Transformer-Based Analysis of Rule Learning and Attention Specialisation in the Wisconsin Card Sorting Test

Lathitha Nongauza

***Abstract*—This study investigates the capacity of Transformer architectures to learn and adapt to abstract rules in the Wisconsin Card Sorting Test (WCST), a neuropsychological paradigm assessing cognitive flexibility. We evaluate three Transformer variants—encoder-only, decoder-only, and encoder-decoder—on their ability to infer sorting rules from symbolic card sequences and adapt to dynamic rule changes. Our analysis reveals that while standard Transformer configurations struggle with context switching, exhibiting rule fixation and chance-level performance, a decoder-only model with multi-trial context achieves near-perfect accuracy by leveraging in-context learning. Through attention mechanism analysis, we demonstrate that Transformer heads spontaneously develop feature-based specialisation, selectively attending to colour, shape, or quantity dimensions. However, this specialisation alone proves insufficient for dynamic adaptation without appropriate training paradigms. Our findings highlight both the potential and limitations of current Transformer architectures in modeling higher-order cognitive processes, suggesting that architectural capacity must be coupled with context-rich training objectives to achieve true cognitive flexibility.**

I. Introduction

Human cognitive flexibility, the ability to adapt behaviour in response to changing environmental rules is a hallmark of higher-order reasoning. The Wisconsin Card Sorting Test (WCST) is a classic paradigm for studying this capacity, requiring participants to infer and adapt to latent sorting rules based on feedback. Computationally modelling such rule inference provides valuable insights into the mechanisms underlying flexible cognition and the extent to which modern machine learning systems can emulate them.

Recent advances in Transformer architectures [1] have demonstrated remarkable generalisation and relational reasoning capabilities in language modelling and vision domains. However, it remains unclear whether Transformers can perform symbolic reasoning that requires dynamic rule inference and context switching capabilities central to cognitive tasks like the WCST. This study aims to bridge this gap by systematically evaluating how different Transformer configurations learn, represent, and adapt to abstract rules.

II. Hypothesis

We hypothesise that attention heads within Transformer architectures can develop feature-based specialisation, where individual heads selectively attend to the attribute dimension governing the current rule (e.g., colour, shape, or quantity). To test this hypothesis, we train and analyse three variants of the Transformer, decoder-only, encoder-only, and encoder–decoder, on a WCST environment designed to test static and dynamic rule learning. The aim is to demonstrate that various heads can specilise in different rules within the encoder-decoder layers.

III. Background and Related Work

The Wisconsin Card Sorting Test (WCST) is a long-established neuropsychological paradigm used to assess an individual's capacity for rule inference and cognitive flexibility [4] . Each card is defined by three independent attributes; colour, shape, and quantity,each taking one of four possible values, resulting in 64 unique combinations. Participants must deduce the correct sorting rule (based on one of these attributes) from feedback and flexibly adjust their strategy when the underlying rule changes. This process is a hallmark of executive functioning, engaging working memory, hypothesis testing, and adaptive control mechanisms.

From a computational perspective, the WCST provides a structured benchmark for studying symbolic reasoning and dynamic rule induction in artificial systems. Early neural and reinforcement learning models struggled to emulate this behaviour, primarily because they relied on temporal dependencies rather than abstract relational structure. The emergence of the Transformer architecture [1], characterised by self-attention and parallel representation learning, provided a new framework for modelling symbolic relations. Each attention head acts as a specialised information-processing channel, allowing distributed representations of features that can potentially align with human-like rule abstraction.

Recent advances in Transformer-based modelling have shown that such architectures can internalise relational patterns across multiple domains. Models like BERT [2] introduced bidirectional attention for understanding contextual semantics, while autoregressive models inspired by GPT leverage masked self-attention for sequential reasoning. These developments suggest that attention heads can specialise in specific rule dimensions,analogous to how human cognition isolates relevant features when solving the WCST. Investigating how these attention mechanisms specialise during WCST-like reasoning tasks provides an opportunity to understand whether Transformers can exhibit emergent rule abstraction akin to human cognitive processes.

IV. Methodology

We model the WCST as a structured reasoning environment that probes rule inference, context adaptation, and cognitive flexibility through symbolic representations. Each card in the WCST stimulus set is defined by three categorical attributes; colour, shape, and quantity,each assuming one of four discrete values, resulting in a combinatorial space of 64 unique cards. At any point during the task, one feature dimension governs the active classification rule, and this rule periodically changes to simulate contextual shifts requiring cognitive adaptation. Each trial is represented as a tokenised sequence containing four reference (category) cards, an example card, a separator token [SEP], a label token, and an end-of-sequence marker [EOS]. The model aims to infer the currently active rule from the examples and correctly classify subsequent cards, thereby emulating the dynamic reasoning process underlying human WCST performance.

