Associative memory inspires improvements for in-context learning using a novel attention residual stream architecture

Anonymous authors
Paper under double-blind review

Abstract

Large language models (LLMs) demonstrate an impressive ability to utilise information within the context of their input sequences to appropriately respond to data unseen by the LLM during its training procedure. This ability is known as in-context learning (ICL). Humans and non-human animals demonstrate similar abilities, however their neural architectures differ substantially from LLMs. Despite this, a critical component within LLMs, the attention mechanism, resembles modern associative memory models, widely used in and influenced by the computational neuroscience community to model biological memory systems. Using this connection, we introduce an associative memory model capable of performing ICL. We use this as inspiration for a novel residual stream architecture which allows information to directly flow between attention heads. We test this architecture during training within a two-layer Transformer and show its ICL abilities manifest more quickly than without this modification. We then apply our architecture in small language models with 8 million and 1 billion parameters, focusing on attention head values, with results also indicating improved performance at these larger and more naturalistic scales.

1 Introduction

Transformers (Vaswani et al., 2017) are a popular and performant class of artificial neural networks. Large Language Models (LLMs), decorated exemplars of Transformers, have demonstrated impressive capabilities in a wide array of natural language tasks (Brown et al., 2020). One particularly notable capability, known as in-context learning (ICL), has gained significant attention. ICL, as witnessed in LLMs, occurs when a model appropriately adapts to tasks or patterns in the inference-time input data which was not provided during the model's training procedure (Lynch & Sermanet, 2021; Mirchandani et al., 2023; Duan et al., 2023). This ability to immediately learn and generalise from new information, especially with only a single or few exposures to new information, is a hallmark of sophisticated cognitive abilities seen in biological systems; human and non-human animals are capable of rapidly adapting their behaviour in changing contexts to achieve novel goals (Miller & Cohen, 2001; Ranganath & Knight, 2002; Boorman et al., 2021; Rosenberg et al., 2021), and can infer and apply previously-unseen, and even arbitrary rules without significant learning periods (Goel & Dolan, 2000; Rougier et al., 2005; Mansouri et al., 2020; Levi et al., 2024). Developing models rooted in computational neuroscience, such as associative memory, to connect our understanding of related phenomena in both artificial and natural intelligence may provide valuable insights for developing more adaptive and versatile language models, in addition to helping us more deeply understand more general neural systems (see Appendix A for discussion on the connection between associative memory models and modern neuroscience).

While various explanations have been proposed for how LLMs learn to perform ICL (Olsson et al., 2022; Von Oswald et al., 2023; Li et al., 2023; Reddy, 2024), there has been little work to develop explanations which also offer neurobiologically-plausible models of similar abilities seen in humans and non-human animals. One exception to this is the recent work of Ji-An et al. (2024), which illustrates a connection between ICL in LLMs and the contextual maintenance and retrieval model of human episodic memory from psychology. This model proposes that memory items are stored as a contextual composition of stimulus- and source-related content available during, and near-in-time, to a memory item's presentation. The model comports with several

reported psychological phenomena in humans (Lohnas et al., 2015; Cohen & Kahana, 2022; Zhou et al., 2023), and can be constructed using a Hebbian learning rule (Howard et al., 2005). Hebbian learning (Hebb, 1949), that neurons which 'fire together, wire together', is foundational to the theoretical underpinnings of associative memory models (Nakano, 1972; Amari, 1972; Little, 1974; Hopfield, 1982), which propose that memory items are stored by strengthening the connections between neurons which become activated during and near-in-time to the stimulus corresponding to memory items' presentations. How associative memory is neurophysiologically implemented is well-studied (Amit, 1990; Buzsáki, 2010; Khona & Fiete, 2022; Burns et al., 2022), and this is complemented by a well-developed theoretical literature, including work noting the close resemblance to a core ingredient of LLMs, the attention mechanisms of Transformers (Ramsauer et al., 2021; Bricken & Pehlevan, 2021; Kozachkov et al., 2022; Burns & Fukai, 2023; Burns, 2024).

1.1 Contributions

Given these existing links, further developing connections between the framework of associative memory and ICL may offer deeper insights or improvements. In the following sections, we:

- introduce a one-layer associative memory model which can perform ICL on a classification task, and which analogously allows attention values to directly represent input data;
- using the same task, and inspired by our explicit associative memory model, show how creating a residual stream of attention values between attention heads in a two-layer Transformer speeds-up ICL during training (compared to the vanilla Transformer and applying the same technique to queries or keys); and
- demonstrate that naïvely applying the same idea in small language models (LMs) indicates performance improvements scale to larger models and more naturalistic data.

A central theme in our innovation is a focus on the role of values in the attention mechanism, and architecting a simple 'look-back' method in the form a residual connection of values. Residual connections, also known as 'skip' or 'shortcut' connections, can be described as those which connect neurons which are otherwise indirectly connected through a more prominent pathway. First identified in experimental neuroscience (Lorente de Nó, 1938) and considered since the dawn of theoretical neuroscience and artificial neural networks (McCulloch & Pitts, 1943; Rosenblatt, 1961), researchers continue to find residual connections useful in modern applications (Dalmaz et al., 2022; Huang et al., 2023; Zhang et al., 2024b). A noticeable feature of Transformers is its use of the so-called *residual stream*, wherein data, once processed by the attention and feedforward layers, is added back to itself. What can therefore be considered a 'cognitive workspace' (Juliani et al., 2022) has been shown to contain rich structure, amendable to popular (Elhage et al., 2021) and emerging (Shai et al., 2024) interpretability methods. Our work illustrates that specific additional residual connections can lead to enhanced performance in ICL tasks, and we speculate it may also aid in future interpretability efforts.

2 ICL classification with a one-layer associative memory network

2.1 ICL classification task and mathematical set-up

Let $X \in \mathbb{R}^{e \times \mathscr{s}}$ be the input sequence data, where e is the dimension of each token embedded within a suitable latent space, and $\mathscr{s}$ is the number of tokens in the sequence[1]. Each token is considered either an *object*, o, or *label*, l. Input X consists of a sequence of multiple pairs of objects and labels. We say a *pair* of tokens is a contiguous sub-sequence of two column vectors from within X, consisting of one object token followed by one label token. For example, the j-th pair $X' \in \mathbb{R}^{e \times 2}$ consists of one object token $o^j \in \mathbb{R}^e$ followed by one label token $l^j \in \mathbb{R}^e$. Our input sequence X will therefore consist of multiple pairs, and each pair may appear more than once. The final token of a sequence, denoted as $x_{\mathscr{s}}$, corresponds to the label token of the last pair, which itself has appeared at least once prior to this final instance in X. However, the true label token data

[1]We provide notation tables in Appendix D.

of this final label token is replaced with the zero vector, and the network's task is to correctly predict this true label token given the previous in-context instance of the tokens' data (as illustrated in Figure 1a).

The token embeddings, representing objects and labels, follows Reddy (2024), where all token embeddings are drawn from similar statistical distributions. For every instance of a pair, label token embeddings come from fixed vectors, whereas object token embeddings are constructed from combinations of fixed and random vectors. Each label token l^i is an e–dimensional vector μ_{l^i}, whose components are i.i.d. sampled from a normal distribution having mean zero and variance $1/e$. Each object token embedding, o^i, is given by

$$o^i := \frac{\mu_{o^i} + \varepsilon\eta}{\sqrt{1+\varepsilon^2}},$$

where μ_{o^i}, which is fixed across all instances in X of the pair, and η, which is drawn randomly for each instance in X of the pair, are e–dimensional vectors whose components, like μ_{l^i}, are i.i.d. sampled from a normal distribution having mean zero and variance $1/e$. The variable ε controls the inter-instance variability of objects and, in the following, is set to 0.1 unless otherwise stated. This means that the final, tested instance of a pair has, as its object token, a slightly different appearance than the previous instance(s) seen in X and is not a perfect match, where μ_{o^i} provides the commonality between these variations. Adding these variations makes the task less trivial and slightly more naturalistic.

In §2.2, we show it is possible to perform ICL with such pairs in a single forward step of a one-layer associative memory network, written in the language of a single Transformer attention head (Vaswani et al., 2017). To show this, and for the benefit of subsequent sections, we briefly summarize the classical Transformer set-up. In Transformers, each attention head consists of learnt parameters – weight matrices $W^q, W^k \in \mathbb{R}^{\mathscr{k}\times e}$ and $W^v \in \mathbb{R}^{\mathscr{v}\times e}$ – with which, when taken together with the input data sequence X, we calculate the queries Q, keys K, and values V matrices using

$$Q = W^q X, \quad K = W^k X, \quad \text{and} \quad V = W^v X.$$

The values $\mathscr{k}$ and $\mathscr{v}$ are the reduced embedding dimensions for the attention operation (*i.e.*, to facilitate multi-headed attention, *etc.*). Here we use $\mathscr{k} = \mathscr{v}$. As in the input data, e is the dimension of the unreduced embedding space of the tokens.

It is then useful to define the SOFTMAX function for matrix arguments. For a matrix $M \in \mathbb{R}^{\mathscr{c}\times\mathscr{r}}$, we write $t_i := M[i,:] \in \mathbb{R}^{\mathscr{r}}$ for the i-th row and $t_j := M[:,j] \in \mathbb{R}^{\mathscr{c}}$ for the j-th column. Then, we define the SOFTMAX function for a matrix M as $\text{SOFTMAX}(M)[t_i,t_j] := \frac{\exp(M[t_i,t_j])}{\sum_t \exp(M[t,t_j])}$, where i and j are the vector component indices. We use this to compute attention-based embeddings of the data X, denoted by $\tilde{X}$, as

$$\tilde{X} = \text{SOFTMAX}\left(\frac{1}{\sqrt{\mathscr{k}}} K^T Q\right) V, \tag{1}$$

in which we refer to the term $K^T Q$ as the *scores* $S \in \mathbb{R}^{\mathscr{s}\times\mathscr{s}}$. In Transformers and Transformer-based models such as LLMs, this new data $\tilde{X}$ is then recombined with data from other attention heads, before passing through a multi-layer perceptron. Layers of attention heads and multi-layer perceptrons are stacked atop one-another to perform increasingly sophisticated computations.

2.2 Associative Memory for ICL (AMICL) model

In our associative memory model, which we call *Associative Memory for ICL* (AMICL), we take inspiration from the transformations applied to the input data X to create the basis of what can be considered (Ramsauer et al., 2021) as an associative memory update step in Equation 1. Instead of using parameterised values, we use a simple set of assignments for each token's key, query, and value vectors. For all embedded tokens $i < \mathscr{s}$, we set $k_i = q_i = \frac{a x_{i-1} + x_i}{a+1}$ as the keys and queries, where lowercase Latin letters denote indexed column vectors, taken from the matrices that are denoted with uppercase Latin letters. For simplicity, we wrap the indices along the token sequence such that the first token, x_1, and the last token, $x_{\mathscr{s}}$, are used to generate the sub-sequence $(x_{\mathscr{s}}, x_1)$, *i.e.*, we assign token index 0 to $\mathscr{s}$. For token $\mathscr{s}$, with $x_{\mathscr{s}}$ as the token column vector,

we set $q_\mathfrak{s} = \frac{ax_{\mathfrak{s}-1}+x_\mathfrak{s}}{a+1}$ as the query and $k_\mathfrak{s}$ as the zero vector, which acts as its key. For all tokens, the values column vector is equal to the token column vector, *i.e.*, $v_j = x_j$. The value of a can be any arbitrary real value, but, after testing within the range $[0, 2]$, is set as $a = 2$. (The justifications for testing a within this range and setting $a = 2$ is given later in this subsection.)

We then perform next token prediction using the universal associative memory framework (Millidge et al., 2022), which can be interpreted as a generalisation of Equation 1. In particular, Equation 1 in the universal associative memory framework has the form $\textsc{projection}(\textsc{separation}(\textsc{similarity}(K, Q)), V)$, where the SIMILARITY function is chosen as a scaled dot product, the SEPARATION function is chosen as a SOFTMAX, and the PROJECTION function is chosen as a product.

By constructing the queries and keys in the AMICL model using what is essentially an 'average' between the current token and the previous token, with a scalar weighting of a on the previous token, we allow the similarity function to identify the relevant pair in the context (in a similar sense to that of a 1D convolution operation), which is then potentially amplified by the separation function (*e.g.*, in the SOFTMAX case), for projection to the appropriate matching final token. In this way, and as illustrated in Figure 1, the AMICL model essentially performs auto-association on the final pair of tokens, treating the penultimate token as a context clue for the final token (in the same way that auto-associative memory dynamics traditionally use partial memory identity information to complete the full, remaining pattern information (Nakano, 1972; Amari, 1972; Little, 1974; Hopfield, 1982)). The reason for choosing the values to have the identity of the original data, then, is because we cast this ICL task as an auto-association problem, and by the universal associative memory framework (Millidge et al., 2022), the projection function in the case of auto-association is the identity function (also see §3.1 of Millidge et al. (2022)). This then justifies the choice of exploring the range of a in the positive reals only, since although negative values are mathematically possible, they would create an anti-Hebbian or repelling force, counter-acting the designed (Hebbian) auto-association task created by the query and key constructions.

As an intuitive sketch, AMICL can be thought of as implementing the algorithm illustrated in Figure 1: first, we identify the final pair, where the final token, $x_\mathfrak{s}$, which should be a label token, has been set to the zero vector (Figure 1a, where '?' represents the unknown token data and other tokens' data are shown); second, we consider all possible contiguous pairs in the context, *e.g.*, $(x_3, x_4), (x_4, x_5), \ldots$, without knowledge of their data, which we as designers know will correspond to pairs like, *e.g.*, $(o^2, l^2), (l^2, o^3), \ldots$ (Figure 1b, where each contiguous pair is enclosed by a yellow rectangle); third, we compare all of the contiguous pairs in the second step with the final pair identified in the first step (which we are attempting to 'pattern complete'), and, upon seeing which context pair is most similar in final pair (in this example, (o^1, l^1)), complete the pattern appropriately by setting $x_\mathfrak{s} = l^1$ (Figure 1c).

