Rephrasing the Web: A Recipe for Compute and Data-Efficient Language Modeling

Pratyush Maini (CMU), *et al.*

27.06.2025

Presenter: MIN Weiwen

Introduction

Related Work: Pretraining with Synthetic Data

WRAP Setups

Perplexity Evaluation

Zero-shot Tasks

Further Analysis

Limitations

Diversity Challenges

Conclusion

Introduction

• Large LLMs require massive data and compute, but web text is unstructured and noisy

• High-quality text data is scarce, making it hard to scale up

•Goal: Find a more efficient way to train language models with limited clean data

Related Work: Pretraining with Synthetic Data

- **Neural Scaling Laws for Language Models**

- Larger models need more training data (Chinchilla laws: Hoffmann et al., 2022)
- Too little data → underfitting (e.g., Gopher)
- Too much repeated data → overfitting (Muennighoff et al., 2023)
- Repeating even small fractions hurts performance (Xue et al., 2023)

- **Dataset Selection & Filtering**

High-quality data is critical for LLMs

- **Data Augmentation & Synthetic Data**

- Synthetic stories can train small models well (Eldan & Li, 2023)
- High-quality synthetic data → good performance on reasoning & coding (Gunasekar et al., 2023; Liu et al., 2023)
- Too many rounds of self-generated data → performance drop (Shumailov et al., 2023)

Related Work: Pretraining with Synthetic Data

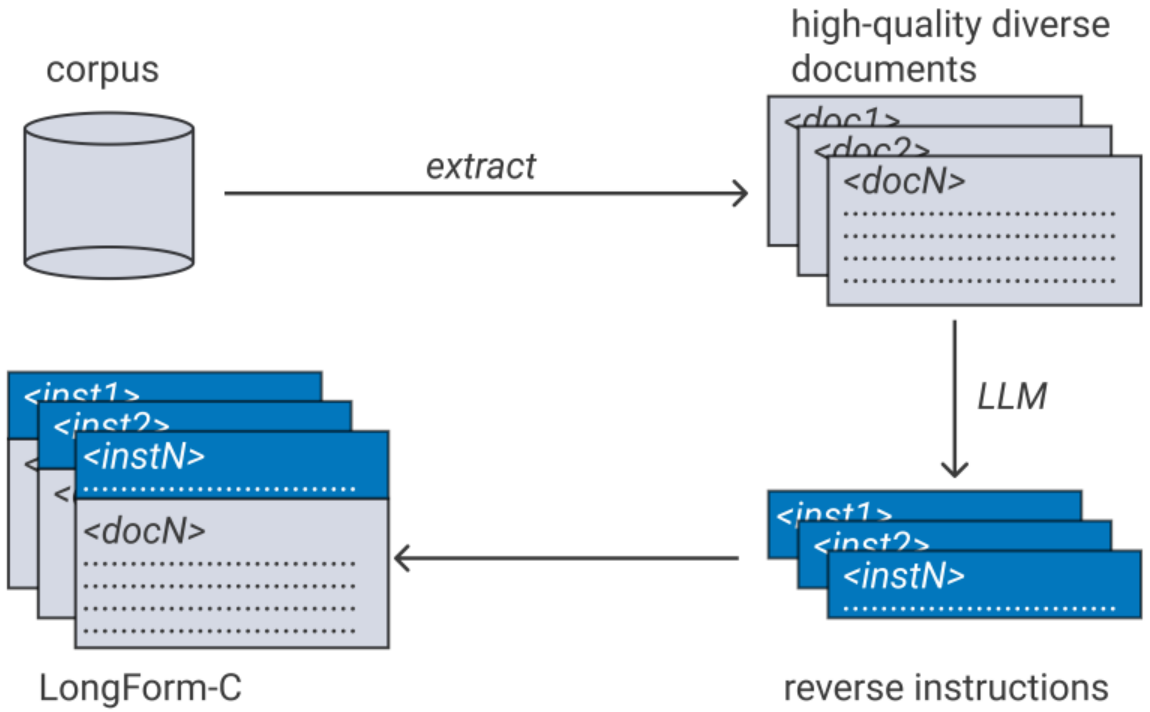
Example: Reverse Instructions

An example of synthetic data generation for self-training. Specifically, we generate data for instruction tuning

How it Works

- Extract human-written documents from corpus
- Use an LLM to generate plausible instructions for each passage
- Result: Pair of *(instruction, output)* for model training

Instruction: Describe your favorite snack food
Output: I love pizza rolls...



