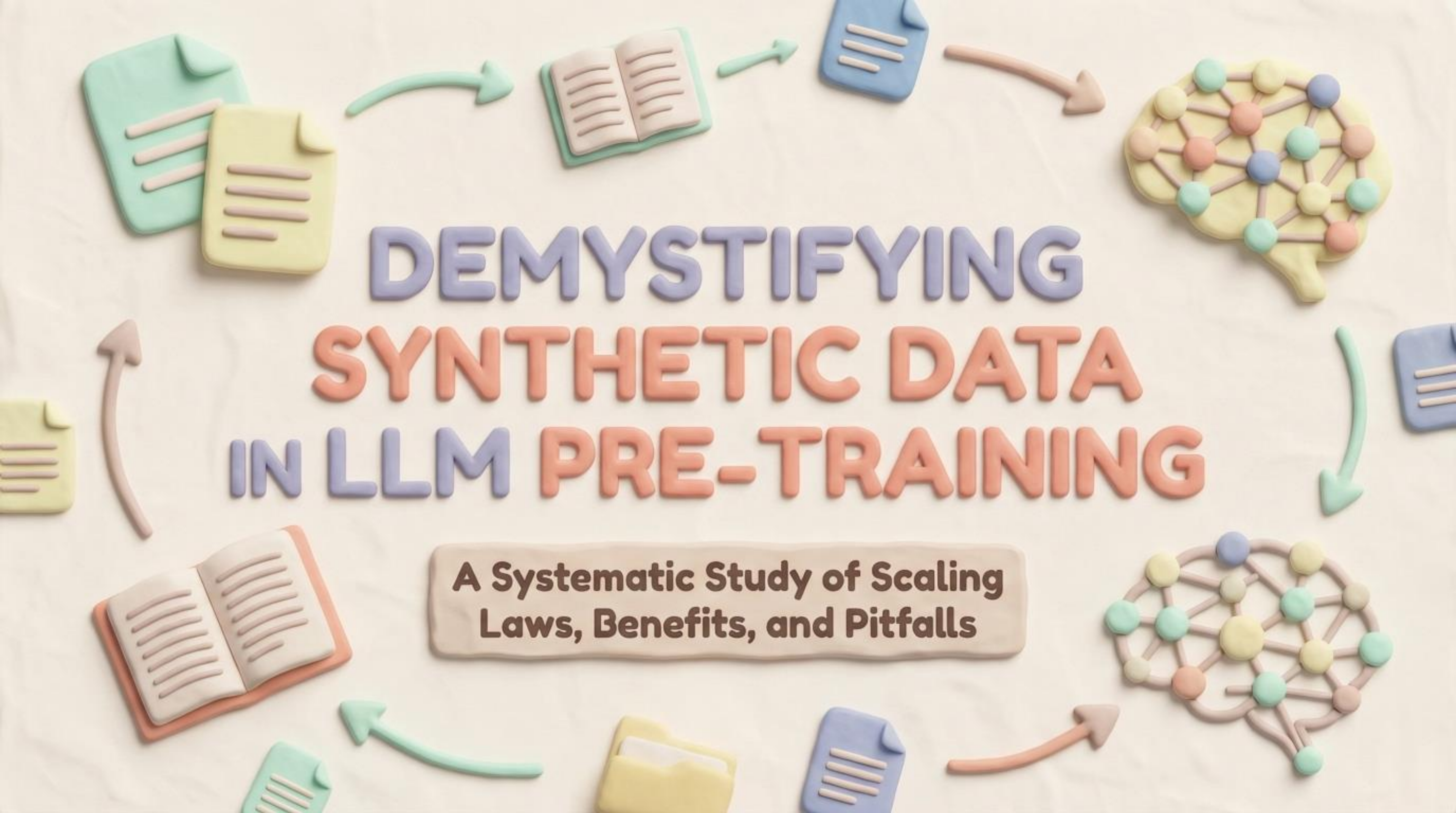
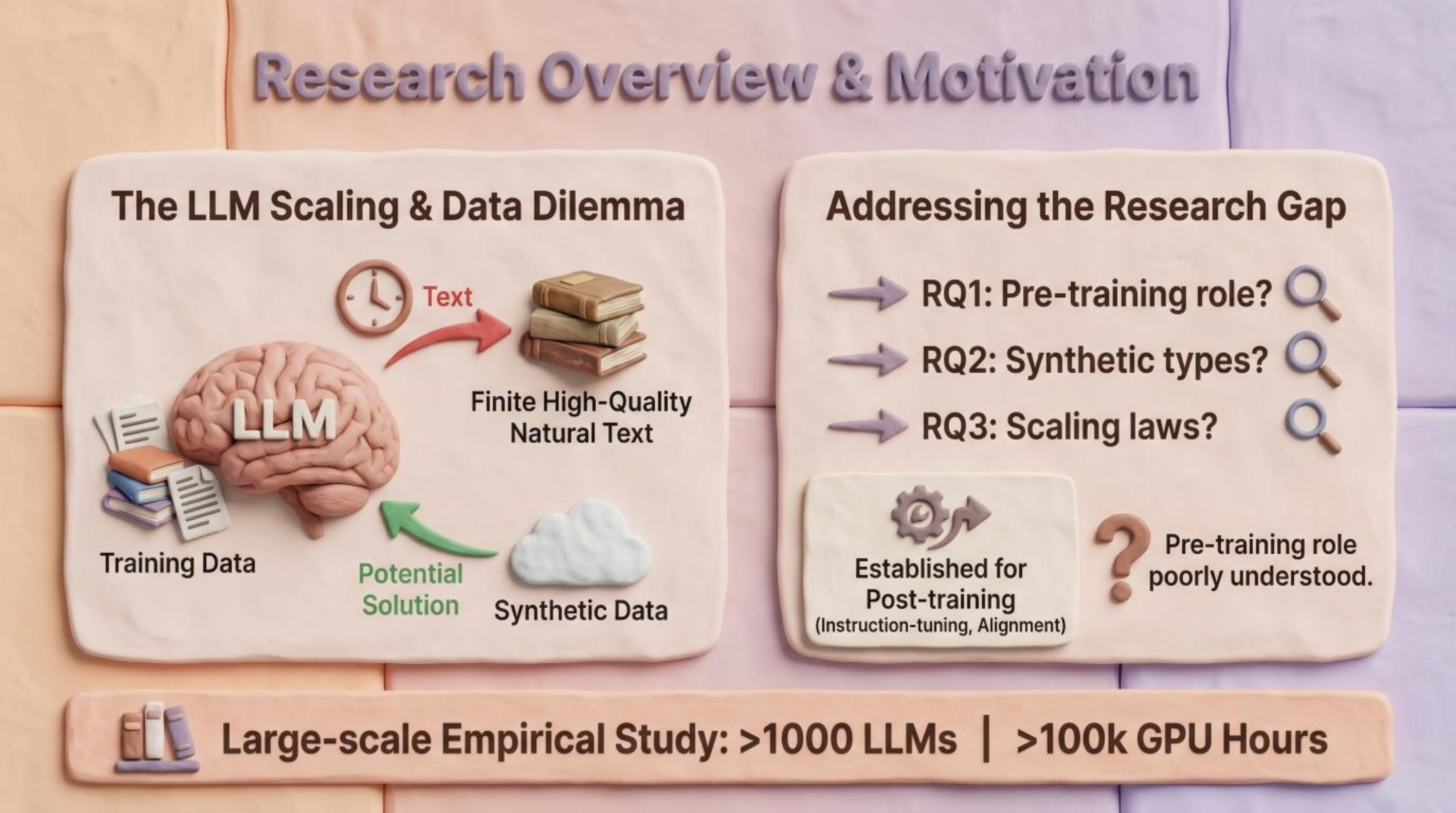

