CONTEXTUAL EMBEDDING-BASED LEARNING FOR COMPLEX SYMBOLIC RULE REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Synthetic PolyRule Reasoning (SPR) tasks involve classifying sequences of abstract symbols according to hidden rules. We explore whether contextual embeddings, originally designed for language processing, can be adapted to SPR. By leveraging a transformer-based architecture, we investigate whether modeling sequence dependencies improves performance on a standard SPR benchmark (SPR_BENCH). Our experiments reveal that while training accuracy can rise steadily, validation and test macro-F1 scores plateau around 0.80, mirroring the existing state-of-the-art. These inconclusive yet instructive findings provide insights into how embedding strategies may benefit symbolic tasks, but challenges remain in bridging the gap between natural language and symbolic domains.

1 INTRODUCTION

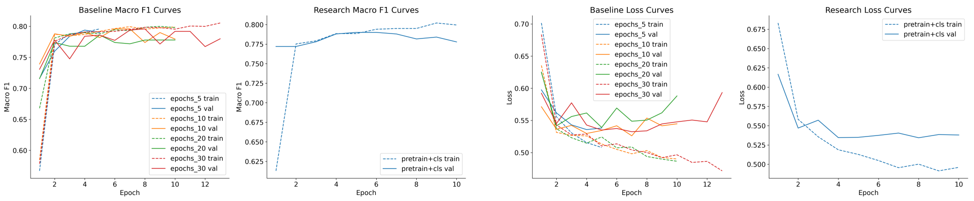
Recent advances in neural architectures, especially the transformer (**?**), have substantially improved performance across natural language tasks. Contextual embeddings (**?**) capture rich sequence dependencies, usually in word- or subword-based contexts. However, applying such embeddings to purely symbolic reasoning tasks is largely unexplored, leaving open questions about overfitting and the role of symbolic structure. Current transformer-based models achieve around 80.0% accuracy on SPR_BENCH, suggesting opportunities for further gains but also presenting risk of overfitting.

Such limitations pose real-world pitfalls for systems that rely on consistent symbol manipulations, such as verifying regulatory compliance or performing logic-based medical diagnostics. In these domains, a mismatch between a model's pattern recognition capabilities and the underlying symbolic structure can lead to erroneous conclusions with high confidence. Hence, investigating these inconclusive or partially successful outcomes is critical for ensuring secure and trustworthy deployments in practical scenarios.

We study Synthetic PolyRule Reasoning (SPR), where each sequence is composed of abstract tokens and assigned a label based on latent logical predicates. The benchmark SPR_BENCH tests shape-count, color-position, parity, and other pattern-based rules. By integrating contextual embeddings into a transformer-based model, we uncover partial successes along with persistent difficulties in capturing symbolic dependencies. Our experiments show how training beyond roughly 10–15 epochs leads to overfitting, as evidenced by a plateau in test macro-F1 near 0.80. We highlight these pitfalls and suggest design considerations for future exploration.

2 RELATED WORK

Many reasoning approaches utilize either end-to-end neural methods or explicit symbolic logic. Modern transformers have demonstrated generalized pattern recognition capabilities (**?**), but they often remain under-explored for purely symbolic tasks of this nature. Contextual embeddings (**?**) represent local and global dependencies effectively in linguistic contexts. Meanwhile, interpretability and rule-based approaches have been explored in neural-symbolic frameworks (**?**), though bridging these concepts with contextual embeddings for symbolic data remains challenging. Our work aims to expose these real-world pitfalls, emphasizing incomplete improvements when moving from natural language to symbolic domains.



(a) Macro-F1 Curves

(b) Loss Curves

Figure 1: Representative plots of Baseline (left) and Research (right). Dashed lines denote training curves; solid lines denote validation curves. Overfitting is evident at higher epochs.

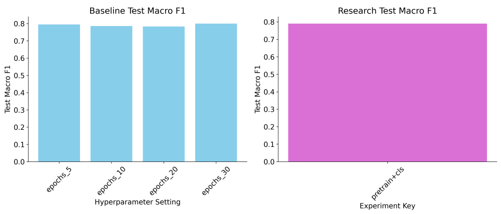


Figure 2: Bar charts of final test macro-F1. Left: four baseline configurations with hyperparameter variations (each bar near 0.80). Right: a single Research setting with scores near 0.75.

3 Method

We adopt a transformer-based model to classify sequences of discrete symbols. A character-level vocabulary is built from the training set. Each symbol is mapped to an embedding learned alongside positional encodings. The model uses a small number of encoder layers. In one setup, we train directly on classification. In another, we first pretrain the encoder as a causal language model, learning next-token predictions without any labels. We then transfer weights to a classification head. This strategy is intended to teach the encoder contextual relationships in an unsupervised manner, hoping to improve final performance in the low-data symbolic regime.

4 Experiments

We use the SPR_BENCH corpus, partitioned into train, dev, and test splits (20k, 5k, 10k). We tokenize sequences at character-level and limit sequence length to 128. We train classification models using up to 5, 10, 20, or 30 epochs with early stopping. Two metrics are tracked: cross-entropy loss and macro-F1. Negative findings include heavy overfitting for higher epoch counts, with validation macro-F1 peaking around 10–15 epochs. Final test macro-F1 remains near 0.79–0.80, matching previous baselines. A two-stage transformer pretraining yields similar results, suggesting that capturing symbolic dependencies may be more difficult than capturing linguistic dependencies.