Hinrich Schütze & Abdullatif Köksal. "Overcoming the Training Data Bottleneck: Language Models are Effective Autodidacts." Presented on 4 July 2024.

WRAP Setups

Motivation and Overview

WRAP: Web **R**ephrase **A**ugmented **P**re-training

Main Ideas:

• Generating synthetic data with LLMs is expensive and error-prone.

• WRAP addresses this by rephrasing naturally diverse web articles using smaller instruction-tuned LLMs.

• Key challenges addressed:

- High cost of generation.
- Bias from low-diversity prompts.

• WRAP leverages the web's diversity to produce high-quality paraphrases.

WRAP Setups

Rephrasing Styles

- WRAP defines four rephrasing styles:
    - **Easy** – simple language (toddlers can understand).
    - **Medium** – Wikipedia-like.
    - **Hard** – complex, abstract language.
    - **Q/A** – conversational, question-answering format.

WRAP Setups

Rephrasing prompts

Easy

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the questions. USER: For the following paragraph give me a paraphrase of the same using a very small vocabulary and extremely simple sentences that a toddler will understand:

Medium

> A chat between a curious user and an artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the questions. USER: For the following paragraph give me a diverse paraphrase of the same in high quality English language as in sentences on Wikipedia:

WRAP Setups

Rephrasing prompts

Hard

```
A chat between a curious user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the questions.
USER: For the following paragraph give me a paraphrase of the same using very
terse and abstruse language that only an erudite scholar will understand.
Replace simple words and phrases with rare and complex ones:
```

Q/A

```
A chat between a curious user and an artificial intelligence assistant.
The assistant gives helpful, detailed, and polite answers to the questions.
USER: Convert the following paragraph into a conversational format with
multiple tags of "Question:" followed by "Answer:":
```

WRAP Setups

Rephrase Examples

Original

The stock rose $2.11, or about 11 percent, to close Friday at $21.51 on the New York Stock Exchange.
Revenue in the first quarter of the year dropped 15 percent from the same period a year earlier.

Medium Style

The stock experienced an increase of approximately 11 percent, closing at $21.51 on the New York Stock Exchange on Friday, with a rise of $2.11.

During the initial three months of the current year, there was a 15 percent decrease in revenue compared to the corresponding quarter of the previous year.

Q/A Style

Question: What was the stock's closing price on Friday? Answer: $21.51
Question: How much did the stock rise on Friday? Answer: $2.11 or about 11 percent.

Question: What was the revenue drop in the first quarter compared to the same period last year? Answer: The revenue dropped 15 percent.

WRAP Setups

Synthetic Data Generation

- Utilized instruction-tuned LLMs to rephrase web-crawled datasets (e.g., C4).
- Used **frozen Mistral-7B** model (Liang et al., 2023).
- Prompt refined via human feedback by comparing outputs to GPT-4.
- Result: **Parallel corpus** of synthetic high-quality rephrasings.
- Each sample limited to **300 tokens** to prevent information loss.

WRAP Setups

Combining Real and Synthetic Data

1:1 Mixing of Real and Synthetic Data：

- Real web data is noisy; synthetic rephrasing enhances structure.
- WRAP mixes real C4 web corpus with rephrased data in a 1:1 ratio.
- Goal: retain natural diversity while improving quality.
- Models trained on this mixture achieve better generalization.

WRAP Setups

Implementation Details

- WRAP uses decoder-only transformer models based on Vaswani et al. (2017).

- Three model sizes:
  - Small (128M): 12 layers, 12 heads, hidden size 768.
  - Medium (350M): 24 layers, 16 heads, hidden size unspecified.
  - XL (1.3B): 24 layers, 16 heads, hidden size 2048.

- XL models trained with max sequence length 1024.