DEMYSTIFYING
SYNTHETIC DATA
IN LLM PRE-TRAINING
A Systematic Study of Scaling
Laws, Benefits, and Pitfalls


Research Overview & Motivation
The LLM Scaling & Data Dilemma
Text
Finite High-Quality Natural Text
LLM
Training Data
Potential Solution
Synthetic Data
Addressing the Research Gap
RQ1: Pre-training role?
RQ2: Synthetic types?
RQ3: Scaling laws?
Established for Post-training
(Instruction-tuning, Alignment)
Pre-training role poorly understood.
Large-scale Empirical Study: >1000 LLMs | >100k GPU Hours

Synthetic Data Generation Paradigms

1. Web Rephrasing (HQ/QA)

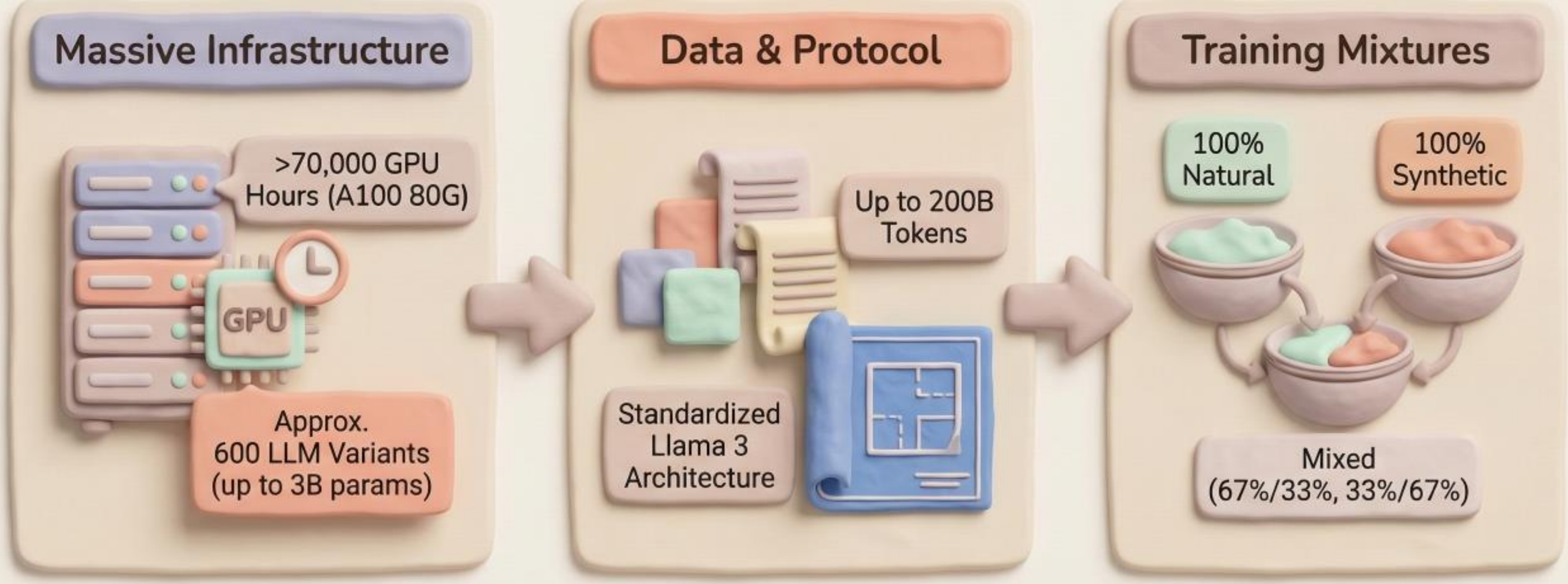
High-Quality (HQ)
Wikipedia-style clarity

