Progressive Learning in LLMs Using Structured Grammar Books

NYU Center for Data Science

By: Airu Liu, Mustafa Poonawala, Devyani Hebbar, Rishabh Patil

Context

Introduction

The development of Large Language Models (LLMs) has significantly advanced natural language processing, enabling machines to perform tasks such as translation, summarization, and question-answering with remarkable proficiency.

Traditional LLM training approaches often rely on vast, unstructured datasets, necessitating substantial computational resources and time. However, recent research suggests that a more structured and efficient learning paradigm—one that mirrors human language acquisition—can enhance model performance while reducing training complexity.

In this paper, we propose a progressive training framework for LLMs that emulates human language acquisition through the pre-training of a sequence of grammar books and incorporating morphological word embeddings. This approach aims to improve model competence and efficiency by optimizing training methodologies through linguistic tagging and curriculum-based training

Related Work

Less is More

- **Minimum Description Length (MDL):** Optimal models balance complexity and generalization (Rissanen, 1978).
- **Data Efficiency:** Selective use of high-quality, diverse data improves generalization and reduces overfitting (Sorscher et al., 2022).
- **Domain-Specific Models:** Smaller, fine-tuned models (e.g., BioBERT, legal LLMs) often outperform large general-purpose ones in specialized tasks.
- **Efficient Architectures:** Techniques like pruning, compression, and parameter sharing (e.g., DistilBERT, ALBERT, DeepSeek) reduce model size without sacrificing performance.

Curriculum Learning

- **Human-Like Learning:** Traditional LLMs process each sequence in isolation, unlike human learners who build from simple sounds to complex sentences, yielding deeper contextual understanding
- **Curriculum Learning Framework:** Bengio et al. (2009) formalized "start small" training—ordering data from easy to hard—later extended by Narvekar et al. (2020) in reinforcement learning and surveyed broadly by Wang et al. (2020)
- **Sample Efficiency:** Warstadt (2023) showed that pretraining on developmentally plausible, child-like corpora achieves strong performance with far fewer data and compute resources
- **Adaptive Pacing Gains:** Introducing complex linguistic phenomena only after mastering simpler structures can halve the number of training epochs needed for comparable results
- **Enhanced Transferability:** Curriculum-trained models demonstrate stronger performance when transferred to diverse downstream tasks beyond their original training objectives
- **Stability and Robustness:** Models trained via curriculum learning exhibit lower variance across training runs and greater resilience to domain shifts

Methodology

Components

Datatset Preparation

1. Constructed from the **New Concept English textbook series** (Volumes 1–4)
2. Digitized and preprocessed one volume using **Tesseract** (OCR)
   - Correction of OCR errors like ligature fixes, punctuation standardization
   - Consistent formatting of dialogue, currency, and number formats
   - Structuring lessons with metadata tags with **Stanza**
3. **Trec** dataset for classification task

18067 sentences extracted

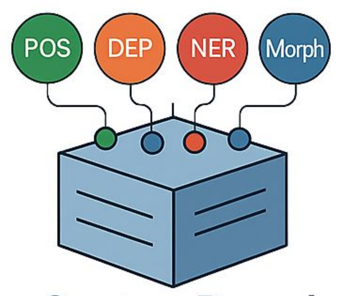
Syntactic Feature Extraction

- Part-of-Speech tags,
- Dependency relations,
- Named Entity Recognition tags, and
- Morphological features (tense, number).

(POS, DEP, NER, Morph)

1704 combinations

('NOUN','root','O','Number=Sing')
('VERB','parataxis','O','Mood=Imp|VerbForm=Fin')



Training Strategy

1. **Training from Scratch (No Weights)**
   - All models were **trained from scratch** with **randomly initialized weights**.
   - This setup removes any influence from pretrained models, isolating the effect of the grammar curriculum.
2. **Curriculum Pre-training**

Progressive training over **345 structured lessons** from advancing from simple to complex grammar

- Each lesson contains **20–70 sentences**
- Lessons are trained **in a loop**, reinforcing earlier structures while introducing new ones.
- Focus: build **syntactic competence** from scratch using pedagogically aligned, grammatically consistent content.

3. **Fine-Tuning on TREC**

Adaptation to the **TREC question classification** task with six coarse-grained labels :

DESC, ENTY, ABBR, HUM, LOC,NUM

- Fine-tunes pretrained models to map input questions to correct classes.

