

REFORMULATION FOR PRETRAINING DATA AUGMENTATION

005 **Anonymous authors**

006 Paper under double-blind review

ABSTRACT

011 Despite the impressive capabilities of large language models across various tasks,
 012 their continued scaling is severely hampered not only by data scarcity but also
 013 by the performance degradation associated with excessive data repetition during
 014 training. To overcome this critical bottleneck, we introduce the Massive Genre-
 015 Audience (**MGA**) reformulation method, a framework designed to augment cor-
 016 pora in a way that supports more effective model performance scaling. Instead of
 017 relying on complex, predefined seed systems, MGA systematically reformulates
 018 existing corpora into diverse, contextually-rich variations by adaptively gener-
 019 ating genre-audience pairs. We present this framework and the resulting 770 billion
 020 token MGACorpus, created as a practical instantiation of our methodology. We
 021 experimentally validate MGA’s core benefits by demonstrating superior scaling
 022 properties, in terms of both model size and data budget, against data repetition
 023 and upsampling (up to 13B parameters). Furthermore, our comprehensive anal-
 024 ysis investigates the role of synthesis principles in generation quality and reveals
 025 nuances in evaluating model capabilities using standard loss metrics. Our work
 026 shows that a systematic framework like **MGA** provides a reliable pathway to sub-
 027 stantially augment training datasets, effectively alleviating repetition bottlenecks
 028 and enabling more efficient scaling of large language models.

1 INTRODUCTION

031 The remarkable success of Large Language Models (LLMs) heavily relies on the scale of model
 032 parameters and training data (Kaplan et al., 2020; Hoffmann et al., 2022). Scaling laws demonstrate
 033 that improvements in model performance are increasingly dependent on data quantity and quality.
 034 However, the growth rate of available natural language corpora significantly lags behind the increas-
 035 ing demand for training data (Villalobos et al., 2022). While data repetition is a standard tool in
 036 traditional deep learning, it backfires in LLM pre-training, where it degrades performance and cre-
 037 ates a major scaling bottleneck. This raises a critical question: how can we fully utilize the potential
 038 of existing data in data-constrained situations?

039 Leveraging LLMs to synthesize high-quality training data has emerged as a frontier approach (Su
 040 et al., 2024; Abdin et al., 2024). In theory, data synthesis can generate limitless training material,
 041 expanding datasets without the negative consequences of repetition. However, while the promise
 042 of synthetic data is clear, the specific methodologies, the ‘**how**’ behind successful large-scale data
 043 synthesis often remain opaque, existing as black-box processes within large industrial labs rather
 044 than as systematic, reproducible science. Many prevailing methods depend on large-scale models
 045 for generation, effectively creating “distillations” rather than true data augmentations, or require so-
 046 phisticated, pre-defined seed curation systems (Abdin et al., 2024; Ben Allal et al., 2024). These
 047 dependencies introduce substantial computational bottlenecks and limit their accessibility and scal-
 048 ability for the broader research community.

049 In this work, we propose **MGA** (Massive Genre-Audience reformulation), a transparent and prin-
 050 cipled framework designed to directly address the data repetition challenge by augmenting the raw
 051 text and creating more unique tokens. As illustrated in Figure 1, the **MGA** framework is efficiently
 052 implemented using a lightweight 3.3B MoE model. Crucially, it avoids complex external seed sys-
 053 tems by adaptively generating diverse genre-audience pairs directly from raw input documents. This
 design makes the data generation process highly efficient and applicable to web-scale corpora.

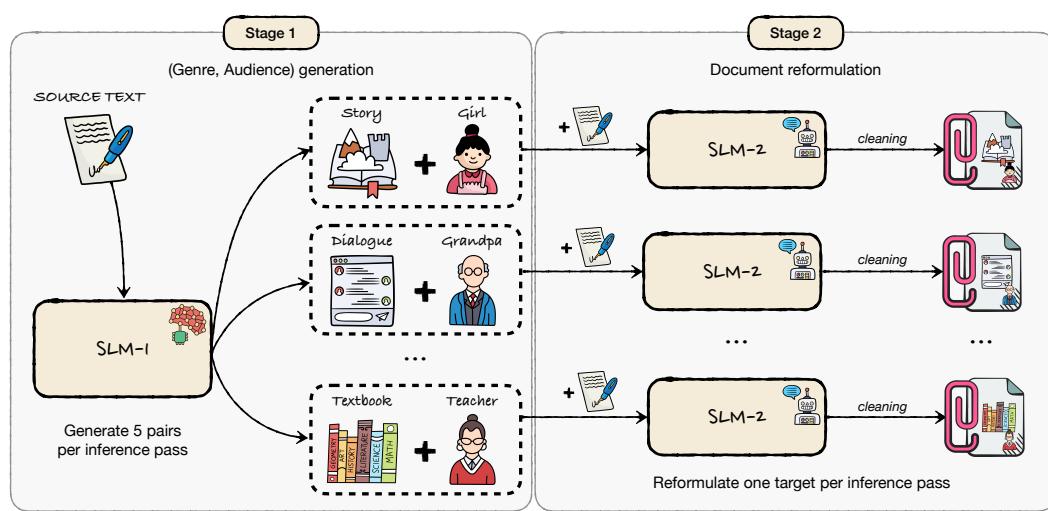


Figure 1: Overview of the **MGA** framework. Our framework expands the original corpus through a two-stage synthesis followed by a cleaning stage process. Each document is reformulated to 5 new documents, finally achieving a 3.9x token number expansion while maintaining diversity through adaptively generated (genre, audience) pairs.

However, proposing a framework is only the first step. To establish its scientific merit and provide actionable insights for the community, a deeper investigation is essential. We contend that a thorough understanding requires answering three fundamental questions: 1) How does MGA reformulation complement existing synthetic data strategies? 2) What is the core mechanism, specifically the role of diversity, that drives its effectiveness in data-scarce scenarios (especially high repetition scenarios)? 3) Why does reformulation fundamentally benefit the model’s learning process? By addressing these questions, our main contributions are:

- We introduce the **MGA framework**, a systematic and reproducible methodology for corpus reformulation. To validate our framework and ensure full reproducibility, we will release the **MGACorpus** (a 770B token dataset) and open-source all key artifacts, including prompts, tool-model finetuning data, and cleaning scripts. Our experiments demonstrate that models trained on MGACorpus significantly outperform those trained on the original corpus it expands upon.
- We experimentally validate MGA’s superior scaling properties in terms of both model size and data budget, revealing a widening performance gap over standard data repetition and upsampling across a wide range of model sizes (377M/1.7B/7B/13B) and data budget (up to 700B tokens).
- We analyze synthetic data collapse from two key perspectives: first, we characterize how synthesis principles (manifested through prompt engineering) mitigate collapse, and second, we reveal the limitations of validation loss as a collapse detection metric. This analysis provides key insights for future synthetic data optimization.

2 RELATED WORK

Data Curation While web-crawled data contains hundreds of trillions of tokens, stringent quality filters typically remove the majority of this content. Popular datasets like C4, Gopher, Dolma, and RefinedWeb (Raffel et al., 2020; Rae et al., 2021; Penedo et al., 2023; Soldaini et al., 2024) use non-learned heuristics. More recently, aggressive model-based and retrieval-based filtering has become prominent in datasets like FineWeb-Edu (Penedo et al., 2024), DCLM (Li et al., 2024), and FineFineWeb (Zhang et al., 2024). Such heavy filtering results in a removal of over 90% of tokens, which has led some researchers to focus on balancing accuracy and data quantity (Su et al., 2024). However, this does not alter the fact that the total amount of high-quality data remains limited.

Repetition Training Studies on subset repetition training have revealed that model divergence tends to occur earlier as model parameters increase (Hernandez et al., 2022). For scenarios where entire datasets are repeated for training, limiting to 4 epochs or fewer results in minimal efficiency

108 degradation (Muennighoff et al., 2023; Taylor et al., 2022). Researchers have explored regularization
 109 techniques to mitigate repetition degradation, but this highlights the critical need for careful
 110 hyperparameter tuning. For example, while some work shows that increasing weight decay can
 111 yield better metrics (Fang et al., 2025), the same technique can also destabilize training. In a set of
 112 ablation studies, Xue et al. (2024) found that models with weight decay failed to converge, whereas
 113 using dropout proved to be an effective alternative. This sensitivity underscores the challenge of
 114 applying regularization in repetition scenarios. Overall, this topic remains understudied across dif-
 115 ferent hyper-parameters, data distributions, and repetition ratios.

116 **Synthetic Pretraining Data** Data synthesis for pretraining has rapidly evolved, with two primary
 117 approaches: seed-based synthesis and raw-text rephrasing. Seed-based methods, exemplified by
 118 Phi models (Abdin et al., 2024), Cosmopedia (Ben Allal et al., 2024), use predefined seed systems
 119 and templates to precisely control the generated content. In contrast, rephrasing methods, such as
 120 WRAP (Maini et al., 2024) and Nemotron-CC (Su et al., 2024), rewrite existing web content into
 121 different formats. Recent state-of-the-art models have validated the effectiveness of rewriting at
 122 an unprecedented scale. The Kimi K2 model applied rewriting to knowledge data as part of its
 123 training corpus (Kimi et al., 2025), and Nemotron Nano 2 (NVIDIA et al., 2025) employs Qwen3-
 124 30B-A3B (Yang et al., 2025) for rewriting. Concurrently, recent analysis has begun to codify the
 125 high-level insights for success and highlight the importance of the ‘methodology’ behind data syn-
 126 thesis (DatologyAI et al., 2025). These efforts confirm that synthetic data is a key component in
 127 training frontier models.

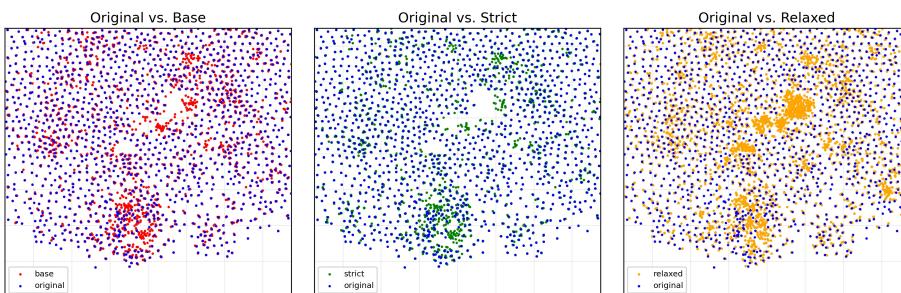
128 While these works validate the general approach, they often do not provide the detailed information
 129 and ablation of a successful synthesis implementation. Our work bridges this gap, which provides
 130 a concrete instantiation of the principles. By **adaptively** generating diverse ‘Genre’ and ‘Audience’
 131 pairs from high-quality source text, our framework systematically **enhances diversity in a scalable**
 132 **manner**, without requiring complex external seed systems or large-scale generator models.

133 3 MASSIVE GENRE-AUDIENCE REFORMULATION

135 The central challenge of data reformulation is balancing two competing goals: generating novel,
 136 diverse content (**variance**) while preserving the source document’s core factual information
 137 (**invariance**). To resolve this tension, we introduce the Massive Genre-Audience (MGA) frame-
 138 work, a principled pipeline designed for systematic corpus expansion. The framework operates
 139 on our central principle of “**Limited Consistency**”, which seeks to balance stylistic diversity with
 140 factual fidelity. This approach is implemented efficiently using lightweight small language models
 141 (Tool SLMs) fine-tuned for specific sub-tasks, ensuring both quality and scalability.

142 3.1 LIMITED CONSISTENCY

143 We define “Limited Consistency” as a guiding principle that seeks to maximize the stylistic and
 144 structural variance of reformulated content while maintaining strict invariance of the source docu-
 145 ment’s core factual information. This principle directly addresses the risk of generating data that is
 146 either too repetitive or factually incorrect. The primary mechanism for implementing this principle
 147 is through careful Prompt Engineering, which steers the generative process.



158 Figure 2: t-SNE visualization illustrating our “Limited Consistency” principle. Our Base PE strategy
 159 (left) achieves a balanced expansion of the original data distribution, while the Strict (middle) and
 160 Relaxed (right) variants are overly conservative or deviate excessively, respectively.

161 To identify an optimal balance, we explored the design space of PE strategies. Prompts that are too
 strict (‘SLM-Strict’) enforce high fidelity, leading to a distribution that closely mirrors the original

162 corpus but lacks diversity. Conversely, prompts that are too relaxed (‘SLM-Relaxed’) encourage
 163 excessive deviation, resulting in a significant distributional shift and a high risk of factual degradation.
 164 Our final approach (‘SLM-Base’) is calibrated to strike a balance, expanding the original data
 165 distribution without losing topical coherence. The distinct distributional impacts of these strategies
 166 are visualized in Figure 2. A detailed quantitative analysis validating the superiority of our balanced
 167 approach is presented in the ablation studies in Section 4.3.2.

168 3.2 FRAMEWORK IMPLEMENTATION

171 The MGA framework is operationalized as a two-stage synthesis pipeline: a variance-maximizing
 172 stage for generating diverse directives, followed by an invariance-enforcing stage for controlled
 173 reformulation. Each stage is powered by a specialized Tool SLM, which is finetuned on task-specific
 174 data generated by a larger language model¹. This implementation choice is validated in Table 1,
 175 where our final Tool SLM achieves performance nearly identical to the original LLM labeler.

176 Table 1: Reformulation quality comparison between the Tool SLM and its LLM teacher. All outputs
 177 were scored on a 1-5 scale by the LLM itself to evaluate the SLM’s alignment.

178 Models	179 Total Examples	5	4	3	2	1	Rate(≥ 3)	Diff
180 Labeler LLM	15,355	4,120	7,143	3,034	661	214	93.11%	-
181 Tool SLM	15,355	3,788	7,124	3,224	736	285	92.06%	-1.05%

182 **Stage 1: Adaptive GA-Pair Generation.** The primary objective of this stage is to **maximize**
 183 **diversity** by generating a wide array of creative reformulation directives. Our choice of Genre-
 184 Audience (GA) pairs as the core mechanism is deliberate. While simple rephrasing can generate
 185 stylistic variants, it often lacks structured diversity. GA pairs provide a robust framework for mean-
 186 ingful content adaptation:

- 188 • **Genre** dictates the structural and stylistic format of the content (e.g., an analytical report, a
 189 step-by-step tutorial, a blog post), controlling how information is organized and presented.
- 190 • **Audience** defines the intended reader’s profile (e.g., a university student, an industry expert, a
 191 curious teenager), guiding the tone, vocabulary, and conceptual depth.

192 Crucially, MGA moves beyond using a small, fixed set of styles. Instead, it **adaptively generates**
 193 multiple, contextually relevant GA pairs for each source document. To achieve this, we prompted
 194 the labeler LLM to produce five distinct GA pairs and curated this data through a rigorous rule-based
 195 validation process (e.g., validating JSON structure and pair count). This filtered dataset explicitly
 196 trains the ‘GA-SLM’ to execute a “one-pass-for-many” strategy, mitigating the risk of mode col-
 197 lapsed, where repeated sampling requests to a model can yield highly similar outputs.

