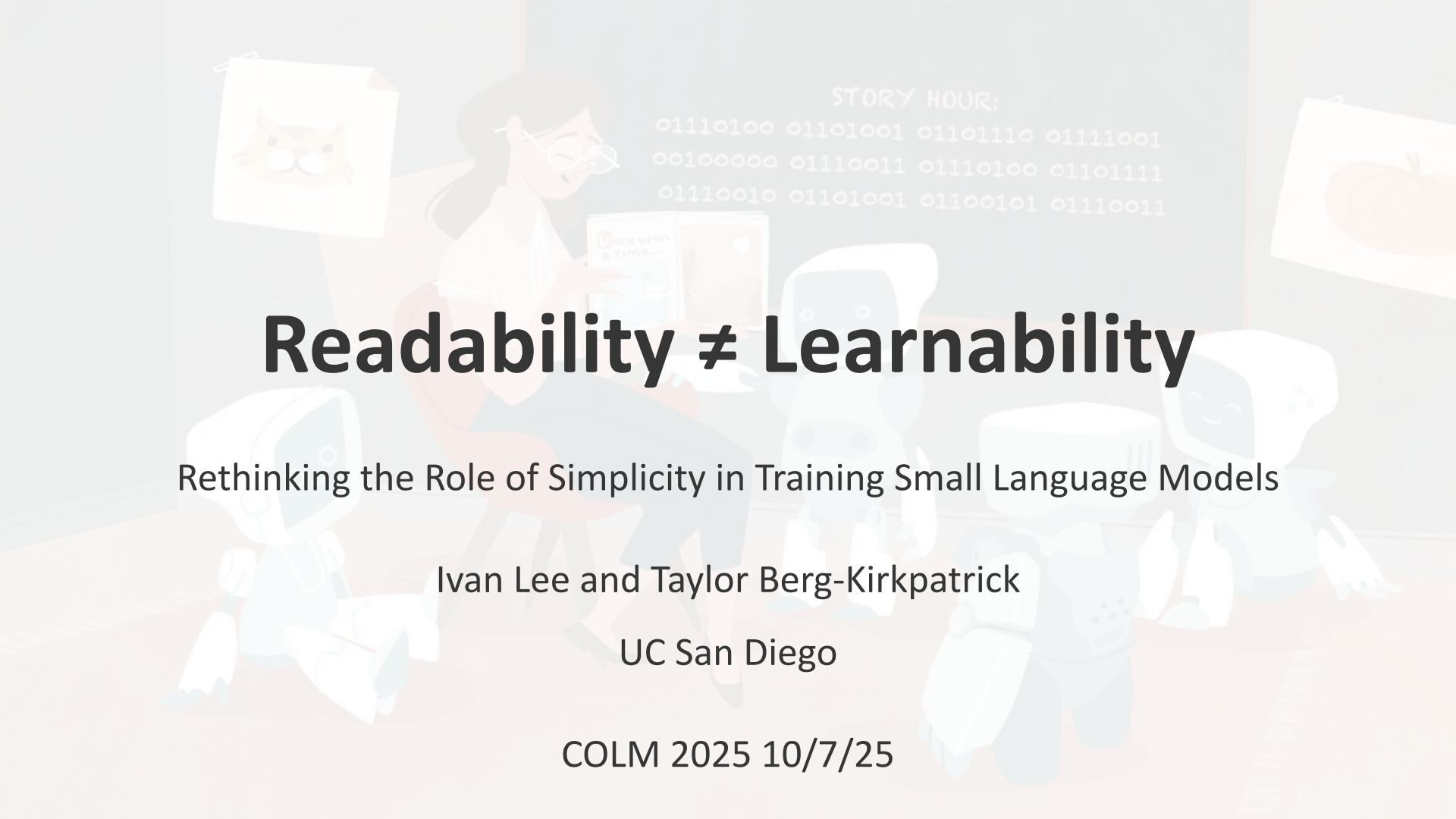


Readability ≠ Learnability

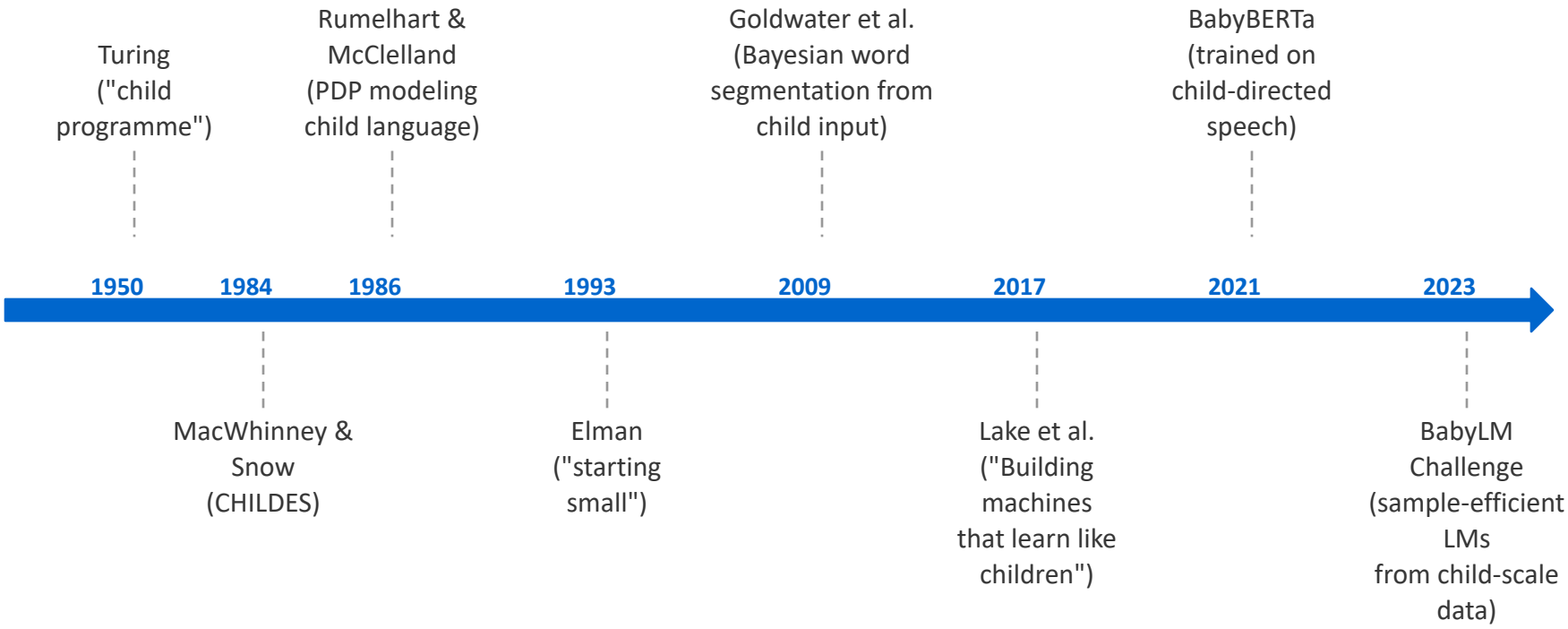
Rethinking the Role of Simplicity in Training Small Language Models

Ivan Lee and Taylor Berg-Kirkpatrick

UC San Diego

COLM 2025 10/7/25

A Longstanding Interest: Child Learning and NLP/AI



TinyStories: Can Simpler Data Enable Smaller Models?

Challenge: Very small models struggle to generate coherent text

Hypothesis: The training data is too complex

Intuition: Children develop language with limited exposure

Proposal: Restrict training data to only children's stories

✓ **Result: 10-100M parameter models generate coherent narratives**

A Natural Interpretation: The Developmental Lens

TinyStories Framing:

- children's stories with 3-4 year old vocabulary
- “children attain abilities with much less exposure to language than an adults”
- “comparable 'understanding' of language to that of children"

Note: Authors made no explicit causal claims about developmental mechanisms.

How this interpretation appears

Academic Literature

- Grouped with child-directed speech (CDS) research
- 'Mimicking human acquisition'
- Developmental frameworks

ACL, EMNLP, NAACL, CoNLL

Media & Corporate

- 'Raise them on a strict diet of children's stories’
- Bedtime story narratives
- 'Tiny Language Models Come of Age’ anthropomorphic framing

Microsoft, Quanta Magazine

Community Discourse

- Casual use of child-learning analogies
- Parallels human development

HackerNews, Reddit, Twitter

The Open Question

What we know:

Small models trained on TinyStories generate fluent, coherent text.