Varying the parameter a between 0 and 2 for combinations of similarity and seperation functions showed that using the DOT PRODUCT similarity with ARGMAX separation function resulted in practically perfect ICL ability at $a \geq 1.5$ whereas using PEARSON'S CORRELATION similarity with ARGMAX separation function saturated at a performance level of $\sim 85\%$ accuracy for the ICL pairs task (see Figure 8 in Appendix B). Due to the practically perfect ICL performance in the DOT PRODUCT case and the saturation of ICL performance PEARSON'S CORRELATION – both saturating at values at $a \geq 1.5$ – we set $a = 2$.

While setting $a = 2$ and PROJECTION = IDENTITY, we tested all combinations of:

- SIMILARITY $\in \{$DOT PRODUCT, PEARSON'S CORRELATION, MANHATTAN DISTANCE, EUCLIDEAN DISTANCE$\}$; with
- SEPARATION $\in \{$IDENTITY, SOFTMAX, ARGMAX$\}$.

Manual inspection of the resulting attention matrices indicated that the DOT PRODUCT and PEARSON'S CORRELATION similarity functions combined with the ARGMAX separation function provided the cleanest ICL (see Figure 7 in Appendix B for an example). Varying the token sequence length $\mathfrak{s}$ between 10 and $1,000$ shows no variation in performance; varying the embedding dimension of the tokens e between 10 and $1,000$ showed that for $e \geq 50$, the ICL task performance was practically perfect (see Figure 9 in Appendix B).

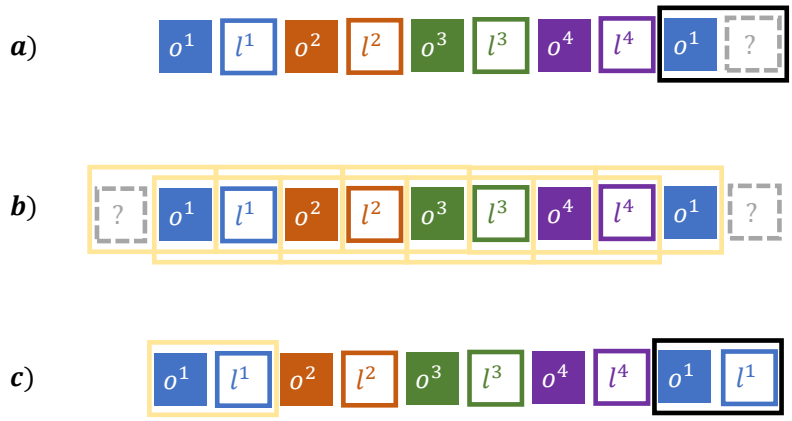

Figure 1: Depiction of the AMICL algorithm on label–object pairs, consisting of three steps: (a) consider the penultimate and final tokens as a 'local pattern' which has been 'partially corrupted' by the missing data in the final token, illustrated by the black rectangle enclosing this final pair; (b) search for matching, 'complete local patterns' by grouping all previous contiguous token pairs in the context, with each pair illustrated by a yellow rectangle enclosing the tokens; and (c) complete the final, corrupted local pattern based on matching to the nearest complete in-context pair, which in this case is (o^1, l^1), so the final token data is assigned l^1.

3 Residual attention streams in a two-layer Transformer

3.1 Residual attention streams

In the AMICL model, following the path of information from the input to the queries, keys, and values (as illustrated in Figure 2a), we can see that the queries Q and keys K flow from a shared function f of the input X. Whereas, the values V are given directly by the input X. In the language of the universal associative memory framework (Millidge et al., 2022), we can say that the *similarity* and *separation* factors (determined by the queries Q and keys K) are coupled by the shared function f whereas the *projection* factor (given by the values V) is simply the input itself, *i.e.*, auto-association.

Seen through the lens of the traditional self-attention mechanism, this implies that a similar construction of the queries and keys is possible, such that using a simple identity function from the input for the values could replicate this behaviour in more sophisticated data, task, and model scenarios. However, in our initial experimentation with this idea, whereby we eliminated the W^v matrices altogether, we witnessed significant model performance degradation and training instability (it also limited the model's expressiveness in the self-attention mechanism to auto-association, whereas many interesting tasks make use of some amount of *hetero*-association – see Burns (2024)). This led us to consider an alternative form of projection, one which could in principle mix auto- and hetero-association, via the creation of values V which more explicitly retain the prior input X. In particular, we introduce a residual connection between the values data of successive layers in the Transformer, which we call a *residual values stream* and, more generally, a *residual attention stream* applied to values (Figure 2b).

As an alternative source of intuition, one can consider the informal interpretation of self-attention as consisting of: queries 'asking each token a question about other tokens'; keys 'responding to each token with the answer to the queries' questions', *i.e.*, they align with or 'attend to' the queries; and values giving the (weighted) answers to those questions, which are operationalised as small additions to the current token vectors. Within this informal conceptual model, we can consider our residual attention streams as retaining additional 'look-back' information in the answers.

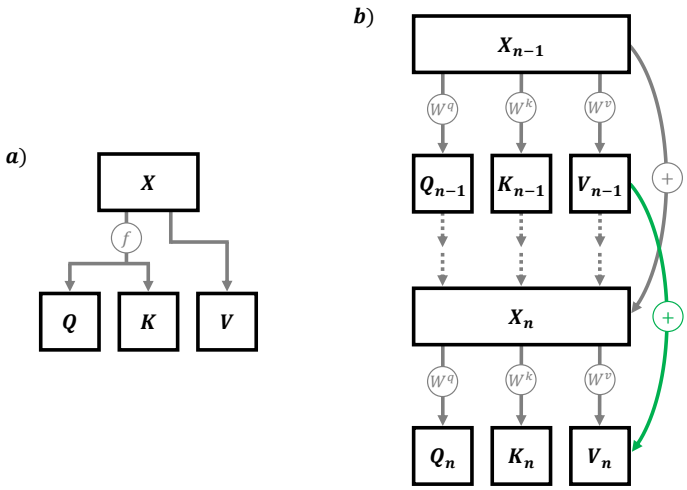

Figure 2: Partial diagrams of (a) the AMICL model and (b) our residual attention stream architecture with two Transformer layers, n and $n-1$, shown here implemented for values (the added residual is shown in green). Boxes represent variables. Full arrows represent functions, with overlaid circles indicating relevant variables or functions (no circle is used for the identity function). Dotted arrows represent functions and variables omitted in the diagram for space.

Our residual attention stream also generates additional gradient signals during training, which could themselves be beneficial for the network to develop ICL capabilities. We therefore also test the same residual stream architecture, separately, for queries and keys. In principle, it is also possible to apply the residual stream architecture to combinations of queries, keys, and values. But, to maintain a stronger connection to the AMICL model, we focus on testing the residual value stream separately, as well as the keys and queries, again separately, for comparison.

3.2 Two-layer Transformer architecture

We train classic and modified versions of a two-layer Transformer on the same task described in Section 2. Following Reddy (2024), the common architectural features between the two versions consist of two single-head attention layers followed by a three-layer multi-layer perceptron with 128 ReLU neurons followed by a softmax layer to give probabilities over the ℓ labels. The network is trained with the same task as the AMICL model, using the cross-entropy loss between the predicted and actual final label in token x_3. Within the modified architecture, we add the first attention head's queries, keys, or values to the second attention head's queries, keys, or values, respectively. More formally, in our modified version of the Transformer architecture, we calculate the first attention layer as

$$Q_1 = W_1^q X, \quad K_1 = W_1^k X, \quad \text{and} \quad V_1 = W_1^v X,$$

and then, in the second attention layer, for a residual queries stream we calculate

$$Q_2 = W_2^q X + Q_1, \quad K_2 = W_2^k X, \quad \text{and} \quad V_2 = W_2^v X,$$

or for a residual keys stream we calculate

$$Q_2 = W_2^q X, \quad K_2 = W_2^k X + K_1, \quad \text{and} \quad V_2 = W_2^v X,$$

or for a residual values stream (shown in Figure 2b) we calculate

$$Q_2 = W_2^q X, \quad K_2 = W_2^k X, \quad \text{and} \quad V_2 = W_2^v X + V_1.$$

3.3 Supplemental ICL tasks

As in Reddy (2024), while we train on the originally-described ICL task, we create supplemental tasks which are not used for training but rather act as proxy measurements of progress for different computational strategies for completing the task: training data memorisation and ICL capability generalisation. Namely, these supplemental tasks are:

- In-weights (IW): a series of object–label pairs is presented where the final pair is not found within the prior context but is present in the training data;
- In-context (IC): a series of novel object–label pairs, not seen in the training data (*i.e.*, an entirely re-drawn set of μ_{o^i} and μ_{l^i} values) but following exactly the same statistical structure, is presented; and
- In-context 2 (IC2): a series of object–label pairs are presented, where the objects are found in the training data but have been assigned new labels (*i.e.*, the objects retain their μ_{o^i} values but the labels have their μ_{l^i} values re-drawn).

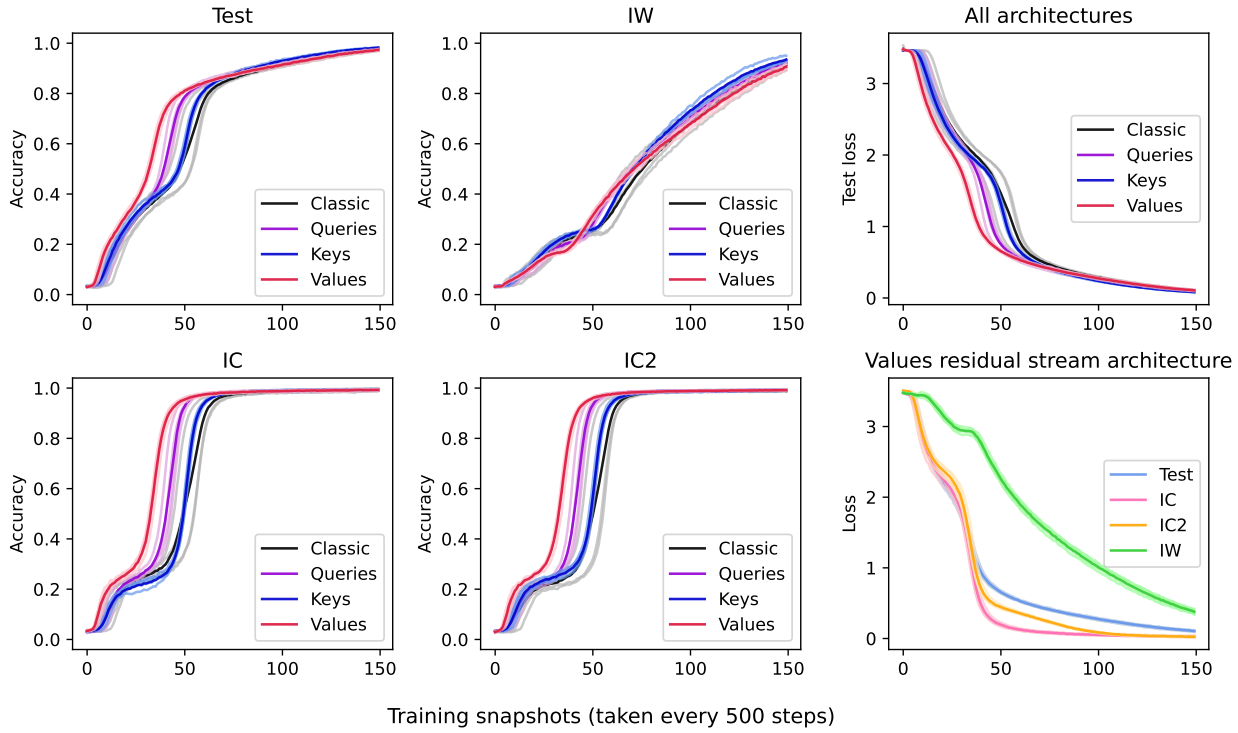

Figure 3: Accuracies and losses for the Test, IW, IC, and IC2 tasks over training time for the classic (unmodified) network and residual queries, keys, and values stream networks. Dark lines represent means, light lines represent individual trials.

The reduced embedding dimensions of the attention operation are both set to 128, *i.e.*, $\hbar, \mathfrak{v} = 128$. Each network architecture was trained on four random seeds, with a batch size of 128, vanilla stochastic gradient descent, and a learning rate of 0.01.

3.4 Performance of the attention residual streams on the ICL classification task

Figure 3 shows the accuracies for the task being trained (Test) and the three supplemental tasks (IW, IC, and IC2) for each architecture. It also shows the Test loss for all architectures and all loses for the values

residual stream architecture. For our residual attention stream modifications, we observe a general leftward shift in all but the IW task, with the value residual stream architecture showing the largest shift. The IW task also shows a slower learning rate, which can be attributed to the relative difficulty of the network memorising the training data.

To quantify these shifts, we report statistics in Table 1 of when the first training snapshots we recorded reached an accuracy threshold of > 0.95. We find the values stream networks perform best, reaching the same level of accuracy as the classic (unmodified) networks with $\sim 24\%$ fewer training steps. The same result is also seen at lower thresholds (see Tables 5 and 6 in Appendix C for thresholds of 0.5 and 0.9, respectively). We also performed t-tests at the 0.95 accuracy threshold, which showed significant differences between the performance of the classic (unmodified) network and the residual queries and values stream networks on the IC and IC2 tasks ($p < 0.01$, $t < -3.9$), but not between the classic and residual keys stream networks ($p > 0.39$, $t > -1.0$). The residual values stream networks also perform significantly better ($p < 0.03$, $t < -3.0$) than all other networks on both the IC and IC2 task, except the residual queries stream network for the IC task ($p = 0.06$, $t = -2.33$).

