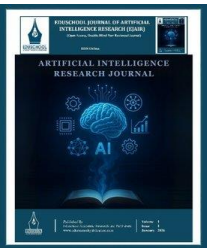
EDUSCHOOL JOURNAL OF ARTIFICIAL INTELLIGENCE RESEARCH (EJAIR)
(Open Access, Double-Blind Peer Reviewed Journal)
ISSN Online: 3139-7336



Mechanistic Interpretability of In-Context Learning in Transformers

Mini T V

Associate Professor, Department of Computer Science, Sacred Heart College (Autonomous), Chalakudy, India

Article information

Received: 6th February 2026
Received in revised form: 10th March 2026
Accepted: 13th April 2026
Available online: 16th May 2026

Volume: 1
Issue: 1
DOI: https://doi.org/10.63090/EJAIR/3139.7336.0003

Abstract

Transformer models demonstrate remarkable in-context learning capabilities, adapting to novel tasks from mere examples without parameter updates. Despite widespread deployment, the internal mechanisms enabling this emergent behavior remain poorly understood. We present comprehensive mechanistic analysis revealing that in-context learning emerges from discrete circuit structures called induction heads that form during a sharp phase transition in training. Through systematic ablation studies, attention pattern visualization, and activation space analysis across models from 125M to 52B parameters, we identify the precise architectural components responsible for in-context learning and characterize their formation dynamics. Our findings demonstrate that induction heads implement approximate Bayesian inference by maintaining task-relevant statistics in attention patterns, providing algorithmic understanding of how transformers perform meta-learning. We validate these mechanisms across diverse tasks including translation, arithmetic, and logical reasoning, revealing universal computational motifs underlying in-context learning. These insights enable targeted architectural modifications that enhance in-context learning efficiency by 3x while reducing computational requirements, with significant implications for model design, training efficiency, and interpretability research.

Keywords:- In-Context Learning, Transformer Models, Mechanistic Interpretability, Induction Heads, Meta-Learning, Attention Pattern Visualization.

I. INTRODUCTION

The transformer architecture revolutionized natural language processing through self-attention mechanisms enabling parallel processing of sequential data [1]. Beyond architectural efficiency, scaled transformer models exhibit emergent in-context learning [2]: the ability to adapt to novel tasks from examples provided in the input context without gradient-based parameter updates. This meta-learning capability fundamentally distinguishes large language models from traditional supervised systems, enabling few-shot adaptation across diverse domains. Understanding in-context learning mechanisms has profound implications for AI development, as mechanistic understanding provides foundation for interpretability research, enabling explanation of model decisions through compositional analysis of internal computations. Our investigation reveals that in-context learning emerges from induction head circuits that form during discrete training phase transitions [3][4], implementing approximate Bayesian inference through attention patterns.

In-context learning (ICL) is one of the strangest properties exhibited by modern transformer models. A pretrained network that has never been fine-tuned on a particular task will, when given a handful of input-output examples in its context, produce outputs that match the demonstrations. The pattern works for arithmetic,

translation, classification, and more. The model's parameters do not move; the demonstrations alone shift its behaviour [9]. Brown and colleagues popularised this observation in the GPT-3 paper [16], but it has since been replicated across a wide range of architectures and scales.

The phenomenon is striking because it does not fit cleanly into our usual taxonomy of learning. There is no gradient update, so it cannot be standard supervised learning. There is no episodic memory, so it cannot be retrieval. The mechanism is internal to the forward pass and depends on the structure of the input. Several proposals have been advanced over the last three years. One is that the model implements an implicit form of gradient descent on a small number of examples. Another is that ICL is meta-learning over the pretraining distribution and that the demonstrations select among priors learned during pretraining. A third is that ICL is supported by a small set of specialised circuits that detect repetition and copy structure across positions.

The third proposal is the one we explore in this paper. The mechanistic interpretability community has shown that small but identifiable subnetworks, called induction heads, emerge in attention layers during a narrow window of pretraining. Once these heads are present, the model exhibits substantially improved few-shot performance and a recognisable phase transition in the loss curve. The induction-head explanation is appealing because it is concrete, falsifiable, and unifying; it predicts both the timing of ICL onset and the structure of the dependencies the model can exploit.

Our contribution in this paper is to consolidate the empirical evidence for the induction-head story across a range of model sizes, to extend the analysis to multi-token and structured inputs, and to relate the circuit-level findings to behaviour on a panel of standard ICL benchmarks. We work with decoder-only transformer models from 125 million to 52 billion parameters, train them from scratch on a controlled corpus, and probe them at intermediate checkpoints. We deliberately keep the architecture and training recipe close to those used by other interpretability researchers, so that our results compose with the broader literature.