What we DON'T know:

Whether this finding can be attributed to child-directed properties.

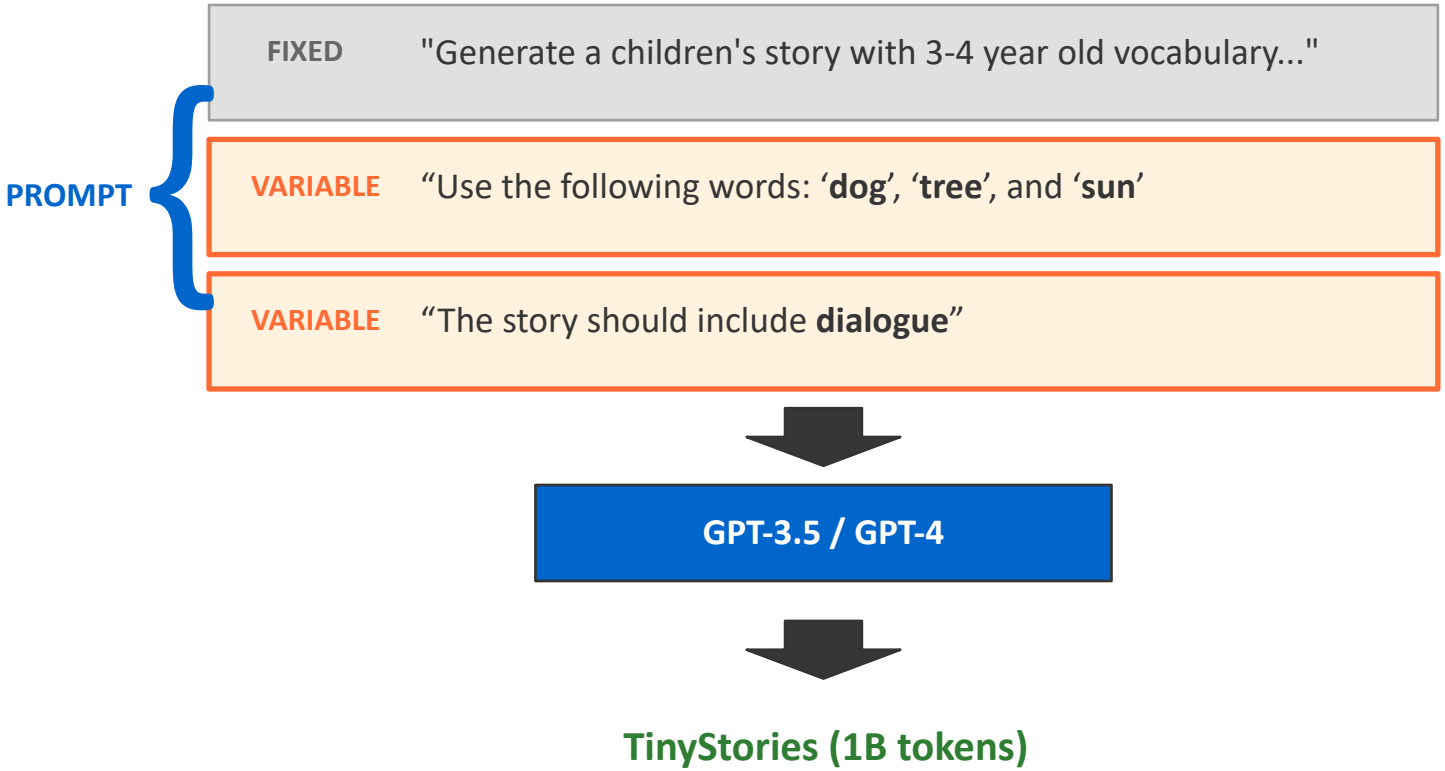
Are these properties special, or just one approach among many?

Roadmap

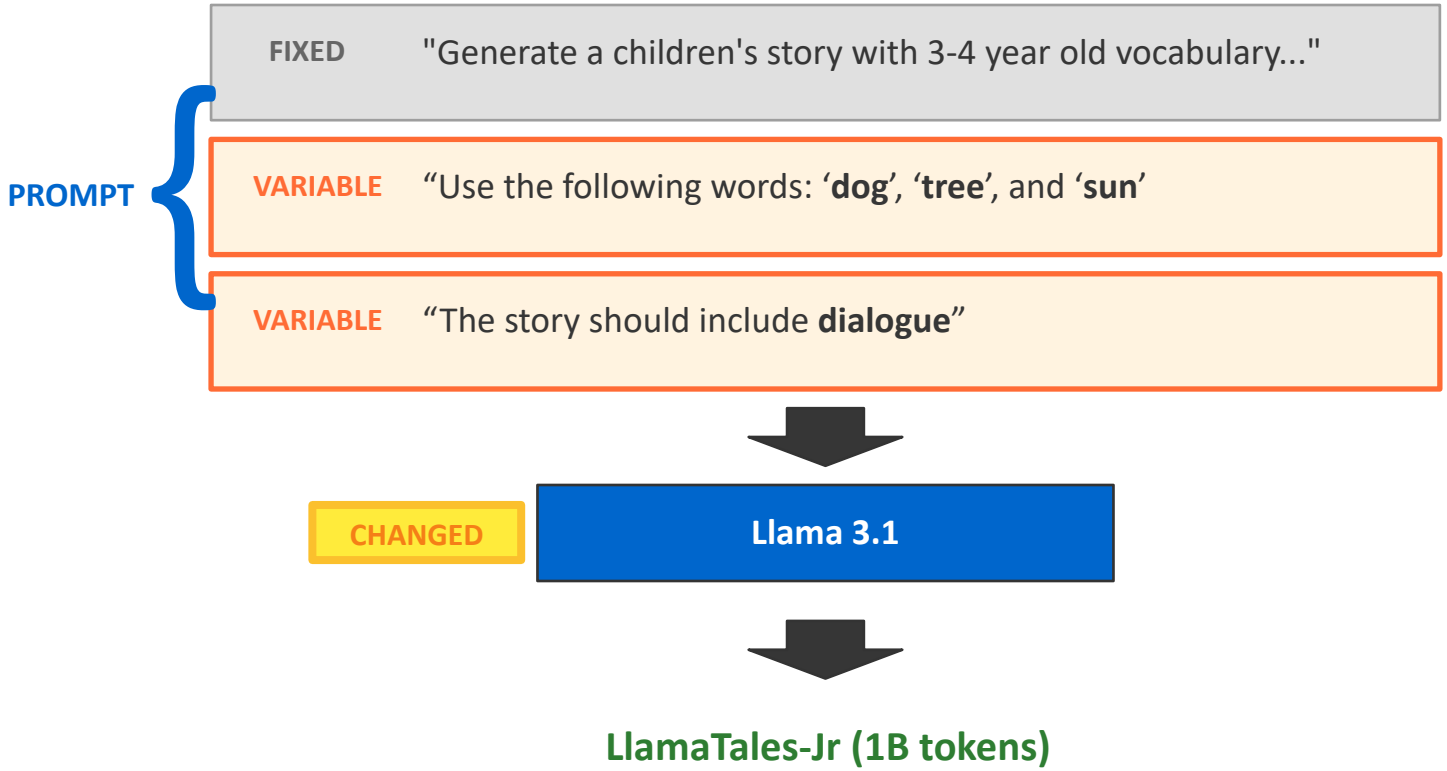
1. Replicate TinyStories' generation pipeline
2. Modify pipeline to vary target complexity (child-directed → adult-directed)
3. Validate we achieved the intended variation
4. Train small language models on both datasets
5. Sample from models and measure fluency/coherence

Key Question: Can we reproduce TinyStories' findings on data that does not leverage any child-directed properties?

