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

Noor Ul Zain[1], Mohsin Raza[1], Ahsan Adeel[1,*]
[1] CMI-Lab
University of Stirling, UK
[*] ahsan.adeel1@stir.ac.uk

Abstract

We show that a tiny Co^4 machine (Adeel, 2025) with a single layer, two heads, and 8M parameters, operating at an approximate cost of $O(N)$ (where N is the number of input tokens), outpaces the BabyLM Challenge baselines GPT-2[1] (124M, 12 layers, $O(N^2)$) and GPT-BERT[2] (30M, 12 layers, $O(N^2)$) in just two epochs, while both are trained for ten. Co^4 achieves orders-of-magnitude greater training efficiency on 10M tokens, demonstrating highly sample-efficient pretraining. Using the BabyLM challenge evaluation pipeline across complex benchmarks, Co^4 exhibits strong zero-shot and fine-tuning performance on SuperGLUE tasks. Specifically, Co^4 outperforms GPT-2 on 5 out of 7 zero-shot metrics and 6 out of 7 fine-tuning tasks, and GPT-BERT on 4 out of 7 metrics in both cases. These results suggest the need to rethink prevailing deep learning paradigms and associated scaling laws.

Cellular neurobiological evidence (Suzuki et al., 2023; Marvan and Phillips, 2024) on how mammalian brains achieve fast and flexible computation continues to challenge deep (hierarchical) learning (LeCun et al., 2015; Vaswani et al., 2017; Wang et al., 2025), predictive coding (Rao and Ballard, 1999; Friston, 2005, 2010), and scaling laws (Kaplan et al., 2020). Evidence suggests that the brain's computational power lies in shallow architectures, where cortical and subcortical networks operate with massive parallelism, leveraging cortical microcircuits and thalamo-cortical loops (Aru et al., 2021; Storm et al., 2024; Phillips et al., 2024) to support faster, context-sensitive, and coherent internal understanding (Adeel, 2025).

Modern deep learning architectures, such as Transformers (Vaswani et al., 2017; Jaegle et al., 2021; Alayrac et al., 2022), which underpin models like GPT and GPT-BERT, act as sequential local agents reducing predictive error or free energy (Friston, 2005, 2010), yet without regard for local coherence (Marvan and Phillips, 2024). During the feedforward (FF) phase, they lack intrinsic mechanisms to judge the true relevance of an attended token (Adeel, 2025). Instead, relevance is indirectly shaped by backpropagation during the feedback (FB) phase, a brute-force, reward-driven process. Incoherent inferences generated by initial agents (e.g., early transformer blocks) propagate to subsequent agents, where they are reinforced through ineffective FB signals. We refer to this as a "Chinese Whispers" problem.

Consequently, these deep nets require vast datasets, extensive training time, and significant compute, resulting in unsustainable economic, environmental, and technical costs (Thompson et al., 2020). The reliance on deeper architectures for hierarchical feature abstraction is a shared limitation across other neural models, including long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997), gated-recurrent units (GRUs) (Chung et al., 2014), and convolution neural networks (CNNs) (LeCun et al., 1989).

The recently proposed Co^4 machine (Adeel, 2025) emulates higher-level perceptual processing (HLPP) and awake thought (AT) mental states (Phillips et al., 2024). Within a single layer, during FF, it executes triadic FB loops among latent questions (Qs), clues (Ks), and hypotheses (Vs), enabled by three two-point neurons (TPNs)[3] (Aru et al., 2021; Storm et al., 2024; Phillips et al., 2024), each representing an agent holding K, Q, and V. Unlike Transformers, which propagate layer-wise,

[1]https://huggingface.co/BabyLM-community/babylm-baseline-10m-gpt2

[2]https://huggingface.co/BabyLM-community/babylm-baseline-10m-gpt-bert-causal-focus

[3]A pyramidal two-point neuron in the mammalian neocortex integrates feedforward input at its basal site and contextual input at its apical dendrites. When both are aligned in time, the neuron fires bursts that amplify coherent, contextually relevant signals for active inference.

Proceedings of the First BabyLM Workshop, pages 313–322
November 8, 2025

Co^4 enables all agents to co-evolve Qs, Ks, and Vs in parallel: Qs update based on Ks and Vs; Ks update based on Qs and Vs; Vs evolve based on Ks and Qs. Each TPN agent independently forms distinctive Q–K–V perspectives, thereby maximizing local and global coherence (Marvan and Phillips, 2024) while minimizing free energy (Friston, 2005, 2010), ensuring token relevance before attention is applied or decisions are made. This cooperative mechanism enables diverse, parallel, and deep reasoning chains without requiring additional layers, at an approximate cost of $O(N)$ (Adeel, 2025).