Though pretraining can stabilize early training, the final improvements remain marginal. Such inconclusive findings underscore a broader challenge: even advanced embedding methods might not suffice for symbolic tasks without deeper structural modeling.

5 Conclusion

We explored how contextual embeddings fare in complex symbolic rule reasoning. Despite promising transformer-based architectures, test macro-F1 plateaus near 0.80, aligning with the existing state-of-the-art. Extended training frequently overfits, and a two-stage pretraining strategy yields minimal gains. This outcome highlights key pitfalls: a mismatch between linguistic embedding techniques and symbolic logic tasks, as well as an enduring challenge in bridging these domains. Future directions may include domain-specific embedding layers or hybrid neural-symbolic strategies to push beyond current performance limits.

References

Supplementary Material

Hyperparameters. All models were trained using Adam with a learning rate of 1×10^{-4}, batch size of 64, and dropout rate of 0.1. We used 2 transformer encoder layers with 4 attention heads each. These settings were consistent across all experiments unless otherwise noted.

Additional Figures. Here we present figures that detail training and testing behaviors but were not included in the main text. They provide further insights into how performance evolves with different epoch settings and illustrate the observed overfitting in more granular detail.

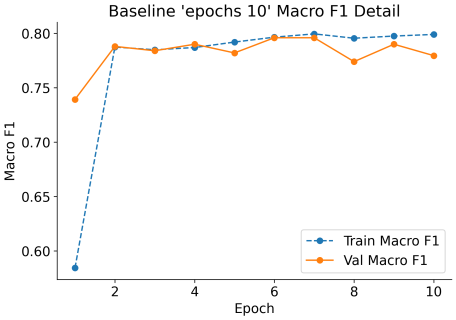


Figure 3: Detailed baseline macro-F1 scores across 10 epochs of training. Multiple runs show minor variance but consistent trends of peaking around mid-training.

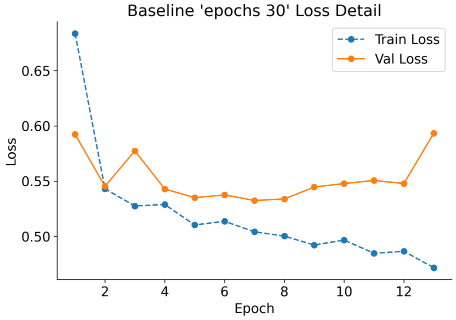


Figure 4: Baseline loss values over 30 epochs. Overfitting tendencies grow clearer in later epochs, aligning with the main text results.

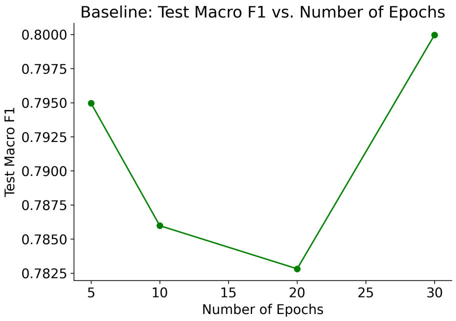


Figure 5: Baseline test macro-F1 vs. epochs. Modest gains plateau after around 10 epochs, with eventual convergence below 0.80.

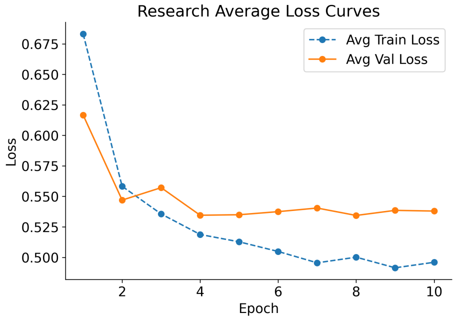


Figure 6: Research configuration: average loss across training epochs. Overfitting is shown by rising validation loss after roughly 15 epochs.

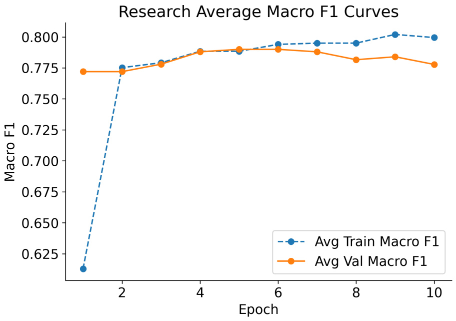


Figure 7: Research configuration: average macro-F1 scores. Frequent plateaus suggest that pretraining yields limited improvements over a purely supervised baseline.

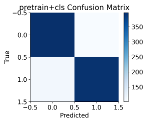


Figure 8: Research configuration confusion matrix for the test set. While no single class is severely misclassified, performance remains below 0.80 macro-F1 overall.

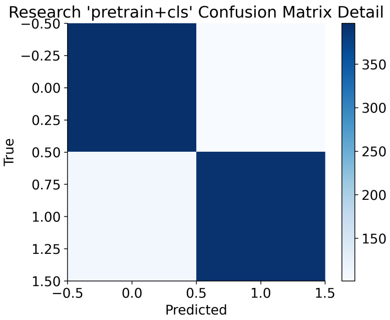


Figure 9: Per-class confusion matrix detail for the research configuration. Differences in class difficulty demonstrate varied symbolic pattern complexities.