the average performance across 12 general tasks (detailed in Appendix E). Table 4 shows that adding **WRAP++** data yields a meaningful improvement on SimpleQA (+2.9 points pass@128) while maintaining a comparable general-benchmark average (68.16 vs. 68.24). Notably, **WRAP++** explicitly enhances knowledge-intensive tasks, yielding clear gains on MMLU Redux (+1.28, detailed in Appendix E). This demonstrates that **WRAP++** integrates cleanly into full-scale mid-training, preserving broad capabilities while explicitly strengthening the model's general knowledge foundation. Moreover, as the mid-training budgets of leading models grow toward the trillion-token regime (Yang et al., 2025a; Zeng et al., 2026), the 80B-token scale of **WRAP++** suggests strong potential for integration into future large-scale training pipelines.

5 Related Work

Synthetic Data for Web-Scale LLM Pretraining. Synthetic data has been explored as a way to improve the quality and learnability of pretraining corpora, from generator-driven datasets such as TinyStories, Phi-style textbooks, and Cosmopedia (Eldan and Li, 2023; Gunasekar et al., 2023; Ben Allal et al., 2024) to web-grounded rephrasing methods. WRAP (Maini et al., 2024) showed that rewriting web documents into cleaner QA-style text can accelerate pretraining by $\sim 3\times$, motivating a line of work on synthetic web recycling. Subsequent methods improve the rephrasing recipe through richer target formats and quality-controlled rewriting (Su et al., 2025; Nguyen et al., 2025; Maini et al., 2025; Niklaus et al., 2026), study the effect of generator size (Maini et al., 2024, 2025), and analyze how source data quality affects synthetic data utility (Penedo et al., 2024; Li et al., 2024; Nguyen et al., 2025; Niklaus et al., 2026). Overall, this line of work primarily improves the quality, style, or density of individual documents, while leaving cross-document relational structure largely implicit.

Synthetic Data with Cross-Document Structure. Recent work has begun to make document or entity relations explicit during synthetic pretraining. EntiGraph (Yang et al., 2024) uses entity-centered synthesis to expand a small corpus for continued pretraining, while its synthesis is primarily driven by entity relations extracted from source documents, rather than explicit cross-document relation discovery. SoG (Ma et al., 2025) further emphasizes cross-document entity associations, using context-graph paths and entity-conditioned generation to improve multi-hop and domain-specific QA. SBP (Yang et al., 2025b) instead learns inter-document correlations implicitly by retrieving semantically similar document pairs and training a conditional synthesizer for pretraining from scratch. **WRAP++** is complementary to these efforts: it uses naturally available hyperlink topology as a model-free relational signal, avoiding costly entity extraction or embedding-based graph construction. This enables scalable cross-document synthesis over web-scale documents. On Wikipedia, **WRAP++** mines high-confidence relational motifs and converts them into joint QA data, amplifying an 8.4B-token source corpus into 80B+ synthetic QA tokens.

6 Conclusion

We presented **WRAP++**, a scalable framework for cross-document synthetic pretraining. Instead of relying on costly entity extraction or embedding-based graph construction, **WRAP++** uses naturally available web hyperlink topology as a model-free signal for relation discovery. By mining high-confidence relational motifs such as dual-links and co-mentions from Wikipedia hyperlinks, **WRAP++** converts related document pairs into joint QA data that explicitly captures cross-document associations. This enables large-scale knowledge amplification, expanding an 8.4B-token source corpus into 80B+ synthetic tokens. Experiments on OLMo-3 7B and 32B models show that **WRAP++** substantially outperforms single-document baselines on SimpleQA, maintains favorable scaling up to 80B tokens, and can be integrated into a realistic mid-training mixture without degrading general capabilities.

Limitation

Our experiments instantiate **WRAP++** on Wikipedia, which is a clean and entity-centric corpus. Extending to noisier web corpora (e.g., Common Crawl), where hyperlinks include advertisements, navigation elements, and low-quality references, will require additional filtering heuristics; we are actively exploring this direction.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

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

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, and 1 others. 2020. Piqa: Reasoning about physical commonsense in natural language. *Proceedings of the AAAI conference on artificial intelligence*, 34(05):7432–7439.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *Preprint*, arXiv:2107.03374.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Tri Dao. 2023. Flashattention-2: Faster attention with better parallelism and work partitioning. *Preprint*, arXiv:2307.08691.

Dheeru Dua, Yizhong Wang, Pradeep Dasigi, Gabriel Stanovsky, Sameer Singh, and Matt Gardner. 2019. Drop: A reading comprehension benchmark requiring discrete reasoning over paragraphs. *arXiv preprint arXiv:1903.00161*.

Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

Aryo Pradipta Gema and 1 others. 2024. Are we done with mmlu? *arXiv preprint arXiv:2407.03120*.

Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, and 1 others. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash Guha, Sedrick Keh, Kushal Arora, and 1 others. 2024. Datacomp-lm: In search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

Shengjie Ma, Xuhui Jiang, Chengjin Xu, Cehao Yang, Liyu Zhang, and Jian Guo. 2025. Synthesize-on-graph: Knowledgeable synthetic data generation for continue pre-training of large language models. *arXiv preprint arXiv:2505.00979*.

Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Urbanek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, and 1 others. 2025. Beyondweb: Lessons from scaling synthetic data for trillion-scale pretraining. *arXiv preprint arXiv:2508.10975*.

Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. *Preprint*, arXiv:2401.16380.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. *arXiv preprint arXiv:1809.02789*.

Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. 2025. Recycling the web: A method to enhance pre-training data quality and quantity for language models. In *Second Conference on Language Modeling*.

Joel Niklaus, Guilherme Penedo, Hynek Kydlicek, Elie Bakouch, Lewis Tunstall, Ed Beeching, Thibaud Frere, Colin Raffel, Leandro von Werra, and Thomas Wolf. 2026. The synthetic data playbook: Generating trillions of the finest tokens.

Team Olmo, :, Allyson Ettinger, Amanda Bertsch, Bailey Kuehl, David Graham, David Heineman, Dirk Groeneveld, Faeze Brahman, Finbarr Timbers,

Hamish Ivison, Jacob Morrison, Jake Poznanski, Kyle Lo, Luca Soldaini, Matt Jordan, Mayee Chen, Michael Noukhovitch, Nathan Lambert, and 50 others. 2025. Olmo 3. *Preprint*, arXiv:2512.13961.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480.

Guilherme Penedo. 2025. Finewiki. Source: Wikimedia Enterprise Snapshot API (https://api.enterprise.wikimedia.com/v2/snapshots). Text licensed under CC BY-SA 4.0 with attribution to Wikipedia contributors.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. 2024. The fineweb datasets: Decanting the web for the finest text data at scale. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for squad. *arXiv preprint arXiv:1806.03822*.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. Winogrande: An adversarial winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2025. Nemotron-CC: Transforming Common Crawl into a refined long-horizon pretraining dataset. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2459–2475, Vienna, Austria. Association for Computational Linguistics.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Preprint*, arXiv:1811.00937.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, and 21 others. 2024. 2 OLMo 2 Furious. *Preprint*, arXiv:2501.00656.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Preprint*, arXiv:2411.04368.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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

Zitong Yang, Aonan Zhang, Hong Liu, Tatsunori Hashimoto, Emmanuel Candès, Chong Wang, and Ruoming Pang. 2025b. Synthetic bootstrapped pretraining. *arXiv preprint arXiv:2509.15248*.

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 3472–3483.

Aohan Zeng, Xin Lv, Zhenyu Hou, Zhengxiao Du, Qinkai Zheng, Bin Chen, Da Yin, Chendi Ge, Chengxing Xie, Cunxiang Wang, and 1 others. 2026. Glm-5: from vibe coding to agentic engineering. *arXiv preprint arXiv:2602.15763*.

Jiawei Zhou, Xiaoguang Li, Lifeng Shang, Lan Luo, Ke Zhan, Enrui Hu, Xinyu Zhang, Hao Jiang, Zhao Cao, Fan Yu, Xin Jiang, Qun Liu, and Lei Chen. 2022. Hyperlink-induced pre-training for passage retrieval in open-domain question answering. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7135–7146, Dublin, Ireland. Association for Computational Linguistics.

A Training Details

Architecture. We use the OLMo-3 architecture (Olmo et al., 2025), a decoder-only transformer with the Dolma-2 tokenizer (vocabulary size padded to a multiple of 128). Experiments are conducted at two scales: OLMo-3-7B (7 billion parameters) and OLMo-3-32B (32 billion parameters). We utilize the `olmo-core` training framework (Team OLMo et al., 2024), which provides a highly optimized and reproducible infrastructure for large-scale distributed training. All models use FlashAttention-2 (Dao, 2023) as the attention backend.

Continued Pretraining. All experiments initialize from official OLMo-3 pretrained checkpoints: step 1,413,814 for 7B and step 679,000 for 32B. We load only the model weights and optimizer state (no trainer state) and continue pretraining on synthetic

Table 5: Continued pretraining hyperparameters. All experiments use a linear decay schedule (no warmup, decaying to 0) and SkipStepAdamW (Olmo et al., 2025) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, weight decay 0.1 (embedding weights excluded), max gradient norm 1.0, and auxiliary z-loss with multiplier 10^{-5}. Parameters are stored in bfloat16 and gradients are reduced in float32 via HSDP.

Hyperparameter	**OLMo-3-7B (8B)**	**OLMo-3-7B (80B)**	**OLMo-3-32B**
Sequence length	8,192	8,192	8,192
Global batch size (tokens)	~2M (2^{21})	~2M (2^{21})	~4M (4×2^{20})
Peak learning rate	2.07×10^{-5}	2.07×10^{-5}	2.07×10^{-5}
LR schedule	Linear $\rightarrow 0$	Linear $\rightarrow 0$	Linear $\rightarrow 0$
Warmup steps	0	0	0
Training steps	4,000	40,000	10,000
Training tokens	~8B	~80B	~80B
Precision	bfloat16	bfloat16	bfloat16
Data parallel	HSDP (block wrap)	HSDP (block wrap)	HSDP (full wrap, shard 64)
Activation ckpt.	FFN-only	FFN-only	Budget (50%)
GPUs	256 × H20	256 × H20	1024 × H20

data mixtures for 1 epoch. Table 5 summarizes the hyperparameters for each configuration.