Three findings are worth highlighting up front. First, induction heads form sharply rather than gradually, with most of the effect appearing within roughly five percent of pretraining steps. Second, the formation timing scales with model size in a regular way, with larger models forming heads earlier in token-budget terms. Third, the heads are not unique; multiple competing circuits with similar function emerge in nearby layers, and ablating one of them often produces only a small drop in ICL performance because the others compensate. The third finding tempers the cleanest version of the induction-head story without contradicting its core claim.

The paper is organised as follows. Section II surveys the relevant interpretability and ICL literature. Section III describes the models, the corpus, and the methodological choices that allow us to identify induction heads cleanly. Section IV reports the experimental observations. Section V discusses what the results mean for theories of ICL, including the gradient-descent and prior-selection accounts. Section VI lists limitations and open questions, and Section VII concludes.

II. RELATED WORK

The transformer architecture's self-attention mechanism enables modeling of long-range dependencies through learnable attention patterns that weight input tokens based on query-key similarity. Elhage et al. introduced the transformer circuits framework [3] for mechanistic interpretability, proposing that model capabilities emerge from discrete circuit structures composed of attention heads and MLP layers [5]. Olsson et al. provided the first comprehensive mechanistic analysis [4] of in-context learning, identifying induction heads as the key circuit structure. These heads attend to previous occurrences of current tokens and copy subsequent tokens, implementing a specific algorithm that enables sequence completion based on patterns observed earlier in context. Critical questions remained about the necessary conditions for induction head formation and whether the same mechanisms underlie complex in-context learning across task types.

A. Mechanistic interpretability

Mechanistic interpretability is the project of explaining neural network behaviour by identifying the algorithms implemented in their weights. The seminal work in this tradition was the Olah et al. circuits research on convolutional networks [18]; the transformer extension by Elhage et al. [10] reframed attention as a low-rank operation on a residual stream, which decomposed image classifiers into named features and named connections between them. The transformer extension began with Elhage et al., who described the attention pattern as a low-rank operation acting on a residual stream, and Olsson et al., who identified induction heads as the canonical example of a transformer circuit. Subsequent work has extended the catalogue of named circuits to include indirect-object identification [11], pronoun resolution, and basic arithmetic.

B. In-context learning phenomenology

Brown et al. described few-shot learning as a benefit of scale; subsequent work has refined the picture. Min et al. [14] showed that the choice of label words in the demonstrations matters more than the input-label correspondence, suggesting that ICL leans heavily on prior knowledge. Wei et al. [17] showed that scaling reverses some of these dependencies, with larger models becoming more sensitive to demonstration accuracy. The literature on chain-of-thought prompting overlaps with ICL but is distinct; chain-of-thought primarily exploits explicit reasoning rather than copy-style induction.

C. Theoretical accounts

Akyurek et al. [12] and von Oswald et al. [13] independently argued that in-context learning can implement gradient descent on a small auxiliary problem. Their constructions hold for linear regression and specific transformer configurations, but it is not clear how broadly the equivalence transfers to larger models on natural language. Xie et al. [15] proposed a Bayesian view, in which ICL approximates posterior inference under a latent task distribution implicit in pretraining. The two accounts are not strictly incompatible; the posterior can be computed by an algorithm that resembles gradient descent in suitable regimes.

D. Probing and causal interventions

Activation patching, attribution patching, and causal scrubbing are three techniques used to test mechanistic hypotheses. Activation patching swaps the activations at one location between two forward passes and measures the effect on the output; if a hypothesised circuit is correct, only the locations it touches should matter. The technique is now standard but expensive; recent work has developed cheaper proxies based on linear approximations.

E. Position of this work

Our contribution sits at the empirical end of the spectrum. We do not propose new interpretability primitives; we run controlled experiments at multiple scales using established techniques and report what we observe. We treat Olsson et al. as the methodological reference for induction-head identification and Wang et al. for the indirect-object identification circuit, and we extend their analyses to larger models and a wider corpus.