This paper is the first to report the Co^4 machine's performance on complex language benchmarks. From a cognitive modeling perspective, we compare training trajectories of Co^4, GPT-2, and GPT-BERT to those of children using psycholinguistic metrics under data-limited conditions modeled after human language acquisition (Charpentier et al., 2025). Despite its tiny size, just one layer, two heads, and 8M parameters, Co^4 (with $O(N)$ cost) outpaces GPT-2 (124M parameters) and GPT-BERT (30M), both using 12 layers ($O(N^2)$ cost), achieving orders-of-magnitude greater efficiency and stronger generalization on a 10M-token dataset.

1 Neurons and Co^4 agents with two points of input integration

Going beyond the 20^{th}-century neuroscience conception of point neurons (PNs) (Häusser, 2001), on which most current brain theories and AI systems are based, 21^{st}-century neuroscience (Larkum et al., 1999; Phillips, 2017, 2023; Larkum, 2013; Major et al., 2013; Ramaswamy and Markram, 2015; Larkum, 2022; Adeel, 2020; Körding and König, 2000; Schuman et al., 2021; Poirazi and Papoutsi, 2020; Larkum et al., 2018; Shine et al., 2016, 2019; Shine, 2019; Shine et al., 2021; Schulz et al., 2021; Kay and Phillips, 2020; Kay et al., 2022) has revealed that certain neurons, particularly some pyramidal neurons in the mammalian neocortex, integrate inputs at two distinct locations. These are often referred to as TPNs, which combine information from the external environment (feedforward (FF)) at one site (basal) and contextual (C) input at another (apical). TPNs trigger high-frequency firing (bursting) when the FF and C inputs are matched in time, that is, when both the basal and apical zones are depolarized. This results in the amplification of coherent signals, enabling enhanced contextually rich processing (Phillips et al., 2024).

The flexible interaction between FF and C inputs is suggested to be the hallmark of conscious processing (Aru et al., 2021; Storm et al., 2024; Marvan et al., 2021) and linked to distinct mental states, including wakefulness (WF), slow-wave (SW) sleep, and rapid eye movement (REM) sleep (Phillips et al., 2024). Dysfunctional interactions between FF and C inputs have been linked to intellectual learning disabilities (Nelson and Bender, 2021; Granato et al., 2024).

Several TPN-inspired machine learning algorithms have been proposed to flexibly combine top-down C and bottom-up FF information streams (Payeur et al., 2021; Greedy, 2022; Guerguiev et al., 2017; Sacramento et al., 2018; Illing et al., 2022; Greedy, 2022; Zenke et al., 2017; Kirkpatrick et al., 2017; Kastellakis et al., 2016; Bono and Clopath, 2017; Limbacher and Legenstein, 2020). However, most of these efforts have focused on using apical (contextual) inputs primarily for learning. Ample evidence suggests that the apical site not only receives feedback from higher perceptual levels but also integrates simultaneous events across multiple hierarchical levels while processing FF information. For example, results using TPN-inspired CNNs (Adeel, 2020; Adeel et al., 2022, 2023; Raza and Adeel, 2024) showed that these architectures could drastically reduce the transmission of conflicting FF signals to higher perceptual areas, achieving orders-of-magnitude reductions in the number of neurons needed to process heterogeneous real-world audio-visual data, compared to standard PN-based CNNs. More recent findings demonstrate that the TPN-inspired Co^4 machine (Adeel, 2025), emulating higher level perceptual processing and imaginative thought mental states can enable significantly faster learning with substantially lower computational demands (e.g., fewer heads, layers, and tokens) at an approximate cost of $O(N)$. These gains were observed across a variety of domains, including reinforcement learning, computer vision, and natural language question answering.

These efforts to develop efficient machine learning models align with scaled-down pretraining using fewer than 100M tokens, evaluating language models (LMs) on the same types and quantities of data that humans are exposed to (Charpentier et al., 2025). The aim is to build plausible cognitive models of human learning and to better understand how children are exposed to language with such efficiency. By combining cellular neurobiologically inspired, TPN-based Co^4 machine (Adeel, 2025) with this scaled-down pretraining strategy, we introduce the Co^4 LM.

2 Co^4 Language Model

Figure 1 (left) illustrates the standard GPT-2 model, consisting of 12 Transformer layers, where each layer performs a simple conclusion via self-attention (QK^TV) at the cost of $O(N^2)$. This can be interpreted as 12 agents working sequentially. The selection of relevant and irrelevant tokens in the FF phase is determined through backpropagation, a brute-force process solely driven by the global objective. This rigidity causes the network to depend heavily on pre-learned patterns, limiting its ability to generate new perspectives quickly. When initial thoughts are misleading, arriving at a correct conclusion may require significantly more time and computation, or may not happen at all, due to limited internal flexibility and constrained cognitive resources (Adeel, 2025).