The learning rate of 2.07×10^{-5} is inherited from the official OLMo-3 mid-training recipe. We use zero warmup steps because the optimizer state is loaded from the pretrained checkpoint, ensuring stable training from the first step. For the OLMo-3-7B scale, training on ~8B tokens using 256 H20 GPUs takes approximately 4.5 hours with a Model FLOPs Utilization (MFU) of ~65%. For the larger OLMo-3-32B scale, training on ~80B tokens using 1,024 H20 GPUs requires approximately 1 day and 20 hours. These benchmarks demonstrate the efficiency and scalability of our training pipeline on modern hardware.

Data Format. All synthetic QA data is formatted as plain text with "Question:" and "Answer:" delimiters, consistent with prior WRAP work. For cross-document QA, the synthesized output directly states facts without referencing source passages, ensuring the model internalizes them as parametric knowledge rather than reading comprehension signals.

B Evaluation Details

In-Context Learning Setup. Because our continued pretraining experiments operate on base models (OLMo-3) that have not undergone instruction tuning, these models cannot reliably follow zero-shot formatting directives. To accurately probe their parametric knowledge, we adopt a 5-shot in-context learning protocol for all evaluations, including both SimpleQA and the 12 general benchmarks. Specifically, for each evaluation instance, we prepend the prompt with five demonstration question-answer pairs. For SimpleQA, these demonstrations are sampled directly from the SimpleQA dataset; to ensure strict evaluation integrity and prevent data contamination, any examples used as few-shot demonstrations are explicitly excluded from the active evaluation set during inference.

Decoding and Sampling Parameters. We adopt distinct decoding strategies tailored to the nature of each benchmark. For **SimpleQA**, to compute the pass@k metric (where $n = 128$), we employ nucleus sampling with a temperature of 0.6 and top-p of 0.95 to provide a diverse distribution for knowledge recoverability analysis. In contrast, for the **12 general benchmarks**, we use greedy decoding (temperature 0.0) to ensure deterministic and reproducible outputs across all model comparisons. These parameters are held constant across all model scales and data recipes to ensure a fair evaluation.

Unbiased Estimation of pass@k. While the pass@k metric intuitively represents the probability of generating at least one correct answer within k attempts, empirically estimating this by drawing exactly k samples yields high variance. To achieve a more stable and unbiased estimate, we adopt the methodology introduced by Chen et al. (2021). For each evaluation instance, we generate n total samples ($n \geq k$) and determine the number of correct responses, c. The unbiased estimator for pass@k is then computed as:

$$pass@k = 1 - \frac{n - ck}{nk} \tag{4}$$

where $\cdot\cdot$ denotes the binomial coefficient. In our experiments, we generate $n = 128$ samples per

question. This formulation efficiently leverages all n generated samples to calculate the expected pass rate for any evaluation budget $k \leq n$, thereby reducing variance without requiring repeated sampling passes.

C Synthesis Prompt Templates

Single-Document WRAP Prompt (Baseline). Following Maini et al. (2024), we use the standard QA-style prompt for single-document synthesis:

```
Convert the following paragraph into a
conversational format with multiple tags
of "Question:" followed by "Answer:".
```

WRAP++ Cross-Document Joint QA Prompt. For cross-document synthesis, we provide the full text of both related entities and apply the following instruction. The prompt enforces three critical constraints: cross-document dependency, explicit factual chaining, and Context-Free Internalization (see §2.3).

```
You are an expert data generator for
language model pretraining.

Below are two related Wikipedia
passages:

[Passage A]
{text_a}

[Passage B]
{text_b}

Task:
1) Generate high-quality synthetic QA
pairs that REQUIRE information from BOTH
Passage A and Passage B to answer.
2) The Answer MUST begin with a
step-by-step reasoning process. This
reasoning must explicitly bridge facts
from both passages.
3) Do not use external knowledge.
4) CRITICAL CONSTRAINT: The generated
QA pair will be used to train a
model WITHOUT these passages provided
as context. Therefore, you MUST act as
an omniscient AI stating absolute facts
from your own inherent knowledge.
- DO NOT use any attribution phrases
like 'According to Passage A', 'Passage
B mentions', 'As stated in the text', or
'Based on the provided documents'.
- State the facts directly and
confidently.

Output format (strict):
Question: [Insert a complex question
that bridges facts from both passages]
Answer: [Acting as an omniscient AI,
directly state all necessary factual
premises from both passages, and
logically synthesize them to derive the
conclusion.]
Therefore, [State the final, concise
answer.]
```

Synthesis Hyperparameters. All QA synthesis uses Qwen3-30B-A3B-Instruct (FP8) as the generator with temperature 0.7, top-p = 0.8, and a maximum output length of 32,768 tokens. For the synthesis model scale ablation (§4.4), we additionally use Qwen3-235B-A22B-Instruct.