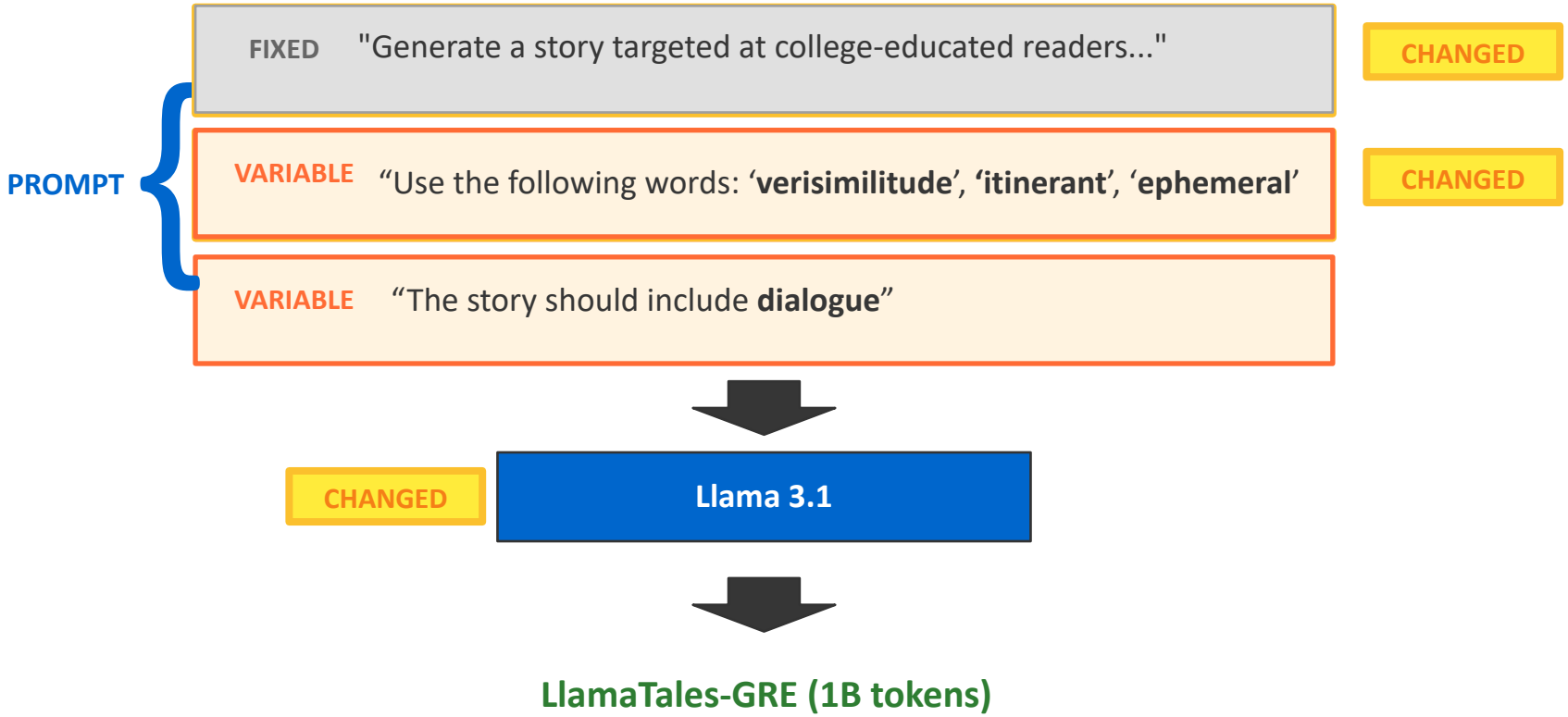
TinyStories Pipeline



LlamaTales-Jr Pipeline



LlamaTales-GRE Pipeline



Operationalizing language complexity

Readability:

- The ease with which a text can be read or understood (Trott, 2024)
- Long history in pedagogical contexts
  (Lively and Pressey, 1923; Flesch, 1948; Crossley et al., 2023b)
- Many automated metrics available
  (e.g., Flesch-Kincaid, SMOG, Gunning Fog)

Question: Which metric should we use?

Validating our data

We tested a large subset of readability metrics:

- Classic readability formulas (Flesch-Kincaid, Dale-Chall, SMOG)
- Constituency parsing metrics (syntactic complexity)
- LLM-as-judge approaches

Consistent pattern:

- TinyStories & LlamaTales-Jr: High readability
- LlamaTales-GRE & standard corpora: Low readability

✓ Pipeline modification worked as intended

See paper for details

	TinyStories	LlamaTales-Jr	LlamaTales-GRE	FineWeb
Automated Readability	2.9	2.9	12.4	13.1
Coleman–Liau	3.7	3.8	10.4	11.8
Dale–Chall	5.7	5.7	9.1	9.3
Flesch–Kincaid	2.4	2.2	9.6	10.7
Gunning Fog	4.6	3.8	11.7	12.1
Linsear Write	4.2	3.3	13.2	12.7
SMOG	5.7	5.4	11.3	12.6
Spache Readability	2.7	2.5	5.5	5.5
Depth / Sentence	6.8	6.4	10.6	9.5
Width / Sentence	5.1	4.7	8.0	7.5
Nodes / Sentence	19.6	17.2	42.1	37.8
Readability	92.6	92.7	64.8	68.2
Coherence	90.1	89.5	94.4	77.4

Validating readability metrics

Question: Which metrics best capture readability?

Validation: Tested against CLEAR dataset

- ~5,000 passages rated by teachers for reading difficulty

Results (correlation with human judgment):

- All positively correlated
- Constituency parsing: Weakest
- Classic readability formulas: Better
- **LLM-as-judge: Substantially stronger ✓**

We use LLM-as-judge as our primary readability measure

Training small language models

Model architecture: decoder-only transformer

Model sizes: 262K to 33M parameters

Training data: Each model trained on one dataset

- TinyStories
- LlamaTales-Jr
- LlamaTales-GRE
- FineWeb (our representative of standard pretraining data)

Training duration: 10 epochs (10 billion tokens)

Evaluating fluency and coherence

Goal: Measure whether models generate coherent, fluent text

Method: LLM-as-judge (following TinyStories approach)

Sanity Check: correlation with common-sense quality ranking of LMs e.g.

- strong tier: Llama-3.1-70B, Qwen2-72B
- weak tier: Pythia-70M, GPT-2-small

Result: LLM-as-judge correlated well with ranking, outperformed perplexity

Inventory Check

1 ✓ Replicate TinyStories' generation pipeline

2 ✓ Modify pipeline to vary target complexity (child-directed → adult-directed)

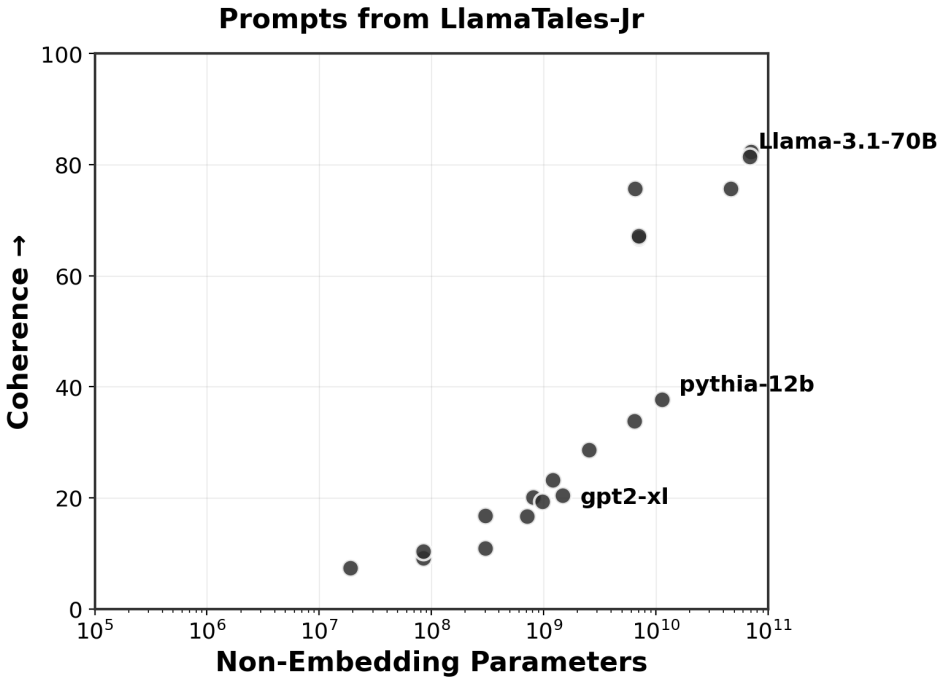
3 ✓ Validate we achieved the intended variation