Question-Answering (QA)
Conversational formats

2. Synthetic Textbooks (TXBK)

- Novel educational content
- Dense information
- Examples & exercises

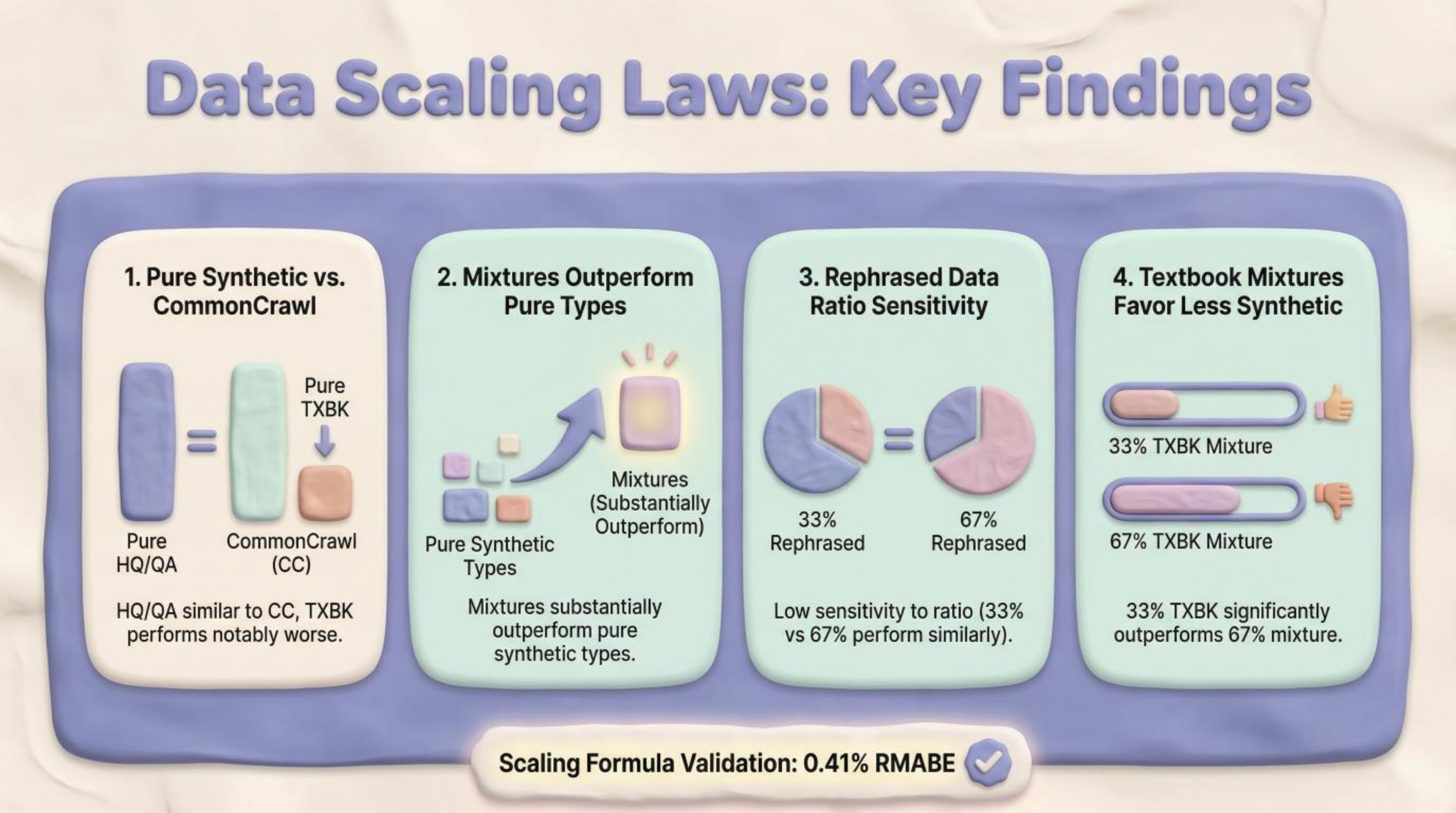