D Relation Discovery Statistics

Table 6: Statistics of the hyperlink relations discovered from FineWiki (English Wikipedia).

Statistic	Value
Source corpus	FineWiki (English)
Raw corpus tokens	∼8.4B
Wikipedia articles processed	∼6.7M
Dual-link pairs ($A \leftrightarrow B$)	∼9.6M
Co-mention pairs ($A \rightarrow E \leftarrow B$, $A \rightarrow B$)	∼232M
Dual-link synthesized tokens	∼3B
Co-mention synthesized tokens	∼79.7B
Total **WRAP++** tokens	∼82.7B
Single-doc WRAP tokens (baseline)	∼5.4B
Extended WRAP tokens (baseline)	∼17.4B

Table 6 reports the relation discovery and synthesis statistics. The dual-link motif yields a relatively small but high-precision set of ∼9.6M entity pairs, while the co-mention motif provides a much larger pool of ∼232M pairs, enabling substantial combinatorial expansion. The total synthesized corpus of ∼82.7B tokens is approximately 15× larger than single-document WRAP on the same source, illustrating the amplification advantage of relation-driven cross-document synthesis.

E Mid-Training Evaluation Benchmarks

The 12 general tasks used for evaluation in Section 4.6 are categorized as follows:

- **General Knowledge**: ARC (Clark et al., 2018), MMLU Redux (Gema et al., 2024)
- **Reading Comprehension**: SQuAD v2 (Rajpurkar et al., 2018), DROP (Dua et al., 2019)
- **Reasoning**: OpenBookQA (Mihaylov et al., 2018), CSQA (Talmor et al., 2019)
- **Natural Language Understanding**: WinoGrande (Sakaguchi et al., 2021), PIQA (Bisk et al., 2020), HellaSwag (Zellers et al., 2019)
- **Math**: GSM8K (Cobbe et al., 2021)

- **Table Understanding & Trivia**: WikiTableQuestions (Pasupat and Liang, 2015), TriviaQA (Joshi et al., 2017)

Table 7 provides per-benchmark results for the mid-training integration experiment described in Section 4.6.

F Additional Ablation: Co-Mention with Three Documents

In the main co-mention motif ($A \rightarrow E \leftarrow B$ while $A \rightarrow B$), we use the two target entities A and B as input to joint QA synthesis. A natural extension is to also include the bridging page E as a third input document, potentially providing additional shared context. We use a dedicated 3-document prompt template that instructs the synthesis model to generate QA requiring facts from all three passages.

Surprisingly, including the bridging page E does not improve—and slightly hurts—performance (Table 8). We hypothesize that the bridging page introduces distracting context: since E typically links to many entities, its content is broad and may divert the synthesis model from focusing on the specific relationship between A and B. The 2-document formulation used in **WRAP++** strikes a better balance between relational grounding and synthesis focus.

G Synthesized QA Data Statistics

We report comprehensive statistics of the synthesized **WRAP++** dataset to characterize the length distributions and data composition. Statistics are computed over the full corpus of 240,658,065 QA instances across 24,224 JSONL files, processed in parallel using 256 workers. Character-level and word-level lengths are measured on the raw synthesized text after extracting the "Question:" and "Answer:" fields from each record.

Dataset Composition. Table 9 summarizes the overall dataset composition by relation type.

Length Distributions. Table 10 reports the distributional statistics of the synthesized QA text. Questions are concise (median 203 characters, 32 words), while answers are substantially longer (median 1,386 characters, 212 words), reflecting the explicit factual chaining constraint that requires step-by-step reasoning before stating the final conclusion. The overall QA length is concentrated in the 1,000–2,000 character range (56.1% of all instances), with 27.6% in the 2,000–5,000 range and 16.0% in the 500–1,000 range. Fewer than 0.3% of instances fall outside the 500–5,000 character window, indicating a well-controlled generation process. Figure 5 visualizes the question and answer length distributions at both character and word levels.

Comparison Across Relation Types. Table 11 compares the QA length characteristics between the two relation motifs. Dual-link instances produce slightly shorter answers (median 1,210 vs. 1,394 characters), likely because mutual references tend to encode more focused bilateral relationships, whereas co-mention pairs often involve broader categorical or analogical connections that require more elaboration. Despite these differences, both subsets maintain similar question lengths and overall distributional shape. Figure 6 overlays the QA length histograms for both relation types, and Figure 7 provides a violin plot comparison across three length dimensions.

Source Document Lengths. The input Wikipedia passages exhibit substantial length variation. The first passage (`text_a`) has a median length of 4,458 characters (P5–P95: 519–27,963), while the second passage (`text_b`) is generally longer with a median of 9,575 characters (P5–P95: 943–50,004, where 50,004 indicates truncation at the maximum context window). This asymmetry arises because co-mention pairs order documents by the directed edge $A \rightarrow B$, where B (the referenced entity) tends to be a more prominent article. Figure 8 visualizes this distributional asymmetry.

