ATTENTION IS NOT ALL YOU NEED WITHOUT STRUCTURE: NEURO-SYMBOLIC MULTI-HEAD ATTENTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Transformer architectures rely on Multi-Head Attention (MHA) as the central mechanism for modelling interactions between tokens in sequence data. Although this is effective for capturing statistical dependencies, the standard MHA lacks explicit structural inductive bias for relational reasoning, which limits systematic generalisation for tasks requiring multi-step inference. Prior neuro-symbolic methods attempt to fix this limitation through external reasoning modules or discrete symbolic pipelines, which can complicate integration with modern transformer architectures and break end-to-end differentiability. In this work, we introduce a role-conditioned extension of MHA that injects structured relational inductive bias directly into the attention computation while preserving full end-to-end differentiability along with transformer compatibility as a drop-in replacement. The proposed mechanism augments standard attention with a parallel role-conditioned projection pathway based on learned role embeddings, combined through a learnable fusion gating mechanism. Additionally, a role-relational attention bias enables symbolic attention weights to depend on interactions between inferred token roles, allowing relational structure to influence attention scores directly. We evaluate the proposed architecture on language modelling, relational reasoning, and compositional generalisation benchmarks, including WikiText-2, LAMBADA, CLUTRR, and SCAN, using multi-seed experiments with paired statistical testing. Results demonstrate consistent reductions in perplexity on language modelling benchmarks, substantial improvements on relational reasoning tasks, and improved probabilistic modelling on compositional tasks. Ablation studies confirm that the observed gains arise from role-conditioned structural bias rather than architectural duplication alone. These findings demonstrate that structured inductive bias can be embedded directly within attention mechanisms, enhancing Multi-Head Attention while preserving end-to-end optimisation and transformer-native design.

1 INTRODUCTION

Neural sequence models progressed from convolutional (LeCun et al., 1998) and recurrent architectures (Hochreiter & Schmidhuber, 1997), including LSTMs, which struggled to train on long sequences and scale poorly, to the Transformer (Vaswani et al., 2017), which replaces recurrence with an attention mechanism over three learnable projections, QKV, letting each token interact directly with all others. At the core of the Transformer is Multi-Head Attention (MHA), a purely neural mechanism computing weighted combinations of input representations via learned similarity scores. MHA does not explicitly encode relational structure or symbolic roles such as predicate–argument relationships, relying instead on statistical patterns learned from data, which hampers generalisation on tasks requiring structured multi-step inference.

A long-standing AI goal is to combine neural learning with symbolic reasoning: symbolic approaches provide explicit structure, while neural models offer scalability through gradient-based learning. This has proven difficult in practice, since most prior neuro-symbolic systems introduce symbolic components as separate, non-differentiable modules (logic engines, rule-based systems, external memory) operating alongside neural networks rather than within them, breaking end-to-end

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

differentiability and making such systems harder to optimise, less scalable, and difficult to integrate into modern Transformer architectures.

In this work, we propose Neuro-Symbolic Multi-Head Attention (NS-MHA), a transformer-native extension of MHA that injects structured relational inductive bias directly into self-attention. We use *neuro-symbolic* to denote role-conditioned structural bias integrated within a fully differentiable attention mechanism, rather than discrete logical operators or external reasoning modules. An equally valid description of the mechanism is role-conditioned structured attention. We retain the neuro-symbolic label to situate this work within the broader tradition of embedding relational and role-based structure into neural computation (Serafini & Garcez, 2016; Dong et al., 2019), without implying discrete symbolic manipulation. NS-MHA embeds structured bias directly inside the QKV projection mechanism, preserving end-to-end differentiability and drop-in compatibility with existing transformer implementations.

This work makes four contributions: (a) **A Transformer-Native Neuro-Symbolic Attention Mechanism**, combining a neural and a role-conditioned stream in one differentiable layer; (b) **Differentiable Role-Conditioned Structural Modulation**, mapping embeddings into a relational subspace learned via gradient descent rather than discrete rules; (c) **Fusion-Gated Integration**, a learnable gate balancing the two streams; and (d) **A Thorough Experimental Evaluation** across WikiText-2, LAMBADA, CLUTRR, and SCAN, showing consistent perplexity reductions and substantial relational-reasoning gains via multi-seed paired testing.

2 Related Work

Modular and rule-based neuro-symbolic systems: Early neuro-symbolic AI combined logic and learning through rule-based systems and modular architectures such as Neural-Symbolic Learning Systems, which embedded logical rules alongside neural networks to guide training (Besold et al., 2017; Garcez et al., 2012). Besold et al. (2017) categorised such systems by how symbolic and neural components interact, but identified a consistent limitation: logic and learning remain distinct subsystems connected through handcrafted, non-differentiable interfaces, unable to share a representational space or train jointly. Later architectures, including Neural-Symbolic Cognitive Integration (Garcez & Lamb, 2023) and cognitive-science-inspired hierarchical models (Mghames et al., 2023), improved interpretability but preserved this separation.

