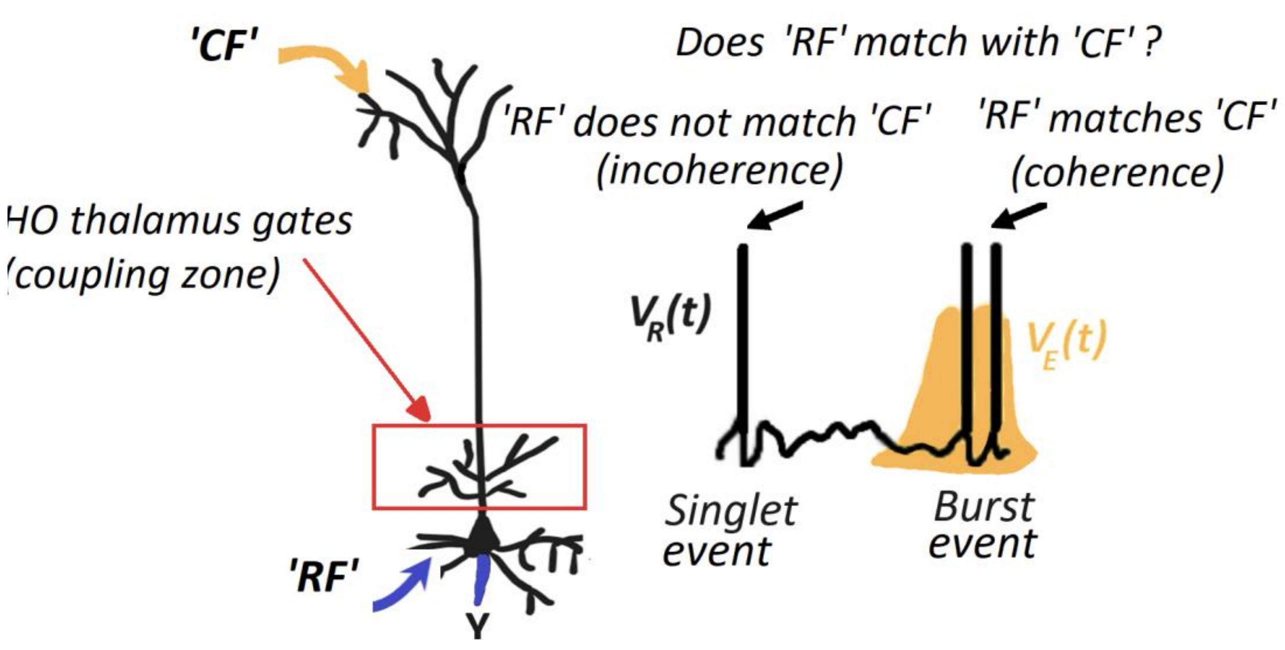
Single layer tiny Co^4 outpaces GPT-2 and GPT-BERT

Noor Ul Zain, Mohsin Raza and Ahsan Adeel
CMI Lab, University of Stirling, UK
Email: ahsan.adeel1@stir.ac.uk

Funded by ARIA (Advanced Research + Invention Agency)

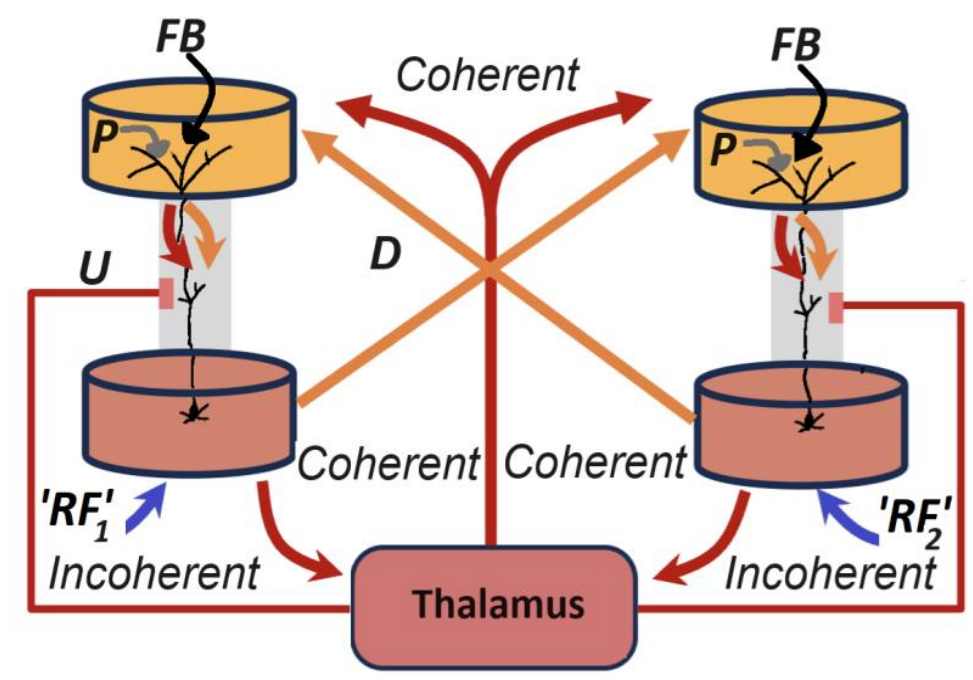
Information processing in Two-Point Neurons in Mammalian Neocortex