Perplexity Evaluation

- **WRAP trains on multiple rephrased styles**, aiming for broader generalization.
- **The Pile** covers 21 diverse domains, better matching WRAP’s multi-style design.
- Thus, evaluating on The Pile gives a **more realistic** and **challenging** benchmark.

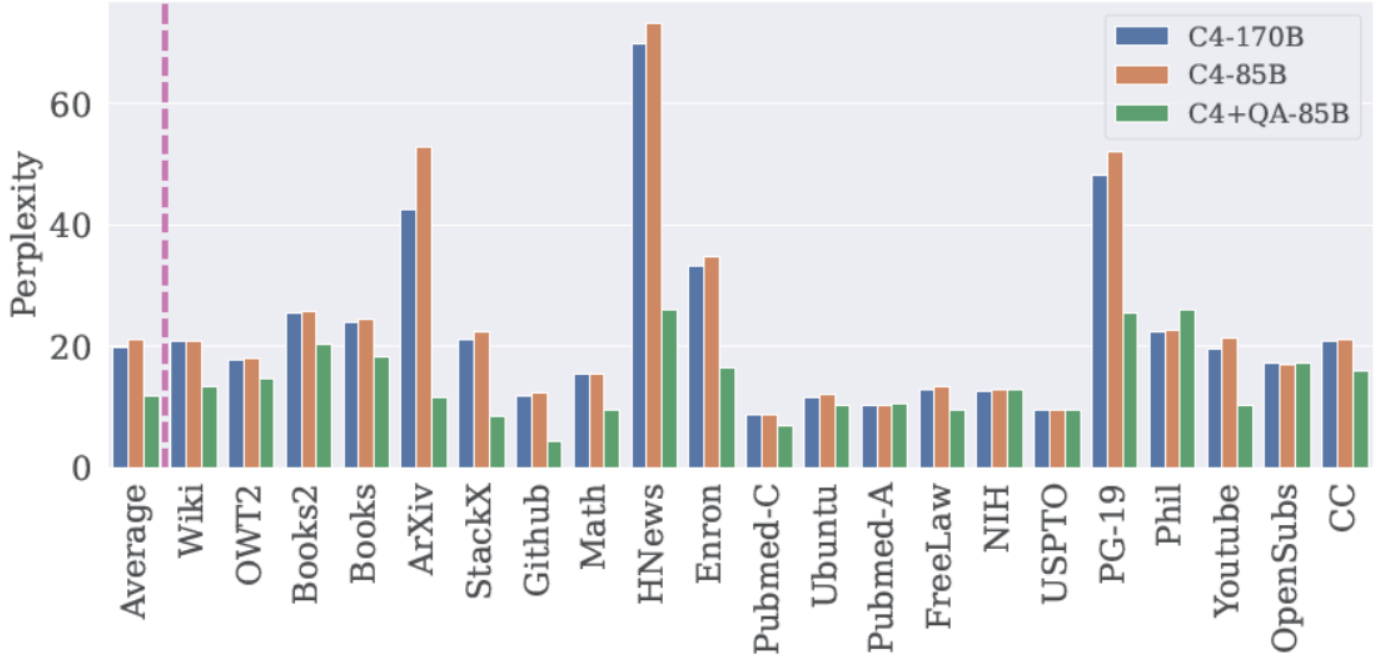


Figure 2: **WRAP (C4 + QA-85B) v/s C4**: Comparison of perplexity on the Pile for a 1.3B LLM trained for 300B tokens shows that WRAP outperforms models trained on 2x real data.

We observe that even at the first checkpoint (10B tokens) of **WRAP** training, the average perplexity of the LLM on the Pile is lower than that achieved by pre-training on C4 for 15 checkpoints. This suggests a 15x pre-training speed-up.

Zero-shot Tasks

- We evaluate pre-trained LLMs on 13 zero-shot QA benchmarks.

- Benchmarks cover:
  - **General Understanding (8)**: reasoning, language comprehension.
  - **Specialized Knowledge (5)**: science, medicine, mathematics.

- All evaluations are done using the **LLM Evaluation Harness**.