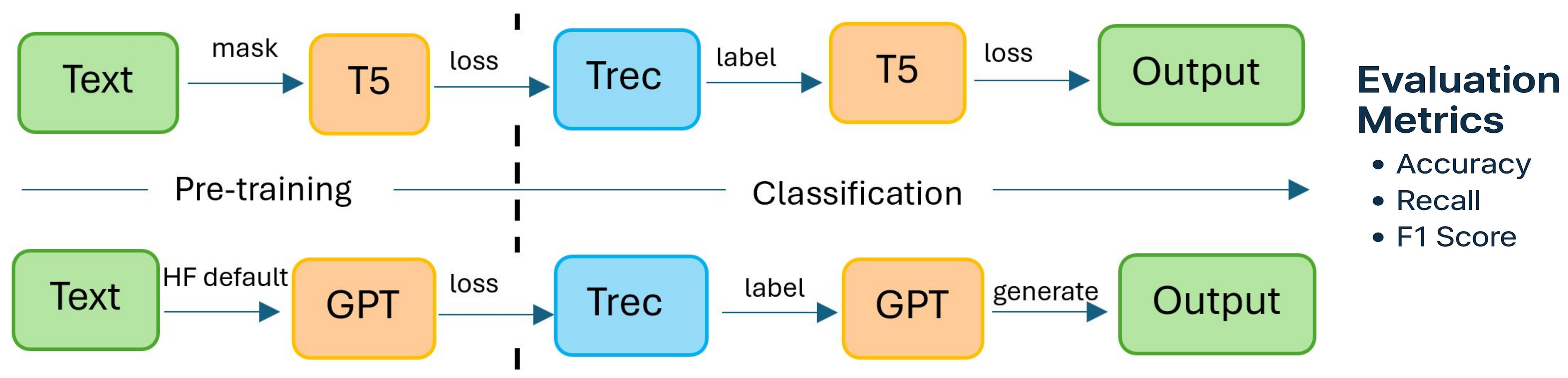
Model Selection: GPT-2 and T5-base

Model	Params (M)	Hidden Size	Layers	Heads
GPT-2	~117	768	12	12
T5-base	~220	768	12	12

Embedding Structure

Model	Original Embedding	With Syntax Embedding
GPT-2	Token + Position	Projected [Token // Syntax] + Position
T5-base	Token only	Projected [Token // Syntax]