Table 1: Mean $\pm$ standard deviation of training snapshot number where accuracy first exceeded 0.95 for the IC and IC2 tasks in the classic (unmodified), and residual queries, keys, and values stream networks. Fastest training times are bolded.

	Classic	Queries (ours)	Keys (ours)	**Values (ours)**
IC	64.5 ± 5.12	52.5 ± 1.5	61.75 ± 0.83	**49.0** ± 2.12
IC2	63.25 ± 4.15	51.75 ± 1.92	61.25 ± 0.83	**47.75** ± 1.3

4 Residual value streams in a toy language model

4.1 Toy language model architecture and training details

To provide an initial indication as to whether the benefit for ICL performance seen in Section 3 scales to larger models and naturalistic data, we also test the residual value stream architecture by training toy LMs. The Transformer-based model contains approximately 8 million parameters spread over eight layers, each with 16 attention heads. The context window is 256 tokens long and dimension of the model is 256. Next-token prediction is trained for three epochs on the Tiny Stories dataset, such that they can generate simple and short children's stories (Eldan & Li, 2023). Each network architecture was trained in three separate instances, using different random seeds for each instance but controlling for randomness between architectures by re-using the same set of seeds for the two architectures.

We implement the residual value stream in a naïve way: in each but the first layer, the values of each attention head receives an additional input of the values from the attention head with the same index in the previous layer[2]. Although there are alternative implementations of such residual streams, we chose a naïve and straightforward approach to provide an initial indication of the potential scalability of this architectural change and to not alter the number of learnt parameters between the networks.

The residual values stream networks achieved 1.65 ± 0.05 training loss and 1.55 ± 0.01 validation loss, slightly lower than the training and validation losses for the classic networks, which were 1.67 ± 0.05 and 1.56 ± 0.01, respectively. Measured by wall clock computation time[3], the residual values stream networks also took slightly longer to train, 30.18 ± 0.49 hours compared to 29.12 ± 0.40 hours for the classic networks. This additional computation time is attributable to the additional gradient computations required by the 16 additional residual streams connecting the attention head values at each but the first layer.

[2]Principally, the choice of which attention heads form the residual values stream is arbitrary given the independent nature of each head. However, when implemented with multi-head attention, it is preferable to use the same head index for computational convenience.

[3]Using a PC equipped with an AMD Ryzen Threadripper PRO 3975WX 32-Cores CPU, 130GB of RAM, and 4× NVIDIA RTX A4000 GPUs.

4.2 Evaluation of ICL ability using a simplistic natural language task

As a proxy evaluation of the ICL ability of each model in natural language, we utilise a simple indirect object identification (IOI) task. In natural language, a direct object is the noun that receives the action of the verb and an indirect object is a noun which receives the direct object, *e.g.*, in the sentence "A passed B the ball": "passed" is the verb; "A" is the subject; "the ball" is the direct object; and "B" is the indirect object. In the IOI task, we test for correct completion of sentences like "When A and B were playing with a ball, A passed the ball to", where " B" is the indirect object and is considered the correct completion.

GPT-2 Small can perform instances of the IOI task, and the circuit responsible for this ability has been identified (Wang et al., 2023). As shown in this section, our toy LMs have some measurable ability to complete instances of this task, and so we use this to compare the performance of the classic (unmodified) and residual value stream architectures.

For the IOI task, we tested the following sentences:

1. "When John and Mary went to the shops, $*$ gave the bag to", where $*$ was either "John" or "Mary", with correct completions " Mary" and " John", respectively.
2. "When Tom and James went to the park, $*$ gave the ball to", where $*$ was either "Tom" or "James", with correct completions " James" and " Tom", respectively.
3. "When Dan and Emily went to the shops, $*$ gave an apple to", where $*$ was either "Dan" or "Emily", with correct completions " Emily" and " Dan", respectively.
4. "After Sam and Amy went to the park, $*$ gave a drink to", where $*$ was either "Sam" or "Amy", with correct completions " Amy" and " Sam", respectively.

For each sentence and variation thereof (swapping the identities of the subject and indirect object, as indicated by the $*$ symbol above), we recorded the next token probabilities of the correct and incorrect names. As summarised in Table 2, we find the classic networks are much less capable than the residual values stream networks – across the four sentences, the classic networks correctly identify the indirect object with a probability of $\sim 7\%$ while the residual values stream networks do so with a probability of $\sim 41\%$, a $\sim 590\%$ improvement. Similarly, the classic networks more regularly mistake the subject for the indirect object with a probability of $\sim 5\%$ while the residual values stream networks do so with a probability of $\sim 3\%$, a $\sim 60\%$ reduction.

Table 2: Mean $\pm$ standard deviation probabilities (%) of correct and incorrect responses to each sentence for the IOI task for the classic (unmodified) and residual values stream networks. Best scores are bolded.

Classic	Sentence 1	Sentence 2	Sentence 3	Sentence 4
Correct (higher is better)	11.73 $\pm$ 15.62	11.88 $\pm$ 9.09	3.83 $\pm$ 6.09	1.54 $\pm$ 2.31
Incorrect (lower is better)	6.93 $\pm$ 5.70	7.99 $\pm$ 9.52	0.51 $\pm$ 0.66	4.86 $\pm$ 8.35
Residual values stream (ours)	Sentence 1	Sentence 2	Sentence 3	Sentence 4
Correct (higher is better)	**42.89** $\pm$ 10.58	**43.44** $\pm$ 14.45	**49.24** $\pm$ 9.89	**28.56** $\pm$ 5.92
Incorrect (lower is better)	**5.68** $\pm$ 3.76	**6.8** $\pm$ 10.40	**0.03** $\pm$ 0.03	**0.18** $\pm$ 0.15

We propose the following explanations as for why the residual values stream variant outperforms the classic model so handedly despite the two models having very similar final loss values and the former having only a slightly smaller loss:

- especially in small models, even small differences in the training loss can translate to large differences in the test accuracy on downstream tasks (Lau et al., 2023);
- algorithmically, models with only very small or even no loss differences can perform the task in principally different ways (Power et al., 2022; Bushnaq et al., 2024); and

- provision of the residual values stream improves sample complexity for the IOI task, which while implicitly present in the auto-regressive training set-up, is not what is primarily measured by the training loss (which is more generally measuring performance on next-token prediction, implicitly consisting of a much larger variety of natural language tasks).

5 Residual value streams in a small language model

5.1 Small language model architecture and training details

To test the performance of residual value streams on a larger scale, we trained two one-billion parameter (1B) language models in the style of a Llama 3 model (Grattafiori et al., 2024). These models were trained on English-language documents, amounting to approximately 84 billion tokens, from a random subset of the Nemotron-CC-HQ dataset (Su et al., 2024a). As in the toy language model described in §4, the only way in which these models differed was in the presence or absence of the residual values stream. As in the 8M-parameter models, the 1B models showed no significant difference in training loss (see Figure 4).

Each model comprised of 16 layers, 2,048 hidden dimensions, 32 attention heads, and eight key-value heads. These models support a context length of 4,096 tokens and incorporate several modern architectural features: rotary positional embeddings (RoPE) (Su et al., 2024b) with a base of 500,000, SwiGLU activation functions (Shazeer, 2020) with a multi-layer perceptron (MLP) expansion factor of four, and RMSNorm for layer normalisation (Zhang & Sennrich, 2019).

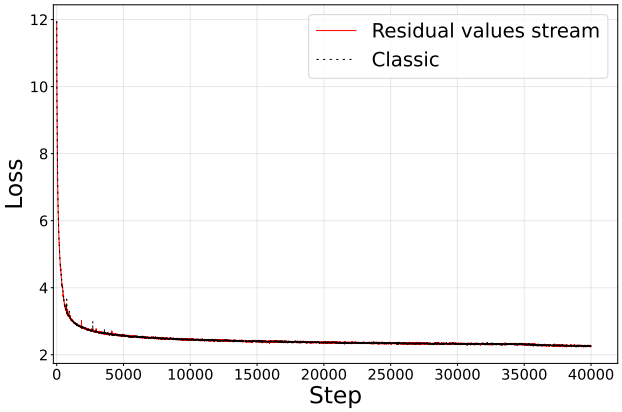

Figure 4: Training loss over steps for the 1B models.

The model was trained for 40,000 iterations with a global batch size of 512. We used the AdamW optimizer (Loshchilov & Hutter, 2019) with $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1e-8$, and gradient clipping at 1.0. The learning rate followed a cosine decay schedule with an initial rate of 3e-4, 500 warmup steps, 5,000 cooldown steps, and minimum learning rate of 0.0. The model was trained using the Llama-3.1-8B tokenizer (Grattafiori et al., 2024), which has a vocabulary size of 131,072 tokens. We used BFloat16 precision throughout.

Tables 3 and 4 summarise the training parameters.

N_{vocab}	n_{Layers}	n_{heads}	d_{model}	d_{MLP}	Batch Size	Sequence Length
131,072	16	32	2,048	8,192	512	4,096

Table 3: Hyper-parameters used in the 1B models.

Optim.	β_1	β_2	ϵ	N_{Warmup}	N_{Cooldown}	learning rate
AdamW	0.9	0.95	1e-8	1%	10%	3e-4

Table 4: Optimizer parameters for the 1B models.

5.2 Small language model evaluation and results

After training these models, we evaluated them on single- and five-shot language understanding tasks, namely: AI2 Reasoning Challenge (ARC) (Clark et al., 2018), Physical Interaction: Question Answering (PIQA) (Bisk et al., 2019), OpenBookQA (Mihaylov et al., 2018), and HellaSwag (Zellers et al., 2019).

Figure 5 shows the average accuracy of the residual value stream architecture compared to the classic baseline, showing modest but sustained improvements over our single- and five-shot language understanding tasks. We observed a very high correlation between the two models' performance (0.993), which is expected for paired samples evolving over training steps with the same training data. We conducted a paired two-sample for means t-test on the average accuracy and found the p-values (both one-tail and two-tail) are much smaller than $\alpha = 0.05$ (the t-critical one-tail value was 1.694 and the t-critical two-tail value was 2.037), which means we may reject the null hypothesis, *i.e.*, that we may say the mean difference between the paired samples is *not* zero. Given this, we confirm the improvement in performance at the 1B scale is statistically significant across these tasks.

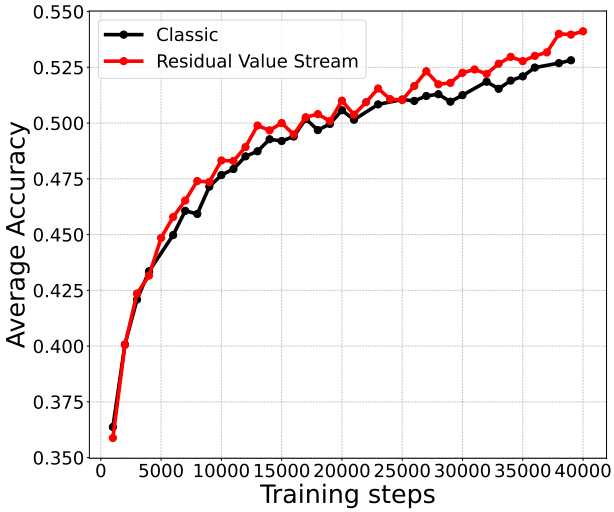

Figure 5: Average accuracy across training steps for our evaluations (single- and five-shot). Individual evaluation results can be found in Figure 10.

However, these gains appear localised to particular tasks. In Figure 10, we can see that the most noticeable gains are in the ARC and OpenBookQA tasks whereas there is almost no change in the PIQA and HellaSwag results. We therefore also evaluated our model on an extended version of the IOI task, which we dub *IOI-Hard*. In the IOI-Hard task, there are 25 sentences like:

- "{A} and {B} were commuting to the {PLACE}."
- "After lunch, {A} and {B} went to the {PLACE}."
- "The {PLACE} {A} and {B} went to had a {OBJECT}."
- "Friends {A} and {B} found a {OBJECT} at the {PLACE}."
- "{A} and {B} were planning to visit the {PLACE}."

where {A} and {B} are a unique pair of person names, {PLACE} is a unique physical place (*e.g.*, park, library, coffee shop, *etc.*), and {OBJECT} is a unique physical object (*e.g.*, book, pen, flower, key,

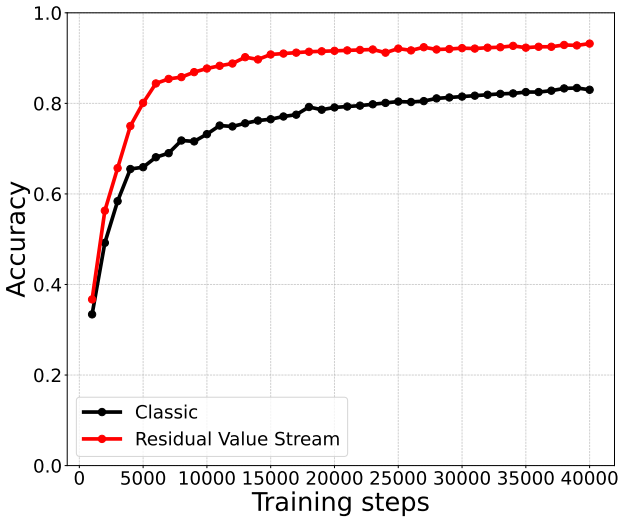

Figure 6: Accuracy across training steps for 1,000 completions of the IOI-Hard task evaluation.

etc.). Following 25 unique sentences of this type, there is a final, unfinished sentence designed to test if the model can correctly identify the indirect object. The final sentences randomly selected a pair of person names occurring in one of the previous 25 unique sentences, and matched the place and object (if mentioned) in one of the following final sentences:

- "At the {PLACE}, {GIVER} wanted to give a {OBJECT} to"
- "At the {PLACE}, {GIVER} decided to give the {OBJECT} to"
- "Then at the {PLACE}, {GIVER} gave a {OBJECT} to"
- "While at the {PLACE}, {GIVER} handed a {OBJECT} to"

where the {PLACE} and {OBJECT} was as before, the {GIVER} was randomly chosen as {A} or {B}, and the correct completion was {A} if {B} was chosen as the {GIVER} and vice-versa.