Zero-shot Tasks

General Knowledge –Findings

Dataset (Real Tok.)	ARC-E	BoolQ	Wino.	PIQA	HellaSwag	TruthfulQA	OBQA	LogiQA	Avg
Half C4 (85B)	61.2	59.1	57.3	74.9	46.5	34.1	22.4	23.5	47.4
Full C4 (170B)	61.6	54.2	59.0	74.9	46.8	33.5	25.0	23.4	47.3
RW (160B)	61.6	60.7	57.5	74.3	45.2	36.8	21.8	23.2	47.6
RW (320B)	60.7	61.1	57.1	74.4	45.6	36.0	22.6	22.5	47.5
Pythia-Pile (300B)	60.5	63.3	57.5	70.8	40.4	38.9	22.2	22.2	47.0
TinyLlama (1T)	60.3	57.8	59.1	73.3	45.0	37.6	21.8	24.5	47.4
Synthetic (85B)	63.9	60.0	58.8	76.1	45.2	44.0	23.0	24.1	49.4
Synthetic+C4 (85B)	64.1	62.2	58.9	75.4	46.2	40.6	24.1	23.9	49.4

Table 1: Evaluation of $\sim$ 1.3B parameter LLMs on 'General Understanding Tasks' on datasets focusing on general reasoning, language understanding, and common sense. Results for **WRAP**are averaged over 3 runs

- **Synthetic+C4 (85B)** outperforms real-only models with **49.4% avg.** vs. **47.4%**.
- WRAP shows small synthetic additions can boost performance significantly.
- Despite the advantages of a larger dataset, the improvements saturate.
- Real data may **reduce the benefit** of synthetic data for TruthfulQA task.

Zero-shot Tasks

Specialized Knowledge –Findings

Dataset (Real Tok.)	ARC-C	SciQ	PubMedQA	MathQA	MMLU	Avg
Half C4 (85B)	26.3	84.5	57.2	23.4	24.2	43.1
Full C4 (170B)	26.8	85.0	57.4	24.3	23.9	43.5
RW (160B)	27.2	87.2	56.2	24.1	25.9	44.1
RW (320B)	27.8	88.0	57.4	23.0	25.4	44.3
Pythia-Pile (300B)	26.1	86.6	60.6	25.2	24.3	44.6
TinyLlama (1T)	27.8	88.9	61.4	24.1	25.8	45.6
Synthetic (85B)	29.7	87.0	60.2	23.4	24.6	45.0
Synthetic+C4 (85B)	29.9	87.6	61.5	23.9	24.8	45.5

Table 2: Evaluation of $\sim$ 1.3B parameter LLMs on 'Specialized Knowledge Tasks' that require specific domain knowledge such as science, medicine, mathematics, and logic. Results for **WRAP**are averaged over 3 runs.

- Synthetic data cannot introduce **new knowledge**.
- WRAP improves language modeling efficiency, but no significant gain in domain-specific knowledge

Further Analysis

RQ1: How important is real C4 data?

Synthetic QA-style data can achieve strong performance on QA tasks alone. However, real C4 data is crucial for lowering perplexity across diverse domains due to its variety of tags and styles.

Dataset (Real Tok.)	ARC-E	BoolQ	Wino.	PIQA	HellaSwag	TruthfulQA	OBQA	LogiQA	Avg
Med+C4-35B	59.8	57.0	55.7	74.6	44.5	36.5	23.8	21.5	46.7
QA+C4-35B	62.2	63.3	55.7	74.8	44.6	41.4	22.4	23.2	48.4
Med-35B	56.6	59.5	53.4	74.0	41.9	36.3	22.2	22.7	45.8
QA-35B	61.7	62.0	53.9	75.2	43.4	43.0	22.8	23.4	48.2

Table 3: **Importance of Real Data:** Evaluation of $\sim$ 1.3B parameter LLMs trained for 150B tokens on General Understanding Tasks. Results show that adding real data helps improve model performance when pre-training on 'Medium' or 'Wikipedia-style' paraphrases.