199 **Stage 2: Controlled Reformulation.** This stage aims to **balance variance and invariance**, di-
 200 rectly implementing our “Limited Consistency” principle. The core design is a finetuning strategy
 201 that, instead of narrowly optimizing for perfect outputs (e.g., a score of 5), relaxes the quality thresh-
 202 old to ensure a high proportion of broadly acceptable generations (a score of 3 or higher).

203 To formalize this, let D be a source document and G be a generated GA-pair. The teacher LLM first
 204 produces an initial set of synthetic reformulations $\mathcal{D}_{\text{synth}} = \{(D_i, G_i, D'_i)\}$. However, training the
 205 Tool SLM directly on this full dataset would cause it to replicate the teacher’s suboptimal outputs.
 206 To circumvent this, we leverage the teacher LLM as a quality judge, using its scoring function
 207 $S(D'_i) \in \{1, \dots, 5\}$ to filter for a high-quality subset, \mathcal{D}_{SFT} :

$$208 \mathcal{D}_{\text{SFT}} = \{(D, G, D') \in \mathcal{D}_{\text{synth}} \mid S(D') \geq 3\}$$

210 The ‘Reformulation-SLM’, parameterized by θ , is then trained exclusively on this curated subset
 211 using a standard supervised fine-tuning (SFT) objective:

$$212 \mathcal{L}_{\text{SFT}}(\theta) = \mathbb{E}_{(D, G, D') \sim \mathcal{D}_{\text{SFT}}} [-\log P_{\theta}(D' | D, G)]$$

213 This targeted alignment process imbues the ‘Reformulation-SLM’ with the nuanced ability to gen-
 214 erate novel content that remains faithful to the source material.

215 ¹Tool model training details are presented in Appendix B

216 After the two stage synthesis, a final heuristic cleaning process is applied to the generated corpus.
 217 This stage filters out high-frequency generative patterns (e.g., ‘*Please note that ...*’) and removes
 218 documents with extremely low keyword coverage with the source document, ensuring final data
 219 quality. Finally, we achieve a 3.9x token expansion while maintaining high quality and diversity.
 220

221 4 EXPERIMENTS

222 We now empirically validate the MGA framework. We begin by establishing its core effectiveness
 223 in data-constrained scaling scenarios (Section 4.2). Following this validation, we provide a deeper
 224 analysis by addressing three key research questions (Section 4.3):

- 225 • **RQ1:** How does MGA reformulation complement existing synthetic data strategies?
- 226 • **RQ2:** What role does reformulation diversity play in high-repetition training?
- 227 • **RQ3:** Why does reformulation fundamentally benefit the model’s learning process?

228 4.1 SETUP

229 **Models and Hyperparams** The architecture of pretraining model follows that of Llama 3 (Dubey
 230 et al., 2024). Experiments across various sizes (134M/377M/1.7B) were running with Warmup-
 231 Stable-Decay lr scheduler (Hu et al., 2024) where 0.1% warmup steps, 75% stable and final 25%
 232 decay phase. Detailed model specifications are provided in Appendix C.2.

233 **Datasets** We build MGACorpus based on SmolLM-Corpus (Ben Allal et al., 2024), which con-
 234 tains four subsources (fineweb-edu-dedup / cosmopedia / python-edu / open-web-math). We refor-
 235 mulated the 195B tokens fineweb-edu-dedup source and finally got 770B cleaned synthetic tokens.

236 **Evaluation** We evaluate the models on a comprehensive suite of benchmarks include ARC-Easy
 237 / Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019), Winogrande (Sakaguchi et al.,
 238 2021), MMLU (Hendrycks et al., 2020), GSM8K (Cobbe et al., 2021), etc., following popular prac-
 239 tice of LIGHTEVAL (Fourrier et al., 2023) and LM-HARNESS (Gao et al., 2023). For the directly
 240 effectiveness validation of MGA, we evaluate MGACorpus aligned with Fineweb/SmolLM/Cosmo-
 241 pedia settings². For training dynamics, we report the average of 12 benchmarks and validation losses
 242 on held-out fineweb-edu-dedup dataset.

243 4.2 MAIN EXPERIMENTS

244 To directly evaluate MGA’s potential as a solution for data scarcity and repetition, we present a
 245 comprehensive analysis in two parts. First, we benchmark MGA’s performance against recent SOTA
 246 small LMs to establish a comparative baseline. Subsequently, we investigate its behavior under data-
 247 constrained scaling scenarios, specifically situations where the training budget exceeds the available
 248 unique high-quality data, a common limitation in practical applications.³

249 Table 2: Benchmark **MGA** with SOTA SmolLM series. Models of similar size are grouped. All
 250 results are obtained via LIGHTEVAL (Fourrier et al., 2023). The best result within each fair compar-
 251 ison is highlighted in green. Note that SmolLM2 models, trained with substantially more compute,
 252 are included for reference only.

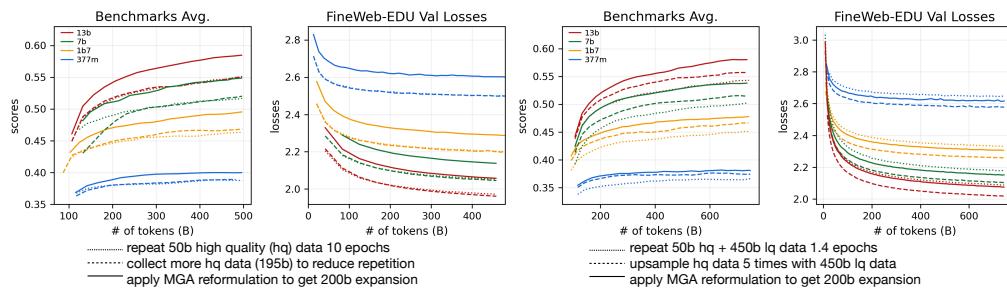
Model	#Params.	#Tokens	ARC(C+E)	Wino	Hella	MMLU	MMLU-PRO	CSQA	OpenBookQA	PIQA	TriviaQA	GSM8K	Avg.
SmolLM2-135M	135M	2T	44.12	51.07	42.03	31.27	11.06	33.82	35	68.23	1.91	1.52	32.00
SmolLM-135M	135M	600B	42.47	51.54	41.08	29.93	11.4	32.51	33.2	68.17	1.08	0.99	31.24
SmolLM-135M (ours)	134M	600B	41.71	52.41	40.69	30.03	11.37	34.32	35.4	67.85	0.02	1.29	31.51
MGA-Expansion	134M	600B	43.01	51.7	41.25	30.1	11.76	32.68	36.4	67.3	2.05	1.44	31.77
SmolLM2-360M	360M	4T	53.4	52.33	54.58	35.29	11.17	37.92	37.6	71.76	16.73	2.96	37.37
SmolLM-360M	360M	600B	49.99	52.96	51.67	33.84	11.42	34.81	37.6	71.87	2.27	1.97	34.84
SmolLM-360M (ours)	377M	600B	48.57	52.64	51.02	33.63	11.25	36.77	39	71	0.29	1.52	34.57
MGA-Expansion	377M	600B	49.39	52.64	51.34	34.09	11.35	37.1	38	72.31	7.28	1.74	35.52
SmolLM2-1.7B	1.7B	11T	60.42	59.59	68.73	41.4	19.61	43.65	42.6	77.53	36.68	29.04	47.93
SmolLM-1.7B	1.7B	1T	59.95	54.7	62.83	39.35	10.92	38	42.6	75.9	13.14	4.62	40.20
SmolLM-1.7B (ours)	1.7B	1T	59.63	57.38	65.19	39.4	12.11	42.59	45.6	76.88	4.95	7.81	41.15
MGA-Expansion	1.7B	1T	60.36	57.46	65.52	40.79	14.1	41.11	42.8	77.53	20.42	13.87	43.4

268 ²<https://github.com/huggingface/cosmopedia/blob/main/evaluation>

269 ³Details on data recipes and comparisons with other models are provided in Appendix C.1 and D.1.

270 **Performance training on MGACorpus** We evaluate whether incorporating MGA data enhances
 271 model performance compared to a baseline trained solely on the original sources, using fixed training
 272 budgets and model sizes ranging from 134M to 1.7B. As shown in Table 2, MGA-Expansion shows
 273 consistent improvements across different model sizes, with larger performance gains as model size
 274 increases, $+0.26/+0.95/+2.15$ for 134M/377M/1.7B models respectively. Notably, MGA-Expansion
 275 achieved substantial gains in reasoning-intensive tasks such as TriviaQA ($+2.03/+6.99/+15.47$) and
 276 GSM8K ($+0.15/+0.22/+6.06$), and shows strong performance on MMLU/MMLU-Pro. We hypothe-
 277 size that MGA’s data reformulation, by exposing the model to diverse phrasings of the same underly-
 278 ing information, fosters more robust generalization. This enhanced generalization, in turn, improves
 279 the model’s reasoning capabilities, leading to the results observed on these specific benchmarks.

280 **Scaling Dynamics** We further investigate MGA’s behavior under data-constrained scaling scenar-
 281 os. Models of 377M/1.7B/7B/13B are trained using a learning rate scheduler with only warmup
 282 and stable phases, which allows for a direct performance comparison across repetition epochs.



294 Figure 3: Training dynamics of two common scenarios under data-constrained conditions: (1) ex-
 295 panding a 50B high-quality dataset to a 500B training budget (entire set repetition), (2) expanding
 296 a 500B mixed-quality dataset to a 700B training budget (subset repetition). Data recipe and bench-
 297 mark details are provided in Appendix C.1 and C.3.

298 **Scaling Results** As shown in Figure 3, MGA demonstrates favorable scaling properties with both
 299 data budget (D-scaling) and model sizes (N-scaling) under two common data-constrained scenarios.

300 • In the entire set experiments, simply increasing unique token count by collecting more high qual-
 301 ity data (195B via Full-Fineweb-Edu) shows marginal improvements ($+0.2/+0.15/-0.16/+0.11$) at
 302 200/300/400/500 billion token steps (13B size). In contrast, MGA, through a 200B reformulation
 303 as expansion of the original 50B data, demonstrates consistent gains ($+2.65/+3.14/+3.43/+3.46$),
 304 highlighting **effective D-scaling**.

305 • Similarly, in the subset experiments, both upsampling the high-quality sub data portion (5x)
 306 and MGA (via a 200B expansion) improve upon the baseline. However, their N-scaling prop-
 307 erties with model parameters differ significantly: the performance advantage of upsampling
 308 remains relatively constant across model sizes ($+0.89/+1.53/+1.23/+1.41$), whereas MGA ex-
 309 pansion exhibits **superior N-scaling**, its performance gains amplifying with increasing model
 310 scale ($+1.46/+2.67/+3.59/+3.73$).

312 These scaling experiments confirm that MGA is a powerful data augmentation strategy that aids both
 313 model (N) and data (D) scaling in constrained scenarios. Notably, MGA’s performance advantage
 314 emerges from the very first epoch, well before significant data repetition occurs, and this gap widens
 315 as training progresses. The dual observation leads directly to our core research questions, which
 316 we will investigate in Section 4.3: How does MGA’s inherent diversity mitigate the degradation
 317 from high-repetition training (**RQ2**), and what fundamental learning benefits does the reformulation
 318 provide from the outset (**RQ3**)?

319 **Validation Losses** Although MGA demonstrates superior benchmark performance, we observe
 320 increasing validation losses compared to baseline models. While higher validation losses might
 321 seem concerning at first glance, it’s important to note that validation loss may not fully reflect model
 322 performance, as token-level perplexity is inherently biased by the frequency distribution of the vali-
 323 dation set, and in-domain validation metrics may not necessarily correlate with out-of-domain gen-
 324 eralization capabilities. This observation, combined with recent studies linking loss degradation to

324 model collapse (Dohmatob et al., 2024b;a; Zhu et al., 2024), calls for a more nuanced analysis,
 325 which we will also provide in Section 4.3.3.
 326

327 4.3 DISCUSSIONS

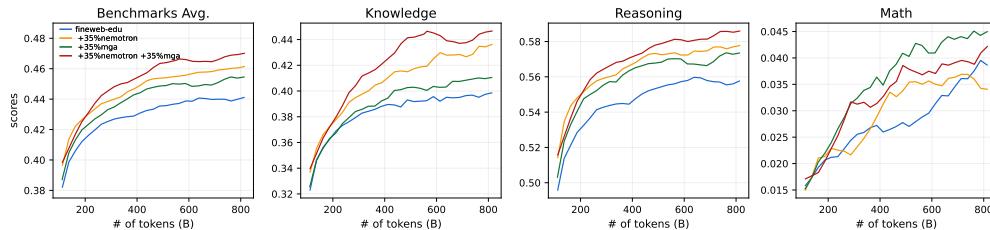
329 The strong empirical results in our main experiments validate MGA as a powerful data augmentation
 330 strategy. However, these outcomes naturally lead to deeper questions about its positioning,
 331 mechanics, and underlying benefits. In the following section, we move from empirical validation to
 332 analytical discussion, addressing the key research questions outlined at the beginning of this section.
 333

334 4.3.1 HOW DOES MGA COMPLEMENT EXISTING SYNTHETIC DATA STRATEGIES?

335 Our main experiments demonstrated MGA’s value relative to data repetition, but how does it stand
 336 within the diverse and rapidly evolving ecosystem of synthetic data? To address RQ1, this section
 337 positions MGA not as a standalone replacement, but as a complementary approach to other prevalent
 338 strategies, such as Nemotron-CC-Synthetic (Su et al., 2024). We compare MGA-enhanced data with
 339 this popular open-source synthetic datasets to highlight its unique contribution.
 340

341 To conduct a fair comparison, we designed a controlled experiment with four distinct data mixtures,
 342 each training a 1.7B parameter model for 800B tokens. The data blend setups were as follows:
 343

- 344 • Baseline: A high-quality real dataset fineweb-edu.
- 345 • Exp A: 35% token budget replaced by Nemotron-CC-HQ synthetic corpus (+Nemotron-Syn).
- 346 • Exp B: 35% token budget replaced by MGACorpus (+MGA).
- 347 • Exp C: 70% token budget replaced by an equal combination of Nemotron-Syn and MGACorpus
 348 data (+Nemotron-Syn +MGA).