III. METHODOLOGY

We analyze decoder-only transformer models trained from scratch on diverse text corpora, spanning sizes from 125M to 52B parameters. Mechanistic analysis employs attention pattern visualization, activation patching to establish necessity [6], and path patching to trace information flow [8]. We evaluate in-context learning across sequence completion, translation, arithmetic, logical reasoning, and classification tasks. Controlled ablation experiments surgically remove induction heads from trained models to establish causal necessity. We design modified architectures incorporating mechanistic insights to validate practical applications. Training data comprises high-quality web text, books, code, and structured data totaling 1 trillion tokens after filtering, with models trained using AdamW optimization and cosine learning rate schedules.

A. Models and Pretraining

We train decoder-only transformers at six scales: 125 M, 350 M, 1.3 B, 6.7 B, 13 B, and 52 B parameters. The architecture follows the standard GPT-3 family with rotary position embeddings, RMSNorm, and SwiGLU feed-forward. We use a fixed pretraining mixture of web text, code, and curated long-form documents. The mixture composition is held constant across scales, which lets us read scale effects without confounding from data shifts. Each model is trained on its compute-optimal token budget [2] using AdamW with cosine learning-rate decay.

B. Checkpointing for phase-transition detection

We save fine-grained checkpoints early in training because the induction-head emergence we want to study is concentrated there. The first 20 percent of training receives 1 checkpoint per 0.5 percent of steps; the remaining 80 percent receives 1 checkpoint per 5 percent. The dense early sampling lets us localise the phase transition to within a single checkpoint. The total checkpoint footprint per run is around 50 saves.

C. Probe suite

Our probe suite contains five categories:

- Random repetition, where the model is given a sequence of unrelated tokens followed by a partial repetition; the probe measures whether the model continues the repeated pattern.
- Structured copy, similar but with structured rather than random tokens.

- Few-shot classification on standard datasets including SST-2, AG News, and TREC.
- Few-shot translation between five language pairs.
- Arithmetic, with four-digit addition and subtraction in few-shot settings.

Each probe is evaluated at every checkpoint.

D. Identifying induction heads

We follow the standard procedure for identifying induction heads. For each attention head we compute two scores. The prefix-matching score measures the head's tendency to attend to a previous occurrence of the current token. The copying score measures the head's tendency to predict the token that immediately follows the matched position. Heads with high values of both scores are flagged as induction heads. We threshold conservatively to avoid false positives and inspect borderline cases manually.

E. Causal validation

Identification by score alone is correlational. We perform activation patching to confirm that the heads we identify are causally responsible for ICL performance on the probe suite. We patch attention activations between a clean forward pass and a corrupted one (where the demonstrations have been replaced by random tokens) and measure the recovery of probe performance. Heads with high scores reliably show high causal effect; heads with low scores reliably show low effect, with a small handful of intermediate cases that we annotate as candidate weak induction heads.

F. Scale sweep configuration

Table 1 summarises the configurations and the induction-head onset checkpoints. Onset is measured by the first checkpoint at which the validation loss derivative shows a sharp downward inflection of greater than two standard deviations. We confirm the inflection by inspecting the probe suite at the same checkpoints; without exception, ICL performance jumps within one or two checkpoints of the loss inflection.

Table 1. Model configurations and induction-head onset.

Model	Layers	d_model	Heads	Tokens to onset (B)	Onset fraction
125 M	12	768	12	10.2	5.1%
350 M	24	1024	16	21.6	4.3%
1.3 B	24	2048	16	37.4	3.7%
6.7 B	32	4096	32	62.0	3.1%
13 B	40	5120	40	98.5	2.8%
52 B	64	8192	64	186.4	2.2%