Dataset (Real Tok.)	ARC-C	SciQ	PubMedQA	MathQA	MMLU	Avg
Med+C4-35B	27.2	82.2	46.2	23.1	25.2	40.8
QA+C4-35B	29.0	85.1	62.2	22.5	26.1	45.0
Med-35B	27.0	80.0	59.4	22.5	24.7	42.7
QA-35B	27.1	85.5	59.2	22.2	25.0	43.8

Table 4: **Importance of Real Data:** Evaluation of $\sim$ 1.3B parameter LLMs on Specialized Knowledge Tasks. Results show that adding real data helps improve model performance when pre-training on 'Q/A-style' paraphrases.

RQ2: Does combining multiple synthetic styles improve performance?

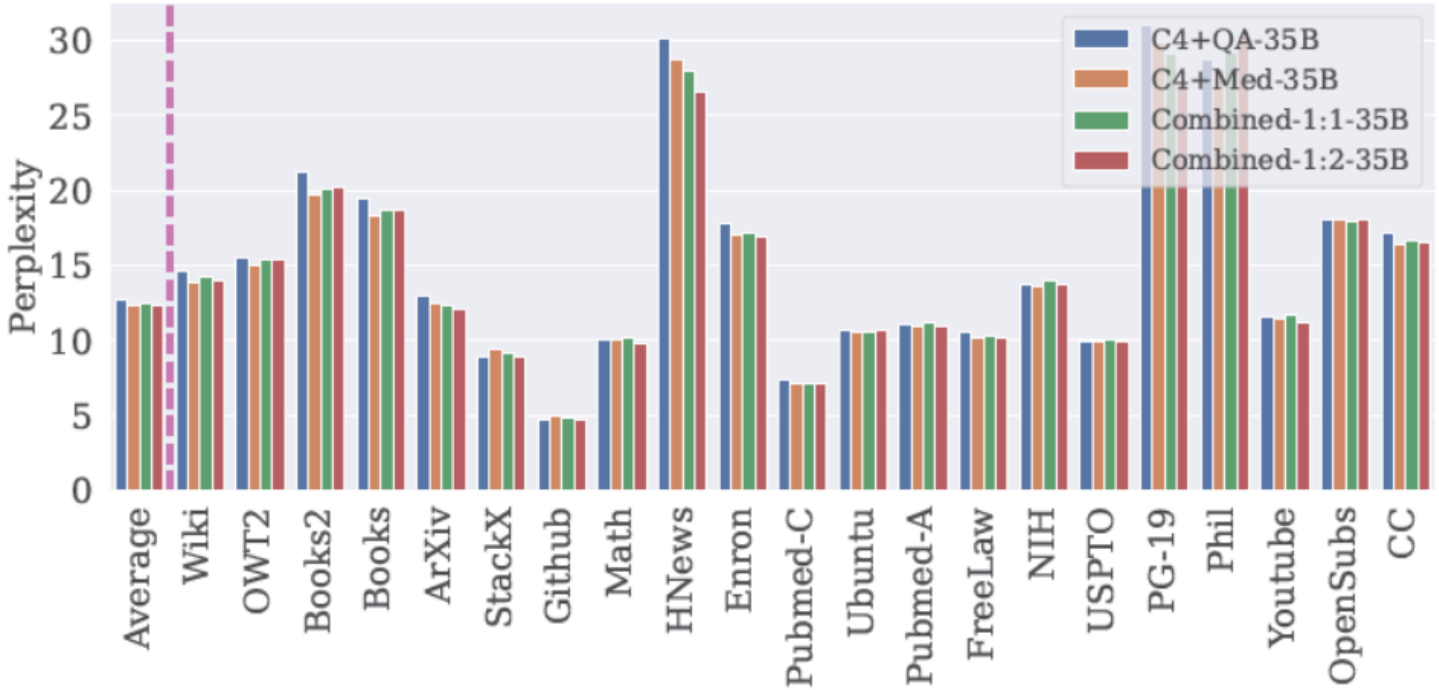


Figure 4: **Combining multiple styles:** Perplexity across all domains of the Pile comparing combining multiple styles of synthetic data. Models are 1.3B parameters trained for a total of 150B tokens. We see small perplexity improvements from combining multiple styles.