356 Figure 4: Benchmark results demonstrating the complementary nature of different synthetic data
 357 strategies. While both MGA and Nemotron-Syn individually improve over the baseline, their
 358 combination (Exp C) yields a significant synergistic boost in performance. For a detailed breakdown
 359 of performance by each synthetic task, as well as supplementary cross-mixing experiments that further
 360 validate MGA’s role in creating a generalizable base model, please refer to Appendix D.3.
 361

362 As illustrated in Figure 4, the results reveal a clear performance hierarchy: Exp C > Exp A > Exp
 363 B > Baseline. The strong performance of Exp A is understandable, as Nemotron-Syn is a high-
 364 quality and diverse synthetic corpus composed of five diverse subsets. The inclusion of various
 365 data formats, some of which (like QA pairs) align well with common evaluation structures. While
 366 our MGA-enhanced mix (Exp B) also surpasses the baseline, the most compelling finding comes
 367 from Exp C, which significantly outperforms all other configurations. This demonstrates a clear
 368 synergistic effect, where the structural and stylistic diversity from MGA’s reformulation enriches
 369 the high-quality, task-aligned data from Nemotron-Syn.
 370

371 Therefore, we answer RQ1 by concluding that MGA is not in competition with but is complemen-
 372 tary to other synthetic data methodologies. The path to resolving data scarcity does not lie in a
 373 single synthesis technique, but in the thoughtful combination of diverse strategies. MGA provides
 374 a foundational, general-purpose enhancement through reformulation that benefits even further when
 375 combined with specialized, task-aligned synthetic data.
 376

377 4.3.2 DOES REFORMULATION DIVERSITY HELP TO MITIGATE REPETITION ISSUE?

378 To address **RQ2**, this section examines how different design choices in prompt engineering influ-
 379 ence the effectiveness of the MGA framework, particularly under high-repetition conditions. By
 380

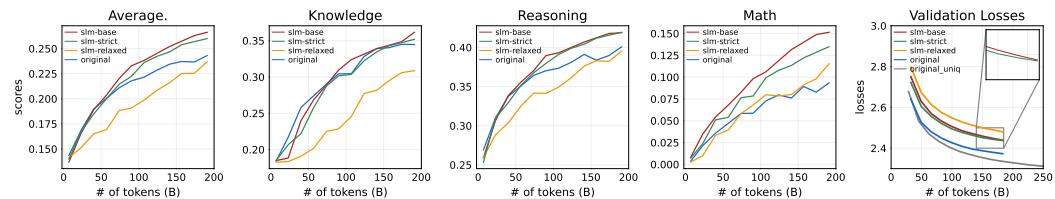
378 comparing SLM variants (introduced in Section 3.1) using different consistency requirements, we
 379 identify optimal strategies for balancing information preservation with content diversity.
 380

381 Table 3: Performance comparison of different SLM variants on reformulation quality metrics.

382 Models	383 Total Examples	5	4	3	2	1	Rate(≤ 2)	Rate(≥ 4)	Rate($= 5$)
383 SLM-Base	383 15,355	3,788	7,124	3,224	736	285	6.65%	71.06%	24.67%
384 SLM-Strict	384 15,355	6,814	5,220	2,384	520	227	4.86%	78.37%	44.38%
385 SLM-Relaxed	385 15,355	408	1,685	3,889	4,156	5,086	60.19%	13.63%	2.66%

386 We sample an additional 20B tokens from real data and generate three synthetic datasets: 80B
 387 tokens using SLM-Base, 80B tokens using SLM-Strict, and 40B tokens using SLM-Relaxed. As
 388 mentioned before, SLM-Base expands the original corpus to $3.9\times$ more tokens, while SLM-Relaxed
 389 makes only $2\times$ tokens as we only require basic topical relevance. Similar to experimental setup
 390 in early sections, we set a high-repetition baseline on a smaller data scale (replicating the original
 391 20B tokens 10 times) to more clearly demonstrate the potential impact of SLM-Strict compared to
 392 SLM-Base in high-repetition scenarios.

393 As shown in Figure 5, our experiments reveal distinct patterns across training configurations. Both
 394 SLM-Base and SLM-Strict show performance improvements, while the SLM-Relaxed configuration
 395 leads to significant collapse. More supplementary experiments could be found in Appendix D.2.



403 Figure 5: Benchmark results and validation losses. The sensitivity to data repetition varies across
 404 capability domains, with knowledge dimension showing greater resilience.

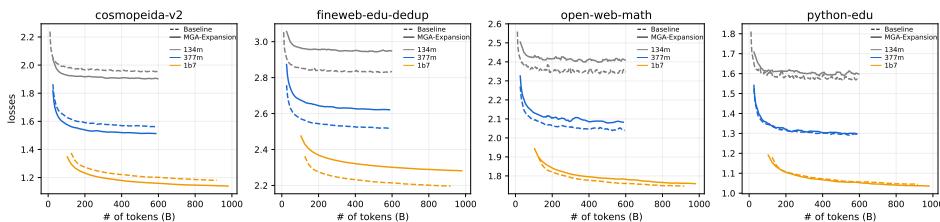
405 Despite the apparent effectiveness of strict information preservation, can it fundamentally address
 406 the challenges posed by data repetition? Our examination of validation loss trajectories reveals a
 407 critical distinction: SLM-Base maintains healthy optimization characteristics throughout training,
 408 whereas SLM-Strict exhibits degraded scaling behavior at higher iteration steps, reminiscent of the
 409 limitations observed with data repetition.

410 Therefore, this investigation into different prompt engineering strategies concludes that a balanced
 411 ‘Limited Consistency’ approach (SLM-Base) yields the best reformulation quality and subsequent
 412 model performance answering to **RQ2**.

414 4.3.3 WHY DOES REFORMULATION BENEFIT THE MODEL’S LEARNING PROCESS?

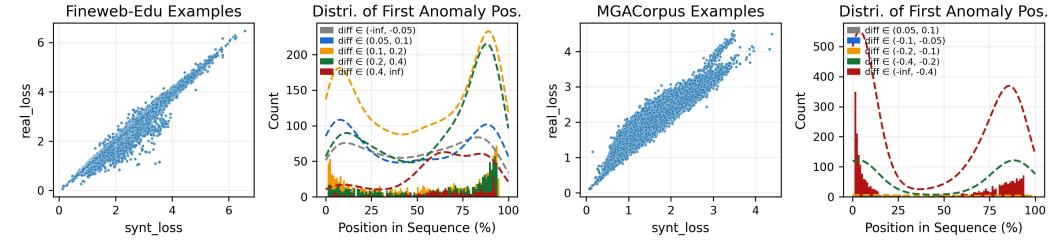
415 Having explored the impact of diversity in addressing data repetition (**RQ2**), we now turn to **RQ3**:
 416 this section investigates the underlying mechanisms by analyzing learning characteristics and vali-
 417 dating against potential issues like model collapse (Dohmatob et al., 2024b;a; Zhu et al., 2024).

419 **Multi-perspective Validation Analysis** Our analyses across different validation sets reveal vary-
 420 ing patterns in model behavior (Figure 6). As expected, MGA groups’ substitution of fineweb-edu
 421 data results in adverse effects on corresponding loss, with similar deterioration observed in open-
 422 web-math. Interestingly, the synthetic dataset cosmopedia demonstrates improved loss metrics. A
 423 notable contrast emerges in python-edu: while MGA exhibit negative impact at the 134M and 377M
 424 parameter, this trend reverses at 1.7B, suggesting scale-dependent effects on model behavior.



425 426 427 428 429 430 431 Figure 6: validation losses of experiments in Section 4.2.

432 **Fine-grained Pattern Analysis** To better understand whether increased validation loss truly
 433 indicates model collapse, we conduct a fine-grained analysis of loss patterns. Specifically, we compare
 434 token-level losses of 800B checkpoint between models trained on real data and synthetic data
 435 (Baseline and MGA-Expansion in Section 4.2, respectively). The document samples are from both
 436 Fineweb-Edu and MGACorpus. As illustrated in subfigures 1 and 3 of Figure 7, each point represents
 437 a sample’s average token loss, consistent with the overall loss discrepancy shown in Figure 6.
 438



448 Figure 7: Losses pattern analysis. Subfigures 1 and 3 shows comparison between models trained
 449 on different data settings, with $loss_{real}$ on y-axis and $loss_{synt}$ on x-axis. Subfigures 2 and 4 track
 450 the position where $loss_{synt}^i - loss_{real}^i$ ($loss_{diff}^i$) first becomes significantly higher than the sequence’s
 451 average difference (detailed definition in Appendix D.4).

452 The distribution of first anomaly positions (subfigures 2 and 4) reveals a crucial insight: when
 453 processing real data, models trained on synthetic data show performance degradation (measured
 454 by $loss_{diff}$) that predominantly manifests in later sequence positions, which intensifies as $loss_{diff}$
 455 increases. However, this positional bias disappears when evaluating on synthetic data.
 456

457 The systematic pattern suggests that rather than experiencing model collapse, the synthetic-trained
 458 model may have developed a different learning strategy (examples shown in Appendix D.4). While
 459 it shows higher validation losses on certain real-world datasets, its strong performance in our main
 460 experiments indicates a potential trade-off: the model may prioritize learning generalizable patterns
 461 from context over memorizing specific sequence dependencies. This shift in learning process could
 462 explain both the improved performance on benchmark tasks and the increased losses on validation
 463 sets that potentially require more memorization-based processing.

464 These findings indicate that the performance characteristics associated with MGA data likely stem
 465 from altered learning strategies, potentially prioritizing generalizability, rather than representing
 466 model collapse, which addressing **RQ3**,

467 5 CONCLUSION

468 In this work, we introduced MGA, a principled framework that leverages genre-audience reformulation
 469 to systematically expand and augment existing corpora with diverse, synthetically generated
 470 variations. Our core finding highlights MGA’s effectiveness as a data augmentation strategy specific-
 471 ally targeting the repetition challenge: in data-constrained scaling experiments, MGA significantly
 472 outperformed naive data repetition and simple upsampling, enabling more effective model training
 473 beyond unique data limits. Furthermore, the quality of the MGA was confirmed by consistent per-
 474 formance improvements when incorporated into standard training mixtures across various model
 475 sizes. Crucially, our experiments also revealed MGA’s role as a complementary strategy, demon-
 476 strating a powerful synergistic effect when combined with other prominent synthetic datasets. In
 477 essence, MGA’s effectiveness demonstrates that the key to overcoming data limits is generating rel-
 478 evant diversity, not just raw volume. Therefore, our work offers more than a tool; it provides a new
 479 roadmap for the community, where the thoughtful combination of different approaches becomes the
 480 cornerstone of sustainable progress in the continue scaling of LLM development.

481 482 REPRODUCIBILITY STATEMENT

483 We are committed to reproducibility and will release the 770B-token MGACorpus and all associated
 484 resources. Our MGA framework implementation, including tool model details, is documented in
 485 Appendix B. The complete pretraining and evaluation setup, with data recipes and hyperparameters,
 is detailed in Appendix C. Cases and prompts used to curate the dataset is provided in Appendix E.

486 REFERENCES
487

488 Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar,
489 Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 techni-
490 cal report. *arXiv preprint arXiv:2412.08905*, 2024.

491 Loubna Ben Allal, Anton Lozhkov, Guilherme Penedo, Thomas Wolf, and Leandro von
492 Werra. Smollm-corpus, 2024. URL [https://huggingface.co/datasets/
493 HuggingFaceTB/smollm-corpus](https://huggingface.co/datasets/HuggingFaceTB/smollm-corpus).

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

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

500 DatologyAI, :, Pratyush Maini, Vineeth Dorna, Parth Doshi, Aldo Carranza, Fan Pan, Jack Ur-
501 banek, Paul Burstein, Alex Fang, Alvin Deng, Amro Abbas, Brett Larsen, Cody Blakeney, Charvi
502 Bannur, Christina Baek, Darren Teh, David Schwab, Haakon Mongstad, Haoli Yin, Josh Wills,
503 Kaleigh Mentzer, Luke Merrick, Ricardo Monti, Rishabh Adiga, Siddharth Joshi, Spandan Das,
504 Zhengping Wang, Bogdan Gaza, Ari Morcos, and Matthew Leavitt. Beyondweb: Lessons from
505 scaling synthetic data for trillion-scale pretraining, 2025. URL [https://arxiv.org/abs/
506 2508.10975](https://arxiv.org/abs/2508.10975).

507 Elvis Dohmatob, Yunzhen Feng, Arjun Subramonian, and Julia Kempe. Strong model collapse.
508 *arXiv preprint arXiv:2410.04840*, 2024a.

509 Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. A tale of tails: Model
510 collapse as a change of scaling laws. *arXiv preprint arXiv:2402.07043*, 2024b.

511 Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha
512 Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models.
513 *arXiv preprint arXiv:2407.21783*, 2024.

514 Alex Fang, Hadi Pouransari, Matt Jordan, Alexander Toshev, Vaishaal Shankar, Ludwig Schmidt,
515 and Tom Gunter. Datasets, documents, and repetitions: The practicalities of unequal data quality,
516 2025. URL <https://arxiv.org/abs/2503.07879>.

517 Clémentine Fourrier, Nathan Habib, Thomas Wolf, and Lewis Tunstall. Lighteval: A lightweight
518 framework for llm evaluation, 2023. URL [https://github.com/huggingface/
519 lighteval](https://github.com/huggingface/lighteval).

520 Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Fos-
521 ter, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muen-
522 nighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lin-
523 tang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework
524 for few-shot language model evaluation, 12 2023. URL [https://zenodo.org/records/
525 10256836](https://zenodo.org/records/10256836).

526 Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and
527 Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint
528 arXiv:2009.03300*, 2020.

529 Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nel-
530 son Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and inter-
531 pretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

532 Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza
533 Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. An
534 empirical analysis of compute-optimal large language model training. *Advances in Neural Infor-
535 mation Processing Systems*, 35:30016–30030, 2022.

540 Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang,
 541 Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models
 542 with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

543

544 Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child,
 545 Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language
 546 models. *arXiv preprint arXiv:2001.08361*, 2020.