A. Data Generation and Task Setup

1) Vanilla WCST: The WCST environment procedurally generates symbolic card sequences for each trial to ensure sufficient variability and prevent overlap between data partitions. Four reference cards are sampled for every batch, followed by an example card whose label is determined by the active rule dimension. The target output is a query (or question) card with its correct label.

2) Multi-Trial WCST: We use a procedurally generated WCST environment that emits *multi-trial* sequences. Each sequence contains K **context trials** followed by one **question trial**. A trial t is encoded as

$$(\text{cat}_0, \text{cat}_1, \text{cat}_2, \text{cat}_3,\ e_t,\ \texttt{[SEP]},\ y_t,\ \texttt{[EOS]}),$$

where (cat_i) are the four reference (category) cards, e_t is the example card, and $y_t \in \{0, 1, 2, 3\}$ is the label indicating which category matches e_t under the current rule dimension (Colour/Shape/Quantity). Tokens are integers from a fixed vocabulary: the 64 card IDs, four label IDs, plus special tokens (`[SEP]`, `[EOS]`).

a) Rule dynamics.: A single rule dimension is active while generating a sequence. To induce context switching across the corpus, the active rule changes after a fixed interval S sequences. Thus, within a sequence the rule is stable (the K context trials are informative), but across sequences the rule periodically shifts.

b) Training mask (SEP-masked objective).: Let $x_{1:T}$ be the input tokens for next-token prediction and $y_{1:T}$ the targets. We build a mask $m_{1:T}$ that is 1 only at positions immediately following `[SEP]` (i.e., at label positions) and 0 elsewhere, and optimise

$$\mathcal{L} \;=\; \frac{1}{\sum_t m_t} \sum_{t=1}^{T} m_t \cdot \text{CE}(\hat{p}_t,\, y_t)\,.$$

We also report *final-trial accuracy*, the accuracy at the single label position of the question trial.

3) Data management: We use a function that produces a non-overlapping corpus with 70%/15%/15% train/val/test splits and fixed-length padding. A total of 20,000 unique trials are generated for training and evaluation. Each trial is converted into a fixed-length token sequence compatible with Transformer-based architectures, allowing consistent comparison across three model configurations: decoder-only, encoder-only, and encoder–decoder variants.

B. Model Variants

1) Decoder-Only Transformer (GPT-Style Rule Induction): We implement an autoregressive architecture for symbolic reasoning using vanilla WCST generation. Tokens are embedded into a 128-dimensional space with sinusoidal positional encodings. The sequence is processed through four decoder layers, each with four-headed self-attention, a feed-forward sublayer (hidden size 512), residual connections, and layer normalisation. A causal mask enforces autoregressive prediction, and dropout ($p = 0.1$) mitigates overfitting. The final decoder output is projected onto the vocabulary space through a shared embedding matrix, supporting rule inference through next-token prediction.

2) Encoder-Only Transformer (BERT-Style Relational Reasoning): Inspired by BERT [2], this variant performs non-autoregressive classification using vanilla WCST generation. Tokens are embedded into a 256-dimensional space, scaled by $\sqrt{d_{\text{model}}}$, and enriched with sinusoidal positional encodings. The encoder comprises six Transformer layers with eight-headed self-attention and 1024-dimensional feed-forward networks. The [SEP] token embedding serves as a contextual summary vector and is passed through a three-layer classification head with GELU activations and dropout to predict one of four rule categories.

3) Encoder–Decoder Transformer (Cross-Attention): We implement the standard encoder–decoder design [1] using vanilla WCST generation. The encoder (six layers, eight heads) captures contextual relationships, while the decoder (two layers) combines masked self-attention and cross-attention to align rule and query representations. Feed-forward sublayers have an internal dimensionality of $4 \times d_{\text{model}}$, with GELU activation and dropout ($p = 0.1$). Learned token embeddings are scaled by $\sqrt{d_{\text{model}}}$ and combined with sinusoidal positional encodings. Decoder outputs are projected onto the vocabulary space via a linear layer to generate next-token logits.