Differentiable logic and relational learning: Logic Tensor Networks (Serafini & Garcez, 2016) expressed symbolic operations as continuous, differentiable tensor functions rather than binary true/-false, though still requiring hand-specified rules. Later work replaced discrete unification with continuous similarity matching (Tsamoura et al., 2021) and, with Neural Theorem Provers (Rocktäschel & Riedel, 2017), learned logical rules via gradient descent. Graph Neural Networks (Lamb et al., 2020; Xu et al., 2019) extended relational reasoning via message passing but need predefined graph topologies unsuited to language, and more recent inductive approaches (Evans & Grefenstette, 2018; Dong et al., 2019) learn symbolic rules directly from data, though a neural-symbolic representational mismatch persists (Zhang & Sheng, 2024).

Transformers and external symbolic augmentation: The Transformer (Vaswani et al., 2017) replaced recurrence with Multi-Head Attention, enabling large-scale models such as BERT (Devlin et al., 2019), GPT (Radford et al., 2019), and T5 (Raffel et al., 2020), which exhibit limited systematic generalisation on relational reasoning. Prior attempts keep symbolic reasoning external to attention: chain-of-thought prompting, semantic parsing with logical inference (Aithal et al., 2022), symbolic functions for arithmetic and spatial reasoning (Demeter & Downey, 2020), and energy-based or generative constraints (Tran et al., 2025; Jacobson & Xue, 2024). A recent review (Colelough & Regli, 2024) confirms such systems remain predominantly modular, trained disjointly rather than integrated within attention.

Structured and relational attention: A closely related line of work embeds structural or relational bias directly into attention rather than relying on an external graph or symbolic module. Graph Attention Networks (Veličković et al., 2018) apply masked self-attention over graph neighbourhoods with fixed topology, while the Structure-Aware Transformer (Chen et al., 2022) injects a subgraph-derived structural kernel into otherwise standard self-attention. The Relational Transformer (Diao & Loynd, 2023) generalises attention to explicitly represent and update edge vectors alongside node

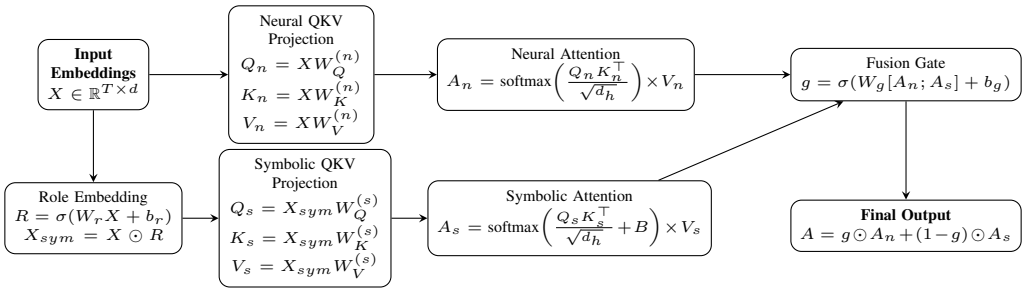

Figure 1: Architecture of the proposed NS-MHA module

vectors, and the TP-Transformer (Schlag et al., 2019) embeds Tensor-Product Representations directly into attention to encode the relation between a query and the cell it retrieves from. Slot Attention (Locatello et al., 2020) learns a fixed set of latent slots that compete via iterative attention to bind to input entities.

Prior approaches rely on modular pipelines or reasoning components alongside the transformer rather than within self-attention itself. Where MHA uses a single Q, K, V set, in this work, we introduce a secondary parallel QKV set for predicate-style role embeddings, learned dynamically via self-attention rather than hand-specified, combined via a learnable fusion gate. This differs from structured-attention mechanisms that inject a single fixed structural signal additively into one score. The proposed NS-MHA maintains two independently parameterised pathways, content-based and role-conditioned, arbitrated per token pair by a gate trained jointly with both, so the bias is dynamically weighted rather than fixed or merged into a single score.

3 METHODOLOGY

NS-MHA extends standard Multi-Head Attention (MHA) (Vaswani et al., 2017) with a parallel role-conditioned symbolic pathway, integrated with the standard pathway via a learnable fusion gate detailed below. An overview of the proposed approach is provided in Figure 1.

3.1 STANDARD MULTI-HEAD ATTENTION (MHA)

The Transformer (Vaswani et al., 2017) replaced recurrent sequence processing with self-attention, letting tokens interact directly with all others in a sequence. Given token embeddings $X \in \mathbb{R}^{T\times d}$, the model projects them into queries ($Q = XW_Q$), keys ($K = XW_K$), and values ($V = XW_V$), where $W_Q, W_K, W_V \in \mathbb{R}^{d\times d_h}$ are learned projection matrices and $d_h = \frac{d}{h}$. Scaled dot-product attention for a single head is

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_h}}\right) V \tag{1}$$

where the scaling factor stabilises gradients. MHA computes attention across h independent heads: $\text{MHA}(X) = \text{Concat}(H_1, H_2, \ldots, H_h)W_O$, with W_O being a learned output projection.