A pyramidal neuron in the neocortex integrates inputs separately in two zones: contextual feedback (CF) at the apical dendrites and feedforward receptive field (RF) signals at the basal compartment. When both are activated at the same time, the neuron produces high-frequency bursts, which amplify coherent, context-relevant signals [1][2].

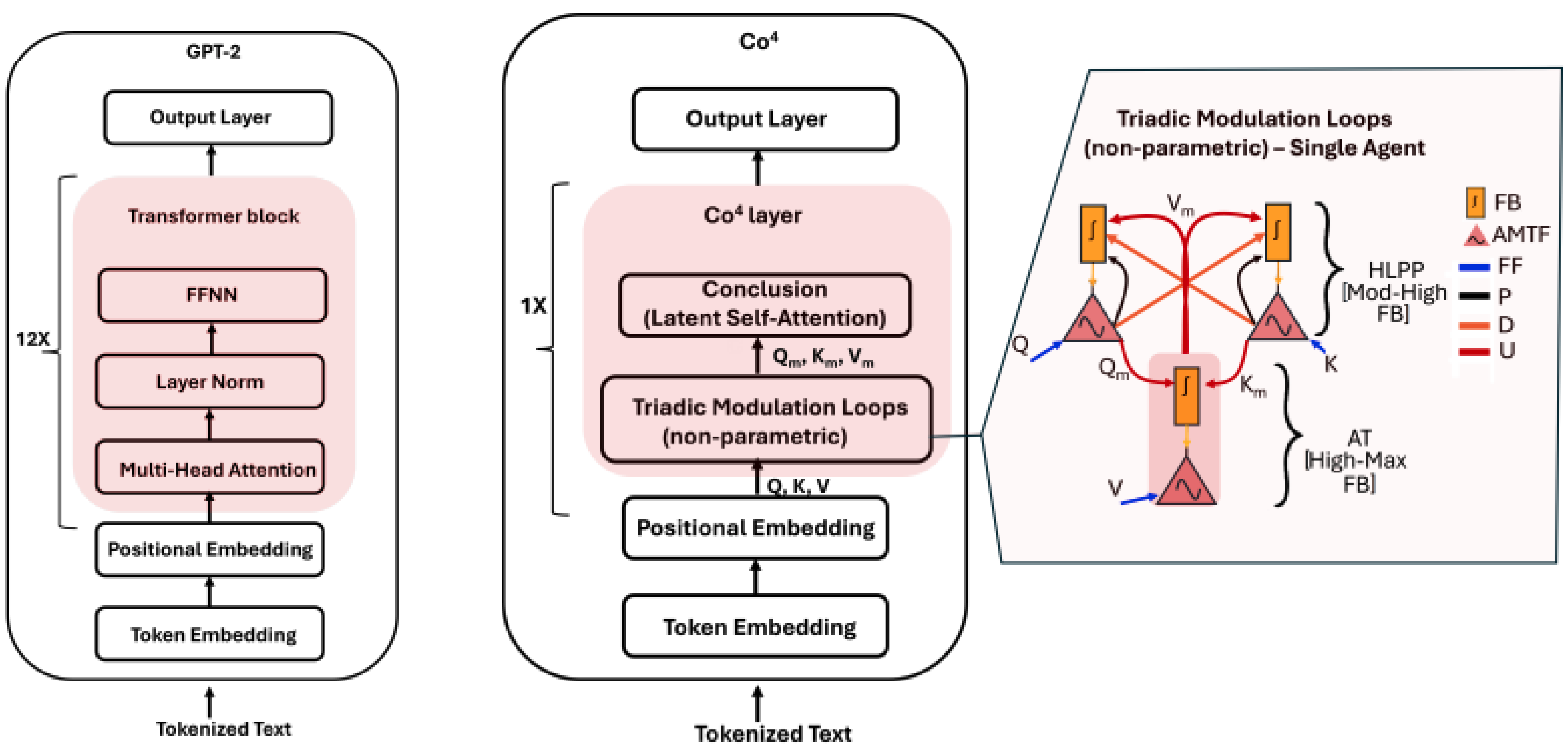
Towards 21st-Century Cellular Neurobiology-Inspired Machines with Intrinsic Higher Mental States