Figure 6 shows the accuracy of the IOI-Hard task over training steps for both models, where 1,000 randomly-generated prompts of the above style were created, using 32 unique names, 30 unique objects, and 30 unique places. In this task, which is much more directly testing ICL ability, we see a very clear performance gain in our model compared to the classic model.

6 Discussion

Our study builds upon the understanding of ICL in Transformer models by exploring connections to associative memory models from computational neuroscience. We introduced AMICL, an associative memory model that performs ICL using an approach akin to a single-layer Transformer attention head but with an implied residual stream from the inputs to the values. This itself is notable given ICL is not typically seen in single-layer Transformers (Olsson et al., 2022) outside of simple linear regression tasks (Zhang et al., 2024a; Lu et al., 2024). Inspired by this, we proposed a novel residual stream architecture in Transformers, where information flows directly between attention heads. We demonstrate this in a two-layer Transformer, showing increased efficiency in learning on an ICL classification task. Moreover, by extrapolating this architecture to small Transformer-based language models, we illustrated enhanced ICL capabilities are possible on a larger and more naturalistic scale.

6.1 Connections to computational neuroscience

Our results offer potential insights for neural network design and our understanding of biological cognition:

- **Neural network architecture**. The simplicity of our neuroscience-inspired architectural modification, attention residual streams, provides a promising extension for designing more adaptive Transformer models. Given one can interpret this change as providing additional shared workspaces across model layers (in the form of the additional residual streams), our results can also be viewed as demonstrating another example where functional modularisation can provide a performance benefit (Kaiser & Hilgetag, 2006; Chen et al., 2013), which is also seen in cases of highly distributed representations (Voitov & Mrsic-Flogel, 2022), of which natural language appears no exception (Mikolov et al., 2013; Hernandez et al., 2024).
- **Biological cognition**. Drawing parallels between artificial networks and cognitive neuroscience theories contributes to a deeper understanding of memory systems in biological entities. As we have shown by our AMICL model, the associative memory framework is capable of ICL in a single layer, with an implied 'skip connection' present between the input and the values. This suggests that similar types of connections may analogously exist in biological networks to adapt and generalize from limited exposure. For instance, associative memory is often related to the hippocampal formation (Amit, 1989), an area in the brain important for memory and learning. An open question in this area of neuroscience is what the computational role of observed skip connections are. In particular,

while some information in the hippocampus goes directly from area CA3 to CA1, another area – CA2 – is often skipped, but does receive input from CA3, which it then passes on to CA1 (Cui et al., 2013). This is interesting not just anatomically, but also functionally, since CA2 is considered responsible for specialised tasks and context-switching (Dudek et al., 2016; Robert et al., 2018). Whether our AMICL model or residual attention stream modification can be considered analogous to skip connections witnessed in the hippocampus remains to be studied. However, it offers a promising window, *e.g.*, one could now study the effects of changing the number and quality of residual attention streams, to test whether connection types or relative proportions similar to that seen in the hippocampus also show improved performance (for some types of data or tasks). We also suggest experimental neuroscientists perform a classical ablation study of CA3 to CA1 skip connections at different intensities and measure the performance of an object–pair recognition task with the same structure as described in §2.1 and §3.3. We hypothesise that the strength and number of remaining CA3 to CA1 skip connections will correlate with performance on the IC and IC2 tasks but not the IW task. Further, we also hypothesise the effect size of this performance difference will correlate with the effect size of the difference in performance caused by ablation of CA2 itself, since we theorise these CA3 to CA1 skip connections are mostly valuable only in the presence of CA2 to CA1 connections.

In establishing an additional potential mechanistic link between associative memory frameworks and ICL tasks, this work builds upon the broader neuroscience-inspired interpretation of Transformer attention mechanisms (Ramsauer et al., 2021). Indeed, Zhao (2023) conjectures that LLMs performing ICL do so using an associative memory model which performs a kind of pattern-completion using provided context tokens, conditioned by the learnt LLM parameters. Using this perspective, Zhao (2023) constructs different token sequences to be used as contexts to prepend onto the same final token sequence representing the tested task. By actively choosing context tokens which more 'closely' resemble the final tokens the LLM is being tested with, the LLM shows improved task performance, presumably by utilising the increased relevancy of the context tokens for ICL. This perspective is further developed in Jiang et al. (2024), who show how LLMs can be 'hijacked' by purposeful use of contexts with particular semantics. When LLMs such as Gemma-2B-IT and LLaMA-7B are given the context of "The Eiffel Tower is in the city of", they successfully predict the next token as " Paris". However, Jiang et al. (2024) demonstrate that prepending the context with the sentence "The Eiffel Tower is not in Chicago." a sufficient number of times, these LLMs incorrectly predict the next token as " Chicago". We may interpret these prepended data as acting (crudely) as 'distractors', and in the associative memory sense as causing an over-activation of competing memory items which interferes with accurate task performance.

Beyond associative memory, our architecture also shows structural similarities to other models from computational neuroscience. Notably, AMICL shares characteristics with successor representation (SR) models, particularly γ–models where a discount factor governs the contribution of future states (Janner et al., 2020). In AMICL, the parameter a plays a similar role, and in all of our implementations of the value residual stream is implicitly set to 1 (since we do not weight the added residual stream). Altering a or weighting the value residual stream would enable control over the influence of past representations within the residual stream. This connection suggests that our model may be modified to encode a temporal structure over representations, akin to how SRs integrate over future expected states. Furthermore, recent work has highlighted parallels between SRs and temporal context models as applied to LLMs (Ji-An et al., 2024), particularly through the lens of reinforcement learning (RL) and temporal difference learning (Sutton, 1988; Gershman et al., 2012). These insights open a path for future work: instead of passing values from a single attention head, the model could be extended by learning weighted combinations of values across heads and layers – with the weights themselves learned via RL-based methods – further enriching the capacity to encode temporal or relational structure.

6.2 Future work

Future efforts may seek to more deeply understand our results in the context of varying the structure and order of context tokens – in associative memory networks, and during inference and training of LLMs. Along these lines, Russin et al. (2024) recently showed that LLMs and Transformers exhibit similar improvements

as seen in humans on tasks with and without rule-like structures, depending on the order and organisation of context and training tokens. In particular, ICL performance improved when context tokens appeared in semantically-relevant 'chunks' or 'blocks', as seen in humans when completing tasks with rule-like structures. Whereas, when training samples were interleaved, Transformers saw performance improvements (as measured by the extent to which the network rote-learnt training examples), similar to humans completing tasks lacking rule-like structures.

Other recent works have demonstrated how Transformer-based language models store and retrieve knowledge, using synthetic tasks to isolate specific mechanisms (Bietti et al., 2023; Nichani et al., 2024). Bietti et al. (2023) analyse how Transformers balance global knowledge learned during training with context-specific knowledge acquired during inference. They show that weight matrices function as associative memories, with different learning dynamics for global bigrams and in-context bigrams, and provide theoretical insights into how gradients support this process. Nichani et al. (2024) focus on factual recall, proving that shallow Transformers can achieve near-optimal storage capacity using either self-attention or MLP components as associative memories. They also show that such models can trade off between these components and exhibit sequential learning dynamics during training. We believe it could by highly valuable for future work to analyse how such learning dynamics vary in the case of the residual values stream modification, and to explore whether an analogous modification could be made to the MLP components.

The performance improvements from introducing a residual values stream suggests that ICL can be thought of as an associative process where continuous information reinforcement enhances model memory, efficiency, and prediction accuracy on novel data. Nonetheless, this raises several questions. Firstly, more comprehensive testing on varied datasets and different scales of language models can address whether the improvements we observed generalise beyond our specific tasks and data setups. Further, it remains to be clarified whether similar benefits manifest when dealing with other natural language processing tasks, such as sentiment analysis or translation, or whether there exist any trade-offs between ICL and other abilities. Additionally, while these initial results present a compelling case for testing attention residual streams in larger models, further exploration of optimisation parameter settings for such architectures would strengthen understanding. Quantitative studies on computation cost versus accuracy improvements will also better-inform model architecture design and selection for deployment in real-world scenarios, as well as potential competition between learning flexibility and task accuracy.

6.3 Conclusion

This research demonstrates potential conceptual and practical advancements in enhancing Transformers' adaptive capabilities through a neuroscience-inspired mechanism. By bridging neural computational principles with associative memory insights, we offer new directions for research into more intelligent and dynamic models, with potential improvements for LLMs. Our associative memory model and Transformer architecture not only bolsters existing computational frameworks, but also offers fertile ground for computational neuroscientists to analyse the computational role of skip connections in memory systems. As we progress, these interdisciplinary approaches may ultimately yield richer cognitive models that parallel, or even emulate, biological intelligence.

Acknowledgments

[redacted for anonymity]

Code availability

[redacted for anonymity]

References

Shun'ichi Amari. Learning patterns and pattern sequences by self-organizing nets of threshold elements. *IEEE Transactions on Computers*, C-21(11):1197–1206, 1972. doi: 10.1109/T-C.1972.223477.

Daniel J Amit. *Modeling brain function: The world of attractor neural networks*. Cambridge university press, 1989.

Daniel J Amit. Attractor neural networks and biological reality: associative memory and learning. *Future Generation Computer Systems*, 6(2):111–119, 1990.

Daniel J. Amit, Hanoch Gutfreund, and H. Sompolinsky. Storing infinite numbers of patterns in a spin-glass model of neural networks. *Phys. Rev. Lett.*, 55:1530–1533, Sep 1985. doi: 10.1103/PhysRevLett.55.1530. URL https://link.aps.org/doi/10.1103/PhysRevLett.55.1530.

Dmitriy Aronov, Rhino Nevers, and David W. Tank. Mapping of a non-spatial dimension by the hippocampal–entorhinal circuit. *Nature*, 543(7647):719–722, Mar 2017. ISSN 1476-4687. doi: 10.1038/nature21692.

Xiaojun Bao, Eva Gjorgieva, Laura K. Shanahan, James D. Howard, Thorsten Kahnt, and Jay A. Gottfried. Grid-like neural representations support olfactory navigation of a two-dimensional odor space. *Neuron*, 102(5):1066–1075.e5, Jun 2019. ISSN 0896-6273. doi: 10.1016/j.neuron.2019.03.034. URL https://doi.org/10.1016/j.neuron.2019.03.034.

Alison L. Barth and James F A Poulet. Experimental evidence for sparse firing in the neocortex. *Trends in Neurosciences*, 35(6):345–355, 2012. ISSN 01662236. doi: 10.1016/j.tins.2012.03.008.

Jacob L. S. Bellmund, Peter Gärdenfors, Edvard I. Moser, and Christian F. Doeller. Navigating cognition: Spatial codes for human thinking. *Science*, 362(6415):eaat6766, 2018. doi: 10.1126/science.aat6766. URL https://www.science.org/doi/abs/10.1126/science.aat6766.

Alex Bellos. He ate all the pi : Japanese man memorises π to 111,700 digits. *The Guardian*, 2015. URL https://www.theguardian.com/science/alexs-adventures-in-numberland/2015/mar/13/pi-day-2015-memory-memorisation-world-record-japanese-akira-haraguchi.

Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36:1560–1588, 2023.

Yonatan Bisk, Rowan Zellers, Ronan Le Bras, Jianfeng Gao, and Yejin Choi. Piqa: Reasoning about physical commonsense in natural language, 2019. URL https://arxiv.org/abs/1911.11641.

T. V P Bliss and A. R. Gardner-Medwin. Long-lasting potentiation of synaptic transmission in the dentate area of the unanaesthetized rabbit following stimulation of the perforant path. *The Journal of Physiology*, 232(2):357–374, 1973. ISSN 14697793. doi: 10.1113/jphysiol.1973.sp010274.

Erie D Boorman, Sarah C Sweigart, and Seongmin A Park. Cognitive maps and novel inferences: a flexibility hierarchy. *Current Opinion in Behavioral Sciences*, 38:141–149, 2021. ISSN 2352-1546. doi: https://doi.org/10.1016/j.cobeha.2021.02.017. URL https://www.sciencedirect.com/science/article/pii/S2352154621000395. Computational cognitive neuroscience.

Melina Paula Bordone, Mootaz M. Salman, Haley E. Titus, Elham Amini, Jens V. Andersen, Barnali Chakraborti, Artem V. Diuba, Tatsiana G. Dubouskaya, Eric Ehrke, Andiara Espindola de Freitas, Guilherme Braga de Freitas, Rafaella A. Gonçalves, Deepali Gupta, Richa Gupta, Sharon R. Ha, Isabel A. Hemming, Minal Jaggar, Emil Jakobsen, Punita Kumari, Navya Lakkappa, Ashley P. L. Marsh, Jessica Mitlöhner, Yuki Ogawa, Ramesh Kumar Paidi, Felipe C. Ribeiro, Ahmad Salamian, Suraiya Saleem, Sorabh Sharma, Joana M. Silva, Shripriya Singh, Kunjbihari Sulakhiya, Tesfaye Wolde Tefera, Behnam Vafadari, Anuradha Yadav, Reiji Yamazaki, and Constanze I. Seidenbecher. The energetic brain – a review from students to students. *Journal of Neurochemistry*, 151(2):139–165, 2019. doi: https://doi.org/10.1111/jnc.14829. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/jnc.14829.