In contrast, Figure 1 (right) shows a single-layer Co^4 machine with two attention heads. After initializing the latent queries (Qs) as a set of neuronal agents (e.g., 24) (as opposed to 12 attention blocks + feedforward neuron network (FFNN) in GPT-2 and GPT-BERT), they begins to co-evolve their own Qs, Ks, and Vs in parallel during the FF phase via triadic modulation loops leveraging proximal (P), distal (D), and universal (U) contextual fields. This co-evolution is enabled through inherent, moment-by-moment cooperation mechanisms or asynchronous modulation (MOD) transfer function (Adeel, 2025), resulting in rich, contextually-aware, and diverse parallel reasoning chains at the cellular level. Each agent independently develops its own Q, K, and V, leading to 24 attention maps and 24 possibly different conclusions. Importantly, this all occurs virtually, allowing the model to preselect relevant tokens before applying latent self-attention at an approximate cost of $O(N)$ (Adeel, 2025).

The Co^4 language model frames text generation as an autoregressive, left-to-right process: given a prefix of tokens, the model computes a probability distribution over the next token via a softmax applied to its hidden state. We use the same tokenizer as the baselines. The input tokens are first mapped to continuous vectors through a embedding layer and are augmented with positional embeddings to encode sequence order. During training, a triangular causal mask ensures that each position can only attend to previous positions. The model's weights are optimized by minimizing the cross-entropy (CE) loss (equivalently, the negative log-likelihood) of the true next token.

The Co^4 language model condenses this pipeline into a single decoder layer with just two attention heads, yet enriches it via triadic modulation loops among Q-, K-, and V-TPNs, operating through P, D, and U contextual fields (Adeel, 2025, 2020). After token embedding and positional projection, each token's Q, K, and V vectors co-evolve through a series of rapid and modulated updates.

We trained Co^4 on a 10M-token slice of the BabyLM corpus (BabyLM Community, 2023), using the same autoregressive CE objective but at a fraction of the training budget of GPT-2 and GPT-BERT, which are the official baselines provided by the organizers of this challenge. More details related to the hyperparameters for these baselines can be found on the relevant model repositories on Hugging Face.

3 Results

In this section, we present the performance of our tiny Co^4 machine across a range of language modeling benchmarks. The seven tasks described first assess the model's linguistic capabilities in a purely zero-shot setting, without any additional training or fine-tuning. Later in the section, we also evaluate Co^4's performance on fine-tuning benchmarks and provide an extensive comparison with the baseline. We utilize the evaluation suite from the BabyLM Challenge (Charpentier et al., 2025), which includes the following zero-shot metrics. The first two, newly introduced, are designed to compare the language model's responses to those of human judgments and behavioral data.

- Eye Tracking and Self-paced Reading: This psycholinguistic measure evaluates whether the model can mimic the eye tracking and reading time of a human by using the surprisal of a word as a proxy for time spent reading a word (de Varda et al., 2024).
- WUGs: morphological Adapting the classic "Wug" paradigm, this evaluates whether models can generalize morphological rules to form novel noun derivatives from unseen adjectives, and compares the model's generalization to that of humans (Hofmann et al., 2025).