MHA underlies modern transformer-based language models, such as BERT, GPT, T5 (Devlin et al., 2019; Radford et al., 2019; Raffel et al., 2020), and similar more recent ones. For input $X = [x_1, x_2, \ldots, x_T]$, the attention output for token i is $A_i = \sum_{j=1}^{T} \alpha_{ij} V_j$, with α_{ij} the normalised attention weight from the softmax distribution. Stacked attention and feed-forward blocks produce hidden representations h_t used to predict the next token, $P(x_{t+1}|x_{\leq t}) = \text{softmax}(W_o h_t)$. Despite strong empirical performance, standard attention captures statistical correlations but does not explicitly encode structured relational roles such as predicate–argument relationships.

3.2 PROPOSED NEURO-SYMBOLIC MULTI-HEAD ATTENTION (NS-MHA)

NS-MHA augments standard attention with a secondary symbolic projection pathway operating alongside the neural projection. While the neural pathway performs conventional MHA, the sym-

bolic pathway introduces role-conditioned representations that inject structured relational inductive bias into the attention computation, with both pathways combined via the fusion gate described in Section 3.6.

3.3 ROLE EMBEDDINGS

To introduce symbolic structure into the attention mechanism, we learn role embeddings capturing functional relationships between tokens. Rather than assigning fixed symbolic roles, we learn a latent set of role prototypes and allow tokens to softly assign to them, following the broader philosophy of continuous neuro-symbolic representation learning (Serafini & Garcez, 2016; Dong et al., 2019), while remaining fully differentiable. Role embeddings are computed as:

$$R = \text{softmax}(XW_r)P \tag{2}$$

where W_r is a learned role assignment matrix, $P \in \mathbb{R}^{k \times d}$ is a set of learned role prototypes, and k denotes the number of latent roles. The resulting role representation $R \in \mathbb{R}^{T \times d}$ captures functional role information for each token.

We choose softmax over discrete assignment because it is fully differentiable, lets a token express multiple roles simultaneously (e.g., 40% agent, 60% predicate) rather than committing to one, and, via the shared prototype bank P, is more parameter-efficient than a full learnable role matrix.

In all experiments, we set $k = 8$, matching the number of attention heads h, allowing each role prototype to potentially specialise for a particular attention head's focus. We found empirically that $k < 8$ underfits the relational reasoning task, while $k > 8$ gives no improvement with increased parameters. We did not conduct a systematic sweep over configurations where k diverges substantially from h; how performance depends on this ratio, and whether roles and heads develop aligned or complementary specialisations remains an open empirical question.

To integrate role information with token semantics we compute a role-conditioned representation:

$$X_{\text{sym}} = X \odot R \tag{3}$$

where $\odot$ denotes the element-wise Hadamard product (element-wise multiplication (Horn & Johnson, 2012)).

We choose element-wise multiplication over concatenation because it acts as a learned gate, where an embedding with low role confidence (small R) is suppressed while a high-confidence one is amplified. It also costs $O(Td)$ rather than $O(2Td)$, and it binds role and semantic information into one representation instead of treating them as separate streams.

3.4 DUAL QKV PROJECTION

The role-conditioned representation is used to compute symbolic attention projections:

$$Q_s = X_{\text{sym}} W_Q^{(s)} \tag{4}$$

$$K_s = X_{\text{sym}} W_K^{(s)} \tag{5}$$

$$V_s = X_{\text{sym}} W_V^{(s)} \tag{6}$$

where $W_Q^{(s)}, W_K^{(s)}, W_V^{(s)}$ are learned symbolic projection matrices.

3.5 ROLE-RELATIONAL ATTENTION BIAS

To capture interactions between token roles, the symbolic attention pathway incorporates a relational bias derived from pairwise role interactions. We compute a role interaction matrix:

$$B = (W_b R)R^\top \tag{7}$$

where W_b is a learned projection matrix.

Here, $W_b R$ first projects the role embeddings into a relational subspace (dimensionality $T \times d_r$), and the outer product with $R^\top$ yields a $T \times T$ matrix where $B[i, j]$ measures the compatibility

Component	Value
Embedding dimension (d)	512
Number of attention heads (h)	8
Maximum sequence length	128
Feed-forward dimension	2048
Dropout	0.1
Optimizer	Adam
Learning rate	5×10^{-5}
Weight decay	0.0
Gradient clipping	1.0
Batch size	8
Training epochs	20
Tokenizer	GPT-2

Table 1: Model and training configuration used for all experiments. Settings were identical for baseline Multi-Head Attention (MHA) and Neuro-Symbolic Multi-Head Attention (NS-MHA), except for the attention module.

between token i's and token j's projected roles. The resulting bilinear form is symmetric and fully differentiable. An attention-based alternative, $\mathrm{softmax}(W_b R R^\top)$, was also tried but performed worse. B is added directly to the symbolic attention logits.

