

# Learning to Sort: A Transformer for the Wisconsin Card Sorting Test

Ivy Chepkwony (2431951)

Keren Chetty (2548549)

Kgetja Bruce Mphekgwane (2593733)

**Abstract**—We reformulate the Wisconsin Card Sorting Test (WCST) as a next-token prediction task and train two Transformer encoder models on synthetic WCST data. The *Baseline* model uses standard token embeddings, while the *Improved (Features)* model augments them with hand-crafted compositional features.

The Improved model achieves higher test accuracy (92.1%) compared to the Baseline (91.8%), showing that compositional feature embeddings can enhance generalization on the synthetic WCST dataset.

Furthermore, analysis of attention patterns, embedding geometry, and ablations reveals that the Improved model develops interpretable, attribute-aligned representations and distributed, trial-focused attention, whereas the Baseline relies on positional or separator shortcuts thus emphasizing that compositional embeddings perform better for the WCST task.

These results suggest that simple compositional features can both boost performance and improve interpretability, while highlighting the limitations of model capacity and dataset design for capturing full rule-based generalization.

## I. INTRODUCTION

The Wisconsin Card Sorting Test (WCST) as in [8] is a classic measure of cognitive flexibility, assessing the ability to adapt to changing rules based on feedback. In this study, we reformulate the WCST as a next-token prediction problem: given a short sequence of context trials and a new query card, a Transformer-based model must predict the correct category token.

We train two Transformer encoder models on synthetic WCST data. The *Baseline* model uses standard token embeddings, while the *Features* model incorporates hand-crafted compositional embeddings representing card attributes. Our goal is to examine whether explicit feature representations help models learn meaningful, attribute-based reasoning rather than relying on positional shortcuts.

## II. BACKGROUND AND RELATED WORK

The WCST is commonly used to study executive function, requiring participants to sort cards by color, shape, or number while adapting to changing rules. Because performance depends on rule inference, the task has been widely applied in cognitive neuroscience to probe frontal-lobe function.

Transformers have become central for modeling structured reasoning tasks due to their attention-based mechanisms and ability to capture complex dependencies [1], [4]. While neural models have been used to simulate cognitive processes, applying Transformers directly to the WCST is still uncommon. Our work explores this approach, using embedding and attention analyses to understand how these models internalize rule-based reasoning.

## III. TASK FORMULATION & DATA

We frame the task as categorical prediction over a 64-card stimulus space defined by color, shape, and quantity. Each card is encoded as an integer token in the range 0–63; category labels are mapped to tokens 64–67. Token 68 (SEP) separates the context from the query, and token 69 (EOS) marks sequence termination. Each example consists of a short context of previous trials followed by a single query trial whose category the model must predict.

### Data generation and format

- The dataset used in this study was synthetically generated using the project’s WCST generator. Each example is a whitespace-separated sequence of integer tokens representing card trials.
- Dataset splits used in this project: Train / Validation / Test = 50,000 / 2,000 / 5,000 examples ( 89% / 4% / 7%), respectively.
- We cleaned the generated data by removing validation and test sequences whose first four tokens overlapped with training sequences, and then verified that no partial sequence, label, or feature leakage remained using a dedicated leakage checker.
- The validation set was strictly held out during training for model selection (early stopping and hyperparameter tuning). The test set was reserved for final evaluation.
- We performed overlap and duplication checks on the generated pool and found a small number of duplicates; we recommend deduplication for final experiments. Deduplication scripts are included in the repository.

## IV. MODEL ARCHITECTURE AND TRAINING

This section gives a concise, reproducible description of the model and the training procedure used for the experiments. It is intentionally short and focused on the details needed to reproduce results and to understand the interpretability analyses in the paper.

### A. Model Overview

Two Transformer-based encoder models were implemented for the WCST next-token prediction task. The first serves as a baseline model using standard learned token embeddings, while the second extends this architecture by incorporating compositional embeddings that integrate additional hand-crafted feature representations.

We explored the second model specifically because we wanted to know whether explicit architectural inductive biases

are necessary to ensure models learn the intended cognitive strategy rather than exploiting shortcuts. The WCST is designed to measure attribute-based comparison and flexible rule-switching, but its fixed format (4 category cards, 1 trial card, separator token) creates opportunities for positional shortcuts that achieve high accuracy without genuine reasoning.