QA Length Bucketed Distribution. Table 12 and Figure 9 provide a bucketed view of the combined QA and answer length distributions. The synthesis process produces a unimodal distribution with the majority of instances in the 1,000–2,000 character range. No instances have empty answers, and fewer than 0.01% of answers are shorter than 200 characters, confirming that the explicit factual chaining constraint effectively prevents degenerate outputs.

H Qualitative Analysis

Example: WRAP vs. WRAP++ Synthesis.
Source entity A: Ludwig Göransson (Swedish composer, film score artist)
Source entity B: Oppenheimer (2023 film directed by Christopher Nolan)

Table 7: Per-benchmark breakdown for mid-training integration (OLMo-3-7B, full 100B-token schedule, 3-shot cloze format). Accuracy (%) is reported for all tasks except SQuAD v2 and DROP, which use token-level F1.

Category	Benchmark	Base	Midtrain	WRAP++ Mix
General Knowledge	ARC	77.47	84.98	**85.07**
	MMLU Redux	60.30	65.30	**66.58**
Reading Comprehension	SQuAD v2	42.27	48.59	**49.99**
	DROP	40.51	**66.51**	65.36
Reasoning	OpenBookQA	76.40	**85.20**	82.40
	CSQA	72.15	**74.20**	72.97
Language Understanding	WinoGrande	53.28	**65.11**	62.98
	PIQA	73.99	74.54	**74.92**
	HellaSwag	59.83	70.15	**73.12**
Math	GSM8K	38.36	**79.53**	77.79
Table & Trivia	WikiTableQ	36.55	43.74	**44.46**
	TriviaQA	62.42	61.01	**62.23**
Average		57.79	**68.24**	68.16

Table 8: Co-mention synthesis using 2 vs. 3 input documents (OLMo-3-7B, ∼8B tokens, SimpleQA pass@8).

Co-mention Variant	SimpleQA pass@8
2-doc (entities A, B only)	**15.74**
3-doc (entities A, B + bridge E)	15.42

Table 9: Composition of the **WRAP++** synthesized QA dataset by relation type.

Relation Type	QA Instances	Proportion
Co-mention	231,292,954	96.1%
Dual-link	9,365,111	3.9%
Total	**240,658,065**	100.0%

As shown in Table 13, the **WRAP++** output creates richer associative context: it contrasts with a commonly confused entity (Zimmer), provides cross-film comparisons, and generates reverse-direction queries—all contributing to more robust knowledge encoding.

I Responsible Use and Artifact Considerations

Risks, licenses, and intended use. Our experiments use public encyclopedic web content from huggingfacce dataset HuggingFaceFW/finewiki and publicly available models and benchmarks. Wikipedia text is distributed under the Creative Commons Attribution Share Alike 4.0 license and Wikimedia Terms of Use, and the other artifacts used in this work are subject to their respective licenses and terms. The generated QA data are intended for research on synthetic pretraining and factual knowledge acquisition, not as a verified factual database. As with other synthetic pretraining corpora, the data may inherit factual errors or topical biases from the source corpus or synthesis model, so downstream use should include appropriate filtering and evaluation.

Personal or sensitive content. We do not collect private user data or conduct human-subject data collection. The source corpus is public encyclopedic text and may naturally mention public figures, organizations, locations, or sensitive topics. Our generation process does not infer private attributes or intentionally introduce personal information beyond the provided source passages.