The symbolic attention therefore becomes:

$$A_s = \mathrm{softmax}\left(\frac{Q_s K_s^\top}{\sqrt{d_h}} + B\right) V_s \tag{8}$$

This formulation allows symbolic attention to depend not only on token similarity but also on interactions between their inferred functional roles.

While we do not provide a formal expressivity proof, the relational bias B can be understood as expanding the attention hypothesis space beyond content-based compatibility alone. Standard attention logits are a bilinear function of token content through $QK^\top$, whereas NS-MHA additionally injects B, a bilinear function of the learned role assignments R. Because R is a differentiable function of X that can specialise independently of the content-based query-key interaction, B can express role-compatibility patterns, for instance consistent attention between tokens assigned complementary roles regardless of their surface content, that $QK^\top$ alone cannot represent unless content embeddings already linearly encode role identity. This is consistent with the larger relative gains observed on CLUTRR, a benchmark defined by explicit relational structure, than on WikiText-2, where content-based similarity is already a strong predictor, as shown in Section 4.

The neural pathway remains unchanged:

$$A_n = \mathrm{softmax}\left(\frac{Q_n K_n^\top}{\sqrt{d_h}}\right) V_n \tag{9}$$

3.6 Fusion Gating Mechanism

The neural and symbolic attention outputs are integrated using a learnable fusion gate:

$$g = \sigma(W_g[A_n; A_s] + b_g) \tag{10}$$

where $[A_n; A_s]$ denotes concatenation and σ is the sigmoid activation function. The final attention representation is:

$$A = g \odot A_n + (1 - g) \odot A_s \tag{11}$$

This gating mechanism is conceptually related to the Gated Multimodal Unit (GMU) (Arevalo et al., 2017), which adaptively combines multiple representation streams.

4 Experimental Evaluation

We compare the proposed NS-MHA against standard MHA across four language-related datasets, each focusing on a different but complementary task. We adopt unmodified MHA as the primary

Metric	WikiText-2		CLUTRR		LAMBADA		SCAN	
	MHA	**NS-MHA (ours)**	**MHA**	**NS-MHA (ours)**	**MHA**	**NS-MHA (ours)**	**MHA**	**NS-MHA (ours)**
Val. Loss (↓)	6.1900 ± 0.0083	$\mathbf{6.1588 \pm 0.0076}$	$0.5817 \pm < 10^{-4}$	$\mathbf{0.4430 \pm < 10^{-4}}$	5.43 ± 0.33	$\mathbf{5.41 \pm 0.34}$	$0.000966 \pm < 10^{-5}$	$\mathbf{0.000957 \pm < 10^{-5}}$
Val. PPL (↓)	487.88 ± 4.08	$\mathbf{472.87 \pm 3.59}$	–	–	201.86 ± 28.47	$\mathbf{198.42 \pm 29.11}$	$1.000966 \pm < 10^{-5}$	$\mathbf{1.000958 \pm < 10^{-5}}$
Val. Acc. (%) (↑)	20.66 ± 0.08	$\mathbf{20.82 \pm 0.07}$	$80.74 \pm < 0.01$	$\mathbf{88.12 \pm < 0.01}$	20.20 ± 0.29	$\mathbf{20.28 \pm 0.34}$	$\mathbf{30.93 \pm 0.10}$	30.74 ± 0.19
p-value (↓)	**0.00390625**		**0.001953125**		**0.000244140625**		0.027344	

Note: PPL: Perplexity, Acc.: Accuracy

Table 2: Validation performance of MHA and NS-MHA on WikiText-2, CLUTRR, LAMBADA, and SCAN. Reported as mean ± standard deviation across 10 paired seeds.

Model	WikiText-2			CLUTRR			LAMBADA			SCAN		
	Loss	**PPL**	**Acc**	**Loss**	**PPL**	**Acc**	**Loss**	**PPL**	**Acc**	**Loss**	**PPL**	**Acc**
Baseline MHA	6.19	489.6	0.205	0.59	–	0.79	5.20	181.54	0.201	1.9742	7.2007	0.3075
NS-MHA (No RE)	6.17	476.2	0.207	0.49	–	0.88	5.21	182.51	0.201	1.9577	7.0831	0.3096
NS-MHA (With RE)	**6.16**	**470.1**	**0.209**	**0.47**	–	**0.88**	**5.19**	**180.13**	**0.201**	**1.9444**	**6.9892**	**0.3105**

Table 3: Ablation study with and without role embeddings (RE) on WikiText-2 (100%, seq. length 128, 20 epochs), CLUTRR (100%, 20 epochs), LAMBADA (100%, seq. length 128, 20 epochs), and SCAN (100%, seq. length 64, 20 epochs).

baseline because our contribution targets the attention mechanism itself, so comparing against the canonical formulation isolates the effect of role-conditioned structure without confounding differences in positional encoding, auxiliary losses, or architectural scale. Comparisons against other structured-attention variants are identified as future work in Section 5.7. Both models are trained and evaluated under identical conditions, i.e. training scripts, optimiser settings, sequence length, batch size, epoch count, GPT-2 BPE tokeniser, and datasets, with the attention layer as the only architectural difference, using the same metrics and experimental configuration. We also conduct an ablation study with the symbolic role embeddings removed while retaining the dual QKV projections and fusion gate, isolating whether gains stem from the role-conditioned structure itself rather than parameter count or the mere presence of a second QKV projection.