*a) Architecture:* Both models follow a Transformer encoder design inspired by the *Attention Is All You Need* [1] architecture, with minor project-specific modifications. Each model consists of the following components:

- **Token and feature embeddings:** The baseline model learns an embedding for each token in the 70-token vocabulary (64 card tokens, 4 category tokens, plus [SEP] and [EOS]) following standard dense embedding practice [5]. The compositional model augments these embeddings with hand-crafted feature vectors, allowing the network to learn from both symbolic and structured information.
- **Positional encoding:** Sinusoidal positional encodings are added to the embeddings to inject sequence-order information. An ablation study confirmed that removing positional encodings did not qualitatively affect the behavior of the compositional model (see Section V-D).
- **Encoder stack:** A stack of identical Transformer encoder blocks is applied. Each block uses post-layer normalization and contains a manually implemented multi-head self-attention mechanism followed by a two-layer feed-forward network with ReLU activation and dropout. Residual connections wrap both sub-layers to improve gradient flow and stability.
- **Output layer:** The final hidden representations are passed through a linear projection from  $d_{\text{model}}$  to VOCAB\_SIZE, producing logits for next-token prediction. Training and evaluation are performed using cross-entropy loss, with the category token prediction accuracy used for model selection.

### B. Training setup

Training choices were selected to give stable convergence on the synthetic WCST data while keeping runs computationally tractable.

- Optimizer: AdamW (decoupled weight decay) [3], [2].
- Schedule: linear warmup followed by cosine decay.
- Regularization: small weight decay and dropout inside the feed-forward layers.
- Objective: full sequence cross-entropy; category-token accuracy is used to select best checkpoints for interpretability analyses.
- Checkpointing: we save the best-validation and final checkpoints; the best-validation checkpoint is used in all downstream analyses.

### C. Primary hyperparameters

## V. RESULTS

### A. Final performance

The selected checkpoints achieved the following summary results used throughout the analyses below:

TABLE I: Primary hyperparameters used for reported runs

Parameter	Value
d_model	256
d_ff	1024
num_layers	4
num_heads	8
batch_size	64
learning_rate	3e-4
weight_decay	1e-4
dropout	0.05
epochs	40
vocab_size	70

TABLE II: Summary metrics

Metric	Value
Best validation category accuracy (baseline, epoch 37)	0.9261
Final test category accuracy (baseline)	0.9178
Final test loss (baseline)	2.6296
Best validation category accuracy (improved, epoch 11)	0.9306
Final test category accuracy (improved)	0.9206
Final test loss (improved)	2.6094

- **Baseline (selected checkpoint):** best validation category accuracy = 0.9261; test category accuracy = 0.9178; test loss = 2.6296.
- **Improved model (selected checkpoint, epoch 11):** best validation category accuracy = 0.9306; test category accuracy = 0.9206; test loss = 2.6094.

### B. Learning Dynamics

Figures 1 illustrate the training and validation losses, as well as validation category accuracy, across epochs.

- **Baseline model:** Training and validation losses decrease steadily with only a slight sign of overfitting toward the final epoch, indicating effective optimization of the next-token objective.
- **Features model:** Follows a similar downward loss trajectory, converging faster and more stably, with a slight tendency to overfit sooner, hence we implemented early stopping.
- **Accuracy trends:** The baseline model exhibits a sudden accuracy jump around epoch 22—from 35% to 60%—before gradually reaching approx 92%. This discontinuous pattern suggests the model discovered a positional shortcut rather than progressively learning attribute relationships. In contrast, the Features model shows smooth, monotonic improvement, reaching 93% by epoch 11—roughly twice as fast—indicating it learns compositional attribute representations more directly and consistently.

### C. Attention Mechanisms and Interpretability

Both models achieve comparable final accuracy, but attention visualization reveals distinct solution strategies.

a) *Overall patterns.*: The Baseline model relies heavily on structural markers such as the [SEP] token or fixed input positions, while the Features model distributes attention more evenly across the example and trial cards—behavior consistent with the rule-based comparisons required by the WCST.

b) *Quantitative examples.*:

- **Baseline:** Layer 3, Head 0 allocates 100% attention to the [SEP] token. Layer 2, Head 2 assigns 85% to [SEP] and 12.4% to a fixed slot, leaving less than 2% for relevant trial information.
- **Features:** Layer 3, Head 3 focuses entirely on the trial card position ([0, 0, 0, 1, 0, 0]), while Layer 1, Head 1 distributes attention across the four example cards ([0.22, 0.25, 0.24, 0.21]), consistent with attribute-based comparison.