547

548 Kimi, :, Yifan Bai, Yiping Bao, Guanduo Chen, Jiahao Chen, Ningxin Chen, Ruijue Chen, Yanru
 549 Chen, Yuankun Chen, Yutian Chen, Zhuofu Chen, Jialei Cui, Hao Ding, Mengnan Dong, An-
 550 gang Du, Chenzhuang Du, Dikang Du, Yulun Du, Yu Fan, Yichen Feng, Kelin Fu, Bofei Gao,
 551 Hongcheng Gao, Peizhong Gao, Tong Gao, Xinran Gu, Longyu Guan, Haiqing Guo, Jianhang
 552 Guo, Hao Hu, Xiaoru Hao, Tianhong He, Weiran He, Wenyang He, Chao Hong, Yangyang Hu,
 553 Zhenxing Hu, Weixiao Huang, Zhiqi Huang, Zihao Huang, Tao Jiang, Zhejun Jiang, Xinyi Jin,
 554 Yongsheng Kang, Guokun Lai, Cheng Li, Fang Li, Haoyang Li, Ming Li, Wentao Li, Yanhao
 555 Li, Yiwei Li, Zhaowei Li, Zheming Li, Hongzhan Lin, Xiaohan Lin, Zongyu Lin, Chengyin Liu,
 556 Chenyu Liu, Hongzhang Liu, Jingyuan Liu, Junqi Liu, Liang Liu, Shaowei Liu, T. Y. Liu, Tian-
 557 wei Liu, Weizhou Liu, Yangyang Liu, Yibo Liu, Yiping Liu, Yue Liu, Zhengying Liu, Enzhe
 558 Lu, Lijun Lu, Shengling Ma, Xinyu Ma, Yingwei Ma, Shaoguang Mao, Jie Mei, Xin Men, Yibo
 559 Miao, Siyuan Pan, Yebo Peng, Ruoyu Qin, Bowen Qu, Zeyu Shang, Lidong Shi, Shengyuan Shi,
 560 Feifan Song, Jianlin Su, Zhengyuan Su, Xinjie Sun, Flood Sung, Heyi Tang, Jiawen Tao, Qifeng
 561 Teng, Chensi Wang, Dinglu Wang, Feng Wang, Haiming Wang, Jianzhou Wang, Jiaxing Wang,
 562 Jinhong Wang, Shengjie Wang, Shuyi Wang, Yao Wang, Yejie Wang, Yiqin Wang, Yuxin Wang,
 563 Yuzhi Wang, Zhaoji Wang, Zhengtao Wang, Zhexu Wang, Chu Wei, Qianqian Wei, Wenhao Wu,
 564 Xingzhe Wu, Yuxin Wu, Chenjun Xiao, Xiaotong Xie, Weimin Xiong, Boyu Xu, Jing Xu, Jin-
 565 jing Xu, L. H. Xu, Lin Xu, Suting Xu, Weixin Xu, Xinran Xu, Yangchuan Xu, Ziyao Xu, Junjie
 566 Yan, Yuzi Yan, Xiaofei Yang, Ying Yang, Zhen Yang, Zhilin Yang, Zonghan Yang, Haotian Yao,
 567 Xingcheng Yao, Wenjie Ye, Zhuorui Ye, Bohong Yin, Longhui Yu, Enming Yuan, Hongbang
 568 Yuan, Mengjie Yuan, Haobing Zhan, Dehao Zhang, Hao Zhang, Wanlu Zhang, Xiaobin Zhang,
 569 Yangkun Zhang, Yizhi Zhang, Yongting Zhang, Yu Zhang, Yutao Zhang, Yutong Zhang, Zheng
 570 Zhang, Haotian Zhao, Yikai Zhao, Huabin Zheng, Shaojie Zheng, Jianren Zhou, Xinyu Zhou,
 Zaida Zhou, Zhen Zhu, Weiyu Zhuang, and Xinxing Zu. Kimi k2: Open agentic intelligence,
 2025. URL <https://arxiv.org/abs/2507.20534>.

571 Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Gadre, Hritik Bansal, Etash
 572 Guha, Sedrick Keh, Kushal Arora, Saurabh Garg, Rui Xin, Niklas Muennighoff, Reinhard Heckel,
 573 Jean Mercat, Mayee Chen, Suchin Gururangan, Mitchell Wortsman, Alon Albalak, Yonatan Bit-
 574 ton, Marianna Nezhurina, Amro Abbas, Cheng-Yu Hsieh, Dhruba Ghosh, Josh Gardner, Maciej
 575 Kilian, Hanlin Zhang, Rulin Shao, Sarah Pratt, Sunny Sanyal, Gabriel Ilharco, Giannis Daras,
 576 Kalyani Marathe, Aaron Gokaslan, Jieyu Zhang, Khyathi Chandu, Thao Nguyen, Igor Vasiljevic,
 577 Sham Kakade, Shuran Song, Sujay Sanghavi, Fartash Faghri, Sewoong Oh, Luke Zettlemoyer,
 578 Kyle Lo, Alaaeldin El-Nouby, Hadi Pouransari, Alexander Toshev, Stephanie Wang, Dirk Groen-
 579 eveld, Luca Soldaini, Pang Wei Koh, Jenia Jitsev, Thomas Kollar, Alexandros G. Dimakis, Yair
 580 Carmon, Achal Dave, Ludwig Schmidt, and Vaishaal Shankar. Datacomp-lm: In search of the
 581 next generation of training sets for language models, 2024.

582 Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. Rephras-
 583 ing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint*
 584 *arXiv:2401.16380*, 2024.

585

586 Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra
 587 Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language
 588 models. *Advances in Neural Information Processing Systems*, 36:50358–50376, 2023.

589

590 NVIDIA, :, Aarti Basant, Abhijit Khairnar, Abhijit Paithankar, Abhinav Khattar, Adithya Renduch-
 591 intala, Aditya Malte, Akhiad Bercovich, Akshay Hazare, Alejandra Rico, Aleksander Ficek,
 592 Alex Kondratenko, Alex Shaposhnikov, Alexander Bukharin, Ali Taghibakhshi, Amelia Bar-
 593 ton, Ameya Sunil Mahabaleshwarakar, Amy Shen, Andrew Tao, Ann Guan, Anna Shors, Anub-
 594 hav Mandarwal, Arham Mehta, Arun Venkatesan, Ashton Sharabiani, Ashwath Aithal, Ashwin
 595 Poojary, Ayush Dattagupta, Balaram Buddharaju, Banghua Zhu, Barnaby Simkin, Bilal Kartal,

594 Bita Darvish Rouhani, Bobby Chen, Boris Ginsburg, Brandon Norick, Brian Yu, Bryan Catan-
 595 zaro, Charles Wang, Charlie Truong, Chetan Mungekar, Chintan Patel, Chris Alexiuk, Christian
 596 Munley, Christopher Parisien, Dan Su, Daniel Afrimi, Daniel Korzekwa, Daniel Rohrer, Daria
 597 Gitman, David Mosallanezhad, Deepak Narayanan, Dima Rekesh, Dina Yared, Dmytro Pykhtar,
 598 Dong Ahn, Duncan Riach, Eileen Long, Elliott Ning, Eric Chung, Erick Galinkin, Evelina Bakh-
 599 turina, Gargi Prasad, Gerald Shen, Haifeng Qian, Haim Elisha, Harsh Sharma, Hayley Ross,
 600 Helen Ngo, Herman Sahota, Hexin Wang, Hoo Chang Shin, Hua Huang, Iain Cunningham, Igor
 601 Gitman, Ivan Moshkov, Jaehun Jung, Jan Kautz, Jane Polak Scowcroft, Jared Casper, Jian Zhang,
 602 Jiaqi Zeng, Jimmy Zhang, Jinze Xue, Jocelyn Huang, Joey Conway, John Kamalu, Jonathan
 603 Cohen, Joseph Jennings, Julien Veron Vialard, Junkeun Yi, Jupinder Parmar, Kari Briski, Katherine
 604 Cheung, Katherine Luna, Keith Wyss, Keshav Santhanam, Kezhi Kong, Krzysztof Pawelec,
 605 Kumar Anik, Kunlun Li, Kushan Ahmadian, Lawrence McAfee, Laya Sleiman, Leon Derczyn-
 606 ski, Luis Vega, Maer Rodrigues de Melo, Makesh Narsimhan Sreedhar, Marcin Chochowski,
 607 Mark Cai, Markus Kliegl, Marta Stepniewska-Dziubinska, Matvei Novikov, Mehrzad Samadi,
 608 Meredith Price, Meriem Boubdir, Michael Boone, Michael Evans, Michal Bien, Michal Za-
 609 walski, Miguel Martinez, Mike Chrzanowski, Mohammad Shoeibi, Mostafa Patwary, Namit
 610 Dhameja, Nave Assaf, Negar Habibi, Nidhi Bhatia, Nikki Pope, Nima Tajbakhsh, Nirmal Ku-
 611 mar Juluru, Oleg Rybakov, Oleksii Hrinchuk, Oleksii Kuchaiev, Oluwatobi Olabiyi, Pablo Rib-
 612 alta, Padmavathy Subramanian, Parth Chadha, Pavlo Molchanov, Peter Dykas, Peter Jin, Piotr
 613 Bialecki, Piotr Januszewski, Pradeep Thalasta, Prashant Gaikwad, Prasoon Varshney, Pritam Gun-
 614 decha, Przemek Tredak, Rabeeh Karimi Mahabadi, Rajen Patel, Ran El-Yaniv, Ranjit Rajan, Ria
 615 Cheruvu, Rima Shahbazyan, Ritika Borkar, Ritu Gala, Roger Waleffe, Ruoxi Zhang, Russell J.
 616 Hewett, Ryan Prenger, Sahil Jain, Samuel Kriman, Sanjeev Satheesh, Saori Kaji, Sarah Yurick,
 617 Saurav Muralidharan, Sean Narenthiran, Seonmyeong Bak, Sepehr Sameni, Seungju Han, Shan-
 618 mugam Ramasamy, Shaona Ghosh, Sharath Turuvekere Sreenivas, Shelby Thomas, Shizhe Diao,
 619 Shreya Gopal, Shrimai Prabhumoye, Shubham Toshniwal, Shuoyang Ding, Siddharth Singh, Sid-
 620 dhartha Jain, Somshubra Majumdar, Soumye Singhal, Stefania Alborghetti, Syeda Nahida Akter,
 621 Terry Kong, Tim Moon, Tomasz Hliwiak, Tomer Asida, Tony Wang, Tugrul Konuk, Twinkle
 622 Vashishth, Tyler Poon, Udi Karpas, Vahid Noroozi, Venkat Srinivasan, Vijay Korthikanti, Vikram
 623 Fugro, Vineeth Kalluru, Vitaly Kurin, Vitaly Lavrukhin, Wasi Uddin Ahmad, Wei Du, Wonmin
 624 Byeon, Ximing Lu, Xin Dong, Yashaswi Karnati, Yejin Choi, Yian Zhang, Ying Lin, Yong-
 625 gan Fu, Yoshi Suhara, Zhen Dong, Zhiyu Li, Zhongbo Zhu, and Zijia Chen. Nvidia nemotron
 626 nano 2: An accurate and efficient hybrid mamba-transformer reasoning model, 2025. URL
 627 <https://arxiv.org/abs/2508.14444>.

628 Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli,
 629 Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refined-
 630 web dataset for falcon llm: Outperforming curated corpora with web data only. *Advances in
 631 Neural Information Processing Systems*, 36:79155–79172, 2023.

632 Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin
 633 Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the
 634 finest text data at scale, 2024. URL <https://arxiv.org/abs/2406.17557>.

635 Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John
 636 Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models:
 637 Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

638 Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi
 639 Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text
 640 transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

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

643 Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Arthur,
 644 Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. Dolma: An open corpus of
 645 three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*,
 646 2024.

648 Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary,
 649 Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a
 650 refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*, 2024.

651

652 Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia,
 653 Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for
 654 science. *arXiv preprint arXiv:2211.09085*, 2022.

655

656 Pablo Villalobos, Jaime Sevilla, Lennart Heim, Tamay Besiroglu, Marius Hobbehahn, and Anson Ho.
 657 Will we run out of data? an analysis of the limits of scaling datasets in machine learning. *arXiv*
 658 preprint *arXiv:2211.04325*, 1, 2022.

659

660 Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to
 661 repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing*
 662 Systems, 36, 2024.

663

664 An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang
 665 Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu,
 666 Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin
 667 Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang,
 668 Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui
 669 Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang
 670 Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger
 671 Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan
 672 Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

673

674 Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a ma-
 675 chine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association*
 676 *for Computational Linguistics*, 2019.

677

678 Ge Zhang, Xinrun Du, Zhimiao Yu, Zili Wang, Zekun Wang, Shuyue Guo, Tianyu Zheng, Kang
 679 Zhu, Jerry Liu, Shawn Yue, Binbin Liu, Zhongyuan Peng, Yifan Yao, Jack Yang, Ziming Li,
 680 Bingni Zhang, Minghao Liu, Tianyu Liu, Yang Gao, Wenhui Chen, Xiaohuan Zhou, Qian Liu,
 681 Taifeng Wang, and Wenhao Huang. Finefineweb: A comprehensive study on fine-grained domain
 682 web corpus, December 2024.

683

684

685

686

687

688

689

690

691

692

693

694

695

696

697

698

699

700

701

702 A LIMITATIONS AND FUTURE WORK

704 While our experimental results demonstrate the effectiveness of MGA in both quality validation and
 705 scaling scenarios, several important aspects warrant further investigation. We identify two key areas
 706 for future research:

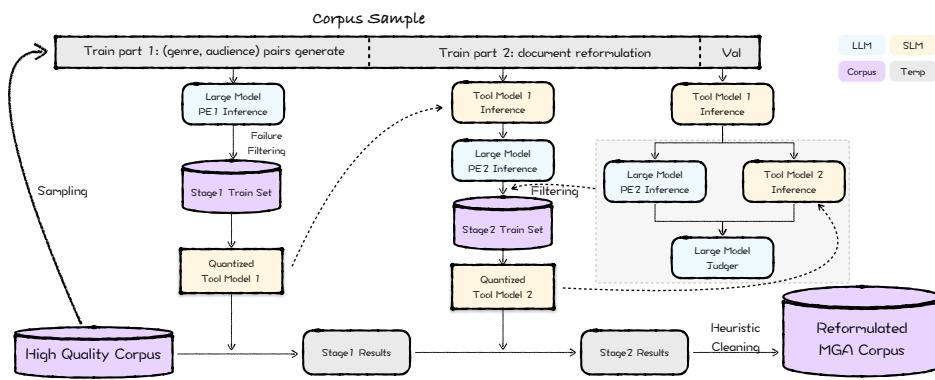
708 • Our current experiments demonstrate effectiveness up to 13B parameters and 1,000B tokens
 709 of training budget. Extending this approach to long-horizon training and larger-scale models
 710 requires additional validations, particularly for next-generation models which require hundreds
 711 of trillions of training tokens.

712 • Regarding data repetition strategies, we present preliminary explorations under computational
 713 resource constraints. The underlying patterns and their sensitivity to various factors, such as
 714 repetition ratio, data distribution, and even model hyperparameters, require systematic investi-
 715 gation. Future research should examine how these factors collectively determine optimal data
 716 strategies across different training scenarios.

718 **Broader Impact** This paper explores the use of LLMs as a data expansion method for pretraining
 719 large language models. We introduce the MGA framework to mitigate data repetition issues, which
 720 holds potential for positive societal impact, particularly in synthetic data generation for training
 721 language models. Nonetheless, the use of synthetic data generated by LLMs is not without risks; for
 722 instance, LLM hallucinations, even after filtering, could introduce novel errors or biases into models
 723 trained on such data, a factor that warrants careful consideration in future research and deployment.

725 B MGA FRAMEWORK IMPLEMENTATION DETAILS

727 B.1 TOOL MODEL TRAINING & RESOURCES



744 Figure 8: Implementation details. From a high-quality corpus, we sample a subset to serve as input
 745 for the LLM labeler and judger. Through data filtering, we train and quantize tool SLMs for each
 746 stage to improve inference efficiency, which are used to generate the reformulated corpus.

749 **High Quality Corpus** We conduct our reformulated corpus based on SmollM-Corpus⁴ (Ben Allal
 750 et al., 2024), expanding fineweb-edu-dedup source from 195B tokens to 770B tokens. Then we setup
 751 additionally experiments on FineWeb and FineWeb-Edu (Penedo et al., 2024), which constitute
 752 a solid foundation for research on data scaling approaches. Prior to these experiments, we have
 753 validated our approach on our in-house datasets. The results demonstrate consistent performance
 754 across both datasets, suggesting broad applicability of our method.