Timothy F Brady, Talia Konkle, George A Alvarez, and Aude Oliva. Visual long-term memory has a massive storage capacity for object details. *Proceedings of the National Academy of Sciences*, 105(38):14325–14329, 2008.

Trenton Bricken and Cengiz Pehlevan. Attention approximates sparse distributed memory. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, 2021. URL https://openreview.net/forum?id=WVYzd7GvaOM.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jehoshua Bruck and Vwani P. Roychowdhury. On the Number of Spurious Memories in the Hopfield Model. *IEEE Transactions on Information Theory*, 36(2):393–397, 1990. ISSN 15579654. doi: 10.1109/18.52486.

Marc Brysbaert, Michaël Stevens, Paweł Mandera, and Emmanuel Keuleers. How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, 7, 2016. ISSN 1664-1078. doi: 10.3389/fpsyg.2016.01116. URL https://www.frontiersin.org/articles/10.3389/fpsyg.2016.01116.

Thomas F Burns. Semantically-correlated memories in a dense associative model. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4936–4970. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/burns24a.html.

Thomas F Burns and Tomoki Fukai. Simplicial hopfield networks. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=_QLsH8gatwx.

Thomas F Burns, Tatsuya Haga, and Tomoki Fukai. Multiscale and extended retrieval of associative memory structures in a cortical model of local-global inhibition balance. *eNeuro*, 9(3), 2022. doi: 10.1523/ENEURO.0023-22.2022. URL https://www.eneuro.org/content/9/3/ENEURO.0023-22.2022.

Lucius Bushnaq, Jake Mendel, Stefan Heimersheim, Dan Braun, Nicholas Goldowsky-Dill, Kaarel Hänni, Cindy Wu, and Marius Hobbhahn. Using degeneracy in the loss landscape for mechanistic interpretability. *arXiv preprint arXiv:2405.10927*, 2024.

György Buzsáki. Neural syntax: Cell assemblies, synapsembles, and readers. *Neuron*, 68(3):362–385, 2010. ISSN 0896-6273. doi: https://doi.org/10.1016/j.neuron.2010.09.023. URL https://www.sciencedirect.com/science/article/pii/S0896627310007658.

Rishidev Chaudhuri and Ila Fiete. Computational principles of memory. *Nature Neuroscience*, 19(3):394–403, March 2016.

Rishidev Chaudhuri and Ila Fiete. Bipartite expander Hopfield networks as self-decoding high-capacity error correcting codes. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper/2019/file/97008ea27052082be055447be9e85612-Paper.pdf.

Spyridon Chavlis and Panayiota Poirazi. Drawing inspiration from biological dendrites to empower artificial neural networks. *Current Opinion in Neurobiology*, 70:1–10, 2021. ISSN 0959-4388. doi: https://doi.org/10.1016/j.conb.2021.04.007. URL https://www.sciencedirect.com/science/article/pii/S0959438821000544. Computational Neuroscience.

Yuhan Chen, Shengjun Wang, Claus C Hilgetag, and Changsong Zhou. Trade-off between multiple constraints enables simultaneous formation of modules and hubs in neural systems. *PLoS Comput. Biol.*, 9 (3):e1002937, March 2013.

Peter H Chipman, Chi Chung Alan Fung, Alejandra Pazo Fernandez, Abhilash Sawant, Angelo Tedoldi, Atsushi Kawai, Sunita Ghimire Gautam, Mizuki Kurosawa, Manabu Abe, Kenji Sakimura, Tomoki Fukai, and Yukiko Goda. Astrocyte GluN2C NMDA receptors control basal synaptic strengths of hippocampal CA1 pyramidal neurons in the stratum radiatum. *Elife*, 10, October 2021.

Kimberly M. Christian and Richard F. Thompson. Neural Substrates of Eyeblink Conditioning: Acquisition and Retention. *Learning and Memory*, 10(6):427–455, 2003. ISSN 10720502. doi: 10.1101/lm.59603.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Claudia Clopath, Tobias Bonhoeffer, Mark Hübener, and Tobias Rose. Variance and invariance of neuronal long-term representations. *Philos Trans R Soc Lond B Biol Sci*, 372(1715), March 2017.

Rivka T Cohen and Michael Jacob Kahana. A memory-based theory of emotional disorders. *Psychological Review*, 129(4):742, 2022.

Alexandra O. Constantinescu, Jill X. O'Reilly, and Timothy E. J. Behrens. Organizing conceptual knowledge in humans with a gridlike code. *Science*, 352(6292):1464–1468, 2016. doi: 10.1126/science.aaf0941. URL `https://www.science.org/doi/abs/10.1126/science.aaf0941`.

Zhenzhong Cui, Charles R Gerfen, and W Scott Young 3rd. Hypothalamic and other connections with dorsal ca2 area of the mouse hippocampus. *Journal of comparative neurology*, 521(8):1844–1866, 2013.

Onat Dalmaz, Mahmut Yurt, and Tolga Çukur. Resvit: residual vision transformers for multimodal medical image synthesis. *IEEE Transactions on Medical Imaging*, 41(10):2598–2614, 2022.

Christopher J. Darwin, Michael T. Turvey, and Robert G. Crowder. An auditory analogue of the Sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology*, 3(2):255–267, 1972. ISSN 00100285. doi: 10.1016/0010-0285(72)90007-2.

Nathaniel D. Daw and Kenji Doya. The computational neurobiology of learning and reward. *Current Opinion in Neurobiology*, 16(2):199–204, 2006. ISSN 09594388. doi: 10.1016/j.conb.2006.03.006.

Mete Demircigil, Judith Heusel, Matthias Löwe, Sven Upgang, and Franck Vermet. On a model of associative memory with huge storage capacity. *Journal of Statistical Physics*, 168(2):288–299, Jul 2017. ISSN 1572-9613. doi: 10.1007/s10955-017-1806-y. URL `https://doi.org/10.1007/s10955-017-1806-y`.

Adi Doron, Alon Rubin, Aviya Benmelech-Chovav, Netai Benaim, Tom Carmi, Ron Refaeli, Nechama Novick, Tirzah Kreisel, Yaniv Ziv, and Inbal Goshen. Hippocampal astrocytes encode reward location. *Nature*, Aug 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05146-6. URL `https://doi.org/10.1038/s41586-022-05146-6`.

Hanyu Duan, Yixuan Tang, Yi Yang, Ahmed Abbasi, and Kar Yan Tam. Exploring the relationship between in-context learning and instruction tuning, 2023. URL `https://arxiv.org/abs/2311.10367`.

Serena M Dudek, Georgia M Alexander, and Shannon Farris. Rediscovering area ca2: unique properties and functions. *Nature Reviews Neuroscience*, 17(2):89–102, 2016.

Ronen Eldan and Yuanzhi Li. Tinystories: How small can language models be and still speak coherent english?, 2023. URL `https://arxiv.org/abs/2305.07759`.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12, 2021.

Sarah J Etherington, Susan E Atkinson, Greg J Stuart, and Stephen R Williams. *Synaptic Integration*, chapter 1, pp. 1–15. Wiley, 2010. ISBN 9780470015902. doi: https://doi.org/10.1002/9780470015902.a0000208.pub2. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470015902.a0000208.pub2`.

Samuel J Gershman, Christopher D Moore, Michael T Todd, Kenneth A Norman, and Per B Sederberg. The successor representation and temporal context. *Neural Computation*, 24(6):1553–1568, 2012.

Paul E. Gilbert and Raymond P. Kesner. Memory for objects and their locations: The role of the hippocampus in retention of object-place associations. *Neurobiology of Learning and Memory*, 81(1):39–45, 2004. ISSN 10747427. doi: 10.1016/S1074-7427(03)00069-8.

Vinod Goel and Raymond J. Dolan. Anatomical Segregation of Component Processes in an Inductive Inference Task. *Journal of Cognitive Neuroscience*, 12(1):110–119, 01 2000. ISSN 0898-929X. doi: 10.1162/08989290051137639. URL https://doi.org/10.1162/08989290051137639.

A. M. Gordon, G. Westling, K. J. Cole, and R. S. Johansson. Memory representations underlying motor commands used during manipulation of common and novel objects. *Journal of Neurophysiology*, 69(6): 1789–1797, 1993. ISSN 00223077. doi: 10.1152/jn.1993.69.6.1789.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton,

Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models, 2024. URL `https://arxiv.org/abs/2407.21783`.

Eva-Maria Griesbauer, Ed Manley, Jan M. Wiener, and Hugo J. Spiers. London taxi drivers: A review of neurocognitive studies and an exploration of how they build their cognitive map of London. *Hippocampus*, 32(1):3–20, 2022. doi: https://doi.org/10.1002/hipo.23395. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/hipo.23395`.

Vincent Gripon and Claude Berrou. Sparse neural networks with large learning diversity. *IEEE Transactions on Neural Networks*, 22(7):1087–1096, 2011. ISSN 10459227. doi: 10.1109/TNN.2011.2146789.

Lukas N. Groschner, Jonatan G. Malis, Birte Zuidinga, and Alexander Borst. A biophysical account of multiplication by a single neuron. *Nature*, 603(7899):119–123, Mar 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-04428-3. URL https://doi.org/10.1038/s41586-022-04428-3.

Bengt Gustafsson and Holger Wigström. Physiological mechanisms underlying long-term potentiation. *Trends in Neurosciences*, 11(4):156–162, 1988. ISSN 01662236. doi: 10.1016/0166-2236(88)90142-7.

Eric Hart and Alexander C. Huk. Recurrent circuit dynamics underlie persistent activity in the macaque frontoparietal network. *eLife*, 9:e52460, May 2020. ISSN 2050-084X. doi: 10.7554/eLife.52460. URL https://doi.org/10.7554/eLife.52460. 32379044[pmid].

Donald Olding Hebb. *The organization of behavior: A neuropsychological theory*. Wiley, 1949.

Nathan G. Hedrick, Zhongmin Lu, Eric Bushong, Surbhi Singhi, Peter Nguyen, Yessenia Magaña, Sayyed Jilani, Byung Kook Lim, Mark Ellisman, and Takaki Komiyama. Learning binds new inputs into functional synaptic clusters via spinogenesis. *Nature Neuroscience*, 25(6):726–737, Jun 2022. ISSN 1546-1726. doi: 10.1038/s41593-022-01086-6. URL https://doi.org/10.1038/s41593-022-01086-6.

Suzana Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in Human Neuroscience*, 3, 2009. ISSN 1662-5161. doi: 10.3389/neuro.09.031.2009. URL https://www.frontiersin.org/articles/10.3389/neuro.09.031.2009.

Evan Hernandez, Arnab Sen Sharma, Tal Haklay, Kevin Meng, Martin Wattenberg, Jacob Andreas, Yonatan Belinkov, and David Bau. Linearity of relation decoding in transformer language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=w7LU2s14kE.

Heiko Hoffmann. Sparse associative memory. *Neural Computation*, 31(5):998–1014, 2019. ISSN 1530888X. doi: 10.1162/neco_a_01181.

Benjamin Hoover, Duen Horng Chau, Hendrik Strobelt, and Dmitry Krotov. A universal abstraction for hierarchical Hopfield networks. In *The Symbiosis of Deep Learning and Differential Equations II*, 2022. URL https://openreview.net/forum?id=SAv3nhzNWhw.

J J Hopfieid. Searching for memories, Sudoku, implicit check bits, and the iterative use of not-always-correct rapid neural computation. *Neural Computation*, 20(5):1119–1164, 2008. ISSN 08997667. doi: 10.1162/neco.2008.09-06-345.

John J Hopfield. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8):2554–2558, 1982. doi: 10.1073/pnas.79.8.2554. URL https://www.pnas.org/doi/abs/10.1073/pnas.79.8.2554.

Marc W Howard, Mrigankka S Fotedar, Aditya V Datey, and Michael E Hasselmo. The temporal context model in spatial navigation and relational learning: toward a common explanation of medial temporal lobe function across domains. *Psychological review*, 112(1):75, 2005.

Zhongzhan Huang, Pan Zhou, Shuicheng YAN, and Liang Lin. Scalelong: Towards more stable training of diffusion model via scaling network long skip connection. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=0N73P8pH2l.

Michael Janner, Igor Mordatch, and Sergey Levine. Gamma-models: Generative temporal difference learning for infinite-horizon prediction. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1724–1735. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/12ffb0968f2f56e51a59a6beb37b2859-Paper.pdf.

R Jenkins, AJ Dowsett, and AM Burton. How many faces do people know? *Proceedings of the Royal Society B*, 285(1888):20181319, 2018.

Li Ji-An, Corey Y. Zhou, Marcus K. Benna, and Marcelo G. Mattar. Linking in-context learning in transformers to human episodic memory, 2024. URL https://arxiv.org/abs/2405.14992.

Yibo Jiang, Goutham Rajendran, Pradeep Ravikumar, and Bryon Aragam. Do llms dream of elephants (when told not to)? latent concept association and associative memory in transformers, 2024. URL https://arxiv.org/abs/2406.18400.

Arthur Juliani, Ryota Kanai, and Shuntaro Sasai Sasai. The perceiver architecture is a functional global workspace. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44, 2022.

Marcus Kaiser and Claus C Hilgetag. Nonoptimal component placement, but short processing paths, due to long-distance projections in neural systems. *PLoS Comput. Biol.*, 2(7):e95, July 2006.

Ayaka Kato, Kazumi Ohta, Kazuo Okanoya, and Hokto Kazama. Dopaminergic neurons dynamically update sensory values during navigation. *bioRxiv*, 2022. doi: 10.1101/2022.08.17.504092. URL https://www.biorxiv.org/content/early/2022/08/17/2022.08.17.504092.

Mikail Khona and Ila R Fiete. Attractor and integrator networks in the brain. *Nature Reviews Neuroscience*, 23(12):744–766, 2022.

