Novelty Assessment Report

Paper: Reformulation for Pretraining Data Augmentation

PDF URL: https://openreview.net/pdf?id=dIOYpj9K8P

Venue: ICLR 2026 Conference Submission

Year: 2026

Report Generated: 2025-12-29

Abstract

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

Disclaimer

This report is **AI-GENERATED** using Large Language Models and WisPaper (a scholar search engine). It analyzes academic papers' tasks and contributions against retrieved prior work. While this system identifies **POTENTIAL** overlaps and novel directions, **ITS COVERAGE IS NOT EXHAUSTIVE AND JUDGMENTS ARE APPROXIMATE**. These results are intended to assist human reviewers and **SHOULD NOT** be relied upon as a definitive verdict on novelty.

Note that some papers exist in multiple, slightly different versions (e.g., with different titles or URLs). The system may retrieve several versions of the same underlying work. The current automated pipeline does not reliably align or distinguish these cases, so human reviewers will need to disambiguate them manually.

If you have any questions, please contact: mingzhang23@m.fudan.edu.cn

Core Task Landscape

This paper addresses: **Pretraining Data Augmentation through Text Reformulation**

A total of **47 papers** were analyzed and organized into a taxonomy with **35 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Systematic Reformulation Frameworks for Pretraining**
- **Multimodal Reformulation**
- **Application-Driven Reformulation for Downstream Tasks**
- **Paraphrasing Techniques and Foundations**
- **Meta-Learning and Optimization for Augmentation**
- **Specialized Pretraining Contexts**

Complete Taxonomy Tree

- Pretraining Data Augmentation through Text Reformulation Survey Taxonomy
- Systematic Reformulation Frameworks for Pretraining
  - Genre-Audience and Multi-Document Reformulation ★ (2 papers)
  - [0] Reformulation for Pretraining Data Augmentation (Anon et al., 2026) View paper
  - [5] Pre-training via paraphrasing (Mike Lewis, 2020) View paper
  - Synthetic Continued Pretraining (1 papers)
  - [16] Synthetic continued pretraining (Yang Zi-tong, 2024) View paper
  - Web Data Recycling and Quality Enhancement (1 papers)
  - [4] Recycling the Web: A Method to Enhance Pre-training Data Quality and Quantity for Language Models (Nguyen Thao, 2025) View paper
- Multimodal Reformulation
  - Vision-Language Contrastive Learning Augmentation (2 papers)
  - [1] Improving clip training with language rewrites (Fan Lijie, 2023) View paper
  - [18] Fine-tuning CLIP text encoders with two-step paraphrasing (Kim HyunJae, 2024) View paper
  - Multimodal Data Synthesis and Cleaning (1 papers)
  - [7] SyntheClean: Enhancing Large-Scale Multimodal Models via Adaptive Data Synthesis and Cleaning (Y Chen, 2025) View paper
  - Audio Captioning Augmentation (1 papers)
  - [13] SLAM-AAC: Enhancing Audio Captioning with Paraphrasing Augmentation and CLAP-Refine through LLMs (Wenxi Chen, 2024) View paper
- Application-Driven Reformulation for Downstream Tasks
  - Classification and Retrieval Task Augmentation
  - General Text Classification Augmentation (5 papers)
    - [3] Large Language Model Data Augmentation for Text-Pair Classification Tasks (Yuyang LI, 2024) View paper
    - [15] Improving text classification with large language model-based data augmentation (Huanhuan Zhao, 2024) View paper
    - [27] Geometry of Textual Data Augmentation: Insights from Large Language Models (Sherry J.H Feng, 2024) View paper
    - [33] Enhancing SLM via ChatGPT and Dataset Augmentation (Ballout, 2024) View paper
    - [36] Text Generation for Dataset Augmentation in Security Classification Tasks (Alexander Paul Welsh, 2023) View paper