Experimental Setup & Scale

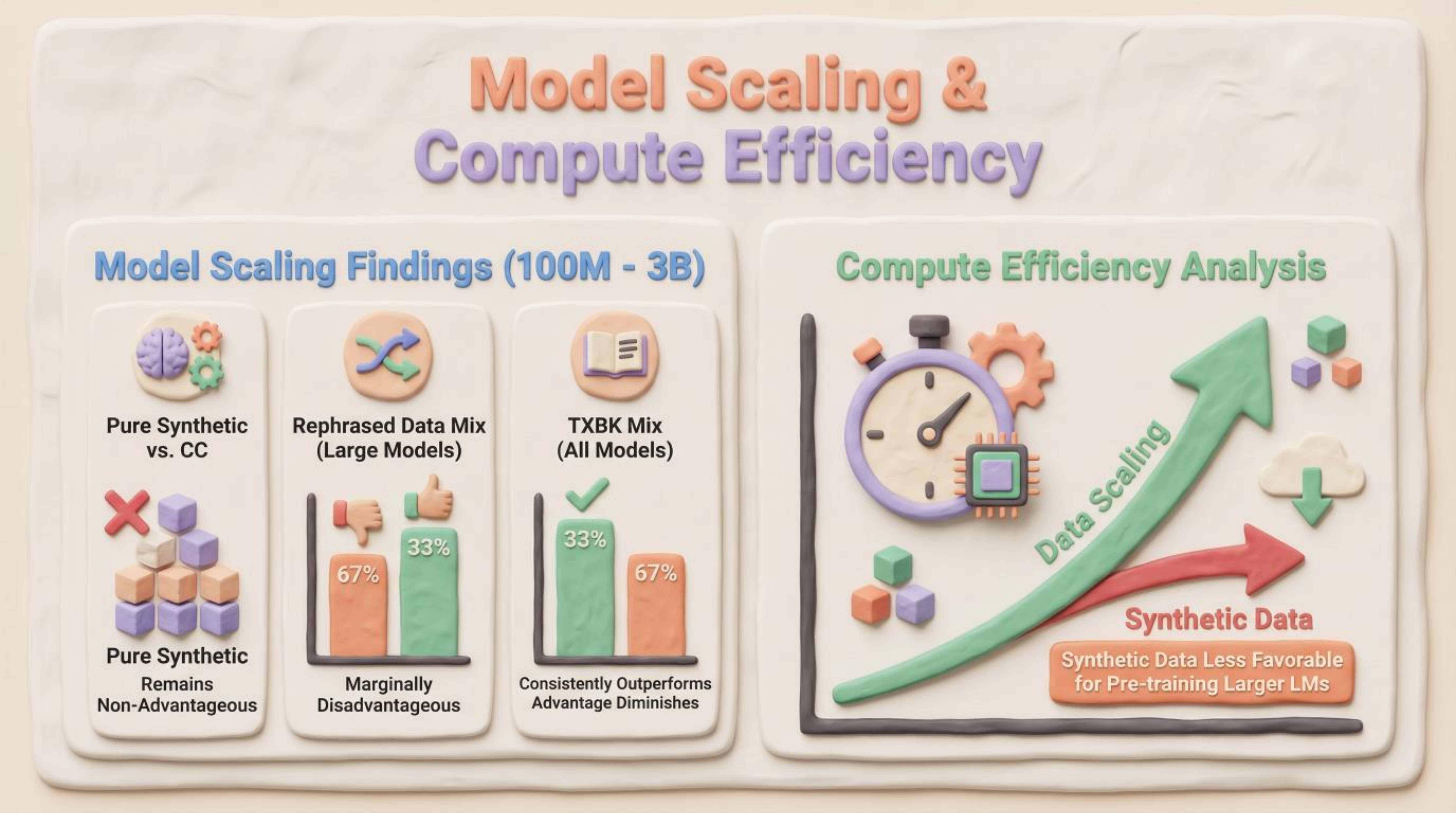


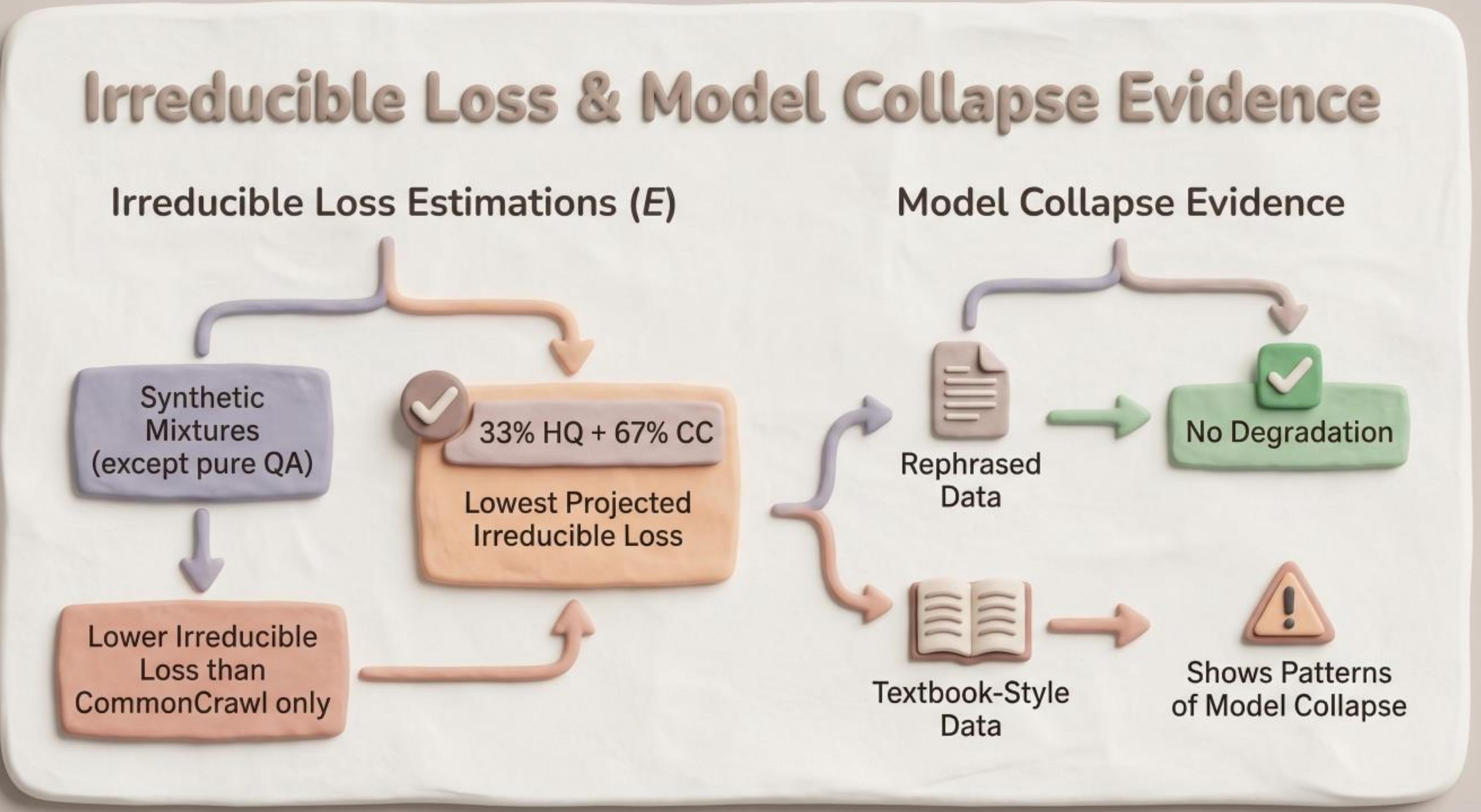
- >70,000 GPU Hours (A100 80G)
- Finite high-qu natural text & synthetic cal factors and lannens