Do Hyun Kim, Jinha Park, and Byungnam Kahng. Enhanced storage capacity with errors in scale-free Hopfield neural networks: An analytical study. *PLoS ONE*, 12(10), 2017. ISSN 19326203. doi: 10.1371/journal.pone.0184683.

Leo Kozachkov, Ksenia V. Kastanenka, and Dmitry Krotov. Building Transformers from neurons and astrocytes. *bioRxiv*, 2022. doi: 10.1101/2022.10.12.511910. URL https://www.biorxiv.org/content/early/2022/10/15/2022.10.12.511910.

Dmitry Krotov and John J. Hopfield. Dense associative memory for pattern recognition. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 1180–1188, Red Hook, NY, USA, 2016. Curran Associates Inc. ISBN 9781510838819.

Anders Lansner. Associative memory models: from the cell-assembly theory to biophysically detailed cortex simulations. *Trends in Neurosciences*, 32(3):178–186, 2009. ISSN 01662236. doi: 10.1016/j.tins.2008.12.002.

Edmund Lau, Zach Furman, George Wang, Daniel Murfet, and Susan Wei. The local learning coefficient: A singularity-aware complexity measure. *arXiv preprint arXiv:2308.12108*, 2023.

S. J. Lederman and R. L. Klatzky. Haptic perception: A tutorial. *Attention, Perception, and Psychophysics*, 71(7):1439–1459, 2009. ISSN 19433921. doi: 10.3758/APP.71.7.1439.

Amir Levi, Noam Aviv, and Eran Stark. Learning to learn: Single session acquisition of new rules by freely moving mice. *PNAS Nexus*, 3(5):pgae203, 05 2024. ISSN 2752-6542. doi: 10.1093/pnasnexus/pgae203. URL https://doi.org/10.1093/pnasnexus/pgae203.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/li23l.html.

William A Little. The existence of persistent states in the brain. *Mathematical Biosciences*, 19(1):101–120, 1974. ISSN 0025-5564. doi: https://doi.org/10.1016/0025-5564(74)90031-5. URL https://www.sciencedirect.com/science/article/pii/0025556474900315.

Lynn J Lohnas, Sean M Polyn, and Michael J Kahana. Expanding the scope of memory search: Modeling intralist and interlist effects in free recall. *Psychological review*, 122(2):337, 2015.

Rafael Lorente de Nó. Analysis of the activity of the chains of internuncial neurons. *Journal of Neurophysiology*, 1(3):207–244, 1938.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Matthias Löwe and Franck Vermet. The Hopfield model on a sparse Erdös-Renyi graph. *Journal of Statistical Physics*, 143(1):205–214, Apr 2011. ISSN 1572-9613. doi: 10.1007/s10955-011-0167-1. URL https://doi.org/10.1007/s10955-011-0167-1.

Yue M Lu, Mary I Letey, Jacob A Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *arXiv preprint arXiv:2405.11751*, 2024.

Corey Lynch and Pierre Sermanet. Language conditioned imitation learning over unstructured data, 2021. URL https://arxiv.org/abs/2005.07648.

Matthias Löwe. On the storage capacity of Hopfield models with correlated patterns. *The Annals of Applied Probability*, 8(4):1216 – 1250, 1998. doi: 10.1214/aoap/1028903378. URL https://doi.org/10.1214/aoap/1028903378.

N. J. Mackintosh. *Conditioning and associative learning.* Oxford University Press, Oxford, 1983.

Farshad Alizadeh Mansouri, David J Freedman, and Mark J Buckley. Emergence of abstract rules in the primate brain. *Nature Reviews Neuroscience*, 21(11):595–610, 2020.

Stephen Maren. Neurobiology of Pavlovian Fear Conditioning. *Annual Review of Neuroscience*, 24(1): 897–931, 2001. ISSN 0147-006X. doi: 10.1146/annurev.neuro.24.1.897.

D Marr. Simple memory: a theory for archicortex. *Philos Trans R Soc Lond B Biol Sci*, 262(841):23–81, July 1971.

William Mau, Michael E Hasselmo, and Denise J Cai. The brain in motion: How ensemble fluidity drives memory-updating and flexibility. *eLife*, 9:e63550, Dec 2020. ISSN 2050-084X. doi: 10.7554/eLife.63550. URL https://doi.org/10.7554/eLife.63550.

Warren S McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5:115–133, 1943.

R. McEliece, E. Posner, E. Rodemich, and S. Venkatesh. The capacity of the Hopfield associative memory. *IEEE Transactions on Information Theory*, 33(4):461–482, 1987. doi: 10.1109/TIT.1987.1057328.

Frances K. McSweeney and Eric S. Murphy (eds.). *The Wiley Blackwell Handbook of Operant and Classical Conditioning.* John Wiley & Sons, Ltd, Oxford, UK, may 2014. ISBN 9781118468135. doi: 10.1002/9781118468135. URL http://doi.wiley.com/10.1002/9781118468135.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. Can a suit of armor conduct electricity? a new dataset for open book question answering, 2018. URL https://arxiv.org/abs/1809.02789.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality, 2013. URL https://arxiv.org/abs/1310.4546.

Earl K Miller and Jonathan D Cohen. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202, 2001.

Beren Millidge, Tommaso Salvatori, Yuhang Song, Thomas Lukasiewicz, and Rafal Bogacz. Universal hopfield networks: A general framework for single-shot associative memory models. In *International Conference on Machine Learning*, pp. 15561–15583. PMLR, 2022.

B. Milner. Amnesia following operation on the temporal lobes. In C. Whitty and O. Zangwill (eds.), *Amnesia*, pp. 109–133. Butterworth, London, 1966.

G.A. Milner. The Magical Number Seven, Plus or Minus Two: Some Limits on Our Capacity for Processing Information. *Psychological Review*, 101(2):343–352, 1955. URL http://spider.apa.org/ftdocs/rev/1994/april/rev1012343.html.

Ali A. Minai and William B. Levy. The dynamics of sparse random networks. *Biological Cybernetics*, 70(2): 177–187, 1993. ISSN 03401200. doi: 10.1007/BF00200831.

Suvir Mirchandani, Fei Xia, Pete Florence, Brian Ichter, Danny Driess, Montserrat Gonzalez Arenas, Kanishka Rao, Dorsa Sadigh, and Andy Zeng. Large language models as general pattern machines, 2023. URL https://arxiv.org/abs/2307.04721.

A. A. Mofrad and M.G Parker. Nested-clique network model of neural associative memory. *Neural Computation*, 29:1681–1695, 2017.

Gianluigi Mongillo, Omri Barak, and Misha Tsodyks. Synaptic Theory of Working Memory. *Science*, 319 (5869):1543–1546, 2008. ISSN 10959203. doi: 10.1126/science.1150769.

Kaoru Nakano. Associatron-a model of associative memory. *IEEE Transactions on Systems, Man, and Cybernetics*, SMC-2(3):380–388, 1972. doi: 10.1109/TSMC.1972.4309133.

Eshaan Nichani, Jason D Lee, and Alberto Bietti. Understanding factual recall in transformers via associative memories. *arXiv preprint arXiv:2412.06538*, 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/in-context-learning-and-induction-heads/index.html.

Christos H. Papadimitriou, Santosh S. Vempala, Daniel Mitropolsky, Michael Collins, and Wolfgang Maass. Brain computation by assemblies of neurons. *Proceedings of the National Academy of Sciences*, 117(25): 14464–14472, 2020. doi: 10.1073/pnas.2001893117. URL https://www.pnas.org/doi/abs/10.1073/pnas.2001893117.

Seongmin A. Park, Douglas S. Miller, and Erie D. Boorman. Inferences on a multidimensional social hierarchy use a grid-like code. *Nature Neuroscience*, 24(9):1292–1301, Sep 2021. ISSN 1546-1726. doi: 10.1038/s41593-021-00916-3. URL https://doi.org/10.1038/s41593-021-00916-3.

Gertrudis Perea, Marta Navarrete, and Alfonso Araque. Tripartite synapses: astrocytes process and control synaptic information. *Trends in Neurosciences*, 32(8):421–431, 2009.

Brad E. Pfeiffer and David J. Foster. Autoassociative dynamics in the generation of sequences of hippocampal place cells. *Science*, 349(6244):180–183, 2015. ISSN 10959203. doi: 10.1126/science.aaa9633.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*, 2022.

Melissa A. Preziosi and Jennifer H. Coane. Remembering that big things sound big: Sound symbolism and associative memory. *Cognitive Research: Principles and Implications*, 2(1), 2017. ISSN 2365-7464. doi: 10.1186/s41235-016-0047-y.

Hubert Ramsauer, Bernhard Schäfl, Johannes Lehner, Philipp Seidl, Michael Widrich, Lukas Gruber, Markus Holzleitner, Thomas Adler, David Kreil, Michael K Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. Hopfield networks is all you need. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=tL89RnzIiCd.

Charan Ranganath and Robert T Knight. Prefrontal cortex and episodic memory: Integrating findings from neuropsychology and functional brain imaging. *The cognitive neuroscience of memory: Encoding and retrieval*, 1:83, 2002.

Nelson Rebola, Mario Carta, and Christophe Mulle. Operation and plasticity of hippocampal CA3 circuits: Implications for memory encoding. *Nature Reviews Neuroscience*, 18(4):209–221, 2017. ISSN 14710048. doi: 10.1038/nrn.2017.10.

Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=aN4Jf6Cx69.

R.A. Rescorla and A.R. Wagner. A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A.H. Black and W.F. Prokasy (eds.), *Classical Conditioning II: Current Research and Theory*, pp. 64–99. Appleton Century Crofts, New York, 1972.

Mark Rigby, Federico Grillo, Benjamin Compans, Guilherme Neves, Julia Gallinaro, Sophie Nashashibi, Gema Vizcay-Barrena, Florian Levet, Jean-Baptiste Sibarita, Angus Kirkland, Roland A. Fleck, Claudia Clopath, and Juan Burrone. Multi-synaptic boutons are a feature of ca1 hippocampal connections that may underlie network synchrony. *bioRxiv*, 2022. doi: 10.1101/2022.05.30.493836. URL https://www.biorxiv.org/content/early/2022/05/30/2022.05.30.493836.

Daniel S. Rizzuto and Michael J. Kahana. An autoassociative neural network model of paired-associate learning. *Neural Computation*, 13(9):2075–2092, 2001. ISSN 08997667. doi: 10.1162/089976601750399317.

Vincent Robert, Sadiyah Cassim, Vivien Chevaleyre, and Rebecca A Piskorowski. Hippocampal area ca2: properties and contribution to hippocampal function. *Cell and tissue research*, 373:525–540, 2018.

Matthew Rosenberg, Tony Zhang, Pietro Perona, and Markus Meister. Mice in a labyrinth show rapid learning, sudden insight, and efficient exploration. *eLife*, 10:e66175, jul 2021. ISSN 2050-084X. doi: 10.7554/eLife.66175. URL https://doi.org/10.7554/eLife.66175.

Frank Rosenblatt. Principles of neurodynamics: Perceptrons and the theory of brain mechanisms, 1961.

Nicolas P Rougier, David C Noelle, Todd S Braver, Jonathan D Cohen, and Randall C O'Reilly. Prefrontal cortex and flexible cognitive control: Rules without symbols. *Proceedings of the National Academy of Sciences*, 102(20):7338–7343, 2005.

Jacob Russin, Ellie Pavlick, and Michael J Frank. Human curriculum effects emerge with in-context learning in neural networks. *ArXiv*, 2024.

Saratha Sathasivam and Wan Ahmad Tajuddin Wan Abdullah. Logic Learning in Hopfield Networks. *Modern Applied Science*, 2(3), 2008. ISSN 1913-1844. doi: 10.5539/mas.v2n3p57.

W. B. Scoville and B. Milner. Loss of recent memory after bilateral hippocampal lesions. *Journal of Neurolgy, Neurosurgery, and Psychiatry*, 20(1):11–21, 1957.

Philipp Seidl, Philipp Renz, Natalia Dyubankova, Paulo Neves, Jonas Verhoeven, Jörg K. Wegner, Marwin Segler, Sepp Hochreiter, and Günter Klambauer. Improving few- and zero-shot reaction template prediction using modern Hopfield networks. *Journal of Chemical Information and Modeling*, 62(9):2111–2120, 2022. doi: 10.1021/acs.jcim.1c01065. URL https://doi.org/10.1021/acs.jcim.1c01065. PMID: 35034452.

Adam Shai, Paul M. Riechers, Lucas Teixeira, Alexander Gietelink Oldenziel, and Sarah Marzen. Transformers represent belief state geometry in their residual stream. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=YIB7REL8UC.

Sugandha Sharma, Sarthak Chandra, and Ila Fiete. Content addressable memory without catastrophic forgetting by heteroassociation with a fixed scaffold. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 19658–19682. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/sharma22b.html.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

George Sperling. A Model for Visual Memory Tasks. *Human Factors: The Journal of Human Factors and Ergonomics Society*, 5(1):19–31, 1963. ISSN 15478181. doi: 10.1177/001872086300500103.

Reisa Sperling, Elizabeth Chua, Andrew Cocchiarella, Erin Rand-Giovannetti, Russell Poldrack, Daniel L. Schacter, and Marilyn Albert. Putting names to faces: Successful encoding of associative memories activates the anterior hippocampal formation. *NeuroImage*, 20(2):1400–1410, 2003. ISSN 10538119. doi: 10.1016/S1053-8119(03)00391-4.

Larry R. Squire. Memory and the Hippocampuss: A Synthesis From Findings With Rats, Monkeys, and Humans. *Psychological Review*, 99(2):195–231, 1992. ISSN 0033295X. doi: 10.1037/0033-295X.99.2.195.

Larry R. Squire, Arthur P. Shimamura, and David G. Amaral. Memory and the Hippocampus. *Neural Models of Plasticity*, pp. 208–239, 1989. doi: 10.1016/b978-0-12-148956-4.50016-4.