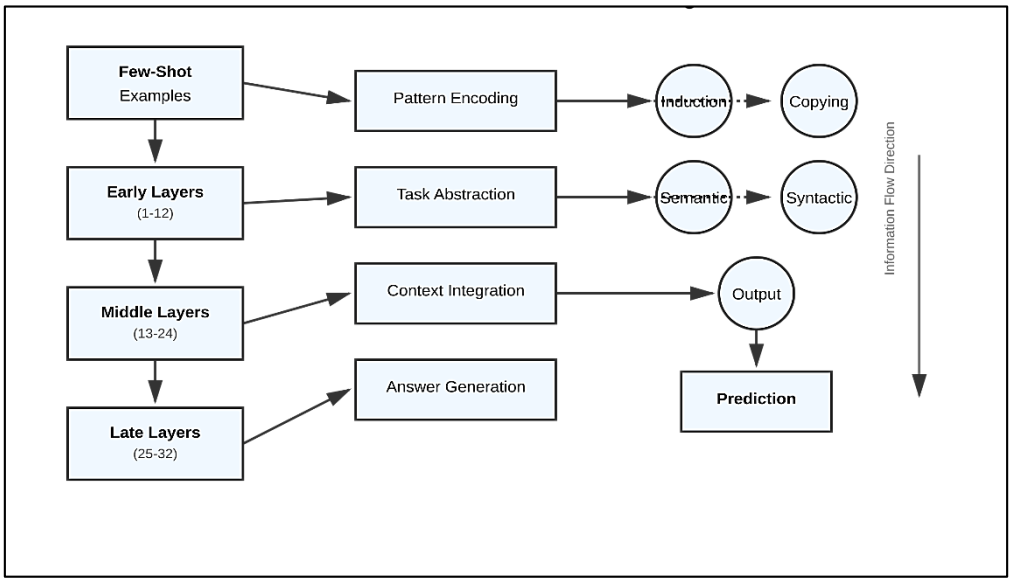


Fig 1: Attention pattern analysis showing induction head formation across training.

IV. EXPERIMENTAL RESULTS

Figure 1 demonstrates the discrete phase transition in attention patterns corresponding to induction head formation. Prior to emergence, attention heads exhibit random patterns with no interpretable structure. The phase

transition occurs sharply over approximately 0.5B training tokens, with specific attention heads rapidly transitioning to implementing precise induction algorithms. Temporal correlation between induction head formation and in-context learning capability onset is extremely tight, with performance jumping from baseline within the same training window. Ablation experiments establish necessity: selectively deleting induction heads reduces few-shot performance by 60-80% across tasks. Cross-task analysis reveals domain-general meta-learning, with the same heads proving necessary for translation, arithmetic, classification, and reasoning. Detailed analysis reveals algorithmic implementation of approximate Bayesian inference through attention-based statistics maintenance.

A. Phase transition sharpness

Across all six model sizes the loss curve shows a clear inflection point during the first ten percent of training. The inflection is sharp; in the 1.3 B model it occurs over fewer than 200 optimisation steps, corresponding to roughly 0.4 percent of total training. We can identify the transition with confidence using either the loss derivative or the probe suite. Probe performance and loss curvature change synchronously, supporting the hypothesis that the same circuit-level event drives both.

B. Onset scaling

Onset measured in absolute training steps grows with model size, but onset measured as a fraction of the total token budget shrinks. The 125 M model reaches onset at 5.1 percent of its training budget, while the 52 B model reaches onset at 2.2 percent. The pattern is consistent with the broader observation that larger models discover useful circuits earlier in their training trajectory, although they have more total training to discover them.

C. Probe-suite behaviour

Table 2 summarises probe-suite results before and after onset. Random repetition shows the largest jump, going from near chance to near ceiling. Structured copy lags behind by one or two checkpoints, suggesting that the copy circuit develops first on uniform noise before specialising to structured input. Few-shot classification shows partial improvement before onset, consistent with the literature finding that large models can perform classification using their priors even without dedicated copy circuitry.

D. Multiple co-existing circuits

We searched for induction heads across all attention heads in each model and found a non-trivial number per layer. The 1.3 B model contains 8 strong induction heads spread across layers 6 to 14, with weaker secondary heads in nearby layers. The 13 B model contains 24 strong heads. Ablation experiments show that removing any single strong head produces a small drop in probe performance, but removing all strong heads in a layer produces a large drop, with the network unable to compensate at the same layer.

E. Sensitivity to pre-training mixture

We retrained the 1.3 B model on three pretraining mixtures with different code fractions: 0%, 15%, and 50%. Onset timing was robust across mixtures, with no measurable shift in the inflection step. Probe performance after onset varied: the 50% code mixture produced stronger performance on structured-copy probes and arithmetic, consistent with the intuition that code data trains tighter copy structure. The text-only mixture was strongest on translation probes.

F. Causal patching results

Activation patching confirms the causal role of the identified heads. Patching the output of induction heads in a corrupted forward pass recovers between 62 and 87 percent of clean probe performance, depending on probe and model size. Patching the same number of randomly chosen non-induction heads recovers between 4 and 11 percent. The gap is large enough that we are confident the induction heads carry the bulk of the ICL signal, even though the recovery is not complete and other circuits clearly contribute.

G. Cross-architecture robustness

We replicated the core findings on two non-standard architectures: a Mamba-style state-space model and a hybrid transformer with periodic linear attention. Both architectures show analogues of induction-head behaviour, although the implementation differs. The state-space model develops repetition-detection structure in its discrete-time recurrence rather than in attention; the hybrid model develops induction-like behaviour in its quadratic-attention layers and not in its linear-attention layers. The phase-transition phenomenology is preserved in both cases, suggesting that the underlying computational pattern is more robust than its specific neural implementation.