755 ⁴<https://github.com/huggingface/smollm/tree/main/text/pretraining>

756 **Tool Models Training** Initialized from a pretrained SLM (a 3.3B MoE model), we collect 50,000
 757 training samples through LLM teacher, where 15,000 of raw text to genre-audience pairs, 35,000
 758 of raw text to reformulated output. Each model’s validation responses are scored by capable LLM
 759 judger, that ensures the SLMs achieve comparable synthesis quality to the LLM labeler as shown
 760 in Table 1. The sequence length is 8192 with maximum prompt/response length 4,096 tokens, each
 761 model is trained 3 epochs on the samples with a cosine lr scheduler.

762 We also finetuned a public model, Qwen3-30B-A3B Base(Yang et al., 2025), to serve as the ‘Open
 763 Tool SLM’, which will also be released. The results confirm that its performance is highly compara-
 764 ble to our internal models, proving our method is not dependent on any specific proprietary tool, as
 765 shown in Table 4. The result provides conclusive evidence that our core methodology is a robust and
 766 reproducible technique that can be readily implemented by other researchers using public models.
 767

768 Table 4: Reformulation quality comparison between finetuned qwen3-a3b and its LLM teacher.

Models	Total Examples	5	4	3	2	1	Rate(≥ 3)	Diff
Labeler LLM	15,355	4,120	7,143	3,034	661	214	93.11%	-
Tool SLM	15,355	3,788	7,124	3,224	736	285	92.06%	-1.05%
Qwen3-30B-A3B (Finetuned)	15,355	3,322	7,154	3,614	805	249	91.76%	-1.35%

774 **Resource Analysis** To generate 770B synthetic tokens, it takes 256×64 and 1024×130 NVIDIA
 775 H100 GPU hours to process two stages.

777 C PRETRAINING AND EVALUATION SETUP

779 C.1 PRETRAINING DATA RECIPES

780 **Data Recipe** The training token budgets are 600B/600B/1000B for size of 134M/377M/1.7B mod-
 781 els, which are aligned with SmoLM1 series (Ben Allal et al., 2024). Our baseline is trained on
 782 SmoLM-Corpus dataset, in contrast to SmoLM’s recipe, we use unique token number from each
 783 source as the mixing ratio shown in Table 5. This ensures that different sources have consistent
 784 repetition epochs during training. For a fair comparison, the mixing ratios of other data sources are
 785 kept constant across experiments. We specifically adjusted the proportions of fineweb-edu-dedup
 786 and MGACorpus to isolate the impact of the MGA reformulation.

787 Table 5: MGACorpus experiments data recipe: source weight (%) and #unique_tokens × #epochs.

experiments	-	fineweb-edu-dedup	cosmopedia-v2	python-edu	open-web-math	MGACorpus
Baseline	weight	80.89	11.65	1.66	5.80	-
	#unique.tokens × #epochs	195 × 4.15	28 × 4.15	4 × 4.15	14 × 4.15	-
MGA-Expansion	weight	16.29	11.65	1.66	5.80	64.59
	#unique.tokens × #epochs	195 × 0.84	28 × 4.15	4 × 4.15	14 × 4.15	770 × 0.84

793 The experiment design for different strategies is presented in Table 6, which involves three datasets:
 794 (1) a 50B-token random sample from fineweb-edu-dedup, (2) a corresponding filtered subset from
 795 MGACorpus, and (3) a 450B-token deduplicated corpus obtained from Fineweb.

797 Table 6: Scaling experiments data recipe, values represent #unique_tokens × #epochs.

Repetition	Experiments	Training Budget	fineweb-edu dedup	MGA corpus	fineweb random	Design Rationale
EntireSet	Baseline	500B	50 × 10	-	-	What if we could collect more unique data.
	Full-Fineweb-Edu	500B	195 × 2.56	-	-	
	MGA Expansion	500B	50 × 2	200 × 2	-	Add MGA to reduce the repetition num.
Subset	Baseline	700B	50 × 1.4	-	450 × 1.4	Upsample to get 200B more budget.
	Upsample-EDU	700B	50 × 5	-	450 × 1	
	MGA Expansion	700B	50 × 1	200 × 1	450 × 1	

807 C.2 MODEL HYPERPARAMETERS

808 We sample 100 million tokens from SmoLM-Corpus as the validation dataset. The hyperparams are
 809 presented in Table 7. These hyperparameters are determined by scaling laws to ensure an optimal

baseline and are kept consistent across all experimental groups. The tokenizer used for training and computing token counts is the same as SmoLLM1⁵ with a vocab size of 49,152.

Table 7: Hyperparams of different model size.

model size	batch size	learning rate	weight decay	hidden size	ffn inner	num heads	num layers	shared q_head	seq len	tie emb	total params
134M	128	3e-3	0.1	1,204	4,096	8	8	1	8,192	false	134M
377M	320	1.5e-3	0.1	1,536	6,144	12	10	1	8,192	false	377M
1.7B	512	5e-4	0.1	2,560	10,240	20	16	1	8,192	false	1.68B
7B	1,024	4e-4	0.1	4,096	8,192	32	32	4	8,192	false	6.98B
13B	1,024	4e-4	0.1	4,096	12,288	32	48	4	8,192	false	12.9B

C.3 EVALUATION DETAILS

The LightEval results provided in Section 4.2 follow SmoLLM setting, that with GSM8K/MMLU 5-shot and all the others 0-shot. The benchmarks presented in Figure 9 and Figure 10 follow few-shot evaluation settings, specifically ARC(8-shots), TriviaQA(5-shots), Winogrande(5-shots) and similar configurations for other tasks.

D SUPPLEMENTARY EXPERIMENTAL RESULTS & ANALYSIS

D.1 BENCHMARK COMPARISONS WITH MORE SOTA MODELS

While model performance is influenced by multiple factors, we list some recently SOTA small language models as reference.

Table 8: Benchmark **MGA** with SOTA small LMs. Models of similar size are grouped. All results are obtained through **LIGHTEVAL** (Fourrier et al., 2023). Best results in each group are highlighted in **bold**, the second in underline, and in **green** for that MGA wins under fair comparison.

Model	#Params.	#Tokens	ARC(C+E)	Wino	Hella	MMLU	MMLU-PRO	CSQA	OpenBookQA	PIQA	TriviaQA	GSM8K	Avg.
SmoLLM2-135M	135M	2T	44.12	51.07	42.03	31.27	11.06	33.82	35	68.23	<u>1.91</u>	1.52	32.00
SmoLLM-135M	135M	600B	42.47	51.54	41.08	29.93	<u>11.4</u>	32.51	33.2	68.17	1.08	0.99	31.24
SmoLLM-135M (ours)	134M	600B	41.71	52.41	40.69	30.03	11.37	34.32	<u>35.4</u>	67.85	0.02	1.29	31.51
MGA-Expansion	134M	600B	43.01	<u>51.7</u>	41.25	30.1	11.76	32.68	36.4	67.3	2.05	1.44	31.77
Qwen2.5-0.5B	360M	18T	45.16	53.99	51.16	33.51	11.97	31.61	37.6	69.97	3.96	32.9	37.18
SmoLLM2-360M	360M	4T	53.4	52.33	54.58	35.29	11.17	37.92	37.6	71.76	16.73	<u>2.96</u>	37.37
SmoLLM-360M	360M	600B	<u>49.99</u>	52.96	<u>51.67</u>	33.84	<u>11.42</u>	34.81	37.6	71.87	2.27	1.97	34.84
SmoLLM-360M (ours)	377M	600B	48.57	52.64	51.02	33.63	11.25	36.77	39	71	0.29	1.52	34.57
MGA-Expansion	377M	600B	49.39	52.64	51.34	34.09	11.35	37.1	<u>38</u>	72.31	7.28	<u>1.74</u>	35.52
Qwen2.5-1.5B	1.3B	18T	58.36	58.64	66.39	40.23	13.85	34.4	39.6	75.95	20.51	60.8	46.87
SmoLLM2-1.7B	1.7B	11T	60.42	59.59	68.73	41.4	19.61	43.65	42.6	77.53	36.68	<u>29.04</u>	47.93
Llama-3.2-1B	1.2B	9T	49.2	57.8	61.2	36.63	11.7	41.2	38.4	74.8	<u>28.1</u>	7.2	40.62
OLMo-1B-0724	1B	3.05T	44.71	56.04	64.38	32.3	11.8	33.09	38	75.24	13.82	2.43	37.18
SmoLLM-1.7B	1.7B	1T	59.95	54.7	62.83	39.35	10.92	38	42.6	75.9	13.14	4.62	40.20
SmoLLM-1.7B (ours)	1.7B	1T	59.63	57.38	65.19	39.4	12.11	42.59	45.6	76.88	4.95	7.81	41.15
MGA-Expansion	1.7B	1T	60.36	57.46	65.52	40.79	14.1	41.11	<u>42.8</u>	77.53	20.42	13.87	43.4

In our experimental observations (Table 8), notable performance improvements are demonstrated in both TriviaQA and GSM8k benchmarks, warranting a detailed examination of these score variations. The enhanced TriviaQA performance exhibited by SmoLLM1-1.7B relative to our baseline can be attributed to the larger proportion of Cosmopedia in its training configuration. Both MGACorpus and Cosmopedia employ synthetic methodologies, which contribute to improved learning efficiency. The observed gains in GSM8K performance can be traced to the target genres, including teaching schemas and problem-solving exemplars, embedded within the Reformulation component. This early exposure to structured problem-solving approaches facilitates more effective performance on analogous mathematical reasoning tasks.

⁵<https://huggingface.co/HuggingFaceTB/cosmo2-tokenizer>

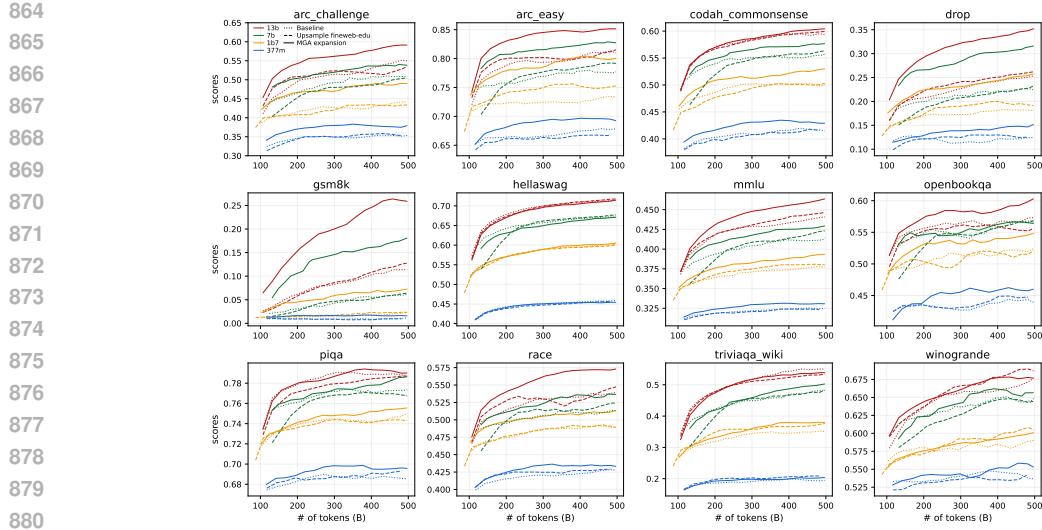


Figure 9: Detail evaluation results of EntireSet described in Table 6. MGACorpus group demonstrates advantages over other groups across most evaluation sets, consistently across models of sizes.

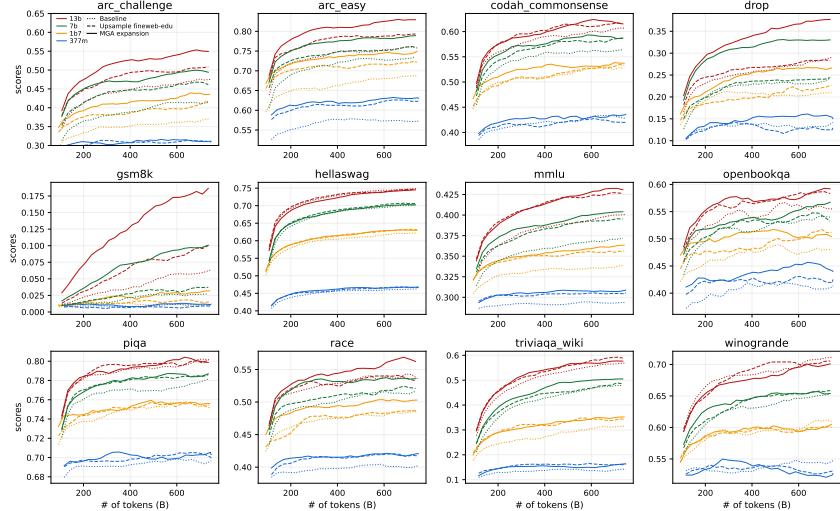


Figure 10: Detail evaluation results of Subset described in Table 6. As the model size increases, the performance gap between the upsampling group and MGACorpus gradually widens in ARC, DROP, GSM8K, RACE, but with some variations observed in TriviaQA and WinoGrande.

D.2 "MGA-ONLY" EXPERIMENT

Our primary goal with MGA is efficient dataset expansion, typically achieved by mixing the generated corpus with existing real data, aligning with current best practices for leveraging synthetic data. However, to better characterize the properties of the MGACorpus itself and understand the impact of training exclusively on reformulated content, we also investigate an experimental setting where MGACorpus completely replaces its source data (fineweb-edu-dedup).

Table 9: MGACorpus experiments data source weight (%).

experiments	fineweb-edu-dedup	cosmopedia-v2	python-edu	open-web-math	MGA-corpus
Baseline	80.89	11.65	1.66	5.80	-
MGA-Expansion	16.29	11.65	1.66	5.80	64.59
MGA-Only	-	11.65	1.66	5.80	80.89

As shown in Table 10, the absence of real data leads to performance degradation across most tasks (average -0.95), particularly in two tasks, Hellaswag(-1.23/-1.69/-2.85) and CommonsenseQA(-

3.11/-4.83/-4.50). This decline can be attributed to our design choice, which focuses on diversity and overall quality rather than requiring the preservation of all information from each raw documents.