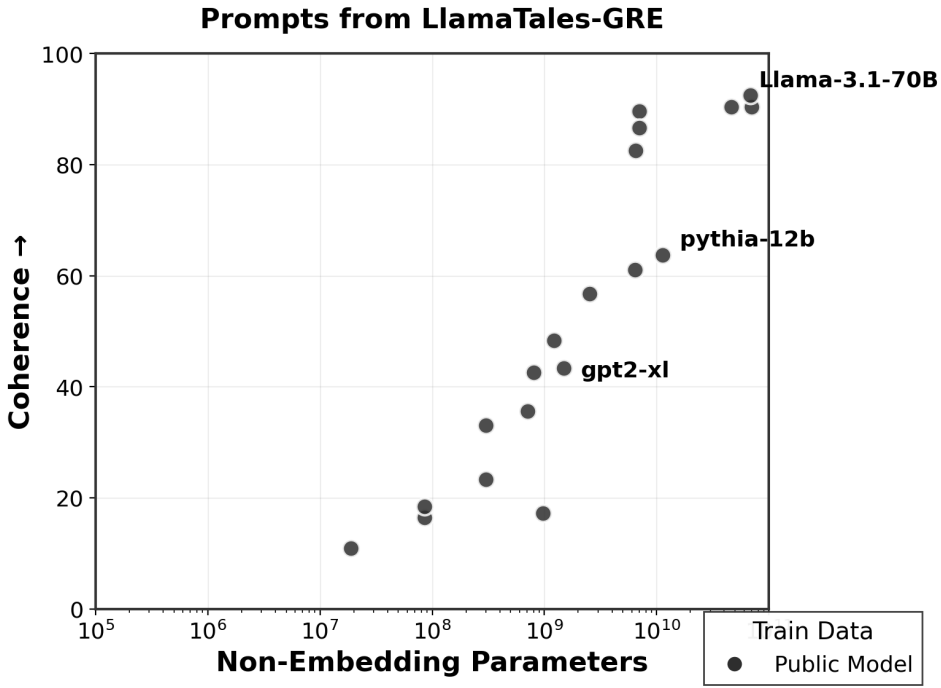
4 ✓ Train small language models on both datasets

5 ✓ Sample from models and measure fluency/coherence

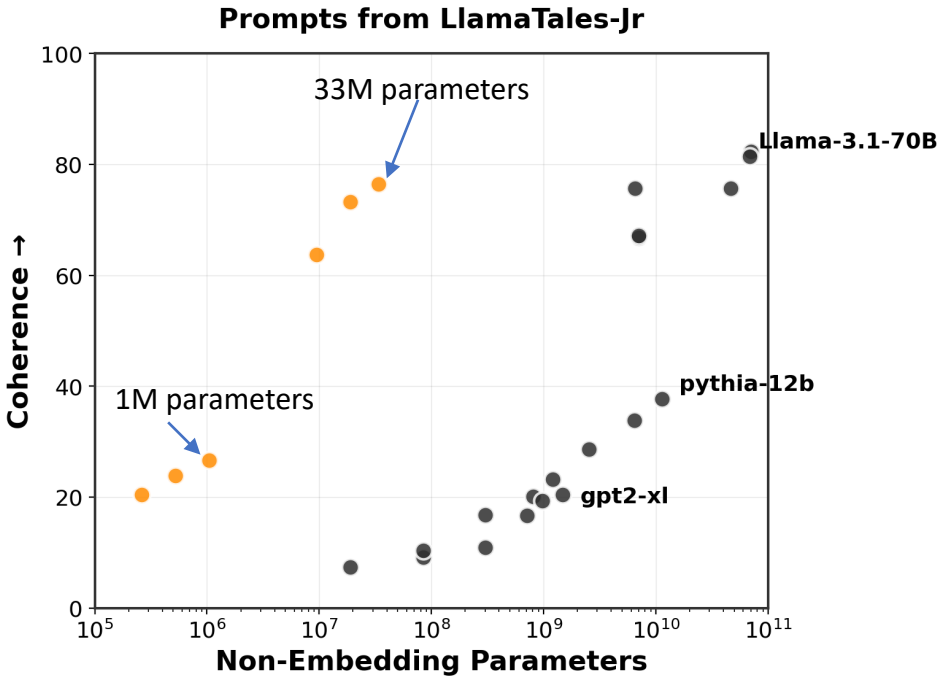
Ready to answer: Can we reproduce TinyStories' findings on data that does not leverage any child-directed properties?