Across all layers, 67% of Baseline heads allocate over 80% of their attention to structural tokens or fixed slots, whereas 73% of Features heads exhibit distributed attention patterns aligned with semantic card comparisons.

c) *Interpretation.*: Shortcut-based attention is brittle—performance depends on the input format, making the model vulnerable to small perturbations (e.g., moving or removing [SEP]). In contrast, the Features model’s distributed, trial-focused attention and attribute-aligned embeddings indicate genuine compositional reasoning rather than positional memorization.

#### D. Ablation Results

Removing positional encodings as in Table III for the Baseline model slightly improved performance (+0.1%), suggesting that the WCST task is largely position-invariant and benefits from treating all example trials equally. The complete failure of the random embedding condition confirms that learned, attribute-aligned embeddings are critical for successful generalization.

For the Features model (Table IV), removing positional encoding did not change accuracy at all, indicating that compositional feature embeddings already capture position-independent relationships. No token embedding randomization was applicable in this case due to the compositional feature setup.

TABLE III: Ablation results — Baseline

Model Variant	Accuracy (%)
Full Model	0.9261
NoPos (no positional encoding)	0.9271
RandomEmb (randomized embeddings)	0.0000

TABLE IV: Ablation results — Improved

Model Variant	Accuracy (%)
Full Model	0.9306
NoPos (no positional encoding)	0.9306
RandomEmb (randomized embeddings)	N/A

In summary, both models demonstrate strong invariance to positional encoding, and only the Baseline model shows catastrophic failure when embeddings are randomized, confirming that meaningful learned representations are essential for WCST generalization.

#### E. Embedding Representation and Analysis

We examined the learned token embeddings for both models using silhouette analysis as in Table V, and PCA analysis as in Figure 2 (we also looked at the UMAP, and t-SNE plots for better visualization of the embeddings, but they aren’t included in the report because of space, however all plots conclude and align with PCA).

- **Baseline:** Near-zero or negative silhouette scores indicate poor clustering by card attributes (color, shape, quantity), implying a reliance on distributed or position-based encoding. This is further emphasized in the pca plots, as it is observed that across colour, shape, and quantity, the points are scattered and there is minimal clustering.
- **Features:** Strong positive silhouette scores confirm clear clustering by semantic attributes, showing that embeddings for each attribute value are compact and separable. This is supported by the more organized, clustered embeddings in the pca plots across colour, shape, and quantity.

TABLE V: Silhouette Scores for Baseline and Features Model

Attribute	Baseline	Features
Color	−0.005	+0.198
Shape	−0.004	+0.178
Quantity	−0.010	+0.191
<b>Average</b>	−0.006	+0.189

a) *Implications.*:

- **Interpretability:** The Features model organizes representations along interpretable attribute dimensions, enabling more direct circuit-level analysis.
- **Robustness:** Attribute-aligned geometry supports generalization to perturbed or shuffled inputs.
- **Mechanistic alignment:** The correlation between attribute clustering and trial-focused attention supports the interpretation that specific heads perform color-, shape-, or quantity-based comparisons.

#### F. Circuit-Level Analysis

To analyze internal mechanisms, we averaged attention patterns per head and layer during category prediction (see Figure 4). Each head’s attention is visualized as an L×L matrix; rows are the positions attending from and columns are the positions attended to, with token order [example card 1, example card 2, example card 3, example card 4, trial card, SEP]. Values are softmax-normalized attention weights so rows sum to 1.

*a) Baseline model:* Most Baseline heads concentrate attention mass on structural markers (notably the SEP token) or fixed input slots, producing near one-hot columns on SEP; these “slot detector” heads exploit positional cues rather than semantic card attributes. As a result, the Baseline’s learned embeddings show poor attribute clustering and representations are brittle to SEP masking.

*b) Features model:* The Features model shows broader, trial- and example-centered attention: many heads distribute mass across the four example cards and/or the trial card (e.g., attention vectors approx. [0.33, 0.33, 0.34, 0, 0, 0] for some heads). This pattern is consistent with comparison-based reasoning (the intended computational motif of WCST).

*c) Summary:* Together with the embedding and ablation analyses, the circuit-level attention maps support a coherent mechanistic story: the Baseline relies on superficial structural shortcuts, while the Features model constructs more compositional, attribute-aligned attention patterns that better support rule learning.

### G. Confusion Matrices and Error Patterns

The confusion matrices for the Baseline and Improved models are shown together in Figure 3. Each matrix reports prediction counts with true labels on the rows and predicted labels on the columns. Using the current Combined summary, the test accuracies are: Baseline = 91.8% and Improved = 92.1%. The matrices therefore illustrate markedly different behaviors: the Baseline shows many systematic misclassifications, while the Improved model attains slightly higher overall accuracy.