- Up to ~200B Tokens
- Standardized Llama 3 Architecture

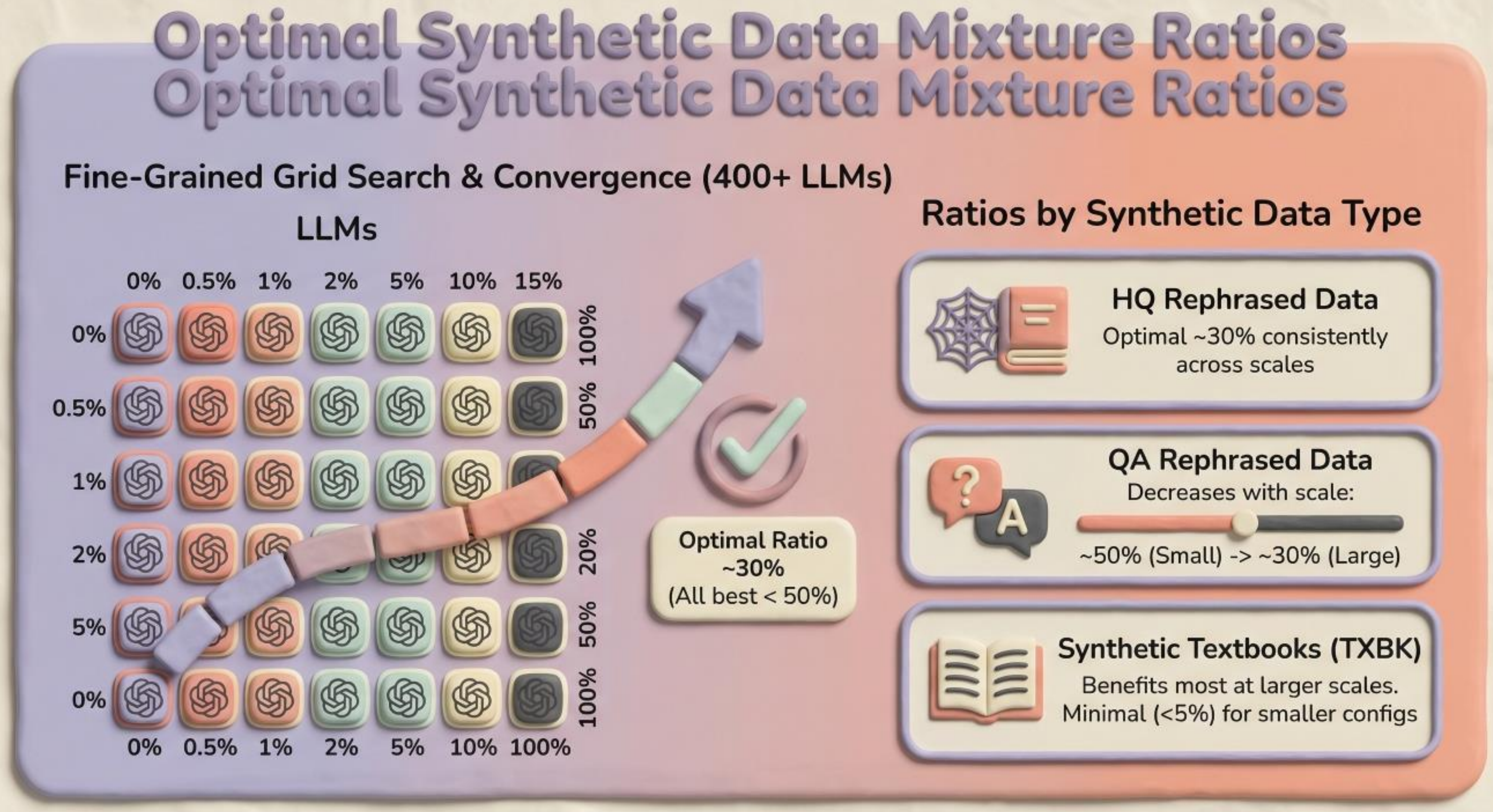
- Syntheticic empirical study
- Finite health quality contents benefits ehectetic concerns