  - Abusive Language Detection (1 papers)
    - [11] Don't augment, rewrite? assessing abusive language detection with synthetic data (Camilla Casula, 2024) View paper
  - Embedding Enhancement through Text Enrichment (2 papers)
    - [10] QZhou-Embedding Technical Report (Yu Peng, 2025) View paper
    - [12] Enhancing embedding performance through large language model-based text enrichment and rewriting (Harris, 2024) View paper
  - Retrieval Task Augmentation (1 papers)
    - [8] Data augmentation for text-based person retrieval using large language models (Li Zheng, 2024) View paper
  - Sequence-to-Sequence Augmentation
  - Sentence Rewriting and Simplification (4 papers)
    - [22] Sequence-to-sequence pre-training with data augmentation for sentence rewriting (Zhang Yi, 2019) View paper
    - [28] Beyond Repetition: Text Simplification and Curriculum Learning for Data-Constrained Pretraining (Velasco, 2025) View paper
    - [37] RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting (Shu, 2023) View paper
    - [45] Large-scale Hierarchical Alignment for Data-driven Text Rewriting (Nikolov, 2018) View paper
  - Summarization (2 papers)
    - [20] Post-training dialogue summarization using pseudo-paraphrasing (Jia Qi, 2022) View paper
    - [46] Dataset construction method of cross-lingual summarization based on filtering and text augmentation. (Hangyu Pan, n.d.) View paper
  - Style Transfer (1 papers)
    - [21] Exploring Large Language Models for Data Augmentation: A Case Study for Text Style Transfer (Martina Toshevska, 2025) View paper
  - Positive Text Rewriting (1 papers)
    - [25] Socratic Reasoning Improves Positive Text Rewriting (Goel, 2024) View paper
  - Multi-Intent Text Revision (1 papers)
    - [47] Editcot: A Novel Multi-Intent Text Revision Modeling Approach Based on Stepwise Reasoning (Xu Li, n.d.) View paper
  - Question-Answering Augmentation
  - In-Context Learning Augmentation (1 papers)
    - [2] Dail: Data augmentation for in-context learning via self-paraphrase (Li Dawei, 2023) View paper
  - Question Rewriting (1 papers)
    - [19] Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space (Dayiheng Liu, 2020) View paper
  - QA-Based Generative Augmentation (1 papers)
    - [14] Leveraging QA datasets to improve generative data augmentation (Dheeraj Mekala, 2022) View paper
  - Utterance Rewriting for Dialogue (1 papers)
    - [39] Improving Utterance Rewriter Based on MMI and Text Data Augmentation (Lina Yang, 2022) View paper
  - Domain-Specific Augmentation
  - Clinical Text Rephrasing (1 papers)
    - [30] Rephrasing Electronic Health Records for Pretraining Clinical Language Models (Liu Jinghui, 2024) View paper
  - Business Text Generation (1 papers)
    - [38] Transforming the Generative Pretrained Transformer Into Augmented Business Text Writer (Faisal Khalil, 2021) View paper
  - Speech-to-Phoneme Augmentation (1 papers)
    - [34] AraS2P: Arabic Speech-to-Phonemes System (Matar, 2025) View paper
- Paraphrasing Techniques and Foundations
  - Paraphrase Generation Models
  - Generative Pretraining for Paraphrasing (1 papers)
    - [24] Generative pre-training for paraphrase generation by representing and predicting spans in exemplars (Tien-Cuong Bui, 2021) View paper
  - Unsupervised Paraphrase Generation (1 papers)
    - [44] Unsupervised Paraphrasing with Pretrained Language Models (Tong Niu, 2021) View paper
  - Image Captioning with Paraphrasing (1 papers)
    - [9] Text-to-text pre-training with paraphrasing for improving transformer-based image captioning (Ryo Masumura, 2023) View paper
  - Paraphrase Evaluation (2 papers)
  - [40] Generative Pretraining for Paraphrase Evaluation (Jack Weston, 2022) View paper
  - [43] Long: Generative Pretraining for Paraphrase Evaluation (Fristed, 2022) View paper
  - Paraphrase Identification and Extraction (2 papers)
  - [31] Neural Paraphrase Identification of Questions with Noisy Pretraining (Gaurav Singh Tomar, 2022) View paper
  - [41] Extracting and filtering paraphrases by bridging natural language inference and paraphrasing (Klemen, 2021) View paper
  - Specialized Paraphrasing Applications
  - Emotion and Sentiment Guided Paraphrasing (1 papers)
    - [17] Emotion and sentiment guided paraphrasing (Xie, 2023) View paper
  - Negation and Contrastive Paraphrasing (1 papers)
    - [26] Contrastive vision-language learning with paraphrasing and negation (Kwun Ho Ngan, 2025) View paper
  - Dense Paraphrasing for Semantic Enrichment (1 papers)
    - [35] Dense Paraphrasing for Textual Enrichment (Tu, 2022) View paper
  - Paraphrasing for AI-Text Detection Evasion (1 papers)
    - [29] TempParaphraser:â□□Heating Upâ□□ Text to Evade AI-Text Detection through Paraphrasing (Junjie Huang, 2025) View paper
  - Paraphrase Generation for Conversational Services (1 papers)
    - [32] Paraphrase generation for automated training of conversational services (Berro, 2024) View paper
- Meta-Learning and Optimization for Augmentation (1 papers)
  - [6] Learning data manipulation for augmentation and weighting (Zhiting Hu, 2019) View paper