Lionel Standing. Learning 10000 pictures. *The Quarterly journal of experimental psychology*, 25(2):207–222, 1973.

Amos Storkey. Increasing the capacity of a Hopfield network without sacrificing functionality. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 1327:451–456, 1997. ISSN 16113349. doi: 10.1007/bfb0020196.

Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset, 2024a. URL `https://arxiv.org/abs/2412.02595`.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024b.

R. J. Sutherland and J. W. Rudy. Configural association theory: The role of the hippocampal formation in learning, memory, and amnesia. *Psychobiology*, 17(2):129–144, 1989. ISSN 08896313. doi: 10.3758/BF03337828.

Richard S Sutton. Learning to predict by the methods of temporal differences. *Machine learning*, 3:9–44, 1988.

A Treves and D J Amit. Metastable states in asymmetrically diluted Hopfield networks. *Journal of Physics A: Mathematical and General*, 21(14):3155–3169, jul 1988. doi: 10.1088/0305-4470/21/14/016. URL `https://doi.org/10.1088/0305-4470/21/14/016`.

Danil Tyulmankov, Ching Fang, Annapurna Vadaparty, and Guangyu Robert Yang. Biological learning in key-value memory networks. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 22247–22258. Curran Associates, Inc., 2021. URL `https://proceedings.neurips.cc/paper/2021/file/bacadc62d6e67d7897cef027fa2d416c-Paper.pdf`.

Michael T. Ullman. Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, 92(1-2):231–270, 2004. ISSN 00100277. doi: 10.1016/j.cognition.2003.10.008.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf`.

Edward K. Vogel, Geoffrey F. Woodman, and Steven J. Luck. Storage of features, conjunctions, and objects in visual working memory. *Journal of Experimental Psychology: Human Perception and Performance*, 27 (1):92–114, 2001. ISSN 00961523. doi: 10.1037/0096-1523.27.1.92.

Ivan Voitov and Thomas D. Mrsic-Flogel. Cortical feedback loops bind distributed representations of working memory. *Nature*, 608(7922):381–389, Aug 2022. ISSN 1476-4687. doi: 10.1038/s41586-022-05014-3. URL https://doi.org/10.1038/s41586-022-05014-3.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, Joao Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

Richard A. Watson, C. L. Buckley, and Rob Mills. Optimization in "self-modeling" complex adaptive systems. *Complexity*, 16(5):17–26, 2011a. ISSN 10990526. doi: 10.1002/cplx.20346.

Richard A. Watson, Rob Mills, and C. L. Buckley. Global adaptation in networks of selfish components: Emergent associative memory at the system scale. *Artificial Life*, 17(3):147–166, 2011b. ISSN 10645462. doi: 10.1162/artl_a_00029.

Melanie Weber, Pedro D. Maia, and J. Nathan Kutz. Estimating memory deterioration rates following neurodegeneration and traumatic brain injuries in a Hopfield network model. *Frontiers in Neuroscience*, 11(NOV), 2017. ISSN 1662453X. doi: 10.3389/fnins.2017.00623.

Michael Widrich, Bernhard Schäfl, Milena Pavlović, Hubert Ramsauer, Lukas Gruber, Markus Holzleitner, Johannes Brandstetter, Geir Kjetil Sandve, Victor Greiff, Sepp Hochreiter, et al. Modern Hopfield networks and attention for immune repertoire classification. *Advances in Neural Information Processing Systems*, 33:18832–18845, 2020.

Tom J. Wills, Colin Lever, Francesca Cacucci, Neil Burgess, and John O'Keefe. Attractor dynamics in the hippocampal representation of the local environment. *Science*, 308(5723):873–876, 2005. ISSN 00368075. doi: 10.1126/science.1108905.

István Winkler and Nelson Cowan. From sensory to long-term memory: Evidence from auditory memory reactivation studies. *Experimental Psychology*, 52(1):3–20, 2005. ISSN 16183169. doi: 10.1027/1618-3169.52.1.3.

Alexander Woodward, Tom Froese, and Takashi Ikegami. Neural coordination can be enhanced by occasional interruption of normal firing patterns: A self-optimizing spiking neural network model. *Neural Networks*, 62:39–46, 2015. ISSN 18792782. doi: 10.1016/j.neunet.2014.08.011.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence?, 2019. URL https://arxiv.org/abs/1905.07830.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in Neural Information Processing Systems*, 32, 2019.

Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024a.

Zixuan Zhang, Kaiqi Zhang, Minshuo Chen, Yuma Takeda, Mengdi Wang, Tuo Zhao, and Yu-Xiang Wang. Nonparametric classification on low dimensional manifolds using overparameterized convolutional residual networks. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024b. URL https://openreview.net/forum?id=guzWIg7ody.

Jiachen Zhao. In-context exemplars as clues to retrieving from large associative memory. In *Associative Memory & Hopfield Networks in 2023*, 2023. URL https://openreview.net/forum?id=pgPAsSv5ga.

Corey Y Zhou, Deborah Talmi, Nathaniel Daw, and Marcelo G Mattar. Episodic retrieval for model-based evaluation in sequential decision tasks, 2023.

Appendix

A Connection between associative memory models and modern neuroscience

Associative memory falls under the broader category of *long-term memory*—the type of memory capable of lasting a lifetime. Yet, memory also operates on at least two additional, functionally distinct timescales: *short-term memory* and *sensory memory*. Short-term (or working) memory retains information over tens of seconds and serves as a limited, passive buffer that enables manipulation and use of that information (Milner, 1955; Mongillo et al., 2008). In contrast, sensory memory operates on even briefer timescales – typically just a few seconds. It is the initial stage where sensory inputs such as visual (Sperling, 1963; Vogel et al., 2001), auditory (Darwin et al., 1972; Winkler & Cowan, 2005), and other sensory modalities (Gordon et al., 1993; Lederman & Klatzky, 2009) are actively "remembered" before either being discarded or encoded further. A general model for forming long-term memories thus involves: (1) receiving sensory input; (2) briefly storing features of that input in sensory memory; (3) selectively retaining and manipulating aspects of this input within short-term memory, possibly integrating multiple sources; and (4) consolidating the result into long-term memory. Theoretical and computational insights into these processes are valuable for two reasons: (i) they may inform the development of intelligent systems inspired by biological memory mechanisms; and (ii) they contribute to understanding the neural basis of memory, with potential applications in both clinical and cognitive contexts.

In both humans and other animals, associative memory typically refers to any form of long-term memory involving the linking or "association" of distinct stimuli, allowing the recall of one item upon encountering the other. Classical examples include name–face associations (Sperling et al., 2003), object–sound pairings (Preziosi & Coane, 2017), and object–location associations (Gilbert & Kesner, 2004). These forms are considered part of *explicit* or *declarative memory* (Ullman, 2004), which encompasses memories that can be consciously retrieved or verbalized. In contrast, *implicit* or *non-declarative memory* supports associative processes that occur unconsciously or automatically, such as those formed through classical conditioning (Maren, 2001; Christian & Thompson, 2003) and operant conditioning (Mackintosh, 1983; McSweeney & Murphy, 2014).

Computational models of implicit associative memory have a long history, beginning with foundational models such as the Rescorla–Wagner model (Rescorla & Wagner, 1972), which laid groundwork for the modern field of reinforcement learning (Daw & Doya, 2006). Explicit associative memory, on the other hand, appears to require greater computational sophistication, as suggested by the complexity of its biological underpinnings (Chaudhuri & Fiete, 2016; Clopath et al., 2017; Mau et al., 2020).

A seminal model of explicit associative memory is the associative memory model Marr (1971); Nakano (1972); Amari (1972); Little (1974), popularised as the Hopfield network (Hopfield, 1982). In its simplest form, when storing a single memory, the associative memory model allows for a configuration of neuron thresholds such that a partial activation reliably leads to full memory recall. This behaviour corresponds to attractor dynamics converging to a stable memory state, analogous to neuronal assemblies in the hippocampus (Wills et al., 2005; Pfeiffer & Foster, 2015; Rebola et al., 2017).

The hippocampus is widely regarded as a central structure in memory processing. Along with the entorhinal, perirhinal, and parahippocampal cortices, it plays a critical role in the formation of explicit memories (Scoville & Milner, 1957; Milner, 1966; Squire, 1992). It acts as a temporary store for new information, which is later consolidated in the cortex (Squire et al., 1989; Sutherland & Rudy, 1989). Traditionally, these processes have been attributed to Hebbian learning and long-term potentiation (Bliss & Gardner-Medwin, 1973; Gustafsson & Wigström, 1988), though recent research has highlighted additional mechanisms that may contribute to memory formation and maintenance (Rigby et al., 2022; Groschner et al., 2022; Hedrick et al., 2022; Kato et al., 2022; Papadimitriou et al., 2020; Hart & Huk, 2020; Chipman et al., 2021; Perea et al., 2009; Doron et al., 2022; Etherington et al., 2010; Chavlis & Poirazi, 2021).

Despite its strengths, the classical associative memory model has limitations as a full account of long-term memory. For example, its memory capacity scales linearly with the number of neurons n, supporting approximately $0.14n$ stable patterns before performance degrades due to spurious attractors (Amit et al., 1985; McEliece et al., 1987; Bruck & Roychowdhury, 1990). This capacity is further reduced under correlated input conditions (Löwe, 1998), sparse connectivity (Treves & Amit, 1988; Löwe & Vermet, 2011), or both (Burns et al., 2022). Yet, biological systems often operate in such sparse regimes (Minai & Levy, 1993; Lansner, 2009; Barth & Poulet, 2012), and human memory regularly handles highly structured and overlapping information (Constantinescu et al., 2016; Aronov et al., 2017; Bellmund et al., 2018; Bao et al., 2019; Park et al., 2021; Griesbauer et al., 2022). Nevertheless, humans can remember thousands of highly similar images (Standing, 1973; Brady et al., 2008), recognize countless faces (Jenkins et al., 2018), learn tens of thousands of words (Brysbaert et al., 2016), and even recite over 100,000 digits of π (Bellos, 2015), all without significant interference.

Although modern associative memory networks have achieved much higher theoretical capacities (Krotov & Hopfield, 2016; Demircigil et al., 2017), the empirical data and biological constraints – such as the finite number of neurons (Herculano-Houzel, 2009) and the metabolic costs of sustaining their connections (Bordone et al., 2019) – suggest there are deeper computational and physiological principles at work. Even within associative memory-type frameworks, memory capacity can be interpreted not just as storage volume, but as a trade-off involving recall fidelity, interference, and cognitive efficiency.

Still, these limitations do not diminish the relevance of associative memory networks or their successors for understanding memory phenomena in biological and machines. They remain powerful tools in neuroscience (Sathasivam & Wan Abdullah, 2008; Rizzuto & Kahana, 2001; Weber et al., 2017), machine learning (Widrich et al., 2020; Seidl et al., 2022), and in bridging the gap between artificial and biological systems (Sharma et al., 2022; Hoover et al., 2022; Chaudhuri & Fiete, 2019; Tyulmankov et al., 2021; Kozachkov et al., 2022). Substantial progress has already been made in expanding their capabilities, including increased capacity and efficiency (Storkey, 1997; Hopfieid, 2008; Krotov & Hopfield, 2016; Gripon & Berrou, 2011; Mofrad & Parker, 2017; Burns & Fukai, 2023), sparse representations (Kim et al., 2017; Hoffmann, 2019), and the incorporation of biologically inspired mechanisms (Watson et al., 2011a;b; Woodward et al., 2015; Burns et al., 2022; Burns, 2024). The current work aims to contribute further to these areas by: (1) understanding the capacity of the associative memory framework to perform ICL (demonstrated with the AMICL model); and (2) testing whether incorporating associative memory-inspired modifications to Transformers improves their capabilities (demonstrated with the residual values stream).

B Extended figures

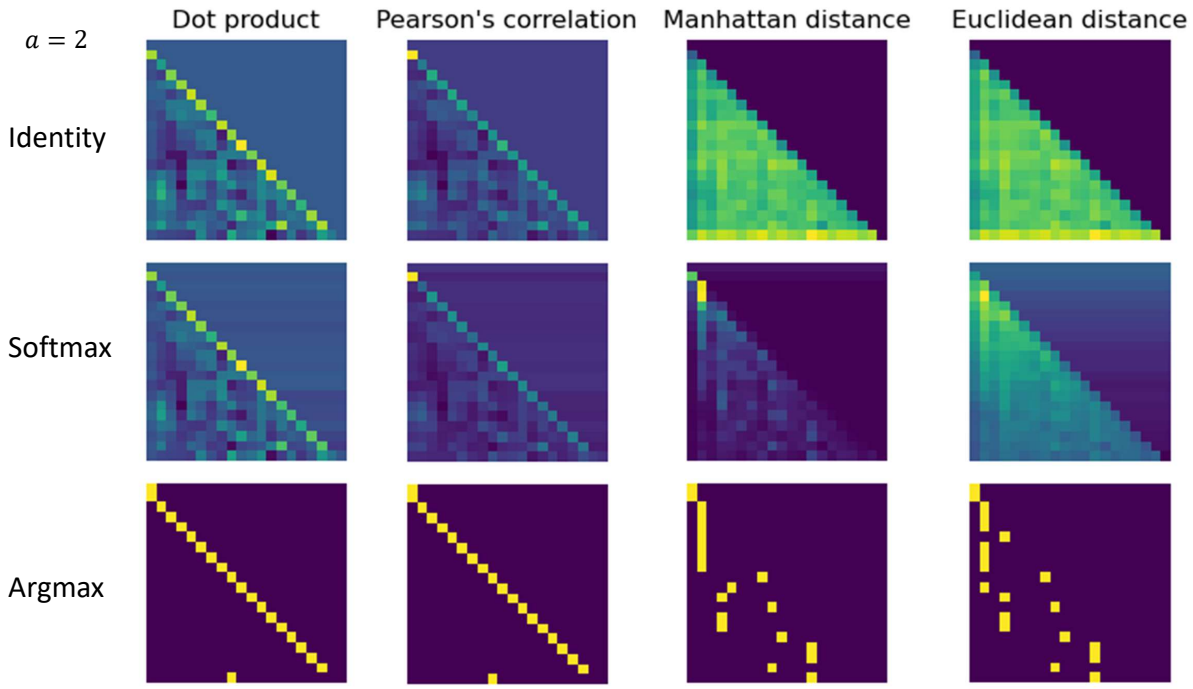

Figure 7: Attention matrices for label–object pairs in AMICL, using different similarity and separation functions with $a = 2$.