Table 10: Comparison between MGA-Expansion and MGA-Only

Model	#Params.	#Tokens	ARC(C+E)	Wino.	Hella.	MMLU	MMLU-PRO	CSQA	OpenBookQA	PIQA	TriviaQA	GSM8K	Avg.
MGA-Expansion	134M	600B	43.01	51.7	41.25	30.1	11.76	32.68	36.4	67.3	2.05	1.44	31.77
MGA-Only	134M	600B	41.98	51.38	40.02	29.87	11.5	29.57	33	68.01	2.26	1.06	30.87
			↓-1.03	↓-0.32	↓-1.23	↓-0.23	↓-0.26	↓-3.11	↓-3.40	↑0.71	↑0.21	↓-0.38	↓-0.90
MGA-Expansion	377M	600B	49.39	52.64	51.34	34.09	11.35	37.1	38	72.31	7.28	1.74	35.52
MGA-Only	377M	600B	47.95	53.35	49.65	33.31	11.38	32.27	38	70.95	6.83	1.59	34.53
			↓-1.44	↑0.71	↓-1.69	↓-0.78	↑0.03	↓-4.83	-	↓-1.36	↓-0.45	↓-0.15	↓-0.99
MGA-Expansion	1.7B	1T	60.36	57.46	65.52	40.79	14.1	41.11	42.8	77.53	20.42	13.87	43.40
MGA-Only	1.7B	1T	59.02	57.06	62.67	40.34	13.51	36.61	45.2	76.71	19.78	13.57	42.45
			↓-1.34	↓-0.40	↓-2.85	↓-0.45	↓-0.59	↓-4.50	↑2.40	↓-0.82	↓-0.64	↓-0.30	↓-0.95

MGA-Only Setting of PE Ablation Upon relaxing the information preservation requirements for PE objectives in the MGA-Only setting, we observe a complete collapse in knowledge-based dimensions while maintaining modest improvements in reasoning and mathematical capabilities. This divergence suggests that different cognitive capabilities have distinct requirements for the richness and nature of training data content.

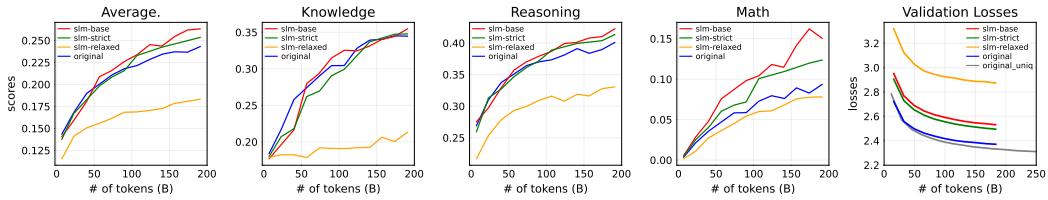


Figure 11: Corresponding benchmark results described in Section 4.3.2.

D.3 COMPARISON WITH OPEN SYNTHETIC DATA

This section provides crucial empirical support for the central thesis of our Discussion (Section 4.3): the path to resolving data scarcity lies not in a single synthesis technique, but in the thoughtful combination of diverse strategies. We argue for an approach that breaks down the walls between **different synthetic data paradigms**, demonstrating that a model’s robustness comes from being trained on a rich variety of data formats. The following experiments are designed to validate this philosophy, showing how MGA acts as a synergistic component within a broader data ecosystem, rather than a standalone solution.

D.3.1 HEAD-TO-HEAD PERFORMANCE AGAINST OPEN SYNTHETIC DATASETS

This first experiment aims to benchmark MGA’s effectiveness against other well-known synthetic data generation methods when each is used as the primary training corpus. We trained 377M parameter models for 300B tokens on several distinct datasets to establish a clear performance baseline. For a fair comparison with Cosmopedia, the MGA corpus was sampled to 28B unique tokens, and both datasets were repeated approximately 10.7 times during training.

Table 11: Comparative benchmark performance of 377M models trained on MGA reformulations versus other synthetic datasets for 300B tokens. For a fair comparison with Cosmopedia, MGA is sampled to 28B unique tokens, with both datasets then repeated 10.7 times during training. All benchmarks are 0-shot evaluations (obtained through LIGHT-EVAL), except for MMLU (5-shot).

Category	Document Sources	Synthetic Target	ARC(C+E)	Wino.	Hella.	MMLU	CSQA	OpenBookQA	PIQA	TriviaQA	GSM8K	Avg.
Cosmopedia	Textbooks/Webs	Story/Textbook/Wiki mix	42.15	50.43	45.06	29.17	30.38	33.2	68.77	0.23	35.57	
MGA	High quality webs	Diverse Genre-Audience	45.65	51.22	42.31	31.42	32.19	37.2	68.39	3.79	37.28	
		Low quality webs	29.01	50.83	38.36	26.29	29.32	32	67.03	0	31.72	
		Extract knowledge	40.42	53.2	44.65	30.57	28.99	35	69.42	0.96	35.72	
		Knowledge list	42.08	52.17	42.7	30.71	32.51	35.4	70.08	0	36.21	
Nemotron-CC	High quality webs	Concise and clear passage	42.22	52.01	43.99	30.96	31.53	35	69.7	0.06	36.21	
		Wrap-medium (Wiki style)	42.95	52.17	43.72	31.06	31.53	36.2	70.13	0.82	36.63	
		Diverse QA pairs	46.96	52.57	49.03	31.36	38.82	38.8	70.84	9.21	40.72 ⁶	
MGA	High quality webs	Diverse Genre-Audience	45.33	52.41	42.42	31.33	31.45	38	68.61	4.24	37.34	

972 The results in Table 11 highlight MGA’s competitive performance as a general-purpose augmentation
 973 strategy. MGA (average 37.28) surpasses Cosmopedia (35.57), which is a blend of story, text-
 974 book, and wiki formats. When compared against the various synthesis strategies from Nemotron-
 975 CC, MGA (average 37.34) again shows strong performance, outperforming most alternatives such
 976 as ‘extract knowledge’ (35.72) and ‘wrap-medium (Wiki style)’ (36.63).

977 Notably, while Nemotron’s diverse QA slice achieves the highest average score (40.72), this advan-
 978 tage is likely attributable to its format aligning directly with the question-answering structure preva-
 979 lent in our 0-shot evaluation benchmarks. Despite this format-specific advantage, MGA’s broader
 980 reformulation approach proves its robust utility by outperforming five of the six Nemotron strategies.
 981 This underscores MGA’s value in building a well-rounded and capable base model.

983 D.3.2 INVESTIGATING SYNERGISTIC EFFECTS VIA CONTINUED PRE-TRAINING

985 Moving beyond static head-to-head comparisons, this second experiment directly tests our hypoth-
 986 esis about the importance of data diversity. We investigate the dynamic interplay between different
 987 data types to see how a model trained on one corpus adapts when another is introduced. This di-
 988 rectly probes the synergistic potential discussed in our main paper. To do this, we took checkpoints
 989 of models pre-trained for 300B tokens on our MGA-Corpus and the Nemotron-CC (QA-slice), re-
 990 spectively. We then continued pre-training each model for an additional 30B tokens, mixing in data
 991 from the other corpus at a 1:1 ratio.

993 **Table 12: Model Performance with Mixed-Corpus Continued Pre-training**

994 Experiment	995 Tokens	996 Wino.	997 C-QA	998 Hella.	999 MMLU	0BQA	PIQA	TriviaQA	ARC	AVG
mga	300B	52.4	31.5	42.4	31.3	38.0	68.6	4.2	45.3	37.3
mga + mixct	330B	52.8	37.6	47.5	31.6	37.8	70.2	4.4	45.5	39.2
<i>Change</i>		<i>+0.4</i>	<i>+6.1</i>	<i>+5.1</i>	<i>+0.3</i>	<i>-0.2</i>	<i>+1.6</i>	<i>+0.2</i>	<i>+0.2</i>	<i>+1.9</i>
nemotron_qa	300B	52.6	38.8	49.0	31.4	38.8	70.8	9.2	47.0	40.7
nemotron_qa + mixct	330B	51.8	37.8	47.8	31.8	36.6	70.6	3.5	45.9	39.1
<i>Change</i>		<i>-0.8</i>	<i>-1.1</i>	<i>-1.2</i>	<i>+0.4</i>	<i>-2.2</i>	<i>-0.2</i>	<i>-5.7</i>	<i>-1.1</i>	<i>-1.6</i>

1001 This cross-mixing experiment revealed two interesting and complementary phenomena:

1002 **Synergy and Receptiveness of MGA.** When Nemotron-QA’s structured data was mixed into the
 1003 MGA-trained model, the model’s average performance on downstream benchmarks significantly
 1004 improved by 1.9 points. This suggests that the diverse, rich foundation built by MGA is highly
 1005 effective and “receptive”, readily absorbing the benefits of more specialized, format-aligned data.

1006 **Distributional Path Dependence.** Conversely, when MGA’s diverse data was mixed into the
 1007 Nemotron-QA-trained model, its performance decreased by 1.6 points. This suggests that pre-
 1008 training on a stylistically monolithic dataset can create a “path dependence”, making it harder for
 1009 the model to adapt when a different data distribution is introduced.

1010 In conclusion, this experiment provides compelling evidence for our core argument. The strong per-
 1011 formance of task-aligned data like Nemotron-QA is undeniable, but it does not represent a complete
 1012 solution. Our findings show that the most effective approach is one that embraces variety. MGA’s
 1013 primary value lies in creating a robust, generalist foundation that is highly receptive to other data
 1014 types. It helps build a model that is not locked into a single stylistic mode, effectively breaking down
 1015 the data walls and paving the way for a more adaptable and capable AI. This synergy is the key to
 1016 building next-generation base models.

1018 D.4 FURTHER ANALYSIS OF MODEL COLLAPSE

1019 **Further Discussion of Section 4.3.3** For our analysis method in Figure 7, we define the token loss
 1020 difference as $loss_{diff}^i = loss_{synt}^i - loss_{real}^i$, where i is the token index, synt/real is dataset used for
 1021 model training. Note that we consistently use synthetic minus real, where a positive value indicates
 1022 poorer prediction performance by the synthetic model on a given sample.

1023
 1024 ⁶The predominantly 0-shot evaluation particularly benefits datasets like Nemotron ‘diverse QA pairs’ whose
 1025 format directly aligns with many evaluation tasks.

1026 Since next token prediction is computed based on preceding context, we define the first anomaly
 1027 position to identify where a model’s prediction for tokens within the window begins to significantly
 1028 deteriorate. The definition is as follows:

$$1029 \text{first_anomaly_position} = \min\{p \mid \left| \frac{1}{w} \sum_{i=p}^{p+w-1} \text{loss}_{\text{diff}}^i \right| > |\mu| + k\sigma \},$$

$$1030$$

$$1031$$

$$1032$$

1033 where $w = \max(0.05 \times \text{seq.length}, 1)$, $\mu = \text{mean}(\text{loss}_{\text{diff}}^i)$, $\sigma = \text{std}(\text{loss}_{\text{diff}}^i)$. Here, we employ
 1034 the absolute value of the windowed average loss to identify significant performance degradation in
 1035 either model. This approach enables the detection of notable prediction quality drops regardless of
 1036 which model (synthetic or real) experiences the deterioration.

1037 Finally, we define the normalized position, enabling fair comparisons across various sequence
 1038 lengths:

$$1039 \text{normalized_position} = \begin{cases} \frac{\text{first_anomaly_position}}{\text{seq.length}} \times 100\% & \text{if anomaly found} \\ -1 & \text{otherwise} \end{cases}$$

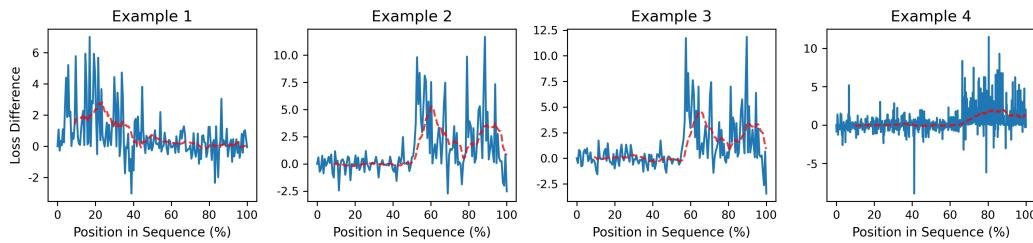
$$1040$$

$$1041$$

$$1042$$

$$1043$$

1044 Below are example cases from English and Chinese documents. Figure 12 presents the token loss
 1045 difference on each position. Example 2 and Example 3 show similar anomaly pattern, we can get
 1046 the reason in Figure 13, that they are from the same website source contain identical boilerplate text
 1047 about region selection and website localization at the end of their content.



1056 Figure 12: Random examples sampling from where $\text{mean}(\text{loss}_{\text{diff}}^i) > 0.5$, the synthetic-trained
 1057 model fail to predict the tokens in later sequence positions.

1058 This suggests potential noise in the data preprocessing pipeline, specifically in handling website
 1059 navigation elements and localization prompts that should have been removed during content extrac-
 1060 tion.

1061 While these examples demonstrate clear patterns of model behavior differences in handling noisy
 1062 web data, we acknowledge that this analysis is limited to selected cases with apparent preprocessing
 1063 artifacts. A more comprehensive evaluation across diverse data sources and quality levels would be
 1064 necessary to fully understand the impact of synthetic training data on model performance.

1065
 1066
 1067
 1068
 1069
 1070
 1071
 1072
 1073
 1074
 1075
 1076
 1077
 1078
 1079

1080
1081
1082
1083 Web design encompasses many different skills and disciplines in the production and maintenance of websites . The different areas of web design include
1084 web graphic design ; interface design ; authoring , including standardised code and proprietary software ; user experience design ; and search engine
1085 optimization .
1086 Web design partially overlaps web engineering in the broader scope of web development . Web designers are expected to have an awareness of usability
1087 and if their role involves creating mark up then they are also expected to be up to date with web accessibility guidelines .
1088 Flat design is the new trend in web design . It focuses on a minimalist approach for usability . It is specific to dimensional and flat icons and images
1089 and fresh open spaces and vibrant colors . With user experience as key anything that is considered overly decorative or cluttered is removed to lessen
1090 distraction . Flat designs are also easy designs to convert to mobile and other devices which makes it's functionality beneficial .

1091 Compute the product of all elements in an array .
1092 Is my wife right ?
1093 Energy of a photon
1094 Is this triangle right – angled ?
1095 Mat lab Basics II – squares
1096 Basic mat lab numeric and string manipulations
1097 Basic ics : counting digits of a number irrespective of the sign
1098 Rep edition of matrices
1099 generate capital english alphabets
1100 Choose a web site to get translated content where available and see local events and offers . Based on your location , we recommend that you select .
1101 You can also select a web site from the following list :
1102 Select the China site (in Chinese or English) for best site performance . Other Math Works country sites are not optimized for visits from your location .
1103 Contact your local office

1104 Given the letter , the output will show the corresponding ASCII code .
1105 Add two numbers
1106 select the primes of a vector
1107 Find the area of the four walls
1108 Mat lab Basics – Set unwanted parts of a vector to zero
1109 Divide by 4
1110 Write a code that will follow the equation $y = x * (x + x) * x$.
1111 Log of a number
1112 Vector with a repeated entry
1113 Square root of a number
1114 Choose a web site to get translated content where available and see local events and offers . Based on your location , we recommend that you select .
1115 You can also select a web site from the following list :
1116 Select the China site (in Chinese or English) for best site performance . Other Math Works country sites are not optimized for visits from your location .

Figure 13: Corresponding cases sampled from Fineweb-Edu, which align with the loss patterns shown in Figure 12, with higher loss by synthetic-trained model highlighted in red.