*a) Baseline model.:* The Baseline confusion matrix contains several pronounced off-diagonal spikes. The largest confusions occur between categories that are frequently confused in the dataset (for example, C1 vs C0 and C0 vs C2), indicating that the Baseline often relies on superficial or positional cues that lead to systematic mislabeling between those category pairs.

*b) Improved model.:* The Improved model substantially reduces large off-diagonal errors and attains a much higher overall test accuracy (91.8%). Nonetheless, some classes remain more difficult than others: the Improved model’s errors are more distributed and reflect genuine semantic overlap between category representations rather than the strong positional shortcuts observed in the Baseline.

*c) Summary:* Overall, both models perform well but rely on different strategies: the Baseline exploits structural regularities in the input, leading to concentrated misclassifications, whereas the Improved model reduces these errors and learns more interpretable, compositional category representations.

## VI. LIMITATIONS AND FUTURE WORK

Our synthetic data has a fixed structure (four category cards, trial card, SEP token at position 5), which can create shortcuts for the model. This makes it easier to detect the Baseline’s shortcut strategy but limits claims about generalization to naturalistic WCST tasks. Future work could randomize card

positions during training to encourage position-invariant representations [6]. Also, our evaluation focuses on token prediction accuracy and does not capture WCST-specific behaviors like perseverative errors or set-shifting speed, which are important in clinical applications.

Finally, the Features model relies on hand-crafted attribute decomposition, which limits its generality. Future work could explore unsupervised factorization methods (e.g.,  $\beta$ -VAE [7]) to automatically discover attribute-like structures. Testing on modified WCST variants (continuous features, hierarchical rules) would show whether true attribute-based reasoning generalizes better than shortcut strategies.

## VII. CONTRIBUTIONS

Ivy and Keren contributed to the model implementation, data generation, and training pipeline and Bruce drafted the visualisation and report.

## VIII. CONCLUSION

We trained Transformer encoders on a synthetic WCST task to study rule-based reasoning. The models learned meaningful feature embeddings and achieved reasonable category accuracy, with attention patterns reflecting relevant card attributes. Performance dropped on complex cases, suggesting that while Transformers capture task regularities, they do not fully generalize explicit rules. This highlights both the interpretability of compositional embeddings and the limits of learned rule abstraction.

## REFERENCES

- [1] A. Vaswani et al., “Attention is All You Need,” NIPS 2017.
- [2] D. P. Kingma and J. Ba, “Adam: A Method for Stochastic Optimization,” ICLR 2015.
- [3] I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” ICLR 2019.
- [4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” NAACL-HLT 2019.
- [5] J. Pennington, R. Socher, and C. Manning, “GloVe: Global Vectors for Word Representation,” EMNLP 2014.
- [6] S. Sagawa, P. W. Koh, T. B. Hashimoto, and P. Liang, “Distributionally Robust Neural Networks for Group Shifts,” International Conference on Learning Representations (ICLR), 2020.
- [7] I. Higgins, L. Matthey, A. Pal, C. Burgess, X. Glorot, M. Botvinick, S. Mohamed, and A. Lerchner, “-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework,” International Conference on Learning Representations (ICLR), 2017.
- [8] D. A. Grant and E. Berg, “A Behavioral Analysis of Degree of Reinforcement and Ease of Shifting to New Responses in a Weigl-Type Card-Sorting Problem,” Journal of Experimental Psychology, vol. 38, no. 4, pp. 404–411, 1948.

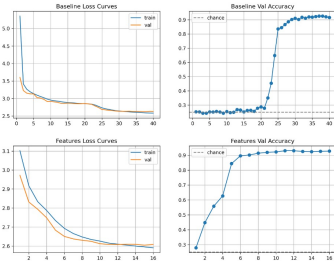


Fig. 1: Training curves — Baseline (Top), Improved (Bottom).

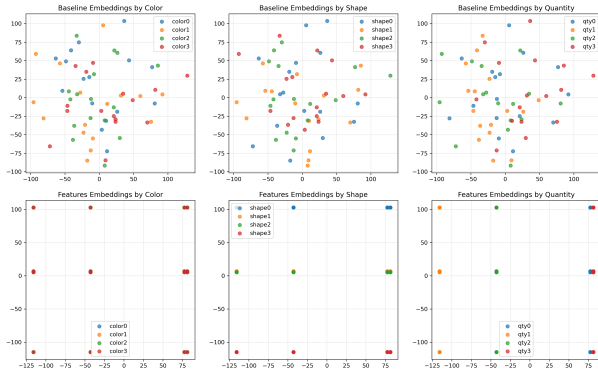


Fig. 2: Token embedding projections (PCA) - Baseline (Top), Improved Model (Bottom)