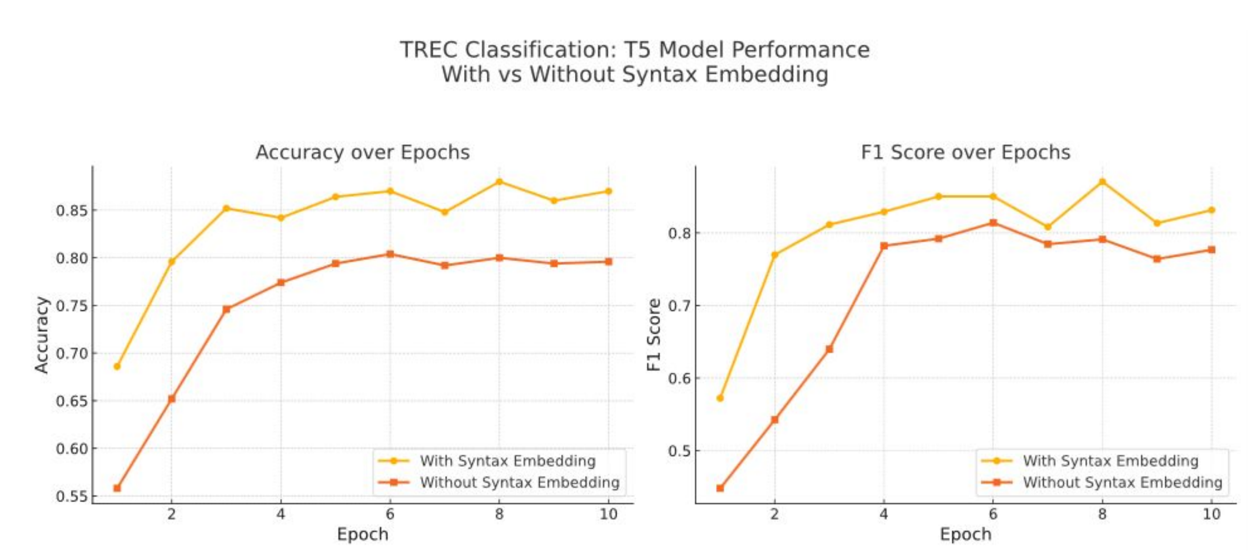
GPT-2 and T5-base Pipelines



Results & Discussion

Model	Accuracy	Recall	F1-score	Pretraining Time	Classification Duration	
					Training	Inference
GPT-2 (Pre-trained)	0.954	0.940	0.922	30 d	387.89 s	32.87 s
T5 (Pre-trained)	0.976	0.8476	0.8387	2.5 d	236.37 s	1.96 s
GPT-2 RAW + GRM + EMBs	0.636	0.201	0.223	1.25 h	288 s	
T5 RAW + GRM + EMBs	0.870	0.846	0.850	3 h	114 s	
Optimal RAW(GPT-2, T5) + GRM	*Details in Section 4.3*					

Table 3: Design of the Evaluation Models



GPT-2

- **SyntaxGPT** (trained on grammar curriculum) achieved **63% accuracy** on TREC, compared to 95.4% for randomly initialized GPT-2 trained directly on TREC.
- During curriculum training, SyntaxGPT showed **elevated eval losses** (5–7), far above typical classification ranges (1.5–2.5).

T5

- Accuracy improved from 68.6% to **88%** by epoch 8, and F1 reached 0.87, reflecting successful curriculum-based learning.
- **Validation loss decreased** steadily from 0.826 to ~0.6, indicating progressive mastery of linguistic structures.

Performance Comparison: TREC Classification

- T5 outperforms GPT-2 when trained from scratch with grammar embeddings and is more efficient

Syntax Embedding Ablation (T5)

- Models with **syntax embeddings outperformed baselines** in both accuracy and F1 on the TREC classification task.
- Syntax-enhanced models **converged more rapidly** and showed smoother improvement during training.
- Explicit syntax enabled the model to **capture deeper grammatical and semantic cues** beyond surface token patterns.

Limitations and Future Works

Limitations

- **Curriculum Complexity Saturation:** Gains plateau as curriculum complexity increases, suggesting diminishing returns from advanced grammar inputs without more semantic diversity
- **Limited Domain and Objective Scope:** Pretraining on structured textbook sentences restricts model exposure to real-world variability, hindering open-domain generalization
- **Data Scale Constraints:** A curriculum dataset of only a few thousand examples limits the learning of robust, transferable representations compared to large-scale corpora
- **Model Architecture Sensitivity:** Decoder-only GPT-2 struggles with limited, structured data, whereas encoder-decoder T5 shows greater resilience to small-scale pretraining

Future Works

- **Expand Linguistic Complexity:** Integrate advanced syntactic structures, semantic nuances, and discourse coherence into the curriculum
- **Incorporate Morphological Enrichment:** Leverage Wiktionary to introduce detailed word-form, inflection, and derivation information
- **Conduct Ablation Studies:** Systematically isolate the impact of grammar embeddings, curriculum stages, and input augmentation strategies
- **Optimize Training Parameters:** Fine-tune learning-rate schedules, optimizer settings, batch sizes, and curriculum granularity to enhance convergence
- **Apply to GEC Tasks:** Extend the progressive curriculum framework to grammatical error correction for deeper linguistic competence acquisition

Conclusion & References

Conclusion

In summary, syntax-augmented progressive pre training provides effective inductive biases that both accelerate convergence and improve classification accuracy of encoder–decoder models in low-resource scenarios. Our ablation experiments of T5-base confirm that integrating syntactic embeddings yields consistent gains, faster training dynamics and higher accuracy and F1 scores, compared to models trained without grammar information. While decoder-only architectures like GPT-2 remain more dependent on large-scale data, our results highlight the complementary role of structured linguistic cues alongside traditional pretraining. This framework offers a promising avenue for developing more data-efficient NLP systems that leverage explicit syntax.

References

[1] Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). Curriculum learning. In Proceedings of the 26th Annual International Conference on Machine Learning (pp. 41–48).

[2] Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. Cognition, 48(1), 71–99.

[3] Hale, J. (2001). A probabilistic Earley parser as a psycholinguistic model. In Proceedings of NAACL (pp. 1–8).

[4] Reali, F., & Christiansen, M. H. (2005). Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. Cognitive Science, 29(6), 1007–1028.