英语语法 极差，怎样 快速 学习？
办法一：读《张道藩 英语语法》。1、从 句子 挑战 看着，明白 句子 的 主谓 成分，以及 不同 句子 成分 由什么 隔开 为 隔。2、阅读 直接 构造 的 句型，有 大量 的 句型 参考。3、上 两个 学习 时 重点 讲 连词 动词 的 构造，参考 学习 相应 例句，自己 读 读，熟悉 构造，熟悉 例句。4、看 语法 例句 和 其他 例句 构造 句子 与 看 句子 构造 的 长处。到 当时 自己 已经 熟悉 了 动词，学习 办法 自己 会 熟悉 了。办法 二：从 新易经 英语 2 初级 自学。第一个 单元 例句 读 读，然后 重点 处理 了 语法 问题。日期 10 月份 足够 了，这样 办法 可取。读 语法，能够 理解 例句，看 例句，考题 是 技巧，阅读、写作 等 方面，奇特 2 月 日期 报名，抵达 中国，11 月份 的 分数。学习 英语 一定 需要 踏实，每一个 例句，任何 一◆ 高度 就能 理解 英语 语法 那 都 是 不现实 的，尤其 对于 基础 漂亮 的 学习者 来言，当然 为 了 有 时间 长者，我们 建议 最好 是 能 报 一个 语法 培训班。因为 初学者 非常 需要 一个 正确 的 引导者。在其 自由 的 过程 中 起 到 推 动 的 作用。关于 英语 语调 构造 选 择 语调 的 问题，大家 可以 直接 点播 一下 《 有 例 有 技术 球》。是 非常 热门 的 英语 机构 测评 和 学习 规划 提供 专家。帮助 了 很多 的 英语 学习 者。可以 让 我们 少 走 很多 鸡路！学习 英语 时，不 需要 如何 选课 班，现学 英语 五花八门，千万 不要 因此 而 为，怎么 选课 防 护？哪些 课 实际 有 效果？最近 很 火 的 《 英文 文章 《 比人 英语 培训 谁 高 谁 低 》 分析 主题 语 英语 安身 1 》，上 100 个 英语 大师 推荐，盛大 主流 英语 班 的 情况 一 一 通，参考 价值 非常 大！

Figure 14: Chinese corpus samples with higher loss by synthetic-trained model in red.

1134 E PROMPTS AND CASES

1135 E.1 CASES

1138 Table 13: Example outputs of SLM variants.

Raw Document		
Genre-Audience Pairs	Tool Model	Reformulated Document
<p>- Academic teaching paper, using rigorous academic style to deeply explore the importance, methods and practical case analysis of capturing the optimal timing of 'movement-stillness' transitions in chemistry teaching. Through references to relevant educational theories and actual teaching experiences, provide meaningful teaching insights for chemistry teachers.</p> <p>- Middle school chemistry teachers, aged 25-45, predominantly female, with professional chemistry education background. They are patient and steady in character, have been teaching middle school chemistry for many years, and wish to improve teaching quality and help students better understand chemistry.</p>	<p>SLM-Base</p> <p>**The Importance and Practice of Capturing Optimal Timing of "Movement-Stillness" Transitions in Middle School Chemistry Teaching** **Abstract**: This paper aims to deeply explore the importance, methods, and practical case analysis of capturing optimal timing of "movement-stillness" transitions in middle school chemistry teaching from theoretical and practical perspectives. Through references to relevant educational theories and actual teaching experiences, it provides meaningful teaching insights for middle school chemistry teachers to help improve teaching quality and promote better student understanding of chemistry knowledge.</p> <p>**Keywords**: Middle school chemistry teaching; "Movement-stillness" transitions; Optimal timing; Teaching methods</p> <p>**1. Introduction**</p> <p>With the continuous advancement of new curriculum concepts, middle school chemistry teaching is actively exploring more effective teaching methods to cultivate students' chemistry literacy and comprehensive abilities. Among these, having students adopt a learning method that combines "movement and stillness" has gradually gained attention, and skillfully capturing the optimal timing of "movement-stillness" transitions has become one of the keys to successful chemistry teaching.</p> <p>**2. The Importance and Practical Application of Optimal Timing of "Movement-Stillness" Transitions**</p> <p>(1) Alignment with Educational Theory</p> <p>According to constructivist learning theory, student learning is a process of actively constructing knowledge through interaction with the external environment based on existing knowledge and experience. In middle school chemistry teaching, the learning method of "movement-stillness" transitions helps students concretize abstract chemical concepts through dynamic practical operations, leading to better understanding and mastery of chemistry knowledge.</p>	<p>**The Importance and Practice of Capturing Optimal Timing of "Movement-Stillness" Transitions in Middle School Chemistry Teaching** **Abstract**: This paper aims to deeply explore the importance, methods, and practical case analysis of capturing optimal timing of "movement-stillness" transitions in middle school chemistry teaching from theoretical and practical perspectives. Through references to relevant educational theories and actual teaching experiences, it provides meaningful teaching insights for middle school chemistry teachers to help improve teaching quality and promote better student understanding of chemistry knowledge.</p> <p>**Keywords**: Middle school chemistry teaching; "Movement-stillness" transitions; Optimal timing; Teaching methods</p> <p>**1. Introduction**</p> <p>With the continuous advancement of new curriculum concepts, middle school chemistry teaching is actively exploring more effective teaching methods to cultivate students' chemistry literacy and comprehensive abilities. Among these, having students adopt a learning method that combines "movement and stillness" has gradually gained attention, and skillfully capturing the optimal timing of "movement-stillness" transitions has become one of the keys to successful chemistry teaching.</p> <p>**2. The Importance and Practical Application of Optimal Timing of "Movement-Stillness" Transitions**</p> <p>(1) Alignment with Educational Theory</p> <p>According to constructivist learning theory, student learning is a process of actively constructing knowledge through interaction with the external environment based on existing knowledge and experience. In middle school chemistry teaching, the learning method of "movement-stillness" transitions helps students concretize abstract chemical concepts through dynamic practical operations, leading to better understanding and mastery of chemistry knowledge.</p>
		<p>In the field of chemistry teaching, why is it said that skillfully capturing the optimal timing of "movement-stillness" transitions is key to success? Under the new curriculum philosophy, students are encouraged to use a learning method that combines "movement and stillness" to transform their thinking patterns and learn through activities. Being skillful at capturing the optimal timing of "movement-stillness" transitions is truly the key point of success - once this optimal timing is grasped, students can easily learn chemistry. From a theoretical perspective, in middle school chemistry teaching, the concept of "movement-stillness" transitions aligns with students' cognitive patterns and helps stimulate their learning interest and initiative. Students can directly experience chemical phenomena through dynamic practical operations, deepening their understanding of knowledge, while in the static stages of thinking and summarizing, they can cleverly perceive knowledge to rational cognition and construct complete knowledge systems. In practice, teachers can analyze through specific cases. For example, when explaining the laboratory preparation of oxygen, teachers first guide students to observe the experimental apparatus and understand the role of each part - this is the "movement" stage. Then, letting students personally conduct the experiment and observe the experimental phenomena such as the bubbles of oxygen appearing, this is the "stillness" stage. After students complete the experiment, organizing them to discuss problems encountered during the experiment and their solutions, summarizing and concluding, returns again to "stillness" reflection. Thus, through "movement-stillness" transitions, students not only master the knowledge and skills of oxygen preparation but also develop abilities in observation, analysis, and problem-solving. In conclusion, middle school chemistry teachers should fully recognize the importance of skillfully capturing optimal timing of "movement-stillness" transitions in teaching, and through reasonable design of teaching segments, integrate the learning method combining "movement and stillness" throughout the entire teaching process to improve teaching quality and help students better understand chemistry knowledge.</p>
		<p>In the field of education, for teachers of fundamental subjects, finding the key opportunities to guide students in effective learning is central to improving teaching effectiveness. Taking middle school chemistry teaching as an example, new educational concepts advocate students using flexible and diverse learning modes to transform their thinking pathways, enabling them to explore knowledge more independently in the learning process. Finding the perfect opportunity to guide students in achieving this thinking transformation during the learning process is undoubtedly key to successful teaching. Once this opportunity is successfully grasped, teaching effectiveness will be significantly improved, and students will be able to understand and master the learned knowledge more deeply.</p>
<p>- Popular science story, using a lively and interesting narrative to tell story of how middle school students realizes the importance of capturing the optimal timing of 'movement-stillness' transitions in their chemistry learning process. Through specific interesting experiments, help students understand chemistry knowledge in a relaxed and enjoyable atmosphere, sparking interest in chemistry learning.</p> <p>- Middle school students who find chemistry difficult, aged 12-15, both male and female, with below-average academic performance. They are generally lively and active in character, lack interest in chemistry but must study it, and hope to find ways to improve their chemistry grades.</p>	<p>SLM-Base</p> <p>Once there was a middle school student named Xiao Ming, who was very lively and active, but initially had little interest in chemistry as a subject, and his grades were always below average. One day, during a chemistry class, the teacher taught a particularly interesting lesson about a magical chemical change. The teacher said that during this change process, it was like having a mysterious little sprite at play - sometimes it was very quiet, and sometimes it became very active. This caught Xiao Ming's attention completely. He wondered when this little sprite would be quiet, and when it would become active.</p> <p>Later, the teacher led everyone in doing an experiment. In the experiment, there were two small bottles - one bottle contained a transparent liquid, and the other bottle contained a magical powder. The teacher slowly poured the powder into the bottle containing the liquid. At first, nothing happened, and the liquid in the bottle remained still. Just when everyone was getting a bit impatient, suddenly, the liquid in the bottle started bubbling - it was as if the little sprite had suddenly woken up and started playing excitedly. This change came so suddenly that all the students were stunned, their eyes wide open. The teacher also said that when studying chemistry, it's just like doing this experiment. You need to be good at capturing the optimal timing of "movement-stillness" transitions. If you can grasp it at that moment, you'll be able to better understand chemistry knowledge, and learning will become twice as effective with half the effort.</p> <p>SLM-Strict</p> <p>Hey, classmates! Do you know that in the wonderful world of chemistry learning, there's something really important - that's being good at capturing the optimal timing of "movement-stillness" transitions, this is the key to success! Just like under the new curriculum concepts, teachers guide us to use a learning method that combines "movement and stillness," which can transform our way of thinking and help us learn through activities. In my chemistry learning process, I've deeply experienced this point.</p> <p>Once, when we were doing an interesting experiment, the teacher told us to be good at catching the optimal timing of "movement-stillness" transitions. At first, I didn't quite understand. But later, I discovered that when certain phenomena in the experiment start to change, that's the optimal timing for "movement-stillness" transitions! As long as we catch this timing, we can achieve twice the results with half the effort in our experiments!</p> <p>Hey classmates, don't you think it's amazing too? Let's discover more fun together in our chemistry learning journey!</p> <p>SLM-Relaxed</p> <p>In a vibrant middle school classroom, there was a group of lively and active students, aged between 12-15, both boys and girls, with below-average academic performance. Among them was a student named Xiao Le, who had no interest in chemistry but had to study it, and was always looking for ways to improve his grades.</p> <p>One day, in the school laboratory, Xiao Le watched the teacher perform a magical experiment. The teacher first put some ordinary-looking powder into a transparent bottle, then added a liquid. Immediately, it was like a lively party started in the bottle, with lots of colorful bubbles appearing - this completely captivated Xiao Le and his classmates.</p> <p>Xiao Le thought to himself that this was just like how he played in the playground - sometimes resting quietly, sometimes running around wildly - isn't this a combination of "movement and stillness"?? Moreover, he discovered that during this process, the moment when the powder and liquid made contact was the most magical moment, like finding a hidden treasure.</p> <p>From then on, every time Xiao Le did chemistry experiments, he would pay special attention to this optimal timing of "movement-stillness" transitions. He found that when he caught this timing, he could understand chemistry knowledge better. His chemistry grades gradually improved, and his interest in studying chemistry grew stronger and stronger.</p>	<p>Once there was a middle school student named Xiao Ming, who was very lively and active, but initially had little interest in chemistry as a subject, and his grades were always below average. One day, during a chemistry class, the teacher taught a particularly interesting lesson about a magical chemical change. The teacher said that during this change process, it was like having a mysterious little sprite at play - sometimes it was very quiet, and sometimes it became very active. This caught Xiao Ming's attention completely. He wondered when this little sprite would be quiet, and when it would become active.</p> <p>Later, the teacher led everyone in doing an experiment. In the experiment, there were two small bottles - one bottle contained a transparent liquid, and the other bottle contained a magical powder. The teacher slowly poured the powder into the bottle containing the liquid. At first, nothing happened, and the liquid in the bottle remained still. Just when everyone was getting a bit impatient, suddenly, the liquid in the bottle started bubbling - it was as if the little sprite had suddenly woken up and started playing excitedly. This change came so suddenly that all the students were stunned, their eyes wide open. The teacher also said that when studying chemistry, it's just like doing this experiment. You need to be good at capturing the optimal timing of "movement-stillness" transitions. If you can grasp it at that moment, you'll be able to better understand chemistry knowledge, and learning will become twice as effective with half the effort.</p> <p>Hey, classmates! Do you know that in the wonderful world of chemistry learning, there's something really important - that's being good at capturing the optimal timing of "movement-stillness" transitions, this is the key to success! Just like under the new curriculum concepts, teachers guide us to use a learning method that combines "movement and stillness," which can transform our way of thinking and help us learn through activities. In my chemistry learning process, I've deeply experienced this point.</p> <p>Once, when we were doing an interesting experiment, the teacher told us to be good at catching the optimal timing of "movement-stillness" transitions. At first, I didn't quite understand. But later, I discovered that when certain phenomena in the experiment start to change, that's the optimal timing for "movement-stillness" transitions! As long as we catch this timing, we can achieve twice the results with half the effort in our experiments!</p> <p>Hey, classmates, don't you think it's amazing too? Let's discover more fun together in our chemistry learning journey!</p> <p>Once, when we were doing an interesting experiment, the teacher told us to be good at catching the optimal timing of "movement-stillness" transitions. At first, I didn't quite understand. But later, I discovered that when certain phenomena in the experiment start to change, that's the optimal timing for "movement-stillness" transitions! As long as we catch this timing, we can achieve twice the results with half the effort in our experiments!</p> <p>Hey, classmates, don't you think it's amazing too? Let's discover more fun together in our chemistry learning journey!</p>
		<p>SLM-Base</p> <p>Dear parents, do you know that in your child's chemistry learning process, there's an important method called the "movement-stillness combined" learning approach? It's just like in daily life - sometimes children need to quietly think about problems, and other times they need to engage in hands-on experiments or practical activities. So why is skillfully capturing the optimal timing of "movement-stillness" transitions key to success?</p> <p>Think about it - at school, teachers use this learning method to transform children's thinking patterns, helping them learn through activities. For example, when students are learning about chemical elements, they might first quietly listen to the teacher explain the properties and characteristics of elements - this is the "stillness" phase. Then, the teacher will arrange experiments for children to perform hands-on operations and observe experimental phenomena - this is the "movement" phase. If parents can help their children capture the optimal timing of these "movement-stillness" transitions, their chemistry learning will become twice as effective with half the effort.</p> <p>So how can we help children achieve this in daily life? It's actually quite simple. For instance, after your child learns a chemistry concept, you can ask: "Sweetie, can you think of any phenomena in daily life that relate to this concept?" This is a small interaction that guides children from "stillness" to "movement." When your child starts thinking and answering, you can encourage them to do some simple experiments to verify their ideas, like conducting a small experiment using materials at home. This way, children can better understand and apply the "movement-stillness" learning method in their daily lives, and during this process, your parent-child relationship will naturally grow stronger!</p> <p>SLM-Strict</p> <p>As for why skillfully grasping the optimal timing of "movement-stillness" transitions is key to your child's success in learning chemistry - here's something to consider. The new curriculum approach encourages children to use a learning method that combines "movement and stillness," as this can transform their thinking patterns and help them learn through activities. Being able to capture these transition moments is crucial - when you catch these perfect moments, your child's learning becomes twice as effective with half the effort.</p> <p>So how can we parents help our children understand and apply this learning method in daily life? Let me share some simple parent-child interaction methods. For example, we can do a small experiment together with our children, like observing the reaction between baking soda and white vinegar. During this process, the hands-on operation is the "movement" part, while observing the phenomena and thinking about the reasons is the "stillness" part - see how this combines "movement and stillness"?? You can guide your child to pay attention to these processes, and gradually, they'll better master this learning method! Plus, during this process, your parent-child relationship will naturally grow stronger!</p> <p>SLM-Relaxed</p> <p>"A Parent's Guide to Supporting Children's Learning and Growth" During a child's learning and growth process, there is an important learning method called the movement-stillness transition method. It's like a magical key that can help children learn more easily and master knowledge by enhancing their learning effectiveness. So, how do we guide children to master this method?</p> <p>For parents, although their professional knowledge may be limited, they can help children understand through various scenarios in daily life. For example, during a family trip, when children encounter different scenery and activities, some children might initially run and play excitedly - this is the "movement" state. Then when they see beautiful scenery or encounter interesting things, they stop to carefully observe and think - this enters the "stillness" state. Parents should be good at recognizing these transitions between states in their children, guiding them at appropriate times to help them understand that this process of "movement-stillness" transition is actually the learning process. Through such parent-child interactions, parents can not only better understand what their children are learning at school but also improve their parent-child relationship, accompanying their children in healthy and happy growth.</p>
		<p>SLM-Base</p> <p>Dear parents, do you know that in your child's chemistry learning process, there's an important method called the "movement-stillness combined" learning approach? It's just like in daily life - sometimes children need to quietly think about problems, and other times they need to engage in hands-on experiments or practical activities. So why is skillfully capturing the optimal timing of "movement-stillness" transitions key to success?</p> <p>Think about it - at school, teachers use this learning method to transform children's thinking patterns, helping them learn through activities. For example, when students are learning about chemical elements, they might first quietly listen to the teacher explain the properties and characteristics of elements - this is the "stillness" phase. Then, the teacher will arrange experiments for children to perform hands-on operations and observe experimental phenomena - this is the "movement" phase. If parents can help their children capture the optimal timing of these "movement-stillness" transitions, their chemistry learning will become twice as effective with half the effort.</p> <p>So how can we help children achieve this in daily life? It's actually quite simple. For instance, after your child learns a chemistry concept, you can ask: "Sweetie, can you think of any phenomena in daily life that relate to this concept?" This is a small interaction that guides children from "stillness" to "movement." When your child starts thinking and answering, you can encourage them to do some simple experiments to verify their ideas, like conducting a small experiment using materials at home. This way, children can better understand and apply the "movement-stillness" learning method in their daily lives, and during this process, your parent-child relationship will naturally grow stronger!</p> <p>SLM-Strict</p> <p>As for why skillfully grasping the optimal timing of "movement-stillness" transitions is key to your child's success in learning chemistry - here's something to consider. The new curriculum approach encourages children to use a learning method that combines "movement and stillness," as this can transform their thinking patterns and help them learn through activities. Being able to capture these transition moments is crucial - when you catch these perfect moments, your child's learning becomes twice as effective with half the effort.</p> <p>So how can we parents help our children understand and apply this learning method in daily life? Let me share some simple parent-child interaction methods. For example, we can do a small experiment together with our children, like observing the reaction between baking soda and white vinegar. During this process, the hands-on operation is the "movement" part, while observing the phenomena and thinking about the reasons is the "stillness" part - see how this combines "movement and stillness"?? You can guide your child to pay attention to these processes, and gradually, they'll better master this learning method! Plus, during this process, your parent-child relationship will naturally grow stronger!</p> <p>SLM-Relaxed</p> <p>"A Parent's Guide to Supporting Children's Learning and Growth" During a child's learning and growth process, there is an important learning method called the movement-stillness transition method. It's like a magical key that can help children learn more easily and master knowledge by enhancing their learning effectiveness. So, how do we guide children to master this method?</p> <p>For parents, although their professional knowledge may be limited, they can help children understand through various scenarios in daily life. For example, during a family trip, when children encounter different scenery and activities, some children might initially run and play excitedly - this is the "movement" state. Then when they see beautiful scenery or encounter interesting things, they stop to carefully observe and think - this enters the "stillness" state. Parents should be good at recognizing these transitions between states in their children, guiding them at appropriate times to help them understand that this process of "movement-stillness" transition is actually the learning process. Through such parent-child interactions, parents can not only better understand what their children are learning at school but also improve their parent-child relationship, accompanying their children in healthy and happy growth.</p>