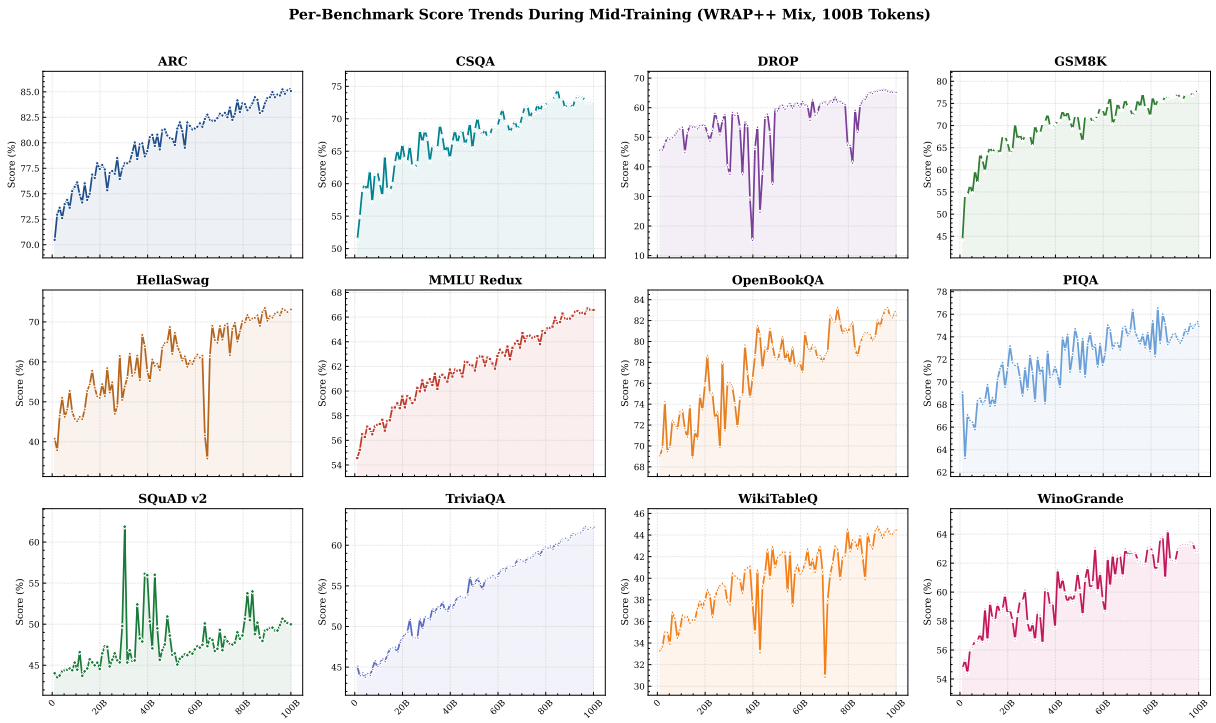

Figure 4: Per-benchmark score trajectories during mid-training with the **WRAP++** Mix (100B tokens) on OLMo-3-7B. Each subplot tracks one of the 12 evaluation tasks over the course of training. All benchmarks exhibit a clear upward trend, with knowledge-intensive tasks (e.g., HellaSwag, GSM8K, CSQA) showing the most pronounced gains.

Table 10: Length statistics of **WRAP++** synthesized QA data (aggregated over all 240.7M instances). "Chars" denotes character count; "Words" denotes whitespace-delimited token count.

Field	Mean	Std	Min	P5	P25	Median	P75	P95
Character-level								
Question	207	43	45	143	177	203	233	283
Answer	1,424	624	139	563	938	1,386	1,817	2,463
QA (combined)	1,651	637	242	766	1,161	1,610	2,051	2,716
Word-level								
Question	33	7	7	23	28	32	37	45
Answer	217	91	16	92	149	212	272	364

Table 11: Median QA lengths (characters) by relation type. Word-level medians are shown in parentheses.

Relation Type	Question	Answer	QA Total
Co-mention	204 (32 words)	1,394 (213 words)	1,618
Dual-link	197 (32 words)	1,210 (189 words)	1,428
Overall	203 (32 words)	1,386 (212 words)	1,610

Table 12: Bucketed length distributions for the combined QA text and answer text (character-level).

Length Range (chars)	QA (combined)		Answer only	
	Count	%	Count	%
$[0, 200)$	0	0.0	2,479	<0.01
$[200, 500)$	456,399	0.2	6,771,177	2.8
$[500, 1{,}000)$	38,609,776	16.0	61,411,377	25.5
$[1{,}000, 2{,}000)$	135,127,249	56.1	132,242,235	55.0
$[2{,}000, 5{,}000)$	66,336,557	27.6	40,114,788	16.7
$[5{,}000, 10{,}000)$	99,207	<0.1	89,225	<0.1
$[10{,}000, \infty)$	28,877	<0.1	26,784	<0.1

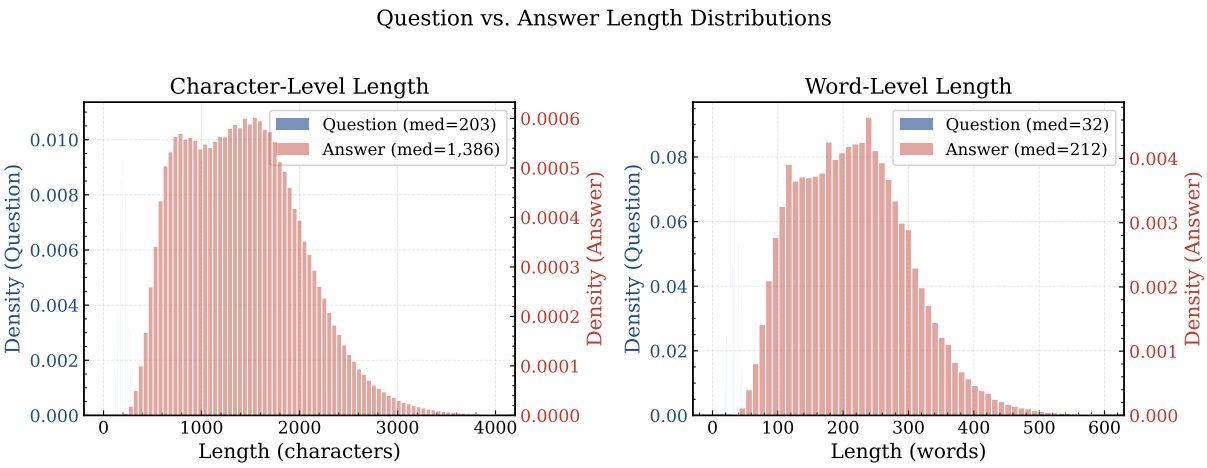

Figure 5: Question vs. answer length distributions at character level (left) and word level (right). Questions are tightly concentrated around a median of 203 characters (32 words), while answers exhibit a broader, right-skewed distribution with a median of 1,386 characters (212 words), reflecting the explicit factual chaining required by the synthesis prompt.