4.1 Benchmark Evaluation

We evaluate MHA and NS-MHA across four benchmarks: WikiText-2 (Merity et al., 2017) and LAMBADA (Paperno et al., 2016) that focus on language modelling and long-range contextual understanding, CLUTRR (Sinha et al., 2019), which focuses on multi-step relational reasoning, and SCAN (Lake & Baroni, 2018), which focuses on compositional generalisation via deterministic command-to-action mappings. Each configuration (100% of data) is evaluated across 10 paired random seeds with identical architectures, optimisation, and data splits, using a fixed sequence length (128 for WikiText-2, CLUTRR, and LAMBADA, 64 for SCAN), and 20 epochs of training on an NVIDIA RTX 2080 Ti GPU. Table 1 summarises the shared configuration. Statistical significance is assessed via a paired Wilcoxon signed-rank test (Wilcoxon, 1945; Dror et al., 2018) at $\alpha = 0.05$, with exact p-values reported. Results for all models and datasets are reported in Table 2 in terms of validation loss, validation perplexity, and validation accuracy as mean ± standard deviation.

From Table 2, it is evident that NS-MHA achieves lower validation loss and perplexity than MHA on every benchmark, with statistically significant differences ($p < 0.01$ on WikiText-2, CLUTRR, and LAMBADA, and $p < 0.05$ on SCAN). CLUTRR shows the largest effect, with accuracy rising from 80.74% to 88.12%, a 7.38 percentage-point absolute improvement corresponding to a reduction in error rate from 19.26% to 11.88% (approximately 38% relative reduction). Its near-zero cross-seed variance reflects the benchmark's deterministic, template-driven construction, where models converge to nearly identical solutions once the underlying reasoning pattern is learned. Accuracy is only slightly improved for NS-MHA on WikiText-2 and LAMBADA, and is slightly lower for SCAN, reflecting limited sensitivity of accuracy on tasks where performance saturates once the mapping is learned. Models are trained from scratch under identical conditions (tokeniser, optimiser, schedule, data splits) so that performance differences arise solely from the attention mechanism. Section 5 discusses these results in detail per benchmark.

4.2 Ablation Study

To isolate the contribution of symbolic role embeddings, we conduct an ablation study (Table 3) removing the role projection while retaining the dual attention pathways and fusion gating. Since

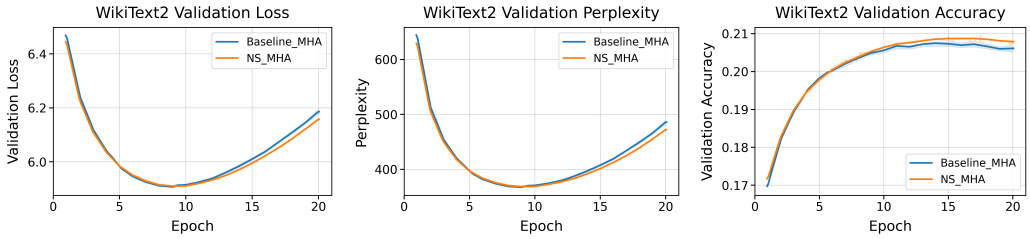

Figure 2: Training dynamics of MHA and NS-MHA on WikiText-2. Curves show mean performance across seeds with shaded regions representing standard deviation.

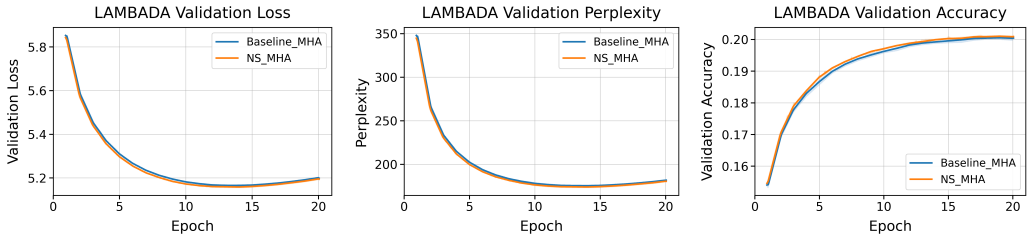

Figure 3: Training dynamics of MHA and NS-MHA on the LAMBADA benchmark. Curves show mean performance across seeds with shaded regions representing standard deviation.

the QKV projection sets, the dominant source of added parameters, are identical between the ablated and full models, and only the comparatively small W_r, P, and W_b are removed, the ablated model is a near-parameter-matched control rather than a reduced-capacity baseline. It still outperforms standard MHA on most benchmarks and metrics, showing that dual attention and gating alone provide a slight inductive bias, but the full NS-MHA consistently performs best, confirming gains stem from role-conditioned structure rather than parameterisation alone. These results use single representative runs rather than the 10-seed protocol used elsewhere, so we treat them as indicative rather than confirmatory.