- Specialized Pretraining Contexts
  - Text-Attributed Heterogeneous Graph Pretraining (1 papers)
  - [23] Pretraining Language Models with Text-Attributed Heterogeneous Graphs (Zou Tao, 2023) View paper
  - Small Language Model Enhancement (1 papers)
  - [42] Smollm2: When smol goes bigâ□□data-centric training of a fully open small language model (A Lozhkov, n.d.) View paper

Narrative

Core task: Pretraining data augmentation through text reformulation. The field explores how systematically rewriting or paraphrasing text can enrich pretraining corpora and improve downstream model performance. The taxonomy reveals several main branches: Systematic Reformulation Frameworks for Pretraining develop principled methods for generating diverse textual variants at scale, often targeting genre shifts or multi-document synthesis; Multimodal Reformulation extends these ideas to vision-language settings, where caption paraphrasing or image-text alignment benefits from textual diversity; Application-Driven Reformulation tailors augmentation strategies to specific downstream tasks such as question answering or sentiment analysis; Paraphrasing Techniques and Foundations provide the algorithmic and evaluation underpinnings, including neural paraphrase generation and identification; Meta-Learning and Optimization for Augmentation investigate how to learn augmentation policies themselves; and Specialized Pretraining Contexts address domain-specific needs like clinical text or low-resource languages. Representative works such as Pretraining via Paraphrasing[5] and Synthetic Continued Pretraining[16] illustrate how reformulation can be integrated into the pretraining pipeline, while CLIP Language Rewrites[1] and Paraphrasing Image Captioning[9] show multimodal applications.

A particularly active line of work focuses on leveraging large language models to generate high-quality paraphrases or style-shifted variants for pretraining, as seen in LLM Text-Pair Augmentation[3] and LLM Style Transfer[21], which trade off generation cost against data diversity. Another contrasting direction emphasizes lightweight, rule-based or model-driven paraphrasing that can scale to massive corpora without expensive LLM calls, exemplified by Recycling Web Pretraining[4] and DAIL Self-Paraphrase[2]. The original paper, Reformulation Pretraining Augmentation[0], sits within the Systematic Reformulation Frameworks branch, specifically targeting genre-audience and multi-document reformulation. Its emphasis on structured, multi-document synthesis aligns it closely with Pretraining via Paraphrasing[5], which also explores paraphrase-driven pretraining, but Reformulation Pretraining Augmentation[0] appears to push further into controlled genre and audience adaptation. This positions it as a bridge between foundational paraphrasing techniques and the emerging trend of using LLMs for targeted, high-fidelity text transformation in pretraining contexts.

Related Works in Same Category

The following **1 sibling papers** share the same taxonomy leaf node with the original paper:

1. Pre-training via paraphrasing

Authors: Mike Lewis, Marjan Ghazvininejad, M. Lewis, GARGI GHOSH, Armen Aghajanyan, et al. (7 authors total) | **Year/Venue**: 2020 | **URL**: View paper

Abstract

We introduce MARGE, a pre-trained sequence-to-sequence model learned with an unsupervised multi-lingual multi-document paraphrasing objective. MARGE provides an alternative to the dominant masked language modeling paradigm, where we self-supervise the reconstruction of target text by retrieving a set of related texts (in many languages) and conditioning on them to maximize the likelihood of generating the original. We show it is possible to jointly learn to do retrieval and reconstruction, given...