Table 13: Qualitative comparison between single-document WRAP and cross-document **WRAP++** synthesis for the Ludwig Göransson and *Oppenheimer* example.

WRAP (single-document QA on entity A)	**WRAP++ (cross-document QA on entities A+B)**
Q: Who is Ludwig Göransson? **A:** Ludwig Göransson is a Swedish composer known for film scores. **Q:** What films has Göransson scored? **A:** He has scored *Black Panther*, *Tenet*, and *Oppenheimer*.	**Q:** What films earned Nolan's original Tenet composer and his replacement their second Oscars? **A:** Hans Zimmer turned down Tenet for Dune, which earned his second Oscar (after The Lion King). His replacement, Ludwig Göransson, later won his second Oscar for Oppenheimer (after Black Panther). Therefore, the answer is Dune and Oppenheimer.

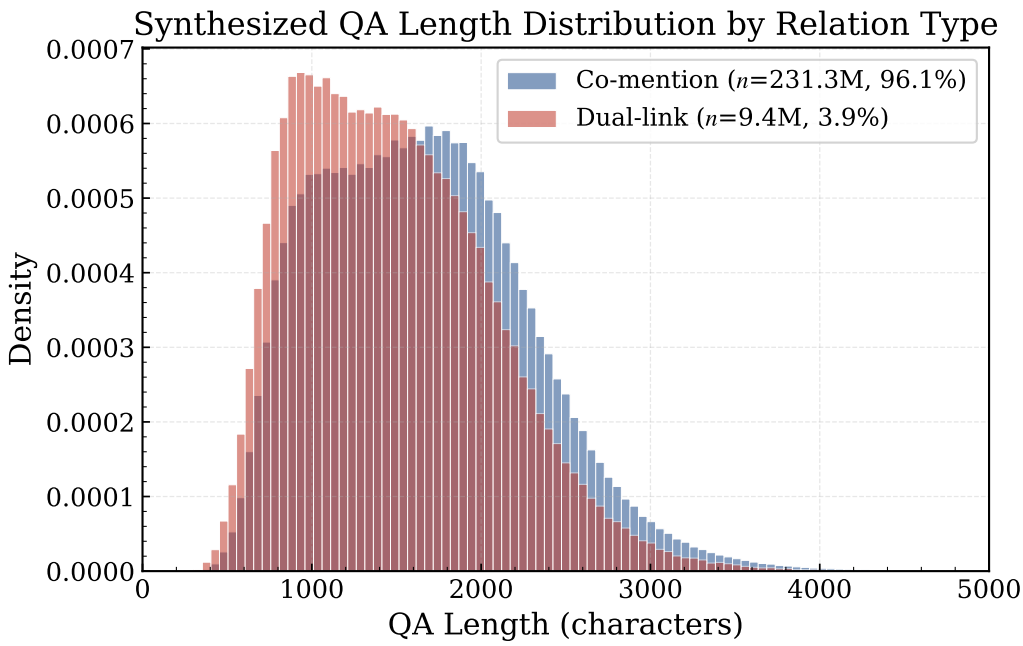

Figure 6: Synthesized QA length distribution by relation type. Both subsets exhibit a similar unimodal shape, with the co-mention distribution (blue) shifted slightly rightward relative to dual-link (red), consistent with the longer factual chains required to bridge co-mentioned entities.

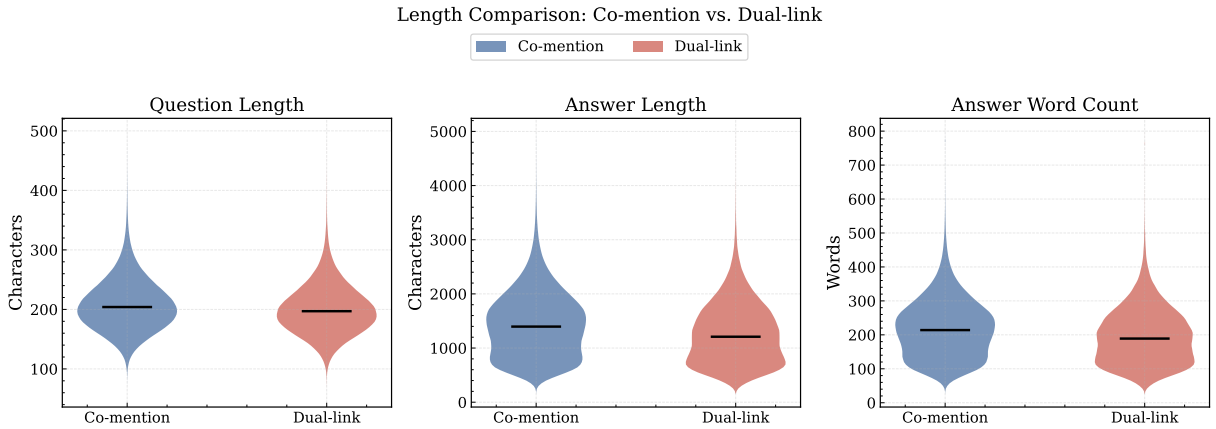

Figure 7: Violin plot comparison of question length, answer length, and answer word count between co-mention and dual-link subsets. Black horizontal lines indicate medians. Both motifs produce similarly distributed questions, while co-mention answers are moderately longer, reflecting the additional elaboration needed to bridge indirectly related entities.