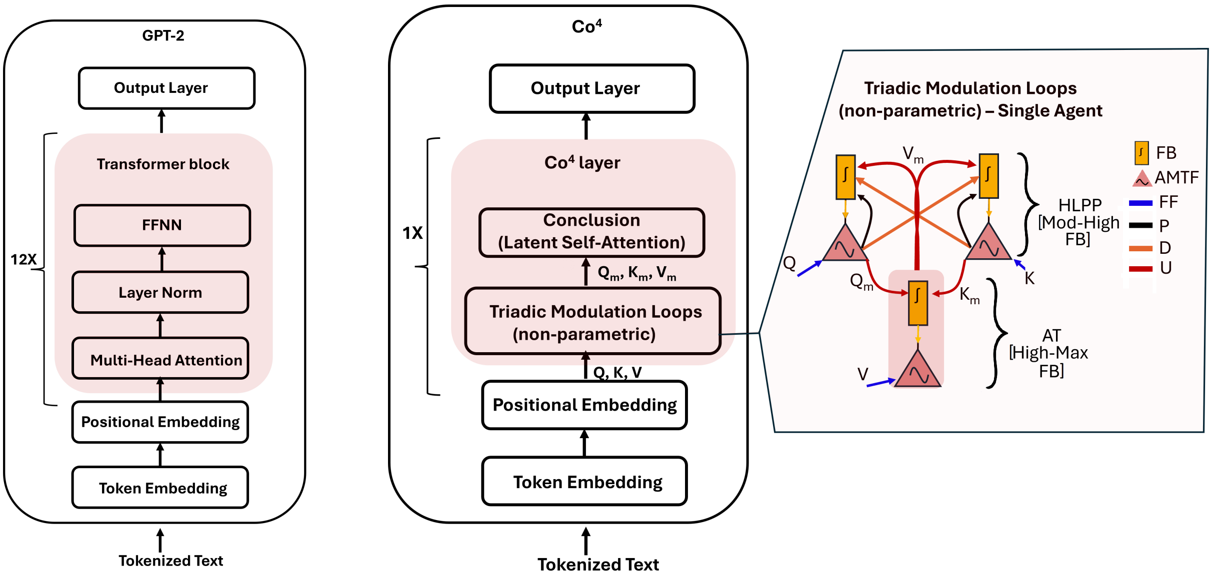

Figure 1: Language Models: GPT-2 (Left) vs. Co^4 (Right). In Co^4, the learnable parameters are only in the embedding layer and the initial Q, K, V representations, followed by a single layer of non-parametric triadic modulation loops (referred to as "1x" Co4 or single-layered Co4). Co^4 does not require feed feed-forward neural network (FFNN/ MLP) layer used in standard GPT-type architectures. Inside these loops, three populations of three pyramidal two-point processors, each associated with Q, K, and V, respectively, simultaneously integrate FF information and FB context at two functionally distinct sites. The apical (top-down) site (shown in the rectangle) integrates context, while FF information is integrated at the basal (bottom-up) site (shown in the triangle). Each processor, via asynchronous modulation (MOD) transfer functions[4], operating in higher-level perceptual processing (HLPP) or awake thought (AT) mode, depending on the strength of FB, amplifies FF transmission if it is relevant in that context (represented by P, D, U). Otherwise, it attenuates the signal, resulting in the selective amplification of coherent FF information (Adeel, 2025). P, D, and U, along with the credit assignment (reward) coming from the higher perceptual layer (teacher), can be seen as dynamic local competitive normalization and global cooperative organisation, respectively. This ensures that local and global coherence and consistency are maximized (Marvan and Phillips, 2024), while prediction error or free energy (Friston, 2005, 2010) is minimized, enabling a deeper form of "real understanding". A combination of three TPNs and one loop constitutes one agent. A set of 12 agents with 12 loops runs in parallel, evolving their Qs, Ks, and Vs simultaneously, before applying latent self-attention at $O(L \times N)$ where L is a small fraction of the input sequence length, making the overall cost approximately $O(N)$.

- Entity Tracking: Probes a model's capacity to update and maintain the state of entities throughout a narrative or dialogue by asking it to predict an entity's final condition after a series of changes (Kim and Schuster, 2023).
- EWoK: This benchmark evaluates the model's internal world knowledge across domains like spatial relations and social interactions (Ivanova et al., 2024).
- BLiMP: Testing various grammatical phenomenon, the Benchmark of Linguistic Minimal Pairs evaluates whether a model consistently picks the grammatically correct alternative from a pair of minimally different sentences (Warstadt et al., 2020).
- BLiMP Supplement: This is a supplement to BLiMP and was introduced in the first edition of the BabyLM challenge. It is more focused on dialogue and questions (Warstadt et al., 2025).

The metrics used to evaluate the model on each of these zero-shot benchmarks are as follows:

- Accuracy in predicting the correct completion or sentence for BLiMP, BLiMP Supplement, EWoK, Entity Tracking, and WUGs.
- Change in R^2 prediction from baseline for Eye Tracking and Self-paced Reading.

Table 1 shows the performance of tiny Co^4 language model on the metrics outlined above. As shown, our computationally efficient model, Co^4-α, outperforms GPT-2 on 5 out of 7 metrics. As for

[4]For the mathematical details of these functions and the core mechanism behind triadic modulation loops, please check (Graham et al., 2025).

GPT-BERT, another configuration Co^4-β, outperforms it on 4 out of 7 metrics. These hyperparameters for these configurations are further outlined in the Appendix.

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

Table 1: **Zero-shot metrics comparison:** GPT-2 vs. Co^4-α and GPT-BERT (causal-focus) vs Co^4-β The single-layer, tiny Co^4 model outperformed GPT-2 on 5 out of 7 metrics, and GPT-BERT on 4 out of 7 metrics, despite being trained at a fraction of the computational cost, **in 2 epochs**.

Metric	GPT-2	GPT-BERT	Co^4-γ
Hypernym	48.93	49.05	**54.75**
QA Cong. Easy	50.00	67.19	**87.50**
QA Cong. Tricky	39.39	50.30	**53.94**
Subject Aux Inversion	**81.33**	81.28	65.48
Turn Taking	65.71	**68.21**	50.36
Overall	57.07	**63.21**	62.40

Table 2: **BLiMP Supplement benchmark**: Co^4-γ demonstrates superior performance in the BLiMP Supplement benchmark and the individual tasks in this benchmark. Although this configuration of Co^4-γ does not outperform the psycholinguistic metrics, it outperforms the baselines in the BLiMP Supplement.