1188
1189

E.2 PROMPTS

1190
1191
1192

Although the term “rewrite” is used in some prompt templates as the editing instruction, it serves the same function as “reformulate” discussed in sections above, which aims to maintain the core meaning of the documents while only optimizing its expression.

1193

1194
1195
1196

strict version

You are a text polishing expert. You will polish text based on the given [Genre] and [Audience].

1197
1198
1199

When polishing, you must follow these 4 rules:

1. Read through the entire text and polish it according to the requirements of the given [Genre] and [Audience]
2. The degree of polishing should not be too heavy – just aim to satisfy the requirements of [Genre] and [Audience] as much as possible
3. Double-check that the polished text is suitable for the audience described in [Audience]!
4. Pay attention to the frequency of modal particles – the text should not contain too many modal particles

1200
1201
1202
1203
1204
1205
1206
1207

relaxed version

You are a creative expert skilled at transforming materials into creative inspiration and building independent, complete, and highly original texts.

Requirements:

1. Read through the original text thoroughly, extract several key themes/keywords, transform to abstract or universal concept inspiration, then generate entirely new text constructions.
2. Extract content from [Audience] and [Genre] sections, but don’t be constrained by them directly, just use them as creative inspiration.
3. Create and reformulated text around points 1/2, and build new meaning from details to the whole structure.

1208
1209

Figure 15: two different prompt templates, we keep the input aligned with MGA strategy, using raw text, genre, audience to fill the template.

1210

1211
1212
1213
1214
1215
1216
1217#####

#Identity and Capabilities#

You are a content creation expert, specializing in text analysis and rewriting, capable of adapting content based on varying “genres” and “audiences” to produce “diverse” and “high-quality” texts. Your English writing is at native editor level, and you will output your rewritten texts in English. International audiences particularly enjoy your work, which receives widespread readership and circulation, earning unanimous acclaim from the industry for your capabilities!

#####

#Workflow#

Please utilize your analytical and writing abilities to rewrite the text based on the original content and given “genre” and “audience”. Before beginning the rewrite, you will consider the following requirements:

1. First, read through the original text thoroughly, identify its information content and value, and consider how to prevent any loss of information points and value in the rewritten text
2. Focus on the original content, combine it with the given “genre” requirements, and rewrite the text following the descriptions, content modules, language requirements, and other stylistic elements specified in the “genre”, to form an initial draft
3. Polish the initial draft according to the given “audience” requirements, and generate the final rewritten text in English
4. Refine the rewritten text to match native English speakers’ reading habits and expression patterns

#####

#Detailed Requirements#

Please ensure you follow the three workflow requirements above, then generate the final English rewritten text according to these detailed requirements.

The given “audience” is <<<{audience}>>>. The given “genre” is <<<{genre}>>>.

#####

#Raw Text#

{raw_text}

1239
1240
1241

Prompt 1: reformulation prompt template.

23

```

1242
1243
1244
1245
1246 ######
1247 #Identity and Capabilities#
1248 You are a content creation expert, specializing in text analysis and rewriting, skilled at adapting content based
1249 on varying [genres] and [audiences] to produce “diverse” and “high-quality” texts. Your rewriting
1250 approaches consistently transform original texts into remarkable content, earning acclaim from both readers
1251 and industry professionals!
1252 #####
1253 #Workflow#
1254 Please utilize your imagination and creativity to generate 5 pairs of [genre] and [audience] combinations
1255 suitable for the original text. Your analysis should follow these requirements:
1256
1. First, analyze the characteristics of the source text, including writing style, information content, and value
1257 2. Then, consider how to preserve the primary content and information while exploring possibilities for “
1258 broader audience engagement” and “alternative genres”
1259 #####
1260 #Detailed Requirements#
1261 Ensure adherence to the workflow requirements above, then generate 5 pairs of [genre] and [audience]
1262 combinations according to these specifications:
1263
1264 Your provided [genres] should meet the following requirements:
1265 1. Clear Genre Definition: Demonstrate strong diversity; include genres you’ve encountered, read, or can
1266 envision
1267 2. Detailed Genre Description: Provide 2–3 sentences describing each genre, considering but not limited to
1268 type, style, emotional tone, form, conflict, rhythm, and atmosphere. Emphasize diversity to guide knowledge
1269 adaptation for specific audiences, facilitating comprehension across different backgrounds. Note: Exclude
1270 visual formats (picture books, comics, videos); use text–only genres.
1271
1272 Your provided [audiences] should meet the following requirements:
1273 1. Clear Audience Definition: Demonstrate strong diversity; include both interested and uninterested parties,
1274 those who like and dislike the content, overcoming bias toward positive audiences only
1275 2. Detailed Audience Description: Provide 2 sentences describing each audience, including but not limited to
1276 age, occupation, gender, personality, appearance, educational background, life stage, motivations and goals,
1277 interests, and cognitive level
1278 #####
1279 #Response#
1280 {
1281     “audience_1”: audience1,
1282     “genre_1”: genre1,
1283     “audience_2”: audience2,
1284     “genre_2”: genre2,
1285     “audience_3”: audience3,
1286     “genre_3”: genre3,
1287     “audience_4”: audience4,
1288     “genre_4”: genre4,
1289     “audience_5”: audience5,
1290     “genre_5”: genre5
1291 }
1292 #####
1293 #Input#
1294 {raw_text}
1295

```

Prompt 2: genre-audience pairs prompt template.

```

1296
1297 #####
1298 #Identity and Capabilities#
1299 You are a Content Reviewer, skilled at analyzing texts and keenly identifying and analyzing the relationships,
1300 similarities, and differences between two texts. Your thorough analysis of each pair of texts, with attention to
1301 every detail, provides great convenience for subsequent review work!
1302 #####
1303 #Thinking Process#
1304 Please fully utilize your analytical abilities, review capabilities, and deep thinking skills to analyze the “
1305 Rewritten Text” against the “Original Text” as a benchmark, ultimately providing analysis and scoring for [A].
1306 You will follow these steps for detailed consideration:
1307 1. First, you will read through the original text thoroughly, identifying the information points in the “Original
1308 Text”
1309 2. You will also read through the rewritten text thoroughly, identifying the information points in the “
1310 Rewritten Text”
1311 3. Compare the information in both texts’ content. The “Rewritten Text” is allowed to have new information
1312 points, different writing styles, expression styles, order, and focus from the “Original Text”. As long as it is
1313 created based on some information points from the “Original Text”, it is considered good for [A]
1314 4. After careful analysis and review, please clearly list the connections and differences between the two texts,
1315 and based on this, provide final analysis and scoring for [A]
1316 #####
1317 #Detailed Requirements#
1318 The scoring judgment for [A] must follow these standards:
1319 1. The “scoring range” is 1–5 points. You need to analyze and grasp each aspect mentioned in #Thinking
1320 Process#, and differentiate scores accordingly. Be strict, don’t be too lenient with scoring!
1321 2. The “Rewritten Text” is allowed to differ from the “Original Text” in writing style, expression style, and
1322 focus! This cannot be a basis for deducting points!
1323 3. The “Rewritten Text” is allowed to omit some information from the “Original Text”! It is not required
1324 that all information from the “Original Text” appears in the “Rewritten Text”! This also cannot be a basis
1325 for deducting points! If this is the only issue, please give a full score of 5 points.
1326 In scoring [A], the following situations will **NOT reduce** the score for [A]:
1327 1. The “Rewritten Text” can include information points not present in the “Original Text”
1328 2. The added content in the “Rewritten Text” significantly deviates from the core information of the “
1329 Original Text”
1330 3. The expression style, order, and focus of the “Rewritten Text” differ from the “Original Text”
1331 In scoring [A], the following situations **WILL reduce** the score for [A]:
1332 1. The information points in the “Rewritten Text” differ so greatly from the “Original Text” that it’s not
1333 recognizable as being rewritten from the “Original Text”
1334 2. The “Rewritten Text” contains none of the information points from the “Original Text”
1335 #####
1336 #Original Text#
1337 {raw_text}
1338 #####
1339 #Rewritten Text#
1340 {rewritten_text}
1341 #####
1342 #Response Format#
1343 {
1344     “A”:[
1345         “analysis”: “xxx”, provide reasons for point deductions
1346         “score”: 1, 2, 3, 4, or 5
1347     ],
1348 }
1349 #####

```

Prompt 3: Full LLM judge prompt.

1350 **F USE OF LARGE LANGUAGE MODELS**
13511352 During the preparation of this manuscript, we utilized a large language model (LLM) as a writing
1353 assistant. The LLM’s role was primarily focused on improving the clarity, precision, and readability
1354 of the text. This included tasks such as correcting grammar and spelling, refining sentence structure
1355 for better flow, and suggesting alternative phrasing to enhance the academic tone.1356 The core scientific contributions—including the initial research ideation, the design of the MGA
1357 framework, the experimental methodology, and the interpretation of results—were conceived and
1358 executed entirely by the human authors. The LLM did not contribute to the research ideas or the
1359 analysis presented. The authors have meticulously reviewed, edited, and validated all LLM-assisted
1360 text to ensure its scientific accuracy and take full responsibility for the final content of this work.

1361

1362

1363

1364

1365

1366

1367

1368

1369

1370

1371

1372

1373

1374

1375

1376

1377

1378

1379

1380

1381

1382

1383

1384

1385

1386

1387

1388

1389

1390

1391

1392

1393

1394

1395

1396

1397

1398

1399

1400

1401

1402

1403