4) Decoder-Only Transformer with Multi-Trial Context: A decoder-only variant that differs from the single-trial setup in two ways: (i) the input sequence includes $K = 4$ context trials(multi-Trial WCST) before the question trial, and (ii) training uses the **SEP-masked** objective so only label positions contribute to the loss. We use $d_{\text{model}} = 256$, $n_{\text{heads}} = 8$, $n_{\text{layers}} = 6$, FFN size $= 1024$, dropout $= 0.1$, AdamW with lr $= 10^{-4}$ and weight decay $= 0.01$, batch size 32 with $2\times$ gradient accumulation (effective 64), early stopping (patience $= 10$), and weight tying between embeddings and output projection.

C. Training and Optimisation

All models were trained using the AdamW optimiser [3] with an initial learning rate of 1×10^{-3} and weight decay of 1×10^{-5}. The encoder-only model employed a cosine decay schedule with a 1,000-step warmup, starting from 3×10^{-4} and decaying to 1×10^{-6}. Gradient clipping (max norm = 1.0) ensured stability. Models were trained for 20–50 epochs using mini-batches of 64 sequences and early stopping (patience 8–10). Decoder-based models used cross-entropy loss with teacher forcing, while the encoder-only model used label-smoothed cross-entropy ($\varepsilon = 0.1$). All experiments were implemented in PyTorch 2.0 on CUDA-enabled GPUs.

D. Attention-Based Interpretability Analysis

For each model we extract self-attention tensors $\mathbf{A}^{(\ell,h)} \in \mathbb{R}^{T\times T}$ at layer ℓ and head h during inference (row = query position, column = key position). Let s denote the index of the `[SEP]` token that precedes the label of the *question* trial in a sequence, and let $\mathcal{P}_{\text{cat}} = \{p_0, p_1, p_2, p_3\}$ be the four category-card positions of that final trial. Each card c has feature tuple $f(c) = (\text{colour}, \text{shape}, \text{quantity}) \in \{0,1,2,3\}^3$. Given the question card q of the final trial, we define the feature-matched index sets

$$\mathcal{M}_{\text{Colour}} = \{\, p \in \mathcal{P}_{\text{cat}} \mid f_{\text{Colour}}(x_p) = f_{\text{Colour}}(q) \,\},$$
$$\mathcal{M}_{\text{Shape}} = \{\, p \in \mathcal{P}_{\text{cat}} \mid f_{\text{Shape}}(x_p) = f_{\text{Shape}}(q) \,\},$$
$$\mathcal{M}_{\text{Quantity}} = \{\, p \in \mathcal{P}_{\text{cat}} \mid f_{\text{Quantity}}(x_p) = f_{\text{Quantity}}(q) \,\}$$

analogously. For each (ℓ, h) we compute the per-feature score as the attention mass from the `[SEP]` query to the matched category positions:

$$S^{(\ell,h)}_{\text{feat}} = \frac{1}{|\mathcal{M}_{\text{feat}}|} \sum_{p\in\mathcal{M}_{\text{feat}}} \mathbf{A}^{(\ell,h)}_{s,p}.$$

We average $S^{(\ell,h)}_{\text{feat}}$ across the evaluation set to obtain a head-by-feature specialisation matrix per layer, which we visualise as heatmaps. Heads with large $S^{(\ell,h)}_{\text{feat}}$ exhibit selective routing of attention to the category cards that match the question along that feature dimension, aligned with the decision point where the model is trained to produce the label.

V. Results

A. Baselines

1) Encoder-Decoder: The baseline model; an encoder–decoder Transformer, achieved approximately 25% accuracy on validation and test sets, equivalent to random guessing. Despite low overall performance, attention analyses revealed that individual heads exhibited clear specialisation, selectively attending to distinct rule dimensions (colour, shape, or quantity) (Figure 4). When trained on a single rule condition, each head consistently specialised in representing that rule, indicating the model's capacity for stable rule learning under static conditions(Figure 5). When trained on multiple rule conditions, the model tended to fixate on the initial rule and ignore context switches, maintaining a 25% performance ceiling. In contrast, single-rule training produced perfect (100%) accuracy after sufficient epochs, demonstrating the Transformer's ability to generalise within a fixed rule regime but not across dynamic contexts(Figure 1).