Data Scaling Laws: Key Findings
1. Pure Synthetic vs. CommonCrawl
Pure TXBK
Pure HQ/QA
CommonCrawl (CC)
HQ/QA similar to CC, TXBK performs notably worse.
2. Mixtures Outperform Pure Types
Mixtures (Substantially Outperform)
Pure Synthetic Types
Mixtures substantially outperform pure synthetic types.
3. Rephrased Data Ratio Sensitivity
33% Rephrased
67% Rephrased
Low sensitivity to ratio (33% vs 67% perform similarly).
4. Textbook Mixtures Favor Less Synthetic
33% TXBK Mixture
67% TXBK Mixture
33% TXBK significantly outperforms 67% mixture.
Scaling Formula Validation: 0.41% RMABE


Model Scaling &
Compute Efficiency
Model Scaling Findings (100M - 3B)
Pure Synthetic
vs. CC
Pure Synthetic
Remains
Non-Advantageous
Rephrased Data Mix
(Large Models)
67%
33%
Marginally
Disadvantageous
TXBK Mix
(All Models)
33%
67%
Consistently Outperforms
Advantage Diminishes
Compute Efficiency Analysis
Data Scaling
Synthetic Data
Synthetic Data Less Favorable
for Pre-training Larger LMs



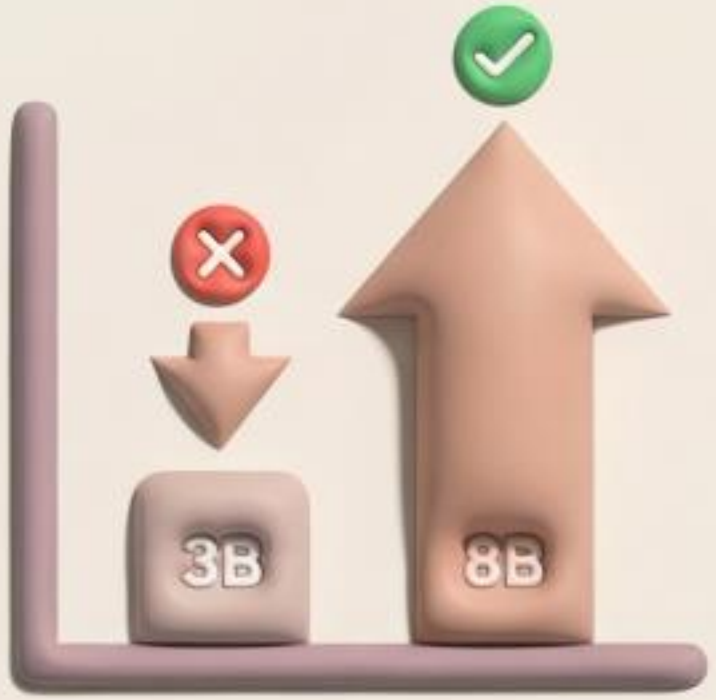
Irreducible Loss & Model Collapse Evidence
Irreducible Loss Estimations (E)
Model Collapse Evidence
Synthetic Mixtures (except pure QA)
33% HQ + 67% CC
Lowest Projected Irreducible Loss
Lower Irreducible Loss than CommonCrawl only
Rephrased Data
No Degradation
Textbook-Style Data
Shows Patterns of Model Collapse



Optimal Synthetic Data Mixture Ratios
Optimal Synthetic Data Mixture Ratios
Fine-Grained Grid Search & Convergence (400+ LLMs)
LLMs
0% 0.5% 1% 2% 5% 10% 15%
0%
0.5%
1%
2%
5%
0%
100%
50%
20%
50%
100%
0% 0.5% 1% 2% 5% 10% 100%
Optimal Ratio
~30%
(All best < 50%)
Ratios by Synthetic Data Type
HQ Rephrased Data
Optimal ~30% consistently across scales
QA Rephrased Data
Decreases with scale:
?
A
~50% (Small) -> ~30% (Large)
Synthetic Textbooks (TXBK)
Benefits most at larger scales.
Minimal (<5%) for smaller configs

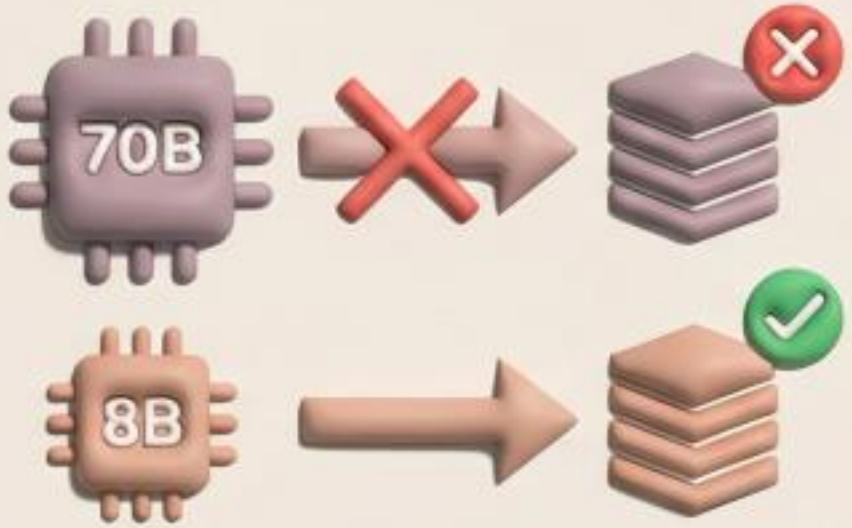
Generator Model Capability Impact

Key Finding: Baseline Matters



- 8B > 3B: Indicates baseline capability is necessary.

Surprising Result: Bigger ≠ Better



Increasing to 70B did not yield superior synthetic data compared to 8B.

- 70B < 8B for HQ Rephrased Data.