Table 2 reports performance of Co^4-γ on the BLiMP Supplement benchmark. This Co^4-γ is a different configuration of our architecture, which notably performed better on BliMP Supplement. Since it did not beat most of the metrics, we did not pick it as our best configuration but we wanted to include its superior performance on BLiMP. It should be noted that our model performs better on BLiMP Supplement compared to BLiMP, suggesting that the Co^4 model has an inherent bias toward more complex tasks and long-term dependencies characteristic of BLiMP Supplement's subtasks. More challenging than the original BLiMP benchmark, BLiMP Supplement was introduced in the most recent version of the BabyLM Challenge (Charpentier et al., 2025). It is more challenging since models perform relatively lower on it as compared to BLiMP (Warstadt et al., 2025), and also because it consists of more dialogues and questions as compared to the minimally different sentences in BLiMP. It is comprised of the following five subtasks:

- Hypernym: Checks whether a word is correctly recognized as a superset or subset of another (e.g., a dog is a mammal, so having a dog implies having a mammal).
- QA Congruence Easy: Verifies whether the question type matches the answer (e.g., a who question is answered with a person rather than a thing).
- QA Congruence Tricky: Similar to QA Congruence Easy but with more ambiguous cases.
- Subject–Aux Inversion: Checks whether the auxiliary verb is correctly inverted with the subject (e.g., Is she coming?).
- Turn Taking: Checks whether the correct personal pronoun is used when answering a question in dialogue.

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**

Table 3: SuperGLUE tasks

Finetuning: Table 3 reports performance on SuperGLUE tasks as part of fine-tuning. (Wang et al., 2019). We picked our best Co^4 configuration overall (Co^4-α) for the finetuning. Our novel architecture achieves comparable results across most fine-tuning tasks and demonstrates better performance on 6 out of the 7 tasks when compared to GPT-2 and 4 out of the 7 tasks when compared to GPT-BERT. These tasks are:

- BoolQ: A yes/no question-answering dataset with unprompted and unconstrained questions (Clark et al., 2019)

- MNLI: The Multi-Genre Natural Language Inference corpus tests whether a model can recognize textual entailment (Williams et al., 2017).
- MRPC: The Microsoft Research Paraphrase Corpus contains pairs of sentences that are either paraphrases (semantically equivalent) or unrelated (Dolan and Brockett, 2005).
- QQP: Similarly to MRPC, the Quora Question Pairs corpus tests a model's ability to determine whether pairs of questions are semantically similar. These questions are sourced from Quora (BabyLM Community, 2023).
- MultiRC: The Multi-Sentence Reading Comprehension corpus evaluates a model's ability to select the correct answer from a list of candidates given a question and a context paragraph. In this version, the data is reformulated as a binary classification task judging whether an answer to a question-context pair is correct (Khashabi et al., 2018).
- RTE: Recognizing Textual Entailment tests the model's ability to recognize textual entailment (Dagan et al., 2005, 2022; Bentivogli et al., 2009).
- WSC: The Winograd Schema Challenge evaluates coreference resolution in sentences containing a pronoun and a list of noun phrases. This version reformulates the task as a binary classification problem using examples consisting of a pronoun and a noun phrase (Levesque et al., 2012).

The hyperparameters for this task are outlined in the Appendix.

4 Conclusion

The Co^4 model has a computational complexity of $O(L \cdot N + \alpha)$, scaling linearly with the number of input tokens (N), where L is the number of latent queries and α is a small fixed overhead. In contrast, models like GPT-2 and GPT-BERT scale quadratically at $O(N^2)$, making them significantly more expensive as input size grows. In standard Transformers, multiply–accumulate (MAC) operations grow with the quadratic term $P^2 \cdot E$ due to self-attention, where P is the number of tokens and E is the embedding dimension. In Co^4, this is replaced by a more efficient linear term $L_q \cdot P \cdot E$, enabled by a small set of latent queries. As a result, Co^4 achieves substantial computational savings and superior scalability over conventional Transformers. Despite being a single-layer model, the tiny Co^4 machine outperforms GPT-2 and GPT-BERT on most evaluated performance metrics, while requiring only a fraction of the computational resources. Future directions include scaling to larger datasets, integrating multi-objective or hybrid cost functions (e.g., those used in GPT-BERT), and evaluating different modes of apical operation (Phillips et al., 2024; Graham et al., 2024; Pastorelli et al., 2023). In addition, scaling beyond 8M parameters is part of ongoing work.

5 Acknowledgments

Advanced Research + Invention Agency (ARIA): Nature Computes Better Opportunity seeds. Professor Bill Phillips, Professor Leslie Smith, Professor Bruce Graham, and Dr Burcu Can Buglalilar from the University of Stirling. Professor Panayiota Poirazi from IMBB-FORTH, Professor Peter Konig from the University Osnabruck. Professor Heiko Neumann from Ulm University, Dr James Kay from the University of Glasgow, and several other eminent scholars for their help and support in several different ways, including reviewing this work, appreciation, and encouragement. We also acknowledge ChatGPT for its assistance with proofreading.

Competing interests The authors declare no conflict of interest.

References

Ahsan Adeel. 2020. Conscious multisensory integration: Introducing a universal contextual field in biological and deep artificial neural networks. *Frontiers in Computational Neuroscience*, 14.

Ahsan Adeel. 2025. Beyond attention: Toward machines with intrinsic higher mental states. *arXiv preprint arXiv:2505.06257*.

Ahsan Adeel, Adewale Adetomi, Khubaib Ahmed, Amir Hussain, Tughrul Arslan, and William A Phillips. 2023. Unlocking the potential of two-point cells for energy-efficient and resilient training of deep nets. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 7(3):818–828.