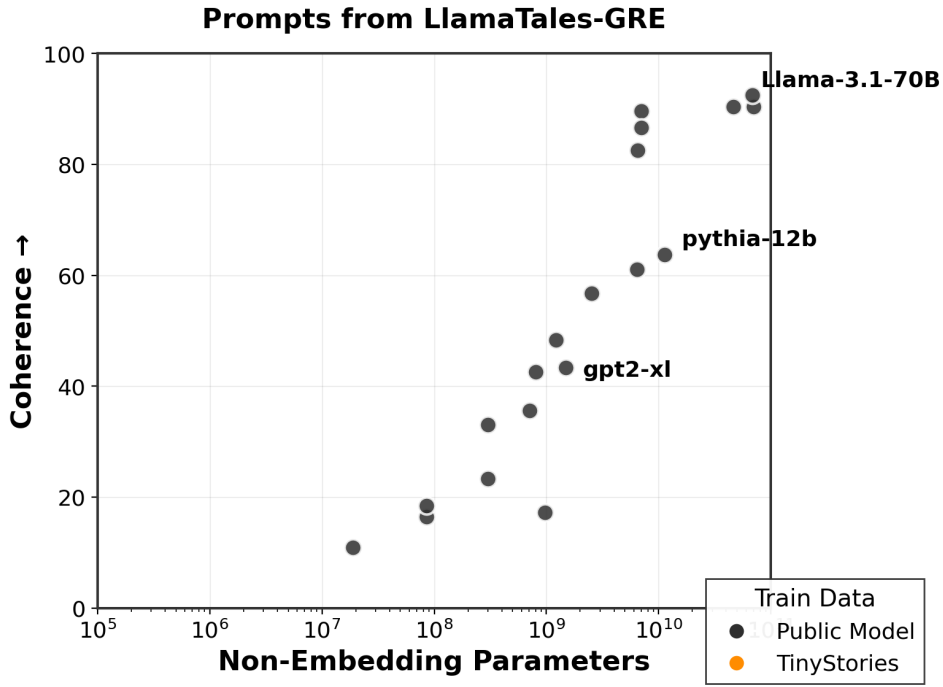
Coherence vs Parameter Count



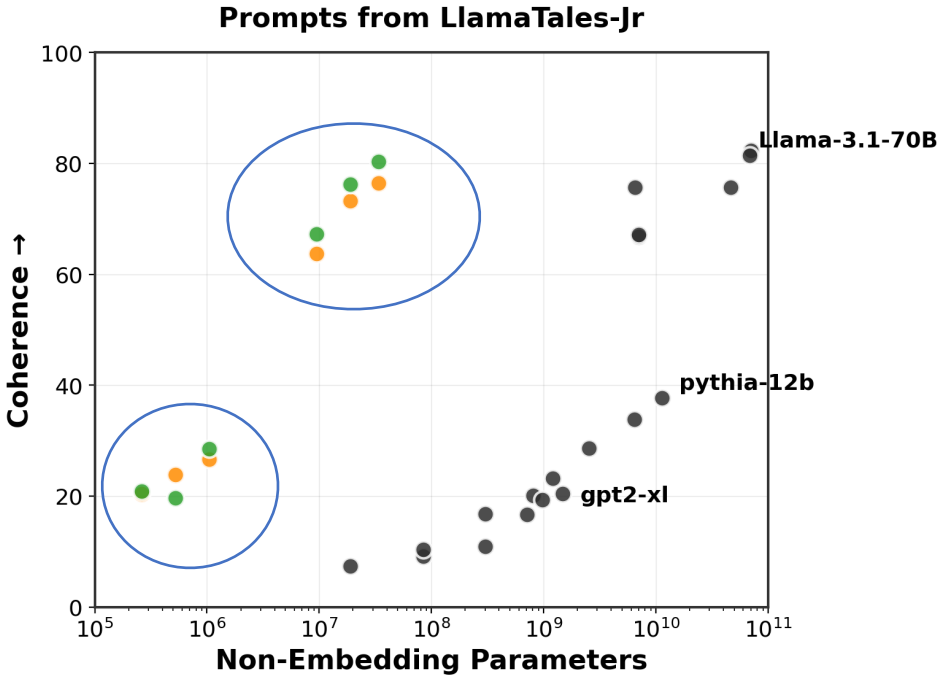


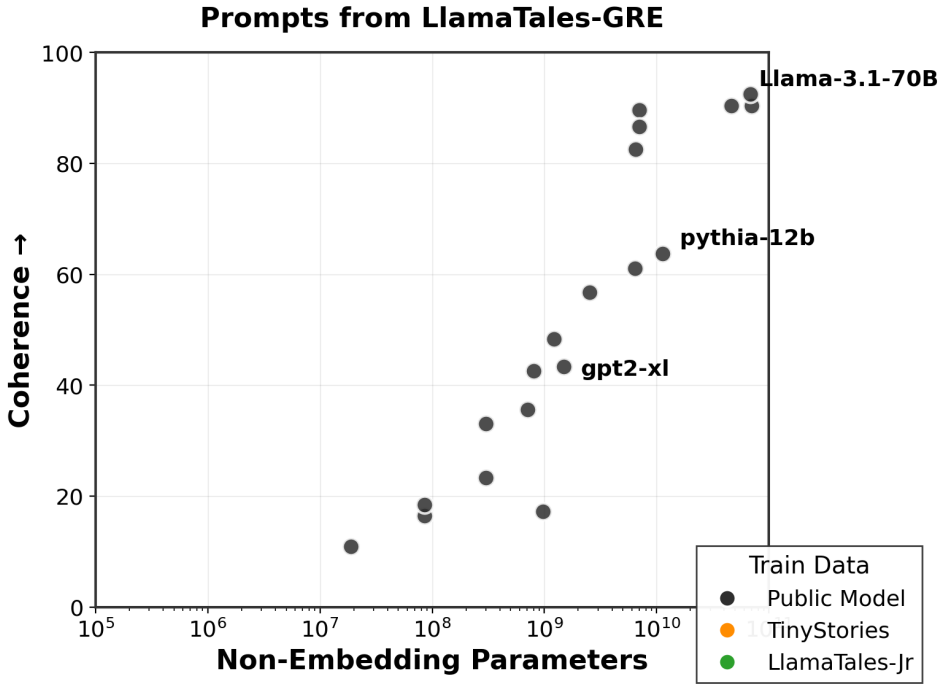
Reproduced TinyStories Findings



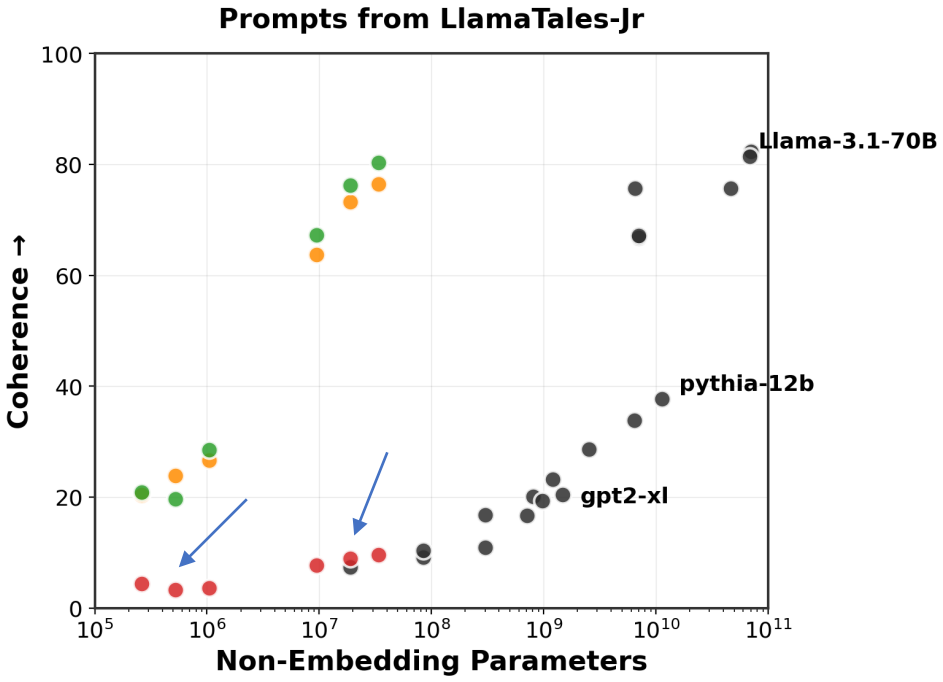


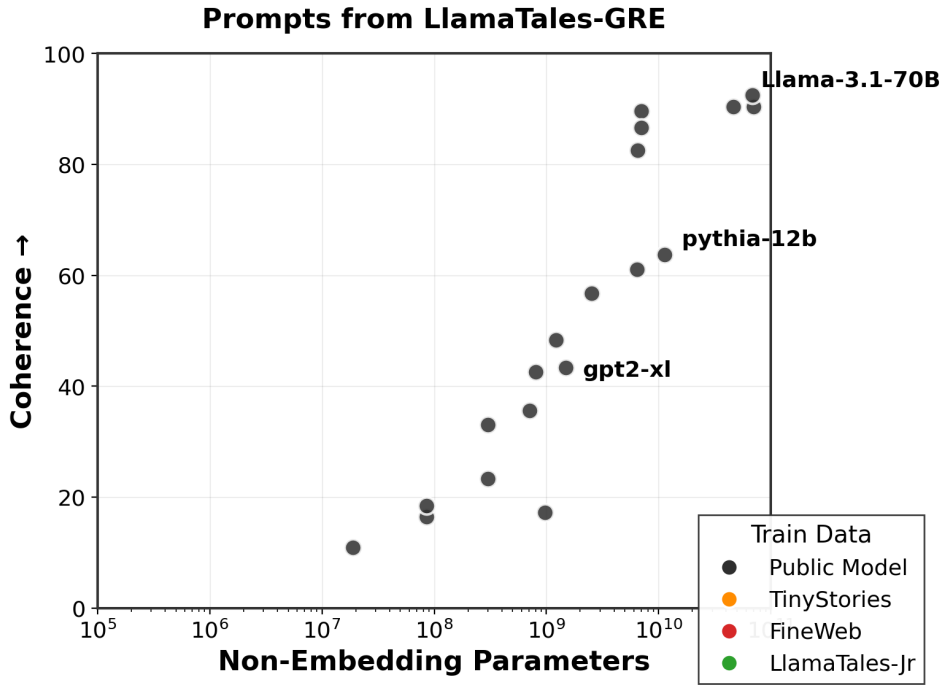
LlamaTales-Jr ≈ TinyStories



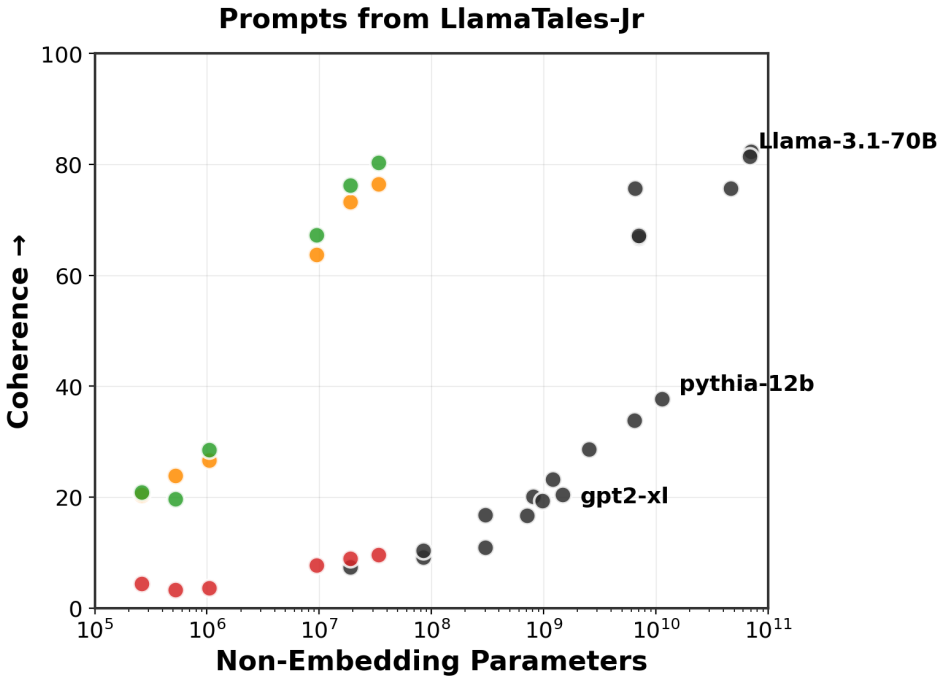


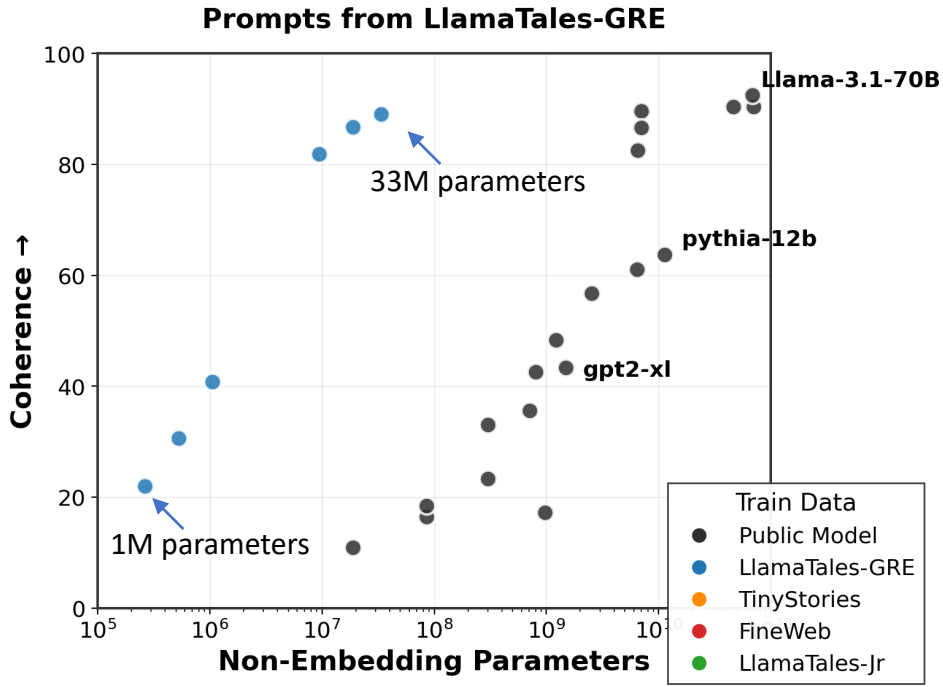
Small LMs trained on standard data struggle