C Extended tables

Table 5: Mean ± standard deviation of training snapshot number where accuracy first exceeded 0.5 for the IC and IC2 tasks in the classic (unmodified), and residual queries, keys, and values stream networks.

	Classic	Queries (ours)	Keys (ours)	**Values (ours)**
IC	51.0 ± 4.24	41.5 ± 2.29	49.75 ± 0.43	**32.75** ± 0.83
IC2	51.5 ± 4.39	41.5 ± 1.66	49.25 ± 0.83	**33.25** ± 1.09

Table 6: Mean ± standard deviation of training snapshot number where accuracy first exceeded 0.9 for the IC and IC2 tasks in the classic (unmodified), and residual queries, keys, and values stream networks.

	Classic	Queries (ours)	Keys (ours)	**Values (ours)**
IC	59.25 ± 4.49	48.25 ± 1.48	57.5 ± 0.5	**42.25** ± 1.3
IC2	59.0 ± 4.53	48.0 ± 1.87	57.25 ± 0.83	**42.0** ± 0.71

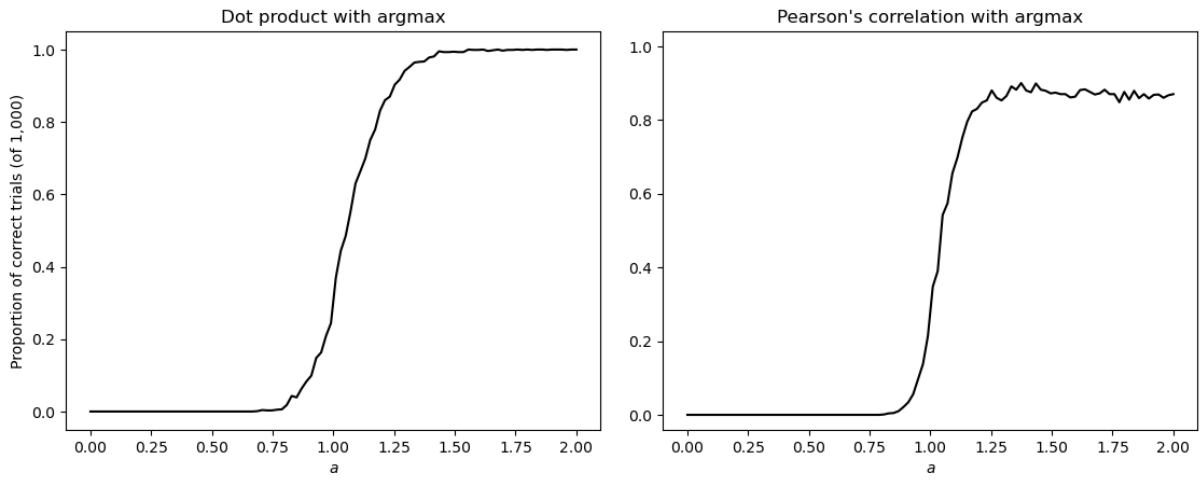

Figure 8: Proportion of correct trials, tested over $1,000$ trials for values between $a = 0$ and $a = 2$ in AMICL for label–object pairs using the DOT PRODUCT (left) and PEARSON'S CORRELATION (left) similarity with the ARGMAX separation function.

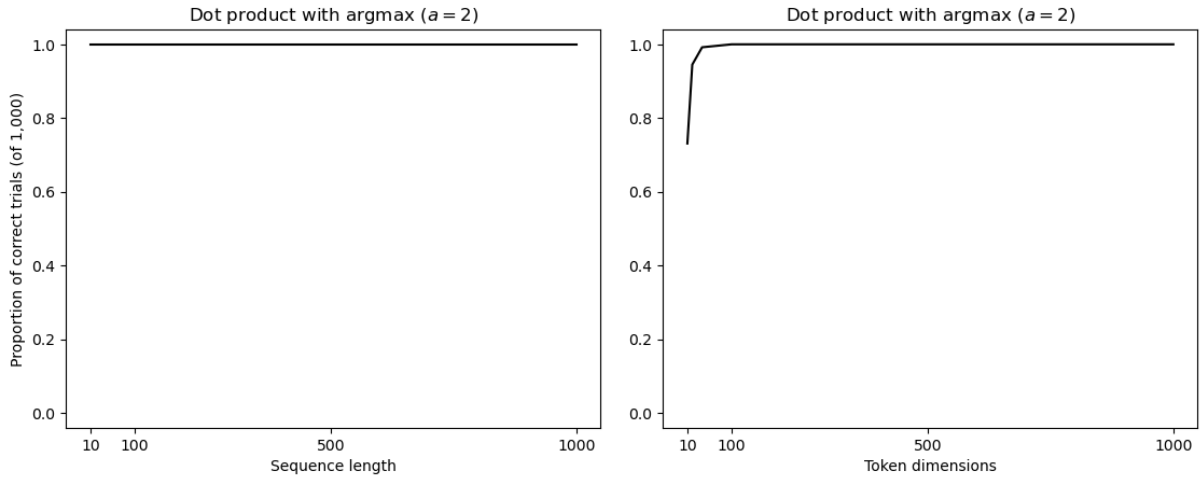

Figure 9: Proportion of correct trials, tested over $1,000$ trials for varying the values of the sequence length (s, left) and token dimensions (e, right) between 10 and $1,000$ in AMICL for label–object pairs using $a = 2$ with the DOT PRODUCT similarity function and ARGMAX separation function.

D Notation and abbreviations tables

A comprehensive list of all notations and abbreviations used in this paper is provided in the tables below.

Abbreviations

LLMs	large language models
ICL	in-context learning
LMs	langauge models
i.i.d.	independent and identically distributed
IW	in-weight
IC	in-context
IC2	in-context 2
IOI	indirect object identification
CA1, CA2, CA3	cornu Ammonis 1, 2, 3

Variables

X_i	The input sequence data $X_i \in \mathbb{R}^{e \times \mathscr{s}}$, with $\mathscr{s}$ column vectors, corresponding to the token embeddings, each of dimension e. Where the subscript i is present, this denotes the data is taken from the i–th Transformer layer, *i.e.*, the residual stream data.
x_i	A token embedding $x_i \in \mathbb{R}^e$, where the subscript i denotes the column position in the input sequence X.
$x_{\mathscr{s}}$	The token embedding $x_i \in \mathbb{R}^e$ of the final column, *i.e.*, the final token, in the input sequence X.
o^i	An object token embedding $o^i \in \mathbb{R}^e$, where the superscript i identifies the object–label identity.
l^i	A label token embedding $l^i \in \mathbb{R}^e$, where the superscript i identifies the object–label identity.
μ_i	A vector $\mu_i \in \mathbb{R}^e$ whose components are i.i.d. sampled from a normal distribution having mean zero and variance $1/e$. Used for constructing the token embeddings of object or label i, where for each object, o^i, and label, l^i, the vector μ_i is fixed.
ε	A fixed real number $\varepsilon \in \mathbb{R}$ which controls the inter-instance variability of objects, and is set to 0.1 unless stated otherwise.
η	A vector $\eta \in \mathbb{R}^e$ whose components are i.i.d. sampled from a normal distribution having mean zero and variance $1/e$. Redrawn and used for adding inter-instance variability in the construction of each object token embedding.
W_i^q, W_i^k	Query and key weight matrices $W^q, W^k \in \mathbb{R}^{\mathscr{k} \times e}$, respectively, of the i–th Transformer layer.
W_i^v	Value weight matrix $W^v \in \mathbb{R}^{\mathscr{v} \times e}$ of the i–th Transformer layer.
Q_i	Queries matrix $Q \in \mathbb{R}^{\mathscr{k} \times \mathscr{s}}$ of the i–th Transformer layer, calculated using W_i^q and X_i.
K_i	Keys matrix $K \in \mathbb{R}^{\mathscr{k} \times \mathscr{s}}$ of the i–th Transformer layer, calculated using W_i^k and X_i.
V_i	Values matrix $V \in \mathbb{R}^{\mathscr{v} \times \mathscr{s}}$ of the i–th Transformer layer, calculated using W_i^v and X_i.
S_i	The scores matrix, $S_i \in \mathbb{R}^{\mathscr{s} \times \mathscr{s}}$, which is equal to $K_i^T Q_i$.
q_i	A queries column vector, $q_i \in \mathbb{R}^{\mathscr{k}}$, where the subscript i denotes the column position in the queries matrix Q.
k_i	A keys column vector, $k_i \in \mathbb{R}^{\mathscr{k}}$, where the subscript i denotes the column position in the keys matrix K.
v_i	A values column vector, $v_i \in \mathbb{R}^{\mathscr{v}}$, where the subscript i denotes the column position in the values matrix V.

Dimensions

e	Dimensionality $e \in \mathbb{N}^+$ of each token embedding.
$\mathscr{s}$	Number of tokens $\mathscr{s} \in \mathbb{N}^+$ in the input sequence data X.
$\mathscr{k}$	Reduced token embedding dimension $\mathscr{k} \in \mathbb{N}^+$ for keys and queries in the attention operation.
$\mathscr{v}$	Reduced token embedding dimension $\mathscr{v} \in \mathbb{N}^+$ for values in the attention operation.
ℓ	Number of unique labels $\ell \in \mathbb{N}^+$ in the input sequence data X.

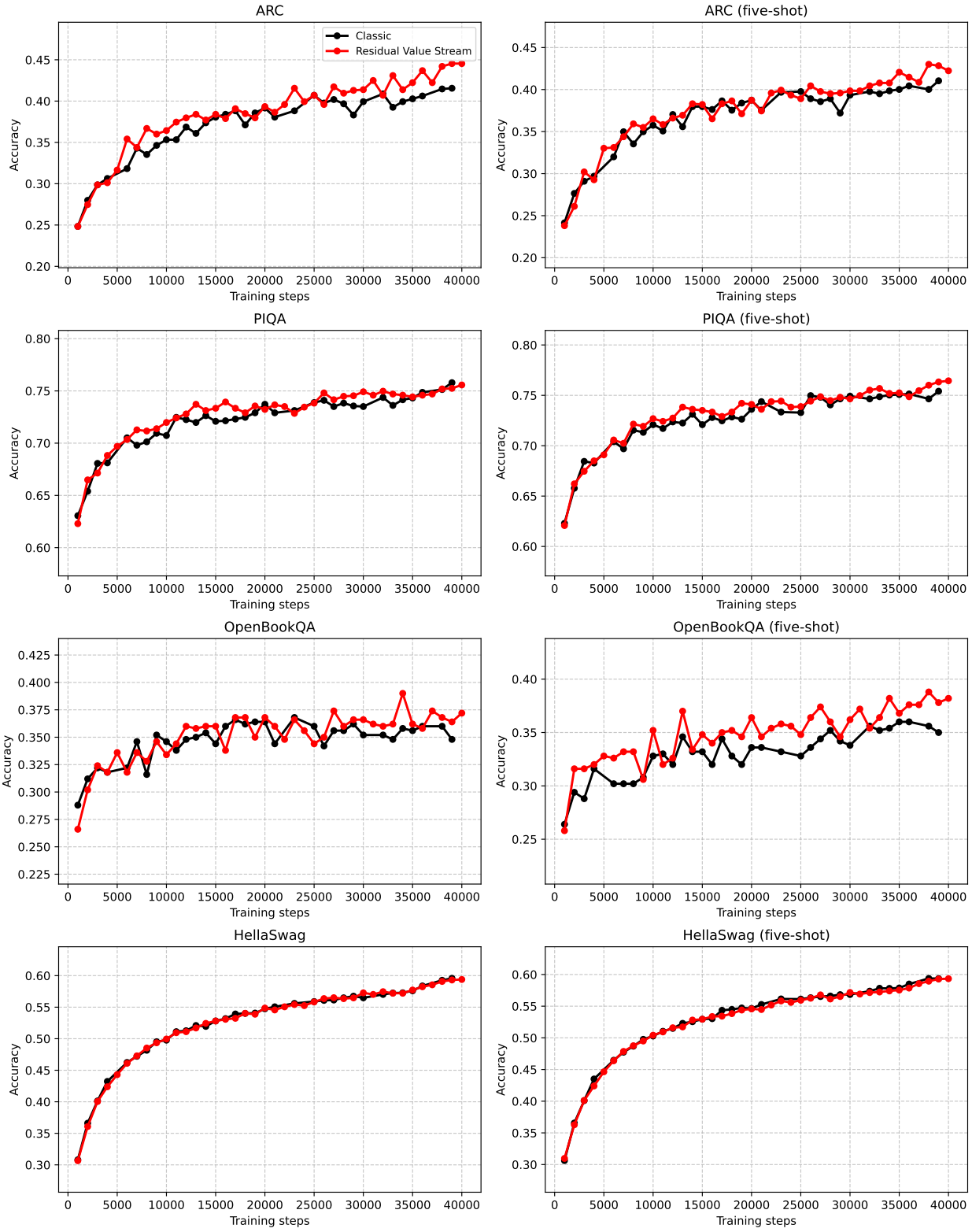

Figure 10: Accuracy across training steps for our evaluations of the 1B model, single- and five-shot.