5 Discussion

NS-MHA integrates symbolic inductive bias directly into the transformer attention pipeline, within a single continuous, gradient-trained representation space (Besold et al., 2017; Serafini & Garcez, 2016). Across the four benchmarks, this bias is not universally beneficial but gives clear advantages when relational structure or multi-step reasoning is required, while remaining competitive on standard language modelling. Training curves are provided in Figure 2, 3, and 4 for WikiText-2, LAMBADA, and SCAN respectively.

5.1 Language Modelling: WikiText-2

WikiText-2 (Merity et al., 2017) consists of long-form Wikipedia text and primarily tests a model's ability to capture lexical, syntactic, and semantic regularities over extended contexts, the standard usage scenario for MHA. NS-MHA is comparable to, and in most cases slightly better than, standard MHA in loss, perplexity, and accuracy ($p < 0.01$), which is expected since WikiText-2 is dominated by surface-level statistical patterns that leave the symbolic pathway less activated. The perplexity reduction nonetheless shows the symbolic projection does not interfere with language modelling and may mildly regularise, confirming NS-MHA is competitive rather than a specialised reasoning-only architecture.

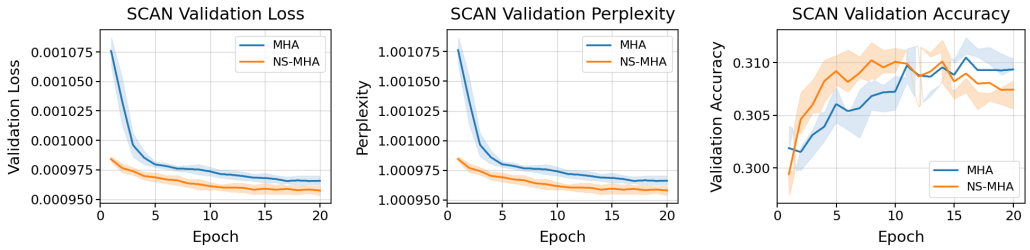

Figure 4: Training dynamics of MHA and NS-MHA on the SCAN benchmark. Curves show mean performance across seeds with shaded regions representing standard deviation.

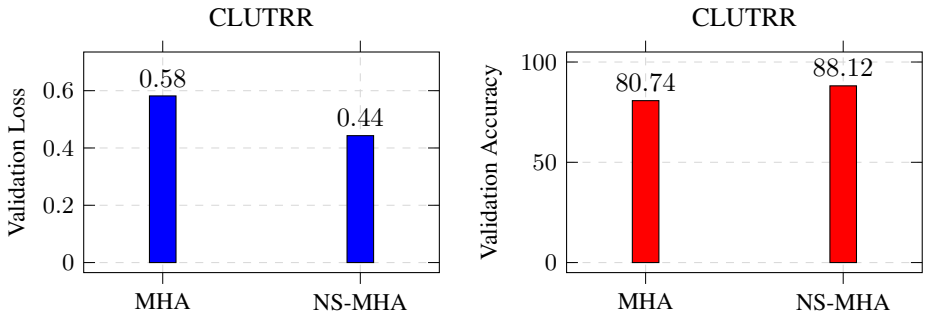

Figure 5: Comparison of MHA and NS-MHA on the CLUTRR relational reasoning benchmark. Bars show mean performance across seeds.

5.2 Relational Reasoning: CLUTRR

The CLUTRR benchmark (Sinha et al., 2019) evaluates multi-step relational and logical reasoning over short natural-language stories via reasoning chains of length 2–10 hops, testing compositional generalisation and predicate–argument structure. NS-MHA clearly and consistently outperforms MHA, converging faster with substantially higher asymptotic accuracy and lower validation loss (Figure 5). The near-zero variance across seeds reflects CLUTRR's template-driven construction. The gap widens as training progresses, suggesting the symbolic pathway supports more stable learning over longer chains, consistent with prior findings that purely neural attention struggles with systematic generalisation beyond seen templates (Lake & Baroni, 2018; Geiger et al., 2019).

The ablation study confirms this, as the ablated (no-role) and full NS-MHA models achieve similar accuracy given CLUTRR's near-saturated performance, but full NS-MHA's lower validation loss shows role embeddings improve representation quality and stability, not just top-1 accuracy, so the gains stem from role-conditioned structure rather than architectural capacity alone.