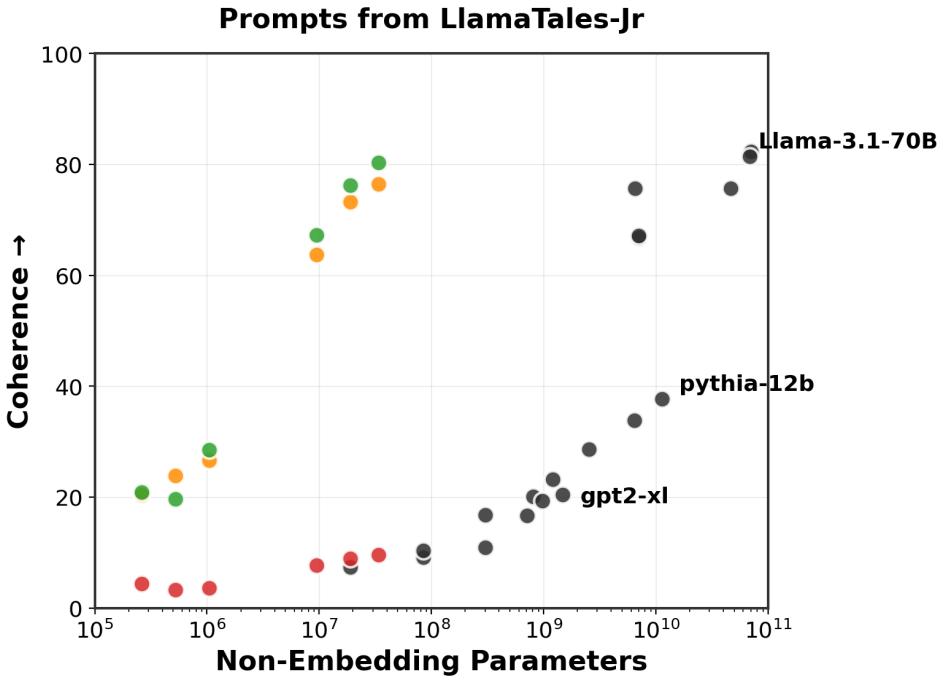


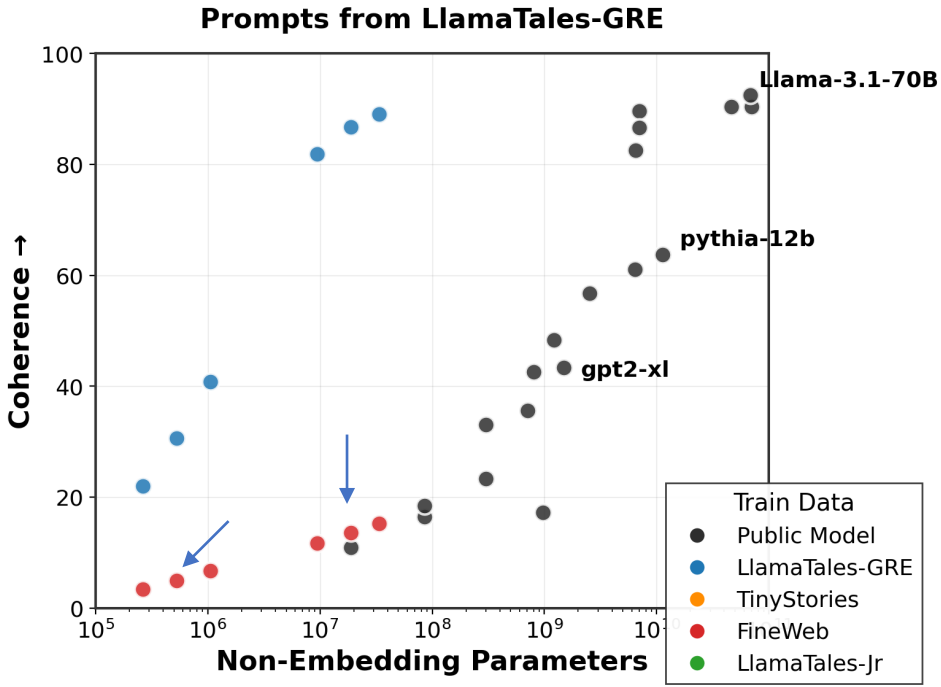
LlamaTales-GRE also coherent



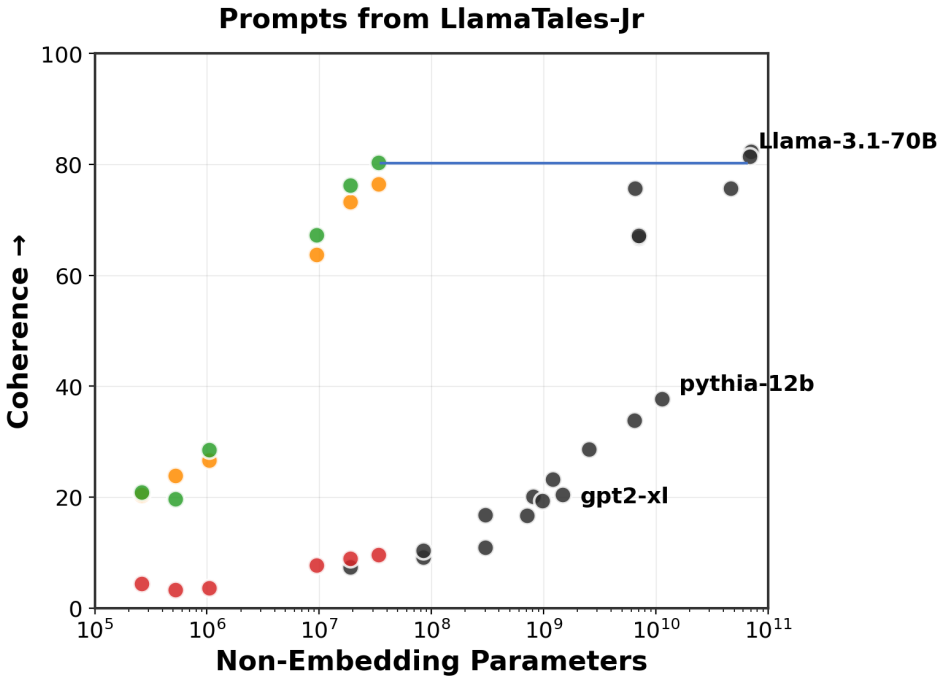


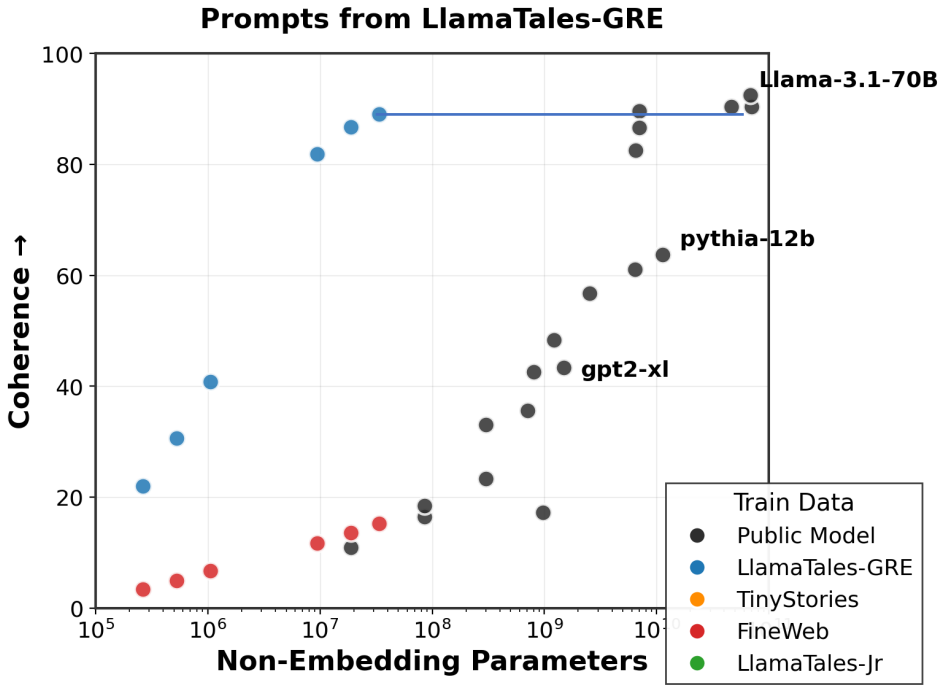
Small LMs trained on standard data struggle



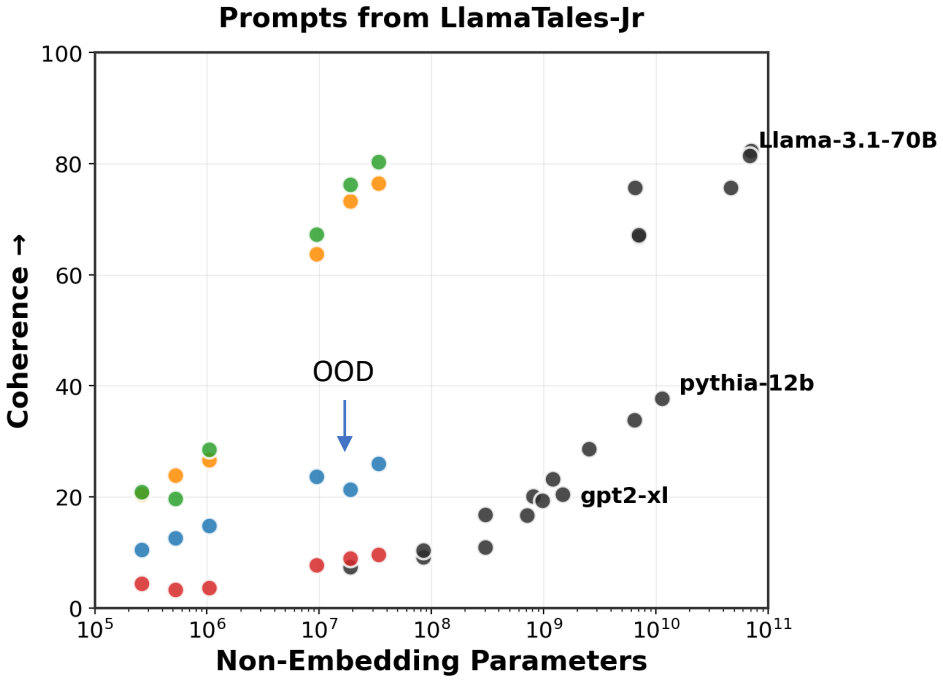