An example of "thinking fast and slow", illustrates how solving a riddle can involve combining rapid, intuitive processing in high-level perceptual state (fast thinking) with more deliberate, reflective refinement in the awake thought state (slow thinking) [1].



Cellular mechanisms of conscious processing: In the conscious state, Two Point Neurons cooperate moment-by-moment via different kinds of context (feedback (FB), proximal (P), distal (D), and universal (U)) [3] [4].

Co^4 Language Model



GPT-2 vs. Co⁴ Architectures. *In Co⁴ , the learnable parameters are only in the embedding layer and the initial Q, K, V representations, followed by a single layer of non-parametric triadic modulation loops (referred to as "1x" or single-layered Co⁴). The triadic modulation loop is composed of three pyramidal two-point processors (Q, K, V), each integrating feedforward input at the basal site and contextual feedback at the apical site. Through asynchronous modulation, relevant signals are selectively amplified while irrelevant ones are attenuated, enforcing local competition and global coherence. Twelve such agents operate in parallel before latent self-attention, reducing complexity toward O(N) while encouraging context-aligned, biologically grounded processing.*

Results

Co⁴ *with a single layer, two attention heads, and O(N) computational cost is an 8M parameter model. Despite its simplicity, Co⁴ outperforms BabyLM baselines—GPT-2 (124M, 12 layers, O(N²)) and GPT-BERT (30M, 12 layers, O(N²))—after just two training epochs (vs. ten). It shows strong zero-shot and fine-tuning performance on SuperGLUE: it surpasses GPT-2 on 5/7 zero-shot and 6/7 fine-tuning tasks, and GPT-BERT on 4/7 in both cases. These findings highlight Co₄ as a lightweight, efficient alternative to traditional large-scale transformers—challenging existing scaling laws and deep learning paradigms.*

Zero-shot metrics comparison

Metric	GPT-2	Co^4-α	GPT-BERT	Co^4-β
Eye Tracking	8.66	**8.67**	**9.89**	8.19
Self-paced Reading	4.34	**4.59**	3.45	**3.62**
WUGs	52.50	**68.00**	43.00	**93.00**
Entity Tracking	13.90	**26.71**	33.96	**41.36**
EWoK	49.90	**50.01**	49.49	**50.11**
BLiMP	**66.36**	53.55	**71.66**	51.20
BLiMP Supplement	**57.07**	52.59	**63.21**	49.82

Finetuning: SuperGLUE tasks comparison

Task	Metric	GPT-2	GPT-BERT	Co^4
MRPC	F1	80.77	83.44	**84.15**
QQP	F1	62.45	**72.03**	62.73
BoolQ	Accuracy	66.91	68.07	**69.05**
MNLI	Accuracy	**51.12**	46.86	44.25
MultiRC	Accuracy	65.72	**68.28**	66.01
RTE	Accuracy	56.83	56.12	**59.71**
WSC	Accuracy	61.54	65.38	**67.31**

[1] Adeel, A., "Beyond Attention: Toward Machines with Intrinsic Higher Mental States", 2025: https://arxiv.org/pdf/2505.06257
[2] Phillips, W. A., et al., "Cellular psychology: relating cognition to context-sensitive pyramidal cells". Trends in Cognitive Sciences 2025: https://doi.org/10.1016/j.tics.2024.09.002
[3] Adeel, A., "Cellular Foundations of Common Sense and Imaginative Thought: Can They Be Intrinsically Embodied in Machines?", 2024 (EMBO) Workshop
[4] Aru, J., et al., "Cellular Mechanisms of Conscious Processing". Trends in Cognitive Sciences 2020: https://doi.org/10.1016/j.tics.2020.07.006