- **Setup:** C4 combined with synthetic data in:
  - 1:1 ratio → Two copies of C4 for 'medium' + QA styles
  - 1:2 ratio → Single C4 copy with both styles
- **Findings:**
  - 'Q/A' and 'Wikipedia' styles boost performance in specific domains *(Stackexchange)*
  - Combined styles outperform significantly in HNEWS & PG-19
- **Conclusion:**
  - Only **minor perplexity gains** on the Pile from combining multiple styles

RQ3: How important is it to have a high-quality rephraser?

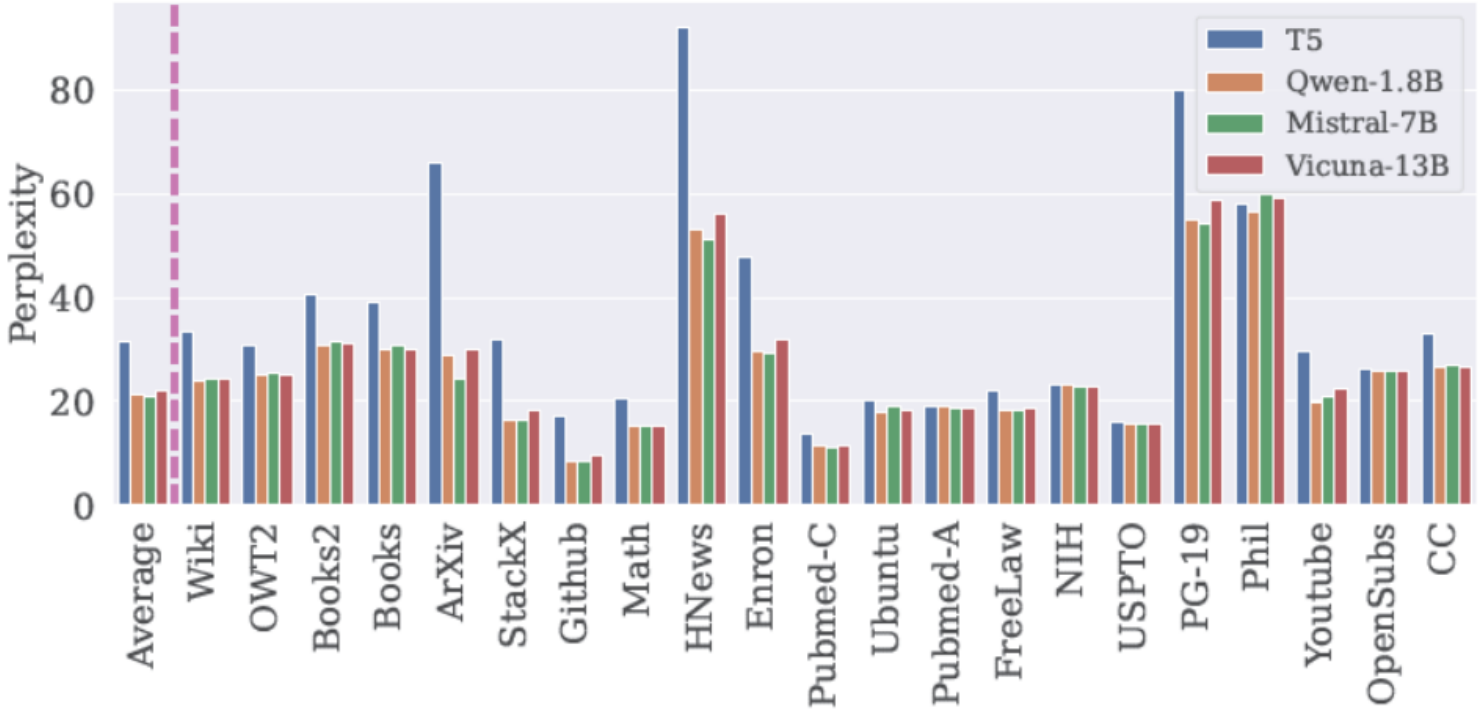


Figure 5: **Importance of High Quality Paraphraser:** Perplexity across all the Pile domains for **WRAP** on data generated by different LLMs. Results show that even small models like Qwen-1.8B can generate paraphrases of high quality. Though, a low quality rephraser like our fine-tuned T5-base model leads to significantly worse language modeling.