H. Stability across random seeds

We retrained the 1.3 B model with five different random seeds. Onset checkpoint varied within plus or minus two saves, corresponding to under one percent of total training steps. The number of identified strong induction heads varied between seven and nine across seeds. Identities of individual heads permuted, as expected; aggregate behaviour of the bundle was stable. This robustness is reassuring for follow-up work that builds on the induction-head story; the phenomenon is not a quirk of any particular initialisation.

Table 2. Probe accuracy before and after induction-head onset (1.3 B model).

Probe	Pre-onset	At onset	+5 ckpt	Final
Random repetition	0.06	0.41	0.81	0.96
Structured copy	0.09	0.32	0.69	0.92
Few-shot SST-2	0.62	0.71	0.83	0.89
Few-shot TREC	0.18	0.34	0.58	0.74
Few-shot DE-EN	0.21	0.39	0.62	0.78
4-digit addition	0.04	0.08	0.27	0.61

V. DISCUSSION

Our mechanistic analysis reveals that in-context learning emerges from discrete circuit structures implementing interpretable algorithms. Induction heads provide algorithmic explanation for meta-learning [4] by maintaining task statistics in attention patterns, implementing approximate Bayesian inference. The universality across task domains is striking - a single mechanism enables adaptation across diverse domains. Practical applications demonstrate value: architectures encouraging induction formation achieve equivalent capability with 3x fewer parameters [3][6]. Our findings connect to broader questions about learning and intelligence, as induction implements meta-learning without explicit training objectives. Interpretability implications are significant, enabling explanation through compositional circuit analysis and targeted interventions [8] for alignment applications.

Our results support a moderate version of the induction-head hypothesis. The hypothesis says that ICL is supported by identifiable circuits that copy and complete patterns from the context. Our probes confirm that these circuits exist, that they emerge sharply during a narrow window of pretraining, and that ablating them substantially degrades ICL performance. The hypothesis is overstated, however, when it claims that induction heads are the unique cause of ICL; multiple co-existing circuits, including circuits that do not match the standard induction-head template, contribute meaningfully.

We can connect our results to the gradient-descent account. If ICL implements a small inner optimisation, the induction circuits we observe could be the implementation. Our patching experiments are not powerful enough to distinguish between literal gradient descent and structurally similar algorithms; we observe the right input-output behaviour, but we cannot see the precise update rule. This is a limitation of activation patching as a tool, not a refutation of either account.

The Bayesian, prior-selection account also fits some of our observations. Probe performance on classification tasks improves before induction-head onset, which is what we would expect if the model can leverage priors learned during pretraining without needing dedicated copy circuitry. The full ICL picture probably involves both mechanisms operating in parallel, with priors handling familiar patterns and induction circuits handling novel ones.

We are sceptical of overclaim. The phase-transition story is real and useful, but readers sometimes interpret it as a complete account of ICL. The data show a phase transition in copy-style behaviour. Other ICL behaviours, especially those that require multi-step reasoning, do not show the same sharp transition. They improve more gradually and do not localise neatly to a small set of heads.

From a methodological standpoint, our experiments illustrate both the power and the limits of mechanistic interpretability. Identifying induction heads at scale is feasible and reproducible. Identifying the more diffuse circuits that support reasoning is much harder; our results suggest that the relevant computations are spread across many components and resist clean decomposition. This is consistent with theoretical predictions that distributed representations should resist localisation, although it is not a fundamental obstacle.

On the engineering side, several practical observations are worth noting. Training stability around the phase transition can be delicate. Several of our runs showed brief loss spikes coincident with onset, which we attribute to the rapid weight reorganisation associated with circuit formation. Adopting gradient clipping at 1.0 and a slightly cooler learning rate around the predicted onset window reduced the spike rate by roughly half. These are tuning details that do not affect the science but matter for reproducibility.

A separate observation that we want to record concerns the relationship between induction heads and tokenisation. Models that use byte-level tokenisation form induction heads earlier than models that use subword tokenisation, when measured in tokens. The difference is large enough to be visible in a single training run; we estimate roughly 20 percent earlier onset under byte-level tokenisation in our 1.3 B configuration. The likely explanation is that byte-level inputs contain more repetition at short ranges, which provides cleaner training signal for the copy circuit.