Conclusion: Beyond Scale

Challenges the “bigger is always better” intuition.

- Suggests factors beyond sheer scale matter.

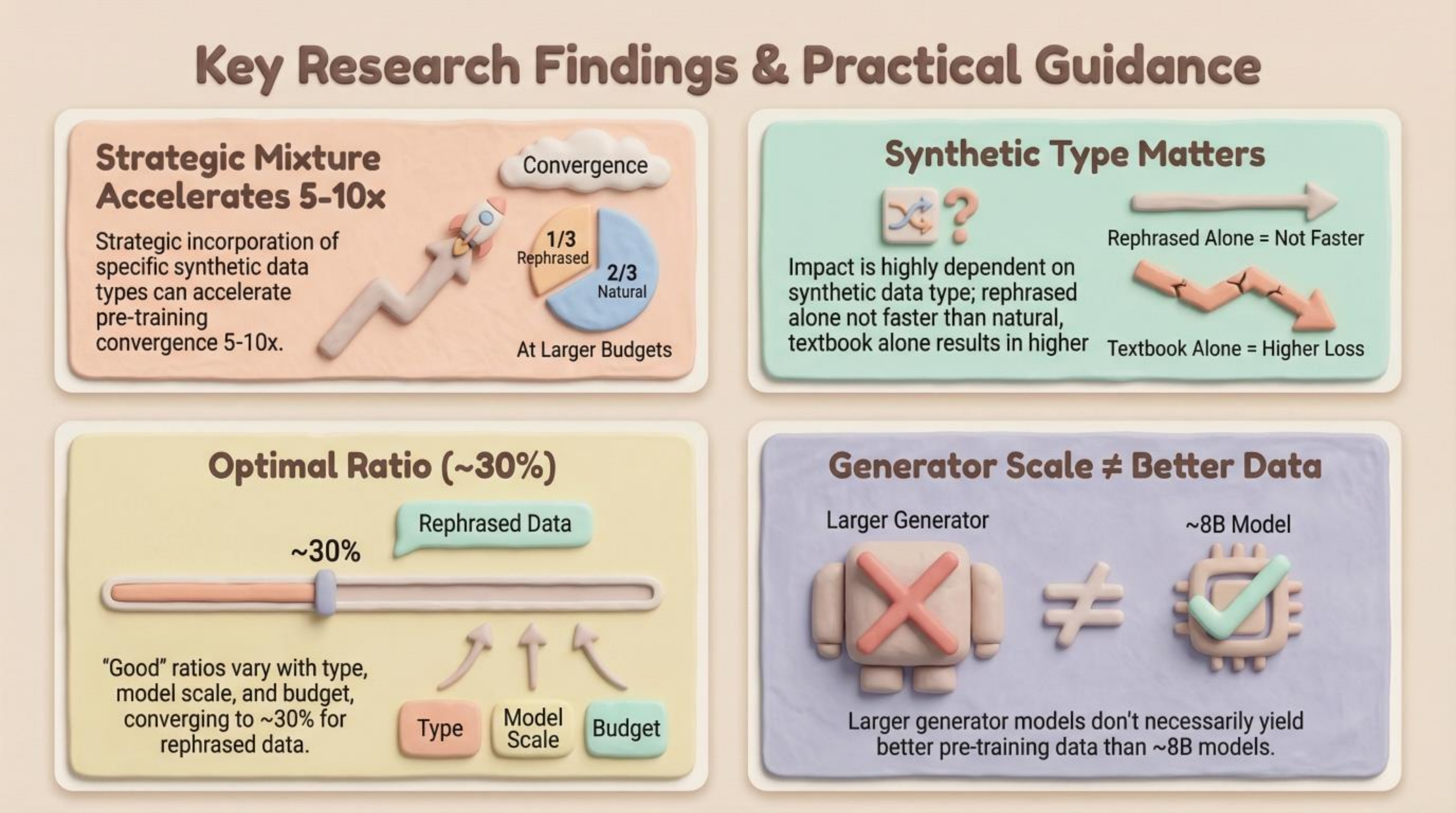
Low-Level Statistical Analysis

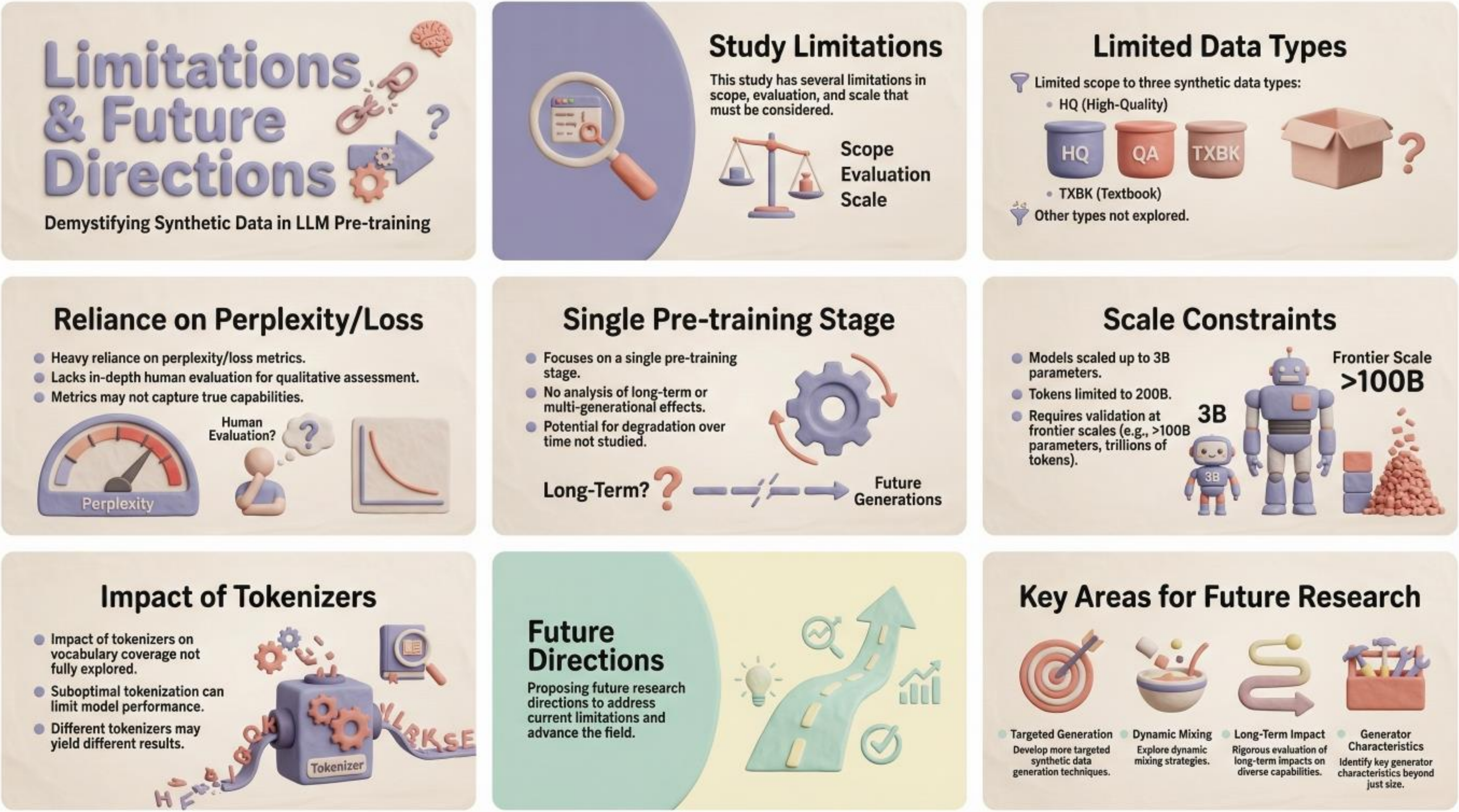
Unigram Frequency Insights

- Training-test mismatch from rare unigrams (e.g., ‘ ’, ‘ hvor’) in training sets causes higher loss.
- No single training set offers complete coverage.
- Synthetic data slightly shrinks unigram distribution vs. CommonCrawl.

Key Takeaways

- CommonCrawl has widest coverage & lowest KL-divergence, but doesn't yield superior performance.
- Good training mixtures depend on complex diversity-quality trade-offs beyond simple similarity.
- **Performance depends on complex diversity-quality trade-offs, not just coverage.**


Key Research Findings & Practical Guidance
Strategic Mixture Accelerates 5-10x
Strategic incorporation of specific synthetic data types can accelerate pre-training convergence 5-10x.
Convergence
1/3
Rephrased
2/3
Natural
At Larger Budgets
Synthetic Type Matters
Impact is highly dependent on synthetic data type; rephrased alone not faster than natural, textbook alone results in higher
Rephrased Alone = Not Faster
Textbook Alone = Higher Loss
Optimal Ratio (~30%)
Rephrased Data
~30%
"Good" ratios vary with type, model scale, and budget, converging to ~30% for rephrased data.
Type
Model Scale
Budget
Generator Scale ≠ Better Data
Larger Generator
~8B Model
Larger generator models don't necessarily yield better pre-training data than ~8B models.


Limitations & Future Directions
Demystifying Synthetic Data in LLM Pre-training
Study Limitations
This study has several limitations in scope, evaluation, and scale that must be considered.