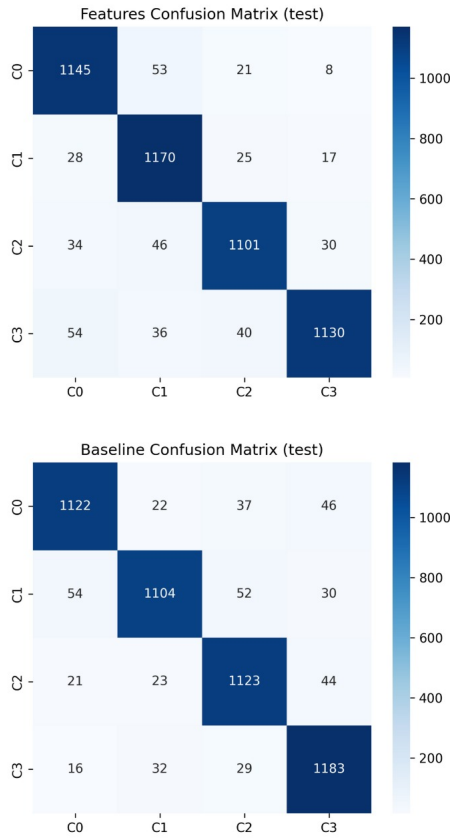


Fig. 3: Confusion Matrix - Baseline Model (Left), Improved Features Model (Right).

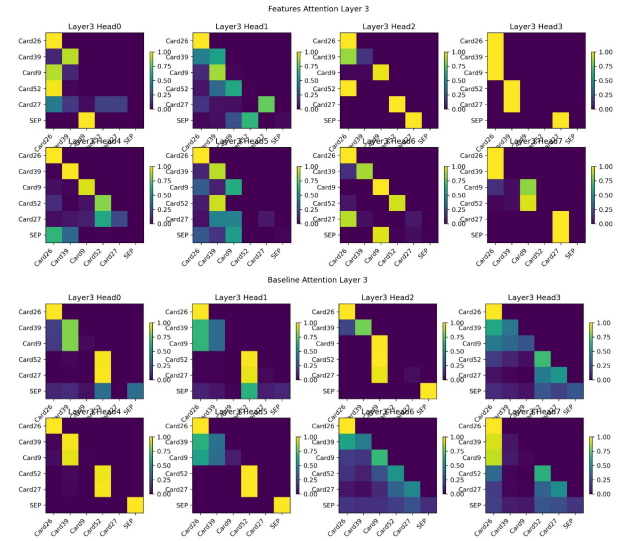


Fig. 4: Attention heatmaps (Layer 3) for each head in the Features model (top block) and Baseline model (bottom block). Each subplot shows the attention matrix for one head (rows = positions attending from; columns = positions attended to).

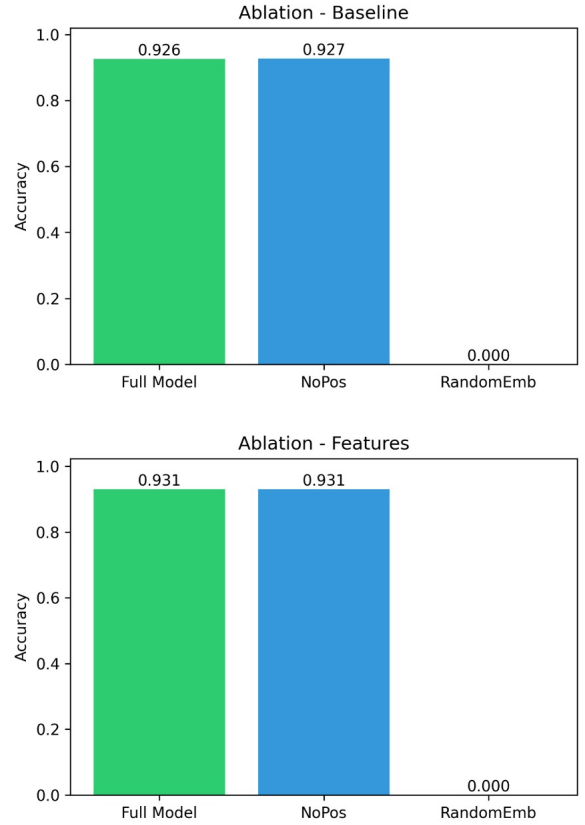


Fig. 5: Per-layer ablation - Baseline (Left), Improved Model(Right)