Relationship Analysis

Both papers belong to the Genre-Audience and Multi-Document Reformulation category, focusing on systematic reformulation frameworks for pretraining data augmentation. The original paper (MGA) generates diverse reformulations by adaptively creating genre-audience pairs from source documents and uses a two-stage synthesis pipeline with lightweight models, while the candidate paper (MARGE) employs multi-document paraphrasing through a retrieval-reconstruction approach that retrieves related documents in multiple languages and conditions on them to reconstruct the target. The key difference is that MGA emphasizes controlled diversity through genre-audience directives within single documents, whereas MARGE leverages cross-document and cross-lingual information through retrieval-based paraphrasing.

Contributions Analysis

Overall novelty summary. The paper introduces the Massive Genre-Audience (MGA) reformulation framework to augment pretraining corpora by generating diverse genre-audience variations, producing a 770 billion token MGACorpus. Within the taxonomy, it resides in the 'Genre-Audience and Multi-Document Reformulation' leaf under 'Systematic Reformulation Frameworks for Pretraining'. This leaf contains only two papers total, including the original work, indicating a relatively sparse research direction. The sibling paper focuses on multi-document paraphrasing, suggesting this area is still emerging rather than crowded.

The taxonomy reveals that neighboring leaves address related but distinct approaches: 'Synthetic Continued Pretraining' synthesizes domain-specific corpora from small documents, while 'Web Data Recycling and Quality Enhancement' focuses on filtering and rewriting web-crawled data. The broader 'Systematic Reformulation Frameworks' branch contrasts with 'Application-Driven Reformulation', which targets downstream tasks rather than pretraining. The MGA framework's emphasis on adaptive genre-audience generation distinguishes it from fixed paraphrasing schemes in the 'Paraphrasing Techniques' branch, positioning it as a structured, pretraining-focused methodology.

Among 24 candidates examined across three contributions, the MGA reformulation framework (Contribution A) shows one refutable candidate out of seven examined, suggesting some prior work overlap in the limited search scope. The MGACorpus dataset (Contribution B) and Limited Consistency principle (Contribution C) examined 10 and 7 candidates respectively, with zero refutable matches, indicating these contributions appear more novel within the examined literature. The statistics reflect a focused semantic search rather than exhaustive coverage, so unexamined work may exist beyond the top-K matches.

Based on the limited search scope of 24 candidates, the framework's core novelty appears moderate given one overlapping prior work, while the dataset and synthesis principle show stronger novelty signals. The sparse taxonomy leaf (two papers) and absence of extensive prior work in genre-audience reformulation suggest this direction is relatively unexplored. However, the analysis covers top semantic matches and immediate citations, not the full field, so definitive novelty claims require broader literature review.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: MGA reformulation framework for corpus augmentation

Description: The authors propose a systematic two-stage framework that reformulates existing text corpora by adaptively generating diverse genre-audience pairs, avoiding complex seed systems and using lightweight models. This framework addresses data scarcity and repetition issues in LLM pretraining by creating contextually-rich variations of source documents.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Sequence-to-sequence pre-training with data augmentation for sentence rewriting

URL: View paper

Brief Assessment

Sentence Rewriting Pretraining[22] focuses on sentence-level rewriting tasks (GEC, FST) using back-translation and feature discrimination for data augmentation, not on large-scale corpus reformulation via genre-audience pairs for LLM pretraining. The candidate addresses different augmentation goals and application contexts than the original's systematic corpus expansion framework.

2. Exploring Large Language Models for Data Augmentation: A Case Study for Text Style Transfer

URL: View paper

Brief Assessment

LLM Style Transfer[21] focuses on text style transfer tasks (formality, sentiment, personal style) using LLMs to generate parallel datasets for specific style adaptation. The ORIGINAL paper's MGA framework targets general corpus augmentation for LLM pretraining through genre-audience reformulation to address data scarcity and repetition, which is a fundamentally different application domain and methodology.

3. Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers

URL: View paper

Brief Assessment

Pearl[56] focuses on personalizing LLM writing assistants through retrieval-augmented generation using user-authored historical documents, not on corpus augmentation for pretraining through genre-audience reformulation.