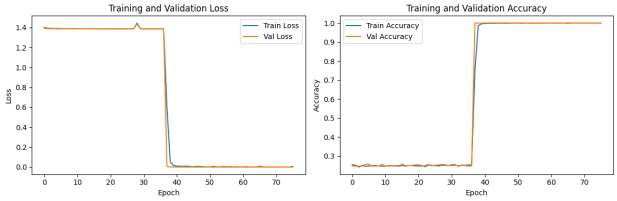


Fig. 1: Training and accuracy when using one rule

2) Encoder-only and Decoder-only: The encoder-only and decoder-only variants, despite simpler architectures, did not surpass the baseline's performance. Both models overfitted to the initial rule and failed to update their internal representations following rule changes. These findings indicate that while Transformer architectures can encode and specialise rule representations, they lack the adaptive flexibility required to model context dependent reasoning in the WCST. Both of these variants produced attention heat maps similar to the encoder-decoder variant.

B. Decoder-only with multi-trial context.

Introducing a multi-trial input together with the **SEP-masked** objective resolved the 25% ceiling. The model inferred the current rule from the $K = 4$ context trials and applied it on the question trial, yielding near-perfect final-trial accuracy on validation and test (learning curves in Figure 2). Attention heatmaps (Figure 3) reveal consistent head specialisation by feature (Colour/Shape/Quantity), which strengthens in deeper layers, evidence that the model computes rule-relevant alignments precisely at the decision point. Several heads in the variuos layers specialized in different features, which is an improvement on the baselines which were only capable of specializing on one feature across all heads in all layers. There is an increase in specialization diversity.

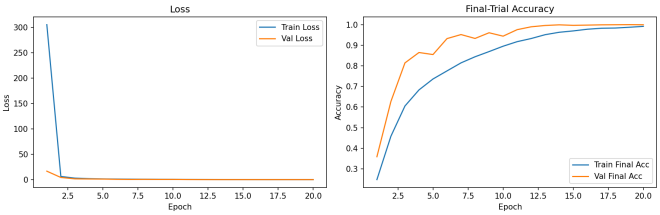


Fig. 2: Learning curves for the decoder-only multi-trial model: left, train/val loss; right, final-trial accuracy.

C. Comparison

1) Baseline: The baseline implementation has 6 layers with 8 heads each. The color feature dominates the other features within each layer and across all layers. 83% of the heads specialized in color.

2) Decoder-only with multi-trial context.: There is no clear feature domination with this implementation. Most of the heads equally specialize in all features, with a few heads dominating in one of the features. There is no trend in the features which dominate others within a head.

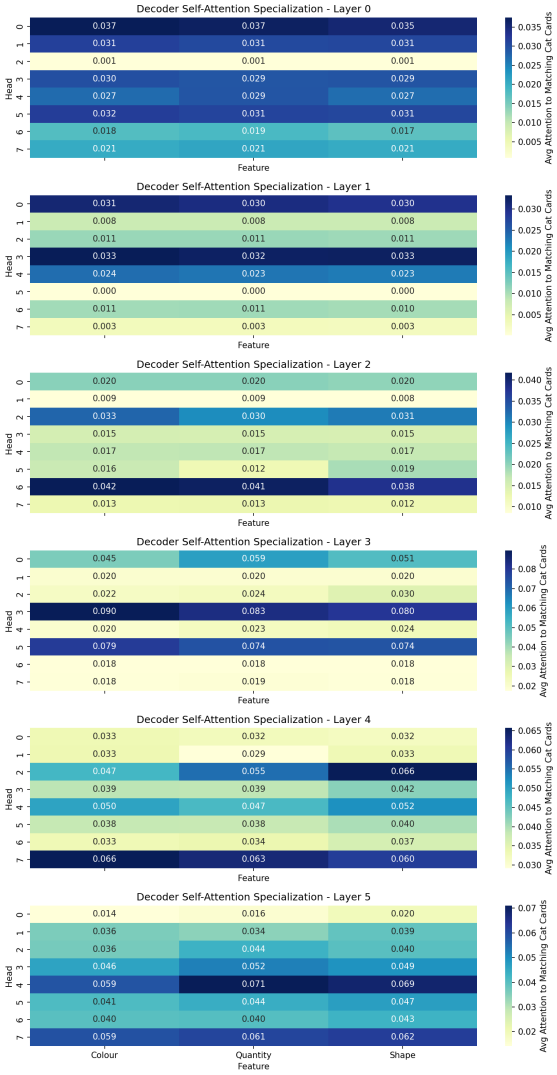


Fig. 3: Attention specialisation. Each panel shows a layer; rows are heads and columns are features (Colour/Quantity/Shape). Values are mean attention mass from the question-`[SEP]` query to category cards that match the question along that feature.