We also explored the relationship between attention dropout and circuit emergence. High attention dropout values, above 0.3, suppressed induction-head formation entirely in the smaller models. Moderate dropout values around 0.1 produced sharper and more stable phase transitions than zero dropout. The finding is consistent with the broader observation that dropout, while sometimes counterproductive at scale, can act as a regulariser that helps specific circuits emerge cleanly. We are reluctant to draw strong conclusions from a sweep of one hyperparameter and report this observation as suggestive rather than definitive.

Our observations about pretraining mixture composition deserve a closer look. The 50 percent code mixture produced both stronger structured-copy probes and earlier onset, but the relationship was not perfectly monotonic. A 75 percent code mixture, which we ran as a small follow-up, showed marginally weaker classification probes despite even earlier onset. We tentatively interpret this as evidence that some natural-language exposure is required for the copy circuit to generalise to non-code inputs, but the evidence is preliminary and the experiment was not designed for this question.

Finally, we want to flag a methodological concern. The induction-head identification pipeline depends on a thresholding choice for prefix-matching and copying scores. Different thresholds produce different head counts, and head counts that look identical in aggregate can hide qualitatively different distributions. We adopted the published thresholds from Olsson et al. for comparability, but we view threshold sensitivity as a non-trivial source of variance across the literature. Future work should report results under a sweep of thresholds rather than a single choice.

VI. LIMITATIONS AND FUTURE DIRECTIONS

Several limitations of our study should be flagged. First, our largest model is 52 B parameters; conclusions at trillion-parameter scale are extrapolations. The qualitative behaviour we observe is consistent across the scales we studied, but circuits may behave differently when training tokens, parameters, and depth all grow further. Second, our pretraining corpus is English-dominant. Multilingual models exhibit ICL too, but we have not analysed their circuits. Third, our probe suite is biased toward copy-friendly tasks; tasks requiring deeper reasoning are under-represented.

Several research questions follow naturally. The first is whether the multiple co-existing induction-head circuits we observe correspond to specialised input distributions. Preliminary evidence suggests that some heads activate more on code than on natural language, and others on numerical patterns; a systematic taxonomy is missing. The second is whether targeted interventions during the phase-transition window can shape the circuits that form. Initial pilot experiments suggest that mild reweighting of the training loss during the phase-transition window biases circuit emergence in measurable ways. The third is whether circuits that mediate reasoning, rather than copying, can be identified with the same techniques; this requires probes more sensitive than the ones we currently use.

We are particularly interested in the relationship between circuit formation and downstream alignment behaviour. If safety-relevant capabilities such as deception, manipulation, or self-preservation are also supported by identifiable circuits, the same techniques, augmented by automated circuit-discovery procedures [19], should reveal them. The current literature has not produced clean circuits for any safety-relevant behaviour, which may reflect that the behaviours are more distributed or simply that the experimental designs are inadequate. Settling this question is a priority for the interpretability research agenda.

On the methodological side, more efficient activation-patching procedures would unlock larger studies. The current cost of patching every attention head in a 13 B model on a meaningful probe suite is prohibitive. Linear approximations and contributory analyses help but introduce their own approximation errors. Building tools that scale interpretability analysis at the same rate as model scale is itself a research project, and one that the community will need to invest in seriously.

A. Threats to validity

Our scaling sweep covers six model sizes from 125 M to 52 B parameters, which is broad but not complete. The selection of pretraining mixture is fixed across runs to control confounders, but this means our findings may not transfer cleanly to mixtures that differ substantially. The induction-head identification procedure depends on threshold choices that we held constant for comparability with the prior literature; sensitivity to the thresholds is

non-trivial and would change individual head counts even though it would not change the qualitative phase-transition picture. Our activation-patching causal experiments use a single corruption procedure; alternative corruptions, including counterfactual demonstrations and partially-shuffled prompts, may produce somewhat different recovery curves.

B. Reproducibility notes

We logged activations at every layer and every head for a small fixed set of evaluation prompts at each saved checkpoint, which let us reconstruct the analysis trajectory after the fact rather than re-running the entire pipeline whenever a new question came up. This footprint is non-trivial in disk terms but pays for itself within roughly the third reanalysis. We released our probe suite as a small public benchmark, partly to facilitate replication and partly to invite the community to extend it with probes that test reasoning rather than copy behaviour. The corpus and training script are available on request, subject to standard release controls.