5.3 Long-Range Language Modeling: LAMBADA

LAMBADA (Paperno et al., 2016) requires predicting a passage's final word from broad discourse context, integrating coherence and entity tracking across entire passages rather than WikiText-2's local next-token prediction. NS-MHA consistently outperforms MHA in loss, perplexity, and accuracy. Differences are small but statistically significant ($p < 0.01$), suggesting that the symbolic pathway helps with implicit relational structure, placing LAMBADA between purely statistical modelling and explicit relational reasoning.

5.4 Compositional Generalisation: SCAN

SCAN evaluates compositional generalisation through deterministic command-to-action mappings (Lake & Baroni, 2018), exhibiting low output entropy and requiring structured, rule-like transformations rather than probabilistic language patterns. NS-MHA achieves consistently lower validation

loss and perplexity than MHA across all epochs ($p < 0.05$, Figure 4), while accuracy is nearly saturated for both, with a small decrease for NS-MHA. This is expected given SCAN's determinism. Once the mapping is learned, accuracy rapidly ceilings, so loss and perplexity are the more sensitive metrics, and both indicate NS-MHA produces more confident, better-calibrated predictions despite similar top-1 accuracy. The ablation performs consistently between MHA and the full model, improving over MHA but not matching NS-MHA, confirming the gains stem from role-conditioning rather than added capacity.

5.5 Computational Complexity and Trade-offs

Standard MHA has asymptotic complexity $O(n^2d+nd^2)$ (Vaswani et al., 2017), where n is sequence length and d hidden dimensionality. The quadratic $O(n^2d)$ term dominates in practice. NS-MHA preserves this complexity, as its two parallel attention streams each cost $O(n^2d)$, giving

$$O(n^2d) + O(n^2d) = 2 \times O(n^2d) = O(n^2d),$$

with the duplicated QKV projections adding only a constant-factor $2 \times O(nd^2)$ and the fusion gate a negligible $O(nd)$. For $d = 512$, $k = 8$ (Table 1), standard MHA requires $4d^2 = 1{,}048{,}576$ parameters per layer. NS-MHA adds a second QKV set ($3d^2 = 786{,}432$), the fusion gate ($2d^2 + d = 524{,}800$), and W_r with P ($2dk = 8{,}192$), a known overhead of $1{,}319{,}424$ before W_b, whose contribution depends on the (unfixed) relational subspace dimensionality.

This overhead is manageable given modern GPU resources and reflects a targeted inductive bias rather than an indiscriminate capacity increase, consistent with observed runtimes showing a slight per-epoch increase but the same quadratic scaling as standard attention. By embedding symbolic bias directly inside attention, NS-MHA avoids external reasoning modules and staged training pipelines, simplifying optimisation.

5.6 Implications for Neuro-Symbolic Attention

Symbolic inductive bias is not universal but targeted. The benefits are largest on CLUTRR, where the ablation confirms role embeddings, not added parameters, drive the gain, modest but consistent on WikiText-2 and LAMBADA, and reflected in calibration rather than accuracy on SCAN. NS-MHA is therefore best viewed as a drop-in replacement that helps where relational structure matters and is neutral elsewhere.

5.7 Limitations and Future Work

Gains on WikiText-2 and LAMBADA are modest, given the surface-level statistical nature of these tasks, and the dual pathways add constant-factor overhead despite unchanged asymptotic complexity. CLUTRR's near-zero cross-seed variance reflects its template-driven construction, so broader relational benchmarks would strengthen generalisation claims. We compare only against standard MHA, not other structured-attention baselines (e.g., relative-position or additive relational-bias variants), so whether the gains are specific to role-conditioning remains open, as does an interpretability or seed-stability analysis of the learned roles. Finally, our controlled small-scale evaluation leaves NS-MHA's behaviour within large-scale pre-trained transformers an open question.

6 Conclusion

This work shows that structured symbolic inductive bias can be embedded directly inside transformer attention, rather than through an external module, without sacrificing differentiability or drop-in compatibility. NS-MHA augments MHA with a role-conditioned pathway alongside the standard neural stream, and across four benchmarks it consistently matches or improves upon standard MHA. Gains are modest on language modelling but substantial on relational reasoning, and on SCAN it improves probabilistic modelling while maintaining comparable accuracy.

These improvements hold across 10 paired seeds with statistically significant differences, and the ablation study indicates they stem from role-conditioned structure rather than architectural duplication. NS-MHA preserves standard MHA's asymptotic complexity with only a small constant-factor overhead, making it a practical drop-in replacement for standard attention.

AI USE STATEMENT

In this work, we used generative AI tools for polishing our writing and to assist us in the discovery of related work. We have not used generative AI tools for any other task.