4. TextSETTR: Few-shot text style extraction and tunable targeted restyling

URL: View paper

Brief Assessment

TextSETTR[57] focuses on few-shot text style transfer using adjacent sentences to extract style vectors for targeted restyling. The original paper proposes a corpus augmentation framework using genre-audience reformulation for LLM pretraining. These are fundamentally different tasks: TextSETTR[57] performs style transfer on individual sentences, while the original work augments entire pretraining corpora through systematic reformulation.

5. Text Style Transfer with Neural Language Models

URL: View paper

Brief Assessment

Neural Style Transfer[58] focuses on text style transfer (modifying stylistic attributes while preserving content) rather than corpus augmentation for pretraining. The candidate addresses different NLP tasks (sentiment transfer, politeness transfer, detoxification) rather than systematic corpus expansion for LLM pretraining.

6. Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space

URL: View paper

Brief Assessment

Question Controllable Rewriting[19] focuses on question data augmentation for machine reading comprehension tasks through controllable rewriting in continuous space, not on general corpus augmentation frameworks for LLM pretraining using genre-audience reformulation.

7. Rephrasing the web: A recipe for compute and data-efficient language modeling

URL: View paper

Prior Art Analysis

Rephrasing the Web[55] demonstrates that a systematic framework for reformulating existing text corpora using genre-audience adaptations was already proposed before the original paper. The candidate paper presents WRAP (Web Rephrase Augmented Pre-training), which uses an off-the-shelf instruction-tuned model to paraphrase documents on the web in specific styles. Like the original paper's MGA framework, WRAP reformulates existing corpora into diverse variations without requiring complex seed systems, uses lightweight models for generation, and addresses data scarcity issues in LLM pretraining. Both frameworks share the core principle of systematically transforming source documents into contextually-rich variations through style-based reformulation.

Evidence

Evidence 1 - **Rationale**: Both papers propose systematic frameworks for reformulating existing corpora. The candidate's WRAP framework uses style-based paraphrasing (like Wikipedia, Q/A format) while the original uses genre-audience pairs, but both achieve the same goal of creating diverse variations from source documents without complex seed systems. - **Original**: we introduce the massive genre-audience (mga) reformulation method, a framework designed to augment corpora in a way that supports more effective model performance scaling. instead of relying on complex, predefined seed systems, mga systematically reformulates existing corpora into diverse, contextu... - **Candidate**: we propose web rephrase augmented pre-training (wrap) that uses an off-the-shelf instruction-tuned model prompted to paraphrase documents on the web in specific styles such as "like wikipedia" or in "question-answer format" to jointly pre-train llms on real and synthetic rephrases.

Evidence 2 - **Rationale**: Both frameworks emphasize using lightweight models to avoid computational bottlenecks. The original uses a 3.3B MoE model while the candidate uses smaller LLMs (1.8B/7B), and both avoid complex seed systems by working directly with raw documents. - **Original**: mga (massive genre-audience reformulation), a transparent and principled framework designed to directly address the data repetition challenge by augmenting the raw text and creating more unique tokens. as illustrated in figure 1, the mga framework is efficiently implemented using a lightweight 3.3b ... - **Candidate**: in this work, we propose wrap that leverages the natural diversity of articles on the web, allowing us to utilize significantly smaller llms (than gpt-3.5) to generate high-quality paraphrases of noisy and unstructured articles on the web.

Evidence 3 - **Rationale**: Both frameworks implement systematic approaches to generating diverse reformulations. The original uses a two-stage pipeline with genre-audience generation followed by reformulation, while the candidate defines specific rephrasing styles and uses prompting to achieve reformulation. Both create diversity through controlled variation mechanisms. - **Original**: the mga framework is operationalized as a two-stage synthesis pipeline: a variance-maximizing stage for generating diverse directives, followed by an invariance-enforcing stage for controlled reformulation. each stage is powered by a specialized tool slm, which is finetuned on task-specific data gen... - **Candidate**: rephrasing styles in lieu of the anecdotal evidence above, we attempt rephrasing documents on the web in four different styles-(i) easy (text that even a toddler will understand); (ii) medium (in high quality english such as that found on wikipedia); (iii) hard (in terse and abstruse language); (iv)...