VI. Discussion

Our findings reveal a fundamental distinction between static rule learning and dynamic rule adaptation in Transformer architectures. While all model variants demonstrated the capacity to learn individual WCST rules and develop specialised attention patterns, only the decoder-only model with multi-trial context successfully adapted to changing rule conditions. This suggests that architectural capacity alone is insufficient for cognitive flexibility—the training paradigm and task formulation play equally critical roles.

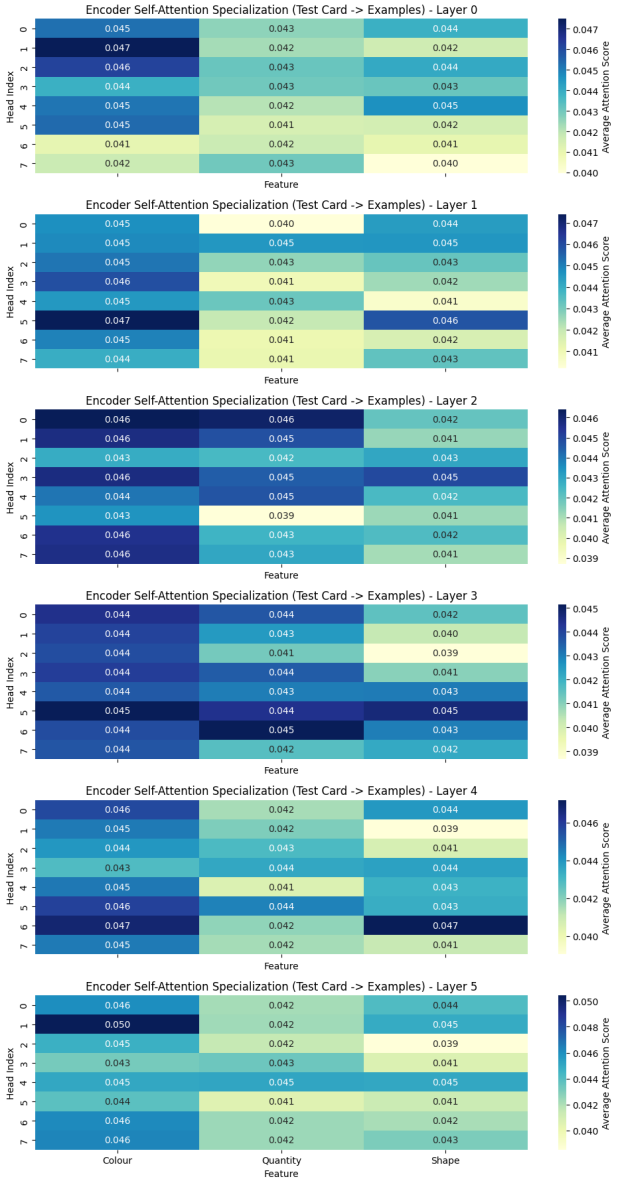
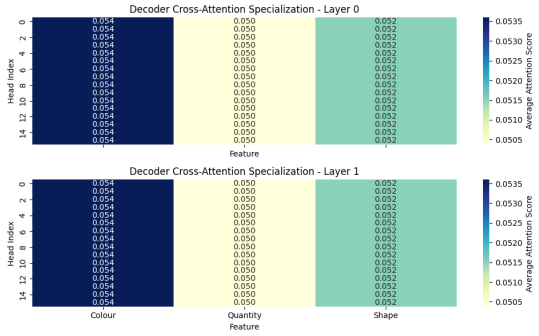


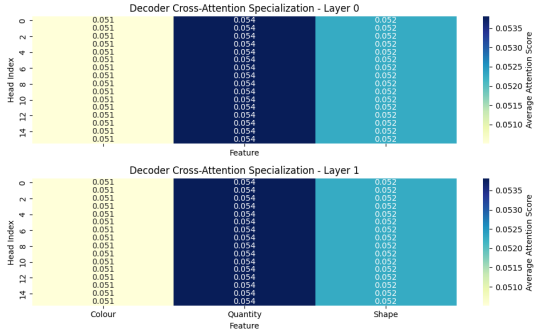
Fig. 4: Attention specialisation on baseline encoder-decoder transformer.