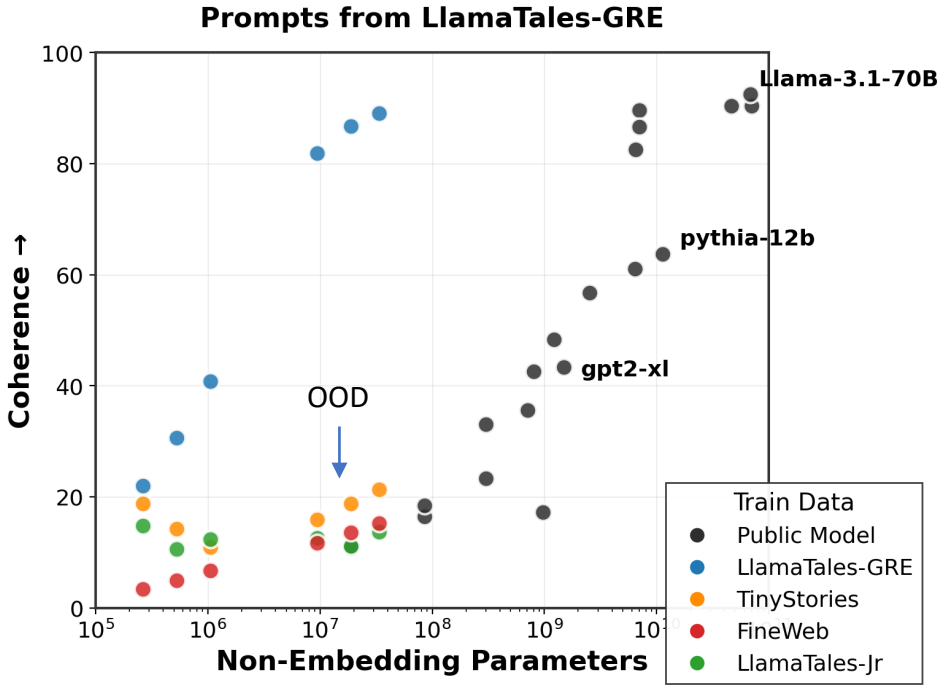
Finding 1: Readability Doesn't Predict Coherence





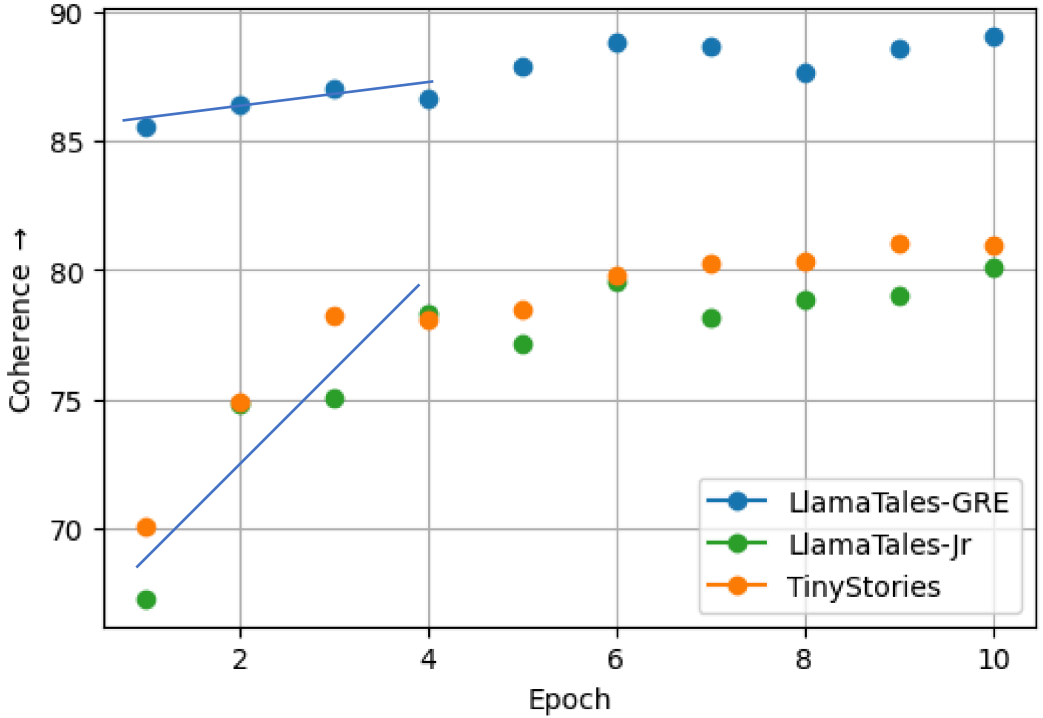
Coherence exists only in-domain





Finding 2: Readability Doesn't Speed Learning

Coherence across training epochs (one model, pattern holds across sizes)