Contribution 2: MGACorpus dataset

Description: The authors release a 770 billion token dataset generated by applying their MGA framework to reformulate the fineweb-edu-dedup corpus, achieving a 3.9x token expansion. This dataset serves as a concrete validation of their methodology and will be made publicly available.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Synthetic continued pretraining

URL: View paper

Brief Assessment

Synthetic Continued Pretraining[16] focuses on synthesizing data from small domain-specific corpora using entity-based augmentation (EntiGraph), not on reformulating large-scale web corpora through genre-audience pairs like MGACorpus.

2. Synthesize-on-Graph: Knowledgeable Synthetic Data Generation for Continue Pre-training of Large Language Models

URL: View paper

Brief Assessment

Synthesize-on-Graph[61] focuses on cross-document knowledge graph construction for synthetic data generation, while the original paper's MGACorpus uses genre-audience reformulation of individual documents. These represent fundamentally different synthetic data generation methodologies with distinct technical approaches and objectives.

3. Scaling laws of synthetic data for language models

URL: View paper

Brief Assessment

Synthetic Data Scaling Laws[59] focuses on generating synthetic data through concept extraction and recombination from web documents for mathematical reasoning, while MGACorpus applies genre-audience reformulation to expand the fineweb-edu-dedup corpus. These represent distinct methodologies for synthetic data generation with different underlying principles and target applications.

4. Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls

URL: View paper

Brief Assessment

Synthetic Data Pretraining[63] focuses on rephrased web text and textbook-style synthetic data for pre-training, not on the MGA framework's genre-audience reformulation methodology or the specific MGACorpus dataset created through this approach.

5. Beyondweb: Lessons from scaling synthetic data for trillion-scale pretraining

URL: View paper

Brief Assessment

BeyondWeb[60] focuses on a different synthetic data generation framework with distinct methodologies (diverse rephrasing strategies, style matching, quality-based selection) rather than the genre-audience reformulation approach of MGACorpus. The datasets serve similar purposes but employ fundamentally different generation principles.

6. Recycling the Web: A Method to Enhance Pre-training Data Quality and Quantity for Language Models

URL: View paper

Brief Assessment

Recycling Web Pretraining[4] focuses on rewriting low-quality web documents to create synthetic data for pre-training, not on genre-audience reformulation. The dataset creation methodology and scale differ fundamentally from MGACorpus.

7. Rephrasing the web: A recipe for compute and data-efficient language modeling

URL: View paper

Brief Assessment

Rephrasing the Web[55] does not release a comparable large-scale reformulated dataset. While the candidate paper describes generating synthetic rephrases of C4, it does not present a 770B token dataset as a concrete contribution or claim to release it publicly.

8. Building Open-Retrieval Conversational Question Answering Systems by Generating Synthetic Data and Decontextualizing User Questions

URL: View paper

Brief Assessment

Conversational QA Synthetic[64] focuses on generating synthetic conversational question-answering dialogs from documents, not on large-scale pretraining corpus reformulation. The candidate addresses a different task (conversational QA) with different methodology (dialog generation from documents) compared to the original's pretraining data augmentation framework.

9. Leveraging large language models for abstractive summarization of Italian legal news

URL: View paper

Brief Assessment

Italian Legal Summarization[62] focuses on abstractive summarization of Italian legal news, not on large-scale synthetic text dataset generation for pretraining language models through document reformulation.

10. T-CLAP: Temporal-Enhanced Contrastive Language-Audio Pretraining

URL: View paper

Brief Assessment

T-CLAP[65] focuses on temporal-enhanced contrastive language-audio pretraining using synthetic temporal-contrastive captions for audio clips. This is fundamentally different from MGACorpus, which is a 770 billion token text dataset for language model pretraining created through document reformulation.

Contribution 3: Limited Consistency principle for data synthesis

Description: The authors introduce a design principle that balances generating diverse stylistic variations (variance) while preserving factual accuracy (invariance) in reformulated text. This principle is implemented through careful prompt engineering and guides the entire reformulation process to avoid both excessive repetition and factual degradation.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

1. Towards Controllable and Explainable Text Generation via Causal Intervention in LLMs

URL: View paper

Brief Assessment

Causal Intervention LLMs[51] focuses on controllable text generation through causal interventions on hidden representations, not on balancing stylistic variance and factual invariance in synthetic data generation for pretraining corpora.