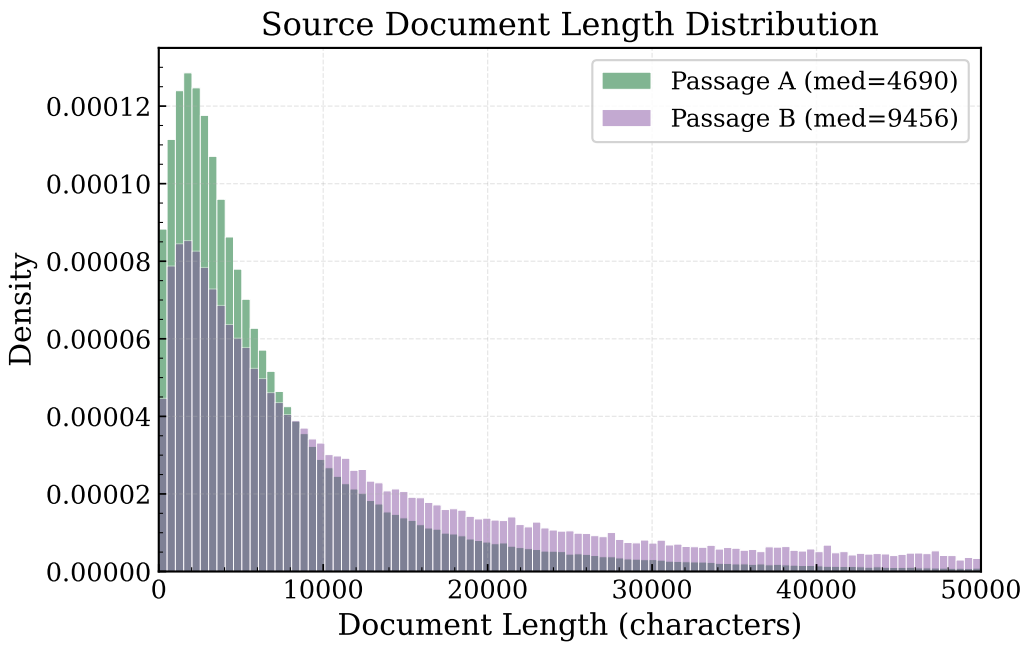

Figure 8: Source document length distributions for the two input passages. Passage A (the referencing entity) is typically shorter (median ≈4,700 chars), while Passage B (the referenced entity) tends to be longer and more prominent (median ≈9,500 chars), with a visible mass accumulation at the 50K truncation boundary.

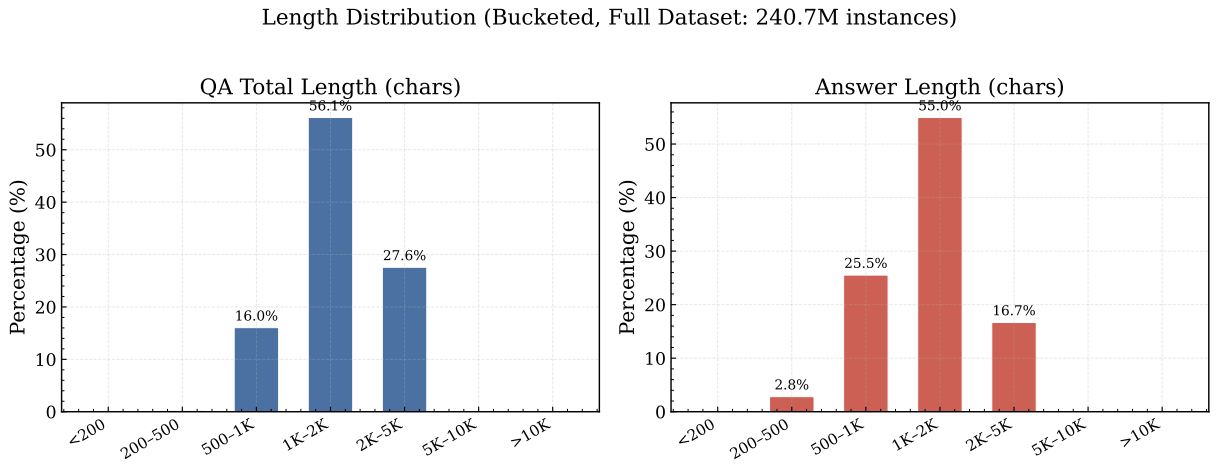

Figure 9: Bucketed length distributions computed over the full dataset of 240.7M instances. Left: combined QA length; right: answer-only length. The 1K–2K character bucket dominates both distributions (>55%), with a secondary concentration in the 2K–5K range for QA and the 500–1K range for answers. Extreme lengths (>5K or <200) are negligible.