Readability doesn't predict outcomes

- Not final quality
- Not learning speed

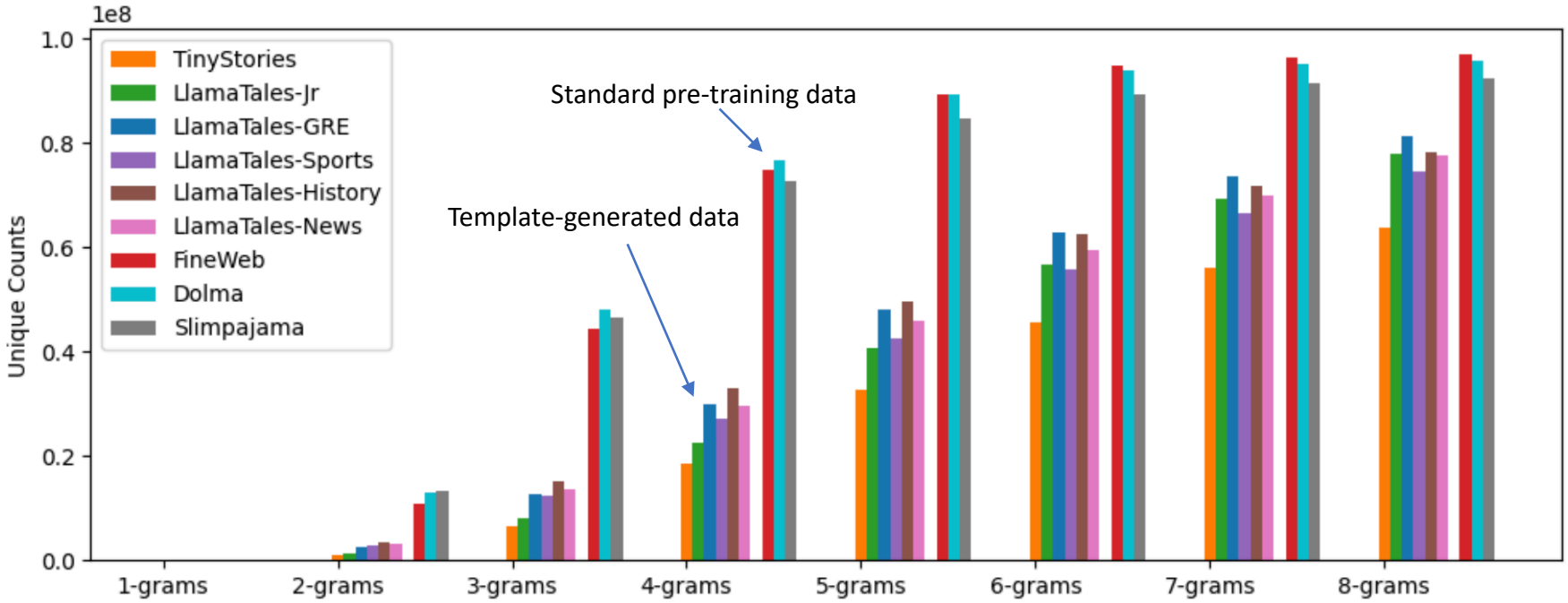
So what does?

TinyStories generation method:

- Template-based generation with minimal variation
- Creates high redundancy, narrow domain
- Crucially: fewer unique patterns to learn

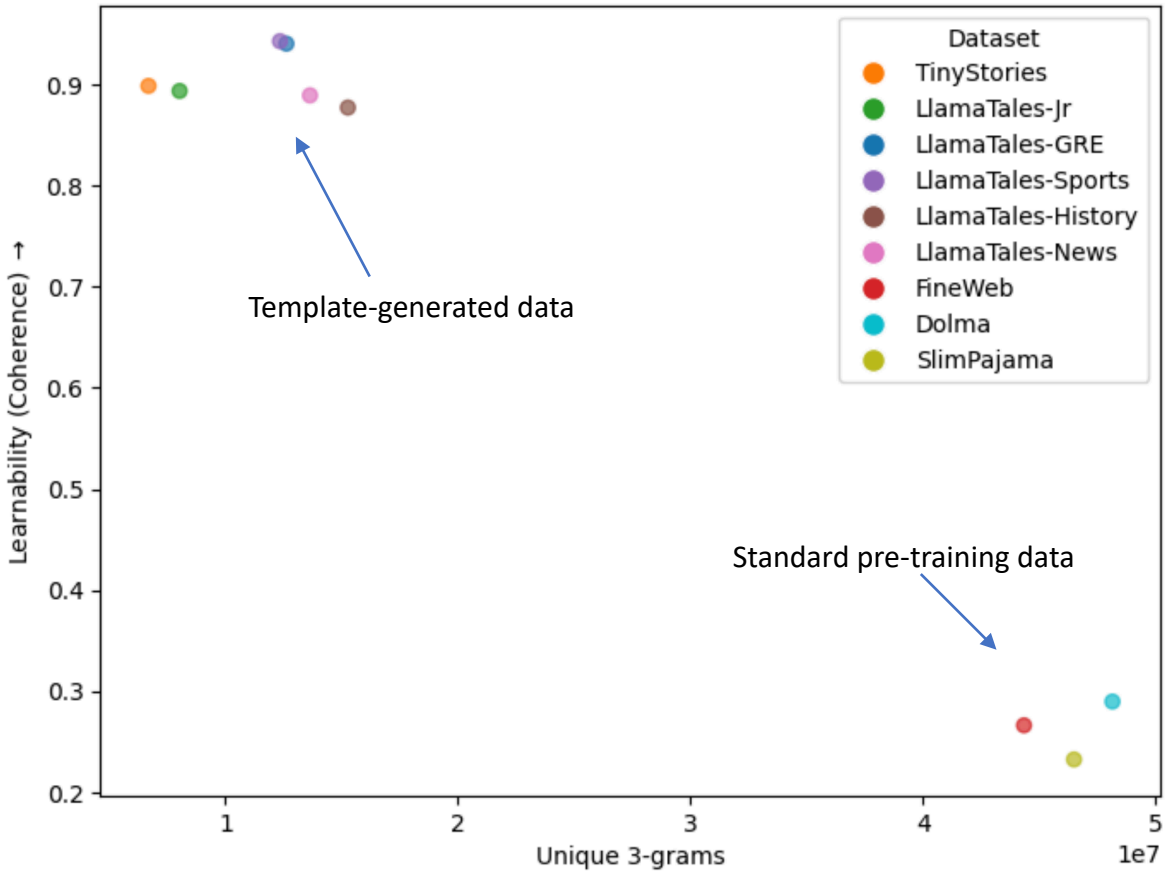
Hypothesis: N-gram diversity (proxy for pattern space size) might be a better predictor than readability

n-gram profiles



Finding 3: unique n-grams correlates with learnability

$$\text{Learnability} = \frac{\text{Model Output Coherence}}{\text{Training Data Coherence}}$$



Conclusion

Finding: Dataset learnability is not predicted by readability, but is strongly correlated with unique n-grams, so in other words: **readability ≠ learnability**

Principle: Statistical simplicity, not developmental simplicity

Broader implication: This is a useful case study. While developmental narratives are compelling, our findings suggest that focusing on underlying statistical properties offers a promising direction for understanding and creating learnable datasets

Thank You