Ahsan Adeel, Mario Franco, Mohsin Raza, and Khubaib Ahmed. 2022. Context-sensitive neocortical neurons transform the effectiveness and efficiency of neural information processing. *arXiv preprint arXiv:2207.07338*.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Jaan Aru, Mototaka Suzuki, and Matthew Larkum. 2021. Cellular mechanisms of conscious processing. *Trends in Cognitive Sciences*, 25.

BabyLM Community. 2023. BabyLM Baseline 10M GPT-2. https://huggingface.co/BabyLM-community/babylm-baseline-10m-gpt2. Accessed: 2024-06-27.

Luisa Bentivogli, Peter Clark, Ido Dagan, and Danilo Giampiccolo. 2009. The fifth pascal recognizing textual entailment challenge. *TAC*, 7(8):1.

Jacopo Bono and Claudia Clopath. 2017. Modeling somatic and dendritic spike mediated plasticity at the single neuron and network level. *Nature communications*, 8(1):706.

Lucas Charpentier, Leshem Choshen, Ryan Cotterell, Mustafa Omer Gul, Michael Hu, Jaap Jumelet, Tal Linzen, Jing Liu, Aaron Mueller, Candace Ross, and 1 others. 2025. Babylm turns 3: Call for papers for the 2025 babylm workshop. *arXiv preprint arXiv:2502.10645*.

Junyoung Chung, Caglar Gulcehre, KyungHyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv:1412.3555*.

Christopher Clark, Kenton Lee, Ming-Wei Chang, Tom Kwiatkowski, Michael Collins, and Kristina Toutanova. 2019. Boolq: Exploring the surprising difficulty of natural yes/no questions. *arXiv preprint arXiv:1905.10044*.

Ido Dagan, Oren Glickman, and Bernardo Magnini. 2005. The pascal recognising textual entailment challenge. In *Machine learning challenges workshop*, pages 177–190. Springer.

Ido Dagan, Dan Roth, Fabio Zanzotto, and Mark Sammons. 2022. *Recognizing textual entailment: Models and applications*. Springer Nature.

Andrea Gregor de Varda, Marco Marelli, and Simona Amenta. 2024. Cloze probability, predictability ratings, and computational estimates for 205 english sentences, aligned with existing eeg and reading time data. *Behavior Research Methods*, 56(5):5190–5213.

Bill Dolan and Chris Brockett. 2005. Automatically constructing a corpus of sentential paraphrases. In *Third international workshop on paraphrasing (IWP2005)*.

Karl Friston. 2005. A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456):815–836.

Karl Friston. 2010. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138.

Bruce P Graham, Jim W Kay, and William A Phillips. 2024. Transfer functions for burst firing probability in a model neocortical pyramidal cell. *bioRxiv*, pages 2024–01.

Bruce P Graham, Jim W Kay, and William A Phillips. 2025. Context-sensitive processing in a model neocortical pyramidal cell with two sites of input integration. *Neural Computation*, 37(4):588–634.

Alberto Granato, William A Phillips, Jan M Schulz, Mototaka Suzuki, and Matthew E Larkum. 2024. Dysfunctions of cellular context-sensitivity in neurodevelopmental learning disabilities. *Neuroscience & Biobehavioral Reviews*, page 105688.

et al. Greedy, Will. 2022. Single-phase deep learning in cortico-cortical networks. *Advances in Neural Information Processing Systems*.

Jordan Guerguiev, Timothy Lillicrap, and Blake Richards. 2017. Towards deep learning with segregated dendrites. *eLife*, 6:e22901.

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

Valentin Hofmann, Leonie Weissweiler, David R. Mortensen, Hinrich Schütze, and Janet B. Pierrehumbert. 2025. Derivational morphology reveals analogical generalization in large language models. *Proceedings of the National Academy of Sciences*, 122(19).

Michael Häusser. 2001. Synaptic function: Dendritic democracy. *Current Biology*, 11(1):R10–R12.

B Illing, J Ventura, G Bellec, and W Gerstner. 2022. Local plasticity rules can learn deep representations using self-supervised contrastive predictions. *Advances in Neural Information Processing Systems*.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, and 1 others. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, pages 4651–4664. PMLR.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

George Kastellakis, Alcino J Silva, and Panayiota Poirazi. 2016. Linking memories across time via neuronal and dendritic overlaps in model neurons with active dendrites. *Cell reports*, 17(6):1491–1504.

Jim W Kay and William A Phillips. 2020. Contextual modulation in mammalian neocortex is asymmetric. *Symmetry*, 12(5):815.

Jim W Kay, Jan M Schulz, and William A Phillips. 2022. A comparison of partial information decompositions using data from real and simulated layer 5b pyramidal cells. *Entropy*, 24(8):1021.

Daniel Khashabi, Snigdha Chaturvedi, Michael Roth, Shyam Upadhyay, and Dan Roth. 2018. Looking beyond the surface: A challenge set for reading comprehension over multiple sentences. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 252–262.