2. Dr Genre: Reinforcement Learning from Decoupled LLM Feedback for Generic Text Rewriting

URL: View paper

Brief Assessment

Dr Genre[52] focuses on text rewriting with decoupled reward modeling for specific objectives (agreement, coherence, conciseness), not on balancing stylistic variance and factual invariance in synthetic data generation for pretraining corpora.

3. An invariant learning characterization of controlled text generation

URL: View paper

Brief Assessment

Invariant Controlled Generation[49] addresses controlled text generation under distribution shift using invariant learning, focusing on maintaining attribute relationships across environments. This differs from the original paper's Limited Consistency principle, which balances stylistic variance with factual invariance in data reformulation for pretraining augmentation.

4. From Style to Facts: Mapping the Boundaries of Knowledge Injection with Finetuning

URL: View paper

Brief Assessment

Style Facts Knowledge[48] focuses on finetuning for knowledge injection versus task customization, examining factors like training data format and information type. It does not address the specific challenge of balancing stylistic variance with factual invariance during synthetic data generation through reformulation.

5. RewriteNet: Reliable Scene Text Editing with Implicit Decomposition of Text Contents and Styles

URL: View paper

Brief Assessment

RewriteNet[54] addresses scene text editing in images, focusing on decomposing visual text content and style features for image synthesis. This is fundamentally different from the original paper's text data reformulation for language model pretraining, which balances stylistic variance and factual invariance in synthetic text generation.

6. Technical Analysis of GPT, Grok, and Gemini Cognitive, Ethical, and Emotional Dimensions What Remains in the Loop Observing the Echoes of Three AIs

URL: View paper

Brief Assessment

GPT Grok Gemini Analysis[50] focuses on analyzing cognitive, ethical, and emotional dimensions of three AI systems. It does not address data synthesis methodologies or principles for balancing stylistic variance and factual invariance in synthetic data generation for language model pretraining.

7. StyleDistance: Stronger Content-Independent Style Embeddings with Synthetic Parallel Examples

URL: View paper

Brief Assessment

StyleDistance[53] focuses on generating synthetic parallel examples for style embeddings by creating near-exact paraphrases with controlled style variations. This is fundamentally different from the original paper's Limited Consistency principle, which balances stylistic variance and factual invariance in reformulated text for pretraining data augmentation. The candidate addresses style representation learning, not general corpus reformulation for LLM pretraining.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Reformulation for Pretraining Data Augmentation View paper
- [1] Improving clip training with language rewrites View paper
- [2] Dail: Data augmentation for in-context learning via self-paraphrase View paper
- [3] Large Language Model Data Augmentation for Text-Pair Classification Tasks View paper
- [4] Recycling the Web: A Method to Enhance Pre-training Data Quality and Quantity for Language Models View paper
- [5] Pre-training via paraphrasing View paper
- [6] Learning data manipulation for augmentation and weighting View paper
- [7] SyntheClean: Enhancing Large-Scale Multimodal Models via Adaptive Data Synthesis and Cleaning View paper
- [8] Data augmentation for text-based person retrieval using large language models View paper
- [9] Text-to-text pre-training with paraphrasing for improving transformer-based image captioning View paper
- [10] QZhou-Embedding Technical Report View paper
- [11] Don't augment, rewrite? assessing abusive language detection with synthetic data View paper
- [12] Enhancing embedding performance through large language model-based text enrichment and rewriting View paper
- [13] SLAM-AAC: Enhancing Audio Captioning with Paraphrasing Augmentation and CLAP-Refine through LLMs View paper
- [14] Leveraging QA datasets to improve generative data augmentation View paper