REFERENCES

Shivani G Aithal, Abishek B Rao, Chandrakala C B, and Sanjay Singh. Application of neuro-symbolic reasoning in natural language processing. In *2022 IEEE 6th Conference on Information and Communication Technology (CICT)*, pp. 1–5, 2022. doi: 10.1109/CICT56698.2022.9997814.

John Arevalo, Thamar Solorio, Manuel Montes-y Gómez, and Fabio A. González. Gated multimodal units for information fusion. In *International Conference on Learning Representations (ICLR) Workshop Track*, 2017.

Tarek R. Besold, Artur d'Avila Garcez, Sebastian Bader, Howard Bowman, Pedro Domingos, Pascal Hitzler, Kai-Uwe Kühnberger, Luis C. Lamb, Daniel Lowd, Priscila Machado Vieira Lima, Leo de Penning, Gadi Pinkas, Hoifung Poon, and Gerson Zaverucha. Neural-symbolic learning and reasoning: A survey and interpretation. *arXiv preprint arXiv:1711.03902*, 2017.

Dexiong Chen, Leslie O'Bray, and Karsten Borgwardt. Structure-aware transformer for graph representation learning. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 3469–3489. PMLR, 2022.

Brandon Colelough and William Regli. Neuro-symbolic AI in 2024: A systematic review. In *Proceedings of the First International Workshop on Logical Foundations of Neuro-Symbolic AI (LNSAI 2024)*, 2024.

David Demeter and Doug Downey. Just add functions: A neural-symbolic language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 7634–7642, 2020.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 4171–4186, 2019.

Cameron Diao and Ricky Loynd. Relational attention: Generalizing transformers for graph-structured tasks. In *International Conference on Learning Representations*, 2023.

Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. Neural logic machines. In *International Conference on Learning Representations*, 2019.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, pp. 1383–1392, 2018.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61(1):1–64, 2018.

Artur d'Avila Garcez and Luis C. Lamb. Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023.

Artur d'Avila Garcez, Krysia Broda, and Dov M. Gabbay. *Neural-Symbolic Learning Systems: Foundations and Applications*. Springer, 2012.

Atticus Geiger, Ignacio Cases, Lauri Karttunen, and Christopher Potts. Posing fair generalization tasks for natural language inference. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 4485–4495, 2019.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8): 1735–1780, 1997.

Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

Maxwell Joseph Jacobson and Yexiang Xue. Integrating symbolic reasoning into neural generative models for design generation, 2024. URL https://arxiv.org/abs/2310.09383.

Brenden M. Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 2873–2882, 2018.

Luís C. Lamb, Artur d'Avila Garcez, Marco Gori, Marcelo O.R. Prates, Pedro H.C. Avelar, and Moshe Y. Vardi. Graph neural networks meet neural-symbolic computing: a survey and perspective. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*, IJCAI'20, pp. 4877–4884, 2020.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Francesco Locatello, Dirk Weissenborn, Thomas Unterthiner, Aravindh Mahendran, Georg Heigold, Jakob Uszkoreit, Alexey Dosovitskiy, and Thomas Kipf. Object-centric learning with slot attention. In *Advances in Neural Information Processing Systems*, volume 33, pp. 11525–11538, 2020.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. In *International Conference on Learning Representations*, 2017.

Sariah Mghames, Luca Castri, Marc Hanheide, and Nicola Bellotto. A neuro-symbolic approach for enhanced human motion prediction. In *2023 International Joint Conference on Neural Networks (IJCNN)*, 2023.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1525–1534, 2016.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. OpenAI Blog, 2019. URL https://cdn.openai.com/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. In *Advances in Neural Information Processing Systems*, volume 30, pp. 3788–3800, 2017.

Imanol Schlag, Paul Smolensky, Roland Fernandez, Nebojsa Jojic, Jürgen Schmidhuber, and Jianfeng Gao. Enhancing the transformer with explicit relational encoding for math problem solving. *arXiv preprint arXiv:1910.06611*, 2019.

Luciano Serafini and Artur d'Avila Garcez. Logic tensor networks: Deep learning and logical reasoning from data and knowledge. In *Proceedings of the 11th International Workshop on Neural-Symbolic Learning and Reasoning (NeSy)*, 2016.

Koustuv Sinha, Shagun Sodhani, Jin Dong, Joelle Pineau, and William L. Hamilton. CLUTRR: A diagnostic benchmark for inductive reasoning from text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pp. 4506–4515, 2019.

Son Tran, Edjard Mota, and Artur d'Avila Garcez. Reasoning in neurosymbolic AI, 2025. URL https://arxiv.org/abs/2505.20313.

Eleni Tsamoura, Timothy Hospedales, and Loizos Michael. Neural-symbolic integration: A compositional perspective. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 5051–5060, 2021.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph attention networks. In *International Conference on Learning Representations*, 2018.

Frank Wilcoxon. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Xin Zhang and Victor S. Sheng. Bridging the gap: Representation spaces in neuro-symbolic AI. *arXiv preprint arXiv:2411.04393*, 2024.