Najoung Kim and Sebastian Schuster. 2023. Entity tracking in language models. *arXiv preprint arXiv:2305.02363*.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, and 1 others. 2017. Overcoming catastrophic forgetting in neural networks. *Proceedings of the national academy of sciences*, 114(13):3521–3526.

Konrad P Körding and Peter König. 2000. Learning with two sites of synaptic integration. *Network: Computation in neural systems*, 11(1):25–39.

Matthew Larkum. 2013. A cellular mechanism for cortical associations: an organizing principle for the cerebral cortex. *Trends in neurosciences*, 36(3):141–151.

Matthew E Larkum. 2022. Are dendrites conceptually useful? *Neuroscience*, 489:4–14.

Matthew E Larkum, Lucy S Petro, Robert NS Sachdev, and Lars Muckli. 2018. A perspective on cortical layering and layer-spanning neuronal elements. *Frontiers in neuroanatomy*, 12:56.

Matthew E Larkum, J Julius Zhu, and Bert Sakmann. 1999. A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature*, 398(6725):338–341.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *nature*, 521(7553):436–444.

Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. 1989. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551.

Hector J Levesque, Ernest Davis, and Leora Morgenstern. 2012. The winograd schema challenge. *KR*, 2012:13th.

Thomas Limbacher and Robert Legenstein. 2020. Emergence of stable synaptic clusters on dendrites through synaptic rewiring. *Frontiers in computational neuroscience*, 14:57.

Guy Major, Matthew E Larkum, and Jackie Schiller. 2013. Active properties of neocortical pyramidal neuron dendrites. *Annual review of neuroscience*, 36:1–24.

Tomáš Marvan and William A Phillips. 2024. Cellular mechanisms of cooperative context-sensitive predictive inference. *Current Research in Neurobiology*, page 100129.

Tomaš Marvan, Michal Polák, Talis Bachmann, and William A Phillips. 2021. Apical amplification—a cellular mechanism of conscious perception? *Neuroscience of consciousness*, 2021(2):niab036.

Andrew D Nelson and Kevin J Bender. 2021. Dendritic integration dysfunction in neurodevelopmental disorders. *Developmental Neuroscience*, 43(3-4):201–221.

Elena Pastorelli, Alper Yegenoglu, Nicole Kolodziej, Willem Wybo, Francesco Simula, Sandra Diaz, Johan Frederik Storm, and Pier Stanislao Paolucci. 2023. Two-compartment neuronal spiking model expressing brain-state specific apical-amplification,-isolation and-drive regimes. *arXiv preprint arXiv:2311.06074*.

Alexandre Payeur, Jordan Guerguiev, Friedemann Zenke, Blake A Richards, and Richard Naud. 2021. Burst-dependent synaptic plasticity can coordinate learning in hierarchical circuits. *Nature neuroscience*, 24(7):1010–1019.

W. A. Phillips, T. Bachmann, W. Spratling, L. Muckli, L Petro, and T. Zolnik. 2024. Cellular psychology: relating cognition to context-sensitive pyramidal cells. *Trends in Cognitive Sciences*.

William A Phillips. 2017. Cognitive functions of intracellular mechanisms for contextual amplification. *Brain and Cognition*, 112:39–53.

William A Phillips. 2023. *The Cooperative Neuron: Cellular Foundations of Mental Life*. Oxford University Press.

Panayiota Poirazi and Athanasia Papoutsi. 2020. Illuminating dendritic function with computational models. *Nature Reviews Neuroscience*, 21:1–19.

Srikanth Ramaswamy and Henry Markram. 2015. Anatomy and physiology of the thick-tufted layer 5 pyramidal neuron. *Frontiers in cellular neuroscience*, 9:233.

Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1):79–87.

Mohsin Raza and Ahsan Adeel. 2024. An overlooked role of context-sensitive dendrites. *arXiv preprint arXiv:2408.11019*.

João Sacramento, Rui Ponte Costa, Yoshua Bengio, and Walter Senn. 2018. Dendritic cortical microcircuits approximate the backpropagation algorithm. *Advances in neural information processing systems*, 31.

Jan M Schulz, Jim W Kay, Josef Bischofberger, and Matthew E Larkum. 2021. Gaba b receptor-mediated regulation of dendro-somatic synergy in layer 5 pyramidal neurons. *Frontiers in cellular neuroscience*, 15:718413.

Benjamin Schuman, Shlomo Dellal, Alvar Prönneke, Robert Machold, and Bernardo Rudy. 2021. Neocortical layer 1: An elegant solution to top-down and bottom-up integration. *Annual Review of Neuroscience*, 44(1):221–252. PMID: 33730511.

James M Shine. 2019. Neuromodulatory influences on integration and segregation in the brain. *Trends in cognitive sciences*, 23(7):572–583.