Scope
Evaluation
Scale
Limited Data Types
Limited scope to three synthetic data types:
HQ (High-Quality)
HQ
QA
TXBK
TXBK (Textbook)
Other types not explored.
Reliance on Perplexity/Loss
Heavy reliance on perplexity/loss metrics.
Lacks in-depth human evaluation for qualitative assessment.
Metrics may not capture true capabilities.
Human Evaluation?
Perplexity
Single Pre-training Stage
Focuses on a single pre-training stage.
No analysis of long-term or multi-generational effects.
Potential for degradation over time not studied.
Long-Term?
Future Generations
Scale Constraints
Models scaled up to 3B parameters.
Tokens limited to 200B.
Requires validation at frontier scales (e.g., >100B parameters, trillions of tokens).
3B
Frontier Scale
>100B
Impact of Tokenizers
Impact of tokenizers on vocabulary coverage not fully explored.
Suboptimal tokenization can limit model performance.
Different tokenizers may yield different results.
Tokenizer
Future Directions
Proposing future research directions to address current limitations and advance the field.
Key Areas for Future Research
Targeted Generation
Develop more targeted synthetic data generation techniques.
Dynamic Mixing
Explore dynamic mixing strategies.
Long-Term Impact
Rigorous evaluation of long-term impacts on diverse capabilities.
Generator Characteristics
Identify key generator characteristics beyond just size.

Ethical Considerations & Conclusion

Ethical Risks & Mitigation

- **Bias Propagation Risks** (from Generator Models)
- **Factual Accuracy Concerns** (Misinformation)
- **Reduced Diversity** (Over-reliance)
- **Importance of Transparency** (Mitigation)

A Nuanced Trade-off & Conclusion