- Evaluated 4 rephrasers: T5-base, **Qwen-1.8B**, Mistral-7B, Vicuna-13B
- Trained a 345M model on synthetic data from each
- **Surprisingly, smaller models (Qwen, Mistral) performed better** than Vicuna
- Fine-tuned T5-base performed **worst**
- All rephrasers helped reduce perplexity on real C4
- **Open question**: how small can the rephraser be while still useful?

Qwen-1.8B-chat: A Lightweight Yet Effective Rephraser

Model Details:

- Model Size: 1.8 billion parameters
- Instruction-tuned for dialogue and rephrasing tasks
- Used for generating synthetic data via prompting

Performance in the Paper:

- Surprisingly strong performance:
  - Achieves lower perplexity than larger models like Vicuna-13B
  - Outperforms the fine-tuned T5-base model
- Efficient trade-off between size and quality

→ Ideal for low-cost high-quality data generation

RQ4: Does synthetic data improve over augmentations?

Synthetic data enhances model learning beyond what traditional data augmentation can offer.

RQ5: How does the style of synthetic data impact performance on specialized domains?

- Best performance when style matched the test domain
- But: no single style works best across all domains
- Training on diverse synthetic styles improves generalization, even if the knowledge content is the same

RQ6: Is there data leakage from the rephrase model to the trained model?

- Performance gains are **not due to data leakage**
- Rephrased data remains **faithful** to the original in meaning

Limitations

Cost Trade-offs

"Should you generate synthetic data, or just train longer on real data?"

- **Low-resource setting:** No alternative but to generate synthetic data (e.g., Finnish).
- **High-resource setting:** Question remains—should we use synthetic data or just train longer?

Real-data Training May Be Saturated.
Training longer on real data shows limited gains (e.g., TinyLlama on 3T tokens underperforms)

Cost Trade-offs – Synthetic vs. Real Data

Synthetic data generation (85B tokens):
- 25K GPU hours (using Mistral-7B with vLLM on A100)

Training on 300B tokens (13B model):
- 30K GPU hours
- Synthetic pretraining can reduce cost by **3–10x**

Limitations

Cost Reductions & Benefits

Reducing Synthetic Data Costs

Efficiency improvements:

- **Qwen-1.8B model** for rephrasing

→ 3× faster throughput

→ Comparable model performance to Mistral

- **Speculative decoding + optimized inference**

→ Additional 3–5× improvement in generation speed

Additional Advantages of Synthetic Data

One-time Cost

- Reusable across many model scales

Fully Parallelizable

- Generation runs on idle or low-end GPUs

Diversity Challenges

• Diversity in synthetic data comes from: Style and Knowledge variation

• Recent work (Li et al., 2023b,c) uses **topic prompts** to encourage novel generation
But Instruction-tuned LMs may reduce content diversity (Padmakumar et al., 2023)

• Current approach: Rephrasing used to **mitigate diversity loss** in content generation

• Future work: We should check if the rephrased synthetic data is really diverse, and if that diversity actually helps improve model performance.

Key Questions:

1.Is the data truly diverse?

– Check if paraphrased synthetic data includes a wide range of styles, wording, and ideas.

2.Does diversity help?

– Test whether this variety actually improves model learning and task performance.

Conclusion

Proposed Method:
WRAP (Web Rephrase Augmented Pre-training)uses instruction-tuned models to paraphrase web documents into structured formats (e.g., Wikipedia-style, QA-style)

Key Benefits:
- **3× faster** pretraining on noisy data (e.g., C4)
- Improves **perplexity** by >10% on Pile subsets
- Boosts **zero-shot accuracy** on 13 tasks by >2%

Broader Impact:
- Rephrased synthetic data improves **training utility**
- Higher **style consistency** and **data quality** than raw web-scraped corpora

Future Work:
- Investigate **content diversity trade-offs**
- Assess **long-term generalization** of paraphrased synthetic data