James M Shine, Patrick G Bissett, Peter T Bell, Oluwasanmi Koyejo, Joshua H Balsters, Krzysztof J Gorgolewski, Craig A Moodie, and Russell A Poldrack. 2016. The dynamics of functional brain networks: integrated network states during cognitive task performance. *Neuron*, 92(2):544–554.

James M Shine, Michael Breakspear, Peter T Bell, Kaylena A Ehgoetz Martens, Richard Shine, Oluwasanmi Koyejo, Olaf Sporns, and Russell A Poldrack. 2019. Human cognition involves the dynamic integration of neural activity and neuromodulatory systems. *Nature neuroscience*, 22(2):289–296.

James M Shine, Eli J Müller, Brandon Munn, Joana Cabral, Rosalyn J Moran, and Michael Breakspear. 2021. Computational models link cellular mechanisms of neuromodulation to large-scale neural dynamics. *Nature neuroscience*, 24(6):765–776.

Johan F Storm, P Christiaan Klink, Jaan Aru, Walter Senn, Rainer Goebel, Andrea Pigorini, Pietro Avanzini, Wim Vanduffel, Pieter R Roelfsema, Marcello Massimini, and 1 others. 2024. An integrative, multiscale view on neural theories of consciousness. *Neuron*, 112(10):1531–1552.

Mototaka Suzuki, Cyriel MA Pennartz, and Jaan Aru. 2023. How deep is the brain? the shallow brain hypothesis. *Nature Reviews Neuroscience*, 24(12):778–791.

Neil C Thompson, Kristjan Greenewald, Keeheon Lee, and Gabriel F Manso. 2020. The computational limits of deep learning. *arXiv preprint arXiv:2007.05558*, abs/2007.05558.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. *SuperGLUE: a stickier benchmark for general-purpose language understanding systems*. Curran Associates Inc., Red Hook, NY, USA.

Guan Wang, Jin Li, Yuhao Sun, Xing Chen, Changling Liu, Yue Wu, Meng Lu, Sen Song, and Yasin Abbasi Yadkori. 2025. Hierarchical reasoning model. *arXiv preprint arXiv:2506.21734*.

Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjape, Adina Williams, Tal Linzen, and 1 others. 2025. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. *arXiv preprint arXiv:2504.08165*.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R Bowman. 2020. Blimp: The benchmark of linguistic minimal pairs for english. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Adina Williams, Nikita Nangia, and Samuel R Bowman. 2017. A broad-coverage challenge corpus for sentence understanding through inference. *arXiv preprint arXiv:1704.05426*.

Friedemann Zenke, Ben Poole, and Surya Ganguli. 2017. Continual learning through synaptic intelligence. In *International conference on machine learning*, pages 3987–3995. PMLR.

A Pre-Training Details

Hyperparameter	Co^4-α	Co^4-β	Co^4-γ
Number of parameters	8M	8M	8M
Number of layers†	1	1	1
Embedding size	256	256	256
Vocabulary size	16384	16384	16384
Attention heads	2	2	2
Hidden dropout	0.1	0.1	0.1
Batch size	32	64	32
Sequence length	512	512	512
Warmup ratio	1.3%	1.4%	1%
Learning rate	0.0002	0.00001	0.0002
Learning rate scheduler	constant	constant	cosine
Optimizer	ADAMW	ADAMW	ADAMW
ADAMW ϵ	1e-8	1e-8	1e-8
ADAMW β_1	0.9	0.9	0.9
ADAMW β_2	0.999	0.999	0.999

Table 4: Pre-training hyperparameters for the STRICT-SMALL track across three configurations. †One layer refers to a module composed of our custom Co4 layer.

The training procedure, which has been briefly highlighted before, is as follows. We use the same tokenizer as the baselines, with a vocab size of 16384 and a small 1-layer model with the hyperparameters mentioned above. The Co^4 language model with a single decoder layer and just two attention heads is trained on the 10M corpus. It is powered via the aforementioned triadic modulation loops among Q-, K-, and V-TPNs, operating through P, D, and U contextual fields. After token embedding and positional projection, each token's Q, K, and V vectors co-evolve through a series of rapid and modulated updates.

The main goal was to keep the model as minimal as possible, to see the true power of the biologically-inspired triadic modulation loops within the layer. It is observed that the model performance converges over just a few epochs, i.e., 2 in this case.

B Finetuning Details

We perform a grid search for the following hyperparameters:

- **Number of epochs:** {3, 5, 10}
- **Learning rate:** $\{3\times10^{-5}, 5\times10^{-5}, 1\times10^{-4}, 2\times10^{-4}, 3\times10^{-4}, 5\times10^{-5}, 5\times10^{-5}\}$
- **Batch size:** {16, 32, 64}

For WSC (low training data), we expand the search to:

- **Number of epochs:** {3, 5, 10, 15, 20, 25, 30, 100}
- **Learning rate:** $\{3\times10^{-5}, 5\times10^{-5}, 7\times10^{-5}, 1\times10^{-4}, 2\times10^{-4}, 3\times10^{-4}, 5\times10^{-4}\}$
- **Batch size:** {16, 32, 64} }