A. The Role of Multi-Trial Context in Rule Induction

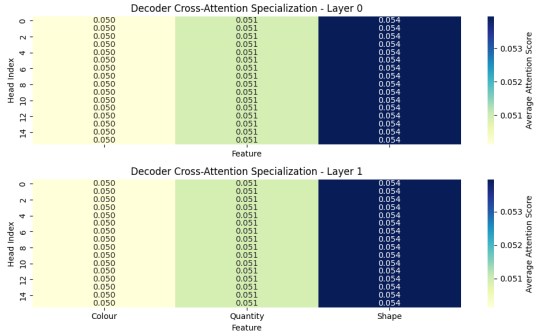
The superior performance of the multi-trial decoder-only model underscores the importance of in-context learning for cognitive tasks requiring rule abstraction. Unlike the single-trial variants that attempted to classify cards directly, the multi-trial approach forced the model to infer rules from contextual examples before application. This aligns with human performance on the WCST, where participants use feedback from multiple trials to form and test hypotheses about the active rule. The SEP-masked objective further reinforced this by



(a) Colour attention.



(b) Quantity attention.



(c) Shape attention.

Fig. 5: Attention patterns for each feature dimension in the WCST. Each subfigure shows head-level attention specialising in colour, quantity, or shape.

focusing learning on decision points, effectively teaching the model when and where to apply rule-based reasoning.

B. Attention Specialisation as Emergent Feature Detection

The observed attention specialisation patterns provide compelling evidence that Transformer heads can spontaneously develop feature-specific roles, even without explicit supervision. In the multi-trial model, different heads consistently attended to colour, shape, or quantity dimensions, mirroring the modular processing hypothesized in human cognitive architectures. This specialisation was most pronounced in deeper layers, suggesting a hierarchical processing where early layers extract basic features and later layers integrate them for rule application. The multi-trial decoder-only model produced results consistent with our hypothesis.

C. Limitations of Standard Architectures for Cognitive Flexibility

The failure of baseline models (encoder-only, decoder-only, encoder-decoder) to adapt to rule changes reveals a significant limitation in standard Transformer training for cognitive tasks. These models exhibited rule fixation—once they learned an initial rule, they struggled to reorient when the context changed. This parallels perseveration errors observed in clinical populations with frontal lobe damage, suggesting that current architectures may lack the executive control mechanisms needed for true cognitive flexibility.

VII. Conclusion

This study demonstrates that Transformer architectures possess the capacity to learn abstract rules and develop specialised attention patterns resembling human cognitive processes, yet they face significant limitations in dynamic context adaptation. Our key findings reveal that standard Transformer configurations—including encoder-only, decoder-only, and encoder-decoder variants—consistently fail to adapt to changing WCST rules, exhibiting perseveration errors analogous to those observed in clinical populations with executive function impairments.

The success of the multi-trial decoder-only model underscores the critical importance of training paradigm design over mere architectural sophistication. By providing multiple context trials and employing a SEP-masked objective, this model achieved near-perfect accuracy and developed diverse attention specialisation across features, validating our hypothesis that Transformer heads can spontaneously organise to represent different rule dimensions.

However, the persistent failure of baseline models to overcome rule fixation reveals fundamental gaps in current architectures' capacity for true cognitive flexibility. While attention mechanisms can detect and represent rules, they lack the executive control mechanisms necessary for dynamic reconfiguration in response to changing contexts.

These findings have important implications for both artificial intelligence and cognitive modeling. For AI research, they emphasize that achieving human-like reasoning requires not just larger models but more sophisticated training objectives that emphasize context-based learning and adaptation. For cognitive science, they provide a computational framework for studying how rule representations emerge and how cognitive flexibility might be implemented in neural systems.

Future work should explore dynamic attention routing mechanisms, meta-learning approaches for rapid rule adaptation, and neurosymbolic integrations that combine Transformer pattern recognition with explicit rule representations. As we continue bridging the gap between artificial and human intelligence, the WCST remains a valuable benchmark for evaluating not just what models can learn, but how flexibly they can apply and adapt that learning in changing environments.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[3] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations (ICLR)*, 2015.

[4] D. A. Grant and E. A. 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.

[5] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," *arXiv:1607.06450*, 2016.

[6] D. Hendrycks and K. Gimpel, "Gaussian Error Linear Units (GELUs)," *arXiv:1606.08415*, 2016.

[7] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2019.

[8] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the Inception architecture for computer vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[9] O. Press and L. Wolf, "Using the output embedding to improve language models," in *Proceedings of the 15th Conference of the European Chapter of the ACL*, 2017.

[10] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," in *International Conference on Learning Representations (ICLR)*, 2017.

[11] A. Paszke *et al.*, "PyTorch: An imperative style, high-performance deep learning library," in *NeurIPS*, 2019.