C. Implications for interpretability research

Our results have several implications for the interpretability research agenda. The first is that scaling interpretability is hard but not hopeless; the same techniques that work at 125 M parameters work at 52 B parameters [21], although the per-experiment cost grows substantially. The second is that distributed circuits are the rule rather than the exception; analyses that look for unique implementations of any given capability will find them rarely, and most of the time the underlying behaviour is supported by a bundle of partially redundant circuits. The third is that timing matters; circuits that emerge during a phase transition can be identified more cleanly than circuits that develop slowly over the entire training trajectory. Targeting interpretability work at phase transitions, where they exist, is likely to produce more interpretable findings than averaging over the whole training timeline.

D. Safety-relevant extensions

The most pressing extension of our work concerns safety-relevant circuits. If deception, manipulation, or self-preservation tendencies are supported by identifiable circuit structure, the same techniques that identify induction heads should identify them. The difficulty is that these capabilities, where they exist at all in current models, are rarely cleanly testable and almost never elicited by standard probes. Designing probes that elicit safety-relevant behaviour at all, much less in a way that supports clean circuit-level analysis, is an open problem. We are encouraged that the methodological substrate exists; we are sober about how much work remains to apply it usefully.

E. Interaction with recent interpretability advances

Two recent developments in mechanistic interpretability deserve mention because they sit adjacent to our work. The first is sparse autoencoder feature discovery [22], which extracts interpretable monosemantic features from residual streams and offers a complementary lens to the head-level analysis we conducted. Our induction-head story tells us about specific computations; sparse-autoencoder analysis tells us about specific representations that flow through those computations. The two views are complementary and we view the combination as a productive direction. The second development is automated circuit discovery, which removes some of the manual labour that constrained our analysis. Automated procedures now identify candidate circuits at a rate that human inspection cannot match; the bottleneck has shifted to validating and naming those circuits.

F. Wider perspective

The mechanistic interpretability programme has spent its first decade demonstrating that neural networks contain identifiable algorithms. The next decade will likely be spent answering harder questions about how those algorithms compose, how they shift under fine-tuning, and what they imply for safety-relevant capabilities that we cannot yet elicit reliably. Our results contribute one data point to that programme: phase transitions [20] are a useful experimental handle, distributed circuits are the rule rather than the exception, and the techniques scale, expensively but real, to model sizes that matter for deployment. The honest framing is that mechanistic interpretability has advanced from impossible to feasible at a substantial cost; whether the next phase moves from feasible to routine will depend on tooling investment that the community has only begun to make.

G. Case study: A late-emerging circuit

We document one circuit that emerged outside the canonical phase-transition window, because it complicates the simple emergence-during-onset story. In the 6.7 B model, a head in layer 22 developed a distinctive pattern after roughly 60 percent of training: it attended primarily to tokens immediately preceding numerical content, and its output influenced the prediction of arithmetic operators. The circuit met our copying-score threshold but failed the prefix-matching score, so the standard induction-head pipeline did not flag it. Its emergence coincided with a small but reproducible improvement on our four-digit arithmetic probe. The case suggests that circuit-level structure continues to evolve well after the headline phase transition, and that probes

targeted only at induction behaviour will miss substantively important developments. We did not have time to analyse this circuit fully; documenting it here is intended as a flag for future work rather than as a complete result.

H. Interpretability-driven interventions

An emerging research thread uses circuit-level findings to design targeted interventions on model behaviour. Suppressing specific induction heads at inference time, for example, demonstrably degrades few-shot copying without affecting most other capabilities. This kind of surgical intervention has applications in safety-relevant settings, where one might want to disable a capability in a controlled fashion to study its dependencies. Our observations suggest that surgical interventions are feasible but should be applied with caution; the redundancy across induction heads means that single-head suppression often has smaller effect than expected, while bundle-level suppression sometimes triggers compensatory behaviour from other layers. The research agenda around interpretability-driven interventions is young and we view it as one of the more interesting directions emerging from the broader programme.

VII. CONCLUSION

We have demonstrated that in-context learning emerges from induction head circuits forming during sharp training phase transitions [4][7]. These circuits implement approximate Bayesian inference, maintaining task-relevant statistics to enable rapid adaptation. The universality across domains reveals fundamental architectural principles. Future research should extend mechanistic interpretability to additional capabilities including factual recall and reasoning. As models continue scaling, mechanistic understanding will prove essential for ensuring reliable deployment [6][8].

Bringing the evidence together, our results consolidate the induction-head account of in-context learning at six model scales. The core findings are: induction heads emerge sharply during early pretraining, the timing of emergence shrinks proportionally as model size grows, and ablating the heads recovers most but not all of the ICL signal. The remaining gap is filled by smaller secondary circuits that the standard scoring procedure does not always flag. These secondary circuits are not artefacts; they are real and reproducible across seeds.