- [15] Improving text classification with large language model-based data augmentation View paper
- [16] Synthetic continued pretraining View paper
- [17] Emotion and sentiment guided paraphrasing View paper
- [18] Fine-tuning CLIP text encoders with two-step paraphrasing View paper
- [19] Tell me how to ask again: Question data augmentation with controllable rewriting in continuous space View paper
- [20] Post-training dialogue summarization using pseudo-paraphrasing View paper
- [21] Exploring Large Language Models for Data Augmentation: A Case Study for Text Style Transfer View paper
- [22] Sequence-to-sequence pre-training with data augmentation for sentence rewriting View paper
- [23] Pretraining Language Models with Text-Attributed Heterogeneous Graphs View paper
- [24] Generative pre-training for paraphrase generation by representing and predicting spans in exemplars View paper
- [25] Socratic Reasoning Improves Positive Text Rewriting View paper
- [26] Contrastive vision-language learning with paraphrasing and negation View paper
- [27] Geometry of Textual Data Augmentation: Insights from Large Language Models View paper
- [28] Beyond Repetition: Text Simplification and Curriculum Learning for Data-Constrained Pretraining View paper
- [29] TempParaphraser:â□□Heating Upâ□□ Text to Evade AI-Text Detection through Paraphrasing View paper
- [30] Rephrasing Electronic Health Records for Pretraining Clinical Language Models View paper
- [31] Neural Paraphrase Identification of Questions with Noisy Pretraining View paper
- [32] Paraphrase generation for automated training of conversational services View paper
- [33] Enhancing SLM via ChatGPT and Dataset Augmentation View paper
- [34] AraS2P: Arabic Speech-to-Phonemes System View paper
- [35] Dense Paraphrasing for Textual Enrichment View paper
- [36] Text Generation for Dataset Augmentation in Security Classification Tasks View paper
- [37] RewriteLM: An Instruction-Tuned Large Language Model for Text Rewriting View paper
- [38] Transforming the Generative Pretrained Transformer Into Augmented Business Text Writer View paper
- [39] Improving Utterance Rewriter Based on MMI and Text Data Augmentation View paper
- [40] Generative Pretraining for Paraphrase Evaluation View paper
- [41] Extracting and filtering paraphrases by bridging natural language inference and paraphrasing View paper
- [42] Smollm2: When smol goes bigâ□□data-centric training of a fully open small language model View paper
- [43] Long: Generative Pretraining for Paraphrase Evaluation View paper
- [44] Unsupervised Paraphrasing with Pretrained Language Models View paper
- [45] Large-scale Hierarchical Alignment for Data-driven Text Rewriting View paper
- [46] Dataset construction method of cross-lingual summarization based on filtering and text augmentation. View paper
- [47] Editcot: A Novel Multi-Intent Text Revision Modeling Approach Based on Stepwise Reasoning View paper
- [48] From Style to Facts: Mapping the Boundaries of Knowledge Injection with Finetuning View paper
- [49] An invariant learning characterization of controlled text generation View paper
- [50] Technical Analysis of GPT, Grok, and Gemini Cognitive, Ethical, and Emotional Dimensions What Remains in the Loop Observing the Echoes of Three AIs View paper
- [51] Towards Controllable and Explainable Text Generation via Causal Intervention in LLMs View paper
- [52] Dr Genre: Reinforcement Learning from Decoupled LLM Feedback for Generic Text Rewriting View paper
- [53] StyleDistance: Stronger Content-Independent Style Embeddings with Synthetic Parallel Examples View paper
- [54] RewriteNet: Reliable Scene Text Editing with Implicit Decomposition of Text Contents and Styles View paper
- [55] Rephrasing the web: A recipe for compute and data-efficient language modeling View paper
- [56] Pearl: Personalizing large language model writing assistants with generation-calibrated retrievers View paper
- [57] TextSETTR: Few-shot text style extraction and tunable targeted restyling View paper
- [58] Text Style Transfer with Neural Language Models View paper
- [59] Scaling laws of synthetic data for language models View paper
- [60] Beyondweb: Lessons from scaling synthetic data for trillion-scale pretraining View paper
- [61] Synthesize-on-Graph: Knowledgeable Synthetic Data Generation for Continue Pre-training of Large Language Models View paper
- [62] Leveraging large language models for abstractive summarization of Italian legal news View paper
- [63] Demystifying synthetic data in llm pre-training: A systematic study of scaling laws, benefits, and pitfalls View paper
- [64] Building Open-Retrieval Conversational Question Answering Systems by Generating Synthetic Data and Decontextualizing User Questions View paper
- [65] T-CLAP: Temporal-Enhanced Contrastive Language-Audio Pretraining View paper