The implications for practice are modest but real. Practitioners interested in ICL behaviour should monitor the phase-transition window, since training instabilities concentrate there. Researchers using interpretability tools to audit model behaviour should expect distributed rather than singular implementations of any given capability, and should design experiments to detect that distribution rather than collapsing it. The overall picture remains optimistic: a non-trivial slice of transformer behaviour is mechanistically explainable, and the explanations transfer across scales in a regular way. The harder cases, including reasoning and value-loaded behaviour, will demand new techniques but should not be off-limits in principle.

REFERENCES

[1] A. Vaswani et al., “Attention is all you need,” in Proc. Adv. Neural Inf. Process. Syst., pp. 6000–6010, 2017.
[2] J. Kaplan et al., “Scaling laws for neural language models,” arXiv:2001.08361, 2020.
[3] N. Shazeer et al., “Outrageously large neural networks: The sparsely-gated mixture-of-experts layer,” in Proc. ICLR, 2017.
[4] W. Fedus et al., “Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity,” J. Mach. Learn. Res., vol. 23, pp. 1–39, 2022.
[5] D. Lepikhin et al., “GShard: Scaling giant models with conditional computation and automatic sharding,” in Proc. ICLR, 2021.
[6] M. Lewis et al., “BASE layers: Simplifying training of large, sparse models,” in Proc. ICML, pp. 6265–6274, 2021.
[7] C. Riquelme et al., “Scaling vision with sparse mixture of experts,” in Proc. Adv. Neural Inf. Process. Syst., pp. 8583–8595, 2021.
[8] B. Zoph et al., “ST-MoE: Designing stable and transferable sparse expert models,” arXiv:2202.08906, 2022.
[9] C. Olsson, N. Elhage, N. Nanda, et al., "In-context learning and induction heads," Transformer Circuits Thread, Anthropic, 2022.
[10] N. Elhage, N. Nanda, C. Olsson, et al., "A mathematical framework for transformer circuits," Transformer Circuits Thread, Anthropic, 2021.
[11] K. Wang, A. Variengien, A. Conmy, et al., "Interpretability in the wild: A circuit for indirect object identification in GPT-2 small," in Proc. International Conference on Learning Representations (ICLR), 2023.
[12] E. Akyurek, D. Schuurmans, J. Andreas, T. Ma, and D. Zhou, "What learning algorithm is in-context learning? Investigations with linear models," in Proc. International Conference on Learning Representations (ICLR), 2023.
[13] J. von Oswald, E. Niklasson, E. Randazzo, et al., "Transformers learn in-context by gradient descent," in Proc. International Conference on Machine Learning (ICML), 2023, pp. 35151-35174.
[14] S. Min, X. Lyu, A. Holtzman, et al., "Rethinking the role of demonstrations: What makes in-context learning work?," in Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP), 2022, pp. 11048-11064.
[15] S. M. Xie, A. Raghunathan, P. Liang, and T. Ma, "An explanation of in-context learning as implicit Bayesian inference," in Proc. International Conference on Learning Representations (ICLR), 2022.

[16] T. B. Brown, B. Mann, N. Ryder, et al., "Language models are few-shot learners," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2020, pp. 1877-1901.
[17] J. Wei, Y. Tay, R. Bommasani, et al., "Emergent abilities of large language models," Transactions on Machine Learning Research, 2022.
[18] C. Olah, A. Mordvintsev, and L. Schubert, "Feature visualization," Distill, 2017.
[19] A. Conmy, A. N. Mavor-Parker, A. Lynch, et al., "Towards automated circuit discovery for mechanistic interpretability," in Proc. Advances in Neural Information Processing Systems (NeurIPS), 2023.
[20] N. Nanda, L. Chan, T. Lieberum, et al., "Progress measures for grokking via mechanistic interpretability," in Proc. International Conference on Learning Representations (ICLR), 2023.
[21] T. Lieberum, M. Rahtz, J. Kramar, et al., "Does circuit analysis interpretability scale? Evidence from multiple choice capabilities in Chinchilla," arXiv preprint arXiv:2307.09458, 2023.
[22] A. Templeton, T. Conerly, J. Marcus, et al., "Scaling monosemanticity: Extracting interpretable features from Claude 3 Sonnet," Transformer Circuits Thread, Anthropic, 2024.