IMPLICIT STATISTICAL INFERENCE IN TRANSFORMERS: APPROXIMATING LIKELIHOOD-RATIO TESTS IN-CONTEXT

Faris Chaudhry & Siddhant Gadkari
Department of Computer Science
Imperial College London
{fc522,svg21}@imperial.ac.uk

ABSTRACT

In-context learning (ICL) allows Transformers to adapt to novel tasks without weight updates, yet the underlying algorithms remain poorly understood. We adopt a statistical decision-theoretic perspective by investigating simple binary hypothesis testing, where the optimal policy is determined by the likelihood-ratio test. Notably, this setup provides a mathematically rigorous setting for mechanistic interpretability where the target algorithmic ground truth is known. By training Transformers on tasks requiring distinct geometries (linear shifted means vs. nonlinear variance estimation), we demonstrate that the models approximate the Bayes-optimal sufficient statistics from context up to some monotonic transformation, matching the performance of an ideal oracle estimator in nonlinear regimes. Leveraging this analytical ground truth, mechanistic analysis via logit lens and circuit alignment suggests that the model does not rely on a fixed kernel smoothing heuristic. Instead, it appears to adapt the point at which decisions become linearly decodable: exhibiting patterns consistent with a voting-style ensemble for linear tasks while utilizing a deeper sequential computation for nonlinear tasks. These findings suggest that ICL emerges from the construction of task-adaptive statistical estimators rather than simple similarity matching.

1 INTRODUCTION

In-context learning (ICL) refers to the remarkable ability of models (particularly Transformers) to adapt to novel tasks at inference time using only a finite context of input-output examples, without explicit parameter updates (Brown et al., 2020; Vaswani et al., 2023). While ICL is now a standard capability of large language models, its underlying algorithmic mechanism remains a subject of debate. Does the model merely retrieve and average similar examples, or does it construct a principled learning algorithm on the fly?

Recent work in controlled synthetic environments has demonstrated that Transformers can recover classical algorithms (such as linear regression, decision trees, and automata) purely from context (Garg et al., 2023; Zhang et al., 2023). These findings suggest that ICL may implement statistically optimal procedures when the task structure allows. However, existing analyses often focus on regression problems with fixed functional forms, emphasizing asymptotic convergence rather than the precise nature of the decision rule applied at the level of individual episodes.

In this work, we adopt a statistical decision-theoretic perspective. We study ICL in binary hypothesis testing, a fundamental framework where optimal decision rules are fully characterized by the Neyman-Pearson lemma (Lehmann & Romano, 2005). For simple hypotheses, the log-likelihood ratio (LLR) is a minimal sufficient statistic, and any Bayes-optimal decision rule must be a monotone function of it. This provides a sharp notion of optimality and identifiability: recovering the LLR up to a monotone (or affine) transformation is both necessary and sufficient for optimal prediction. More importantly, *this establishes a testbed for mechanistic interpretability where the ground truth is known*, addressing a known challenge in mechanistic interpretability (Sharkey et al., 2025).

By training Transformers on dynamic discrimination tasks where the optimal statistic varies across episodes (e.g., linear vs. quadratic), we test whether the model learns to infer and apply the appropriate sufficient statistic from context alone, rather than relying on fixed similarity heuristics. We view this work as a first step toward a broader decision-theoretic understanding of ICL.

1.1 Related Work

ICL as implicit inference. A growing body of literature interprets ICL as a form of implicit Bayesian inference. Xie et al. (2022) propose that ICL can be modeled as Bayesian inference over a hidden variable concept space, while Li et al. (2023) and Zhang et al. (2023) demonstrate that Transformers can approximate posterior predictive distributions for specific function classes. Closest to our work, Bai et al. (2023) analyze Transformers as statisticians in the context of Markov chains, finding that they can approach Bayes-optimal error rates. We extend this perspective by explicitly characterizing the geometry of the decision boundary (linear vs. quadratic) and linking the model's internal representations to the Neyman-Pearson optimal statistic.

Algorithmic induction and optimization. An alternative perspective views ICL as an optimization process., Akyürek et al. (2023), Dai et al. (2023), and von Oswald et al. (2023) have argued that self-attention layers can implement steps of gradient descent (GD) during the forward pass. While the "ICL as GD" hypothesis explains how models improve with more examples, it does not explicitly guarantee statistical optimality in discriminative settings. Our work complements this by focusing on the objective of the induced algorithm: regardless of whether the mechanism resembles GD or exact inference, we ask if it produces the sufficient statistic required for the likelihood-ratio test.

Mechanistic interpretability and task vectors. Finally, our analysis draws on mechanistic interpretability to explain how these statistics are computed (Elhage et al., 2021; Nanda et al., 2023). Olsson et al. (2022) identified induction heads as a primary circuit for copying patterns in ICL. More recently, Hendel et al. (2023) and Todd et al. (2024) have proposed that Transformers compress the context into function vectors or task vectors that modulate downstream processing. This aligns with our finding that the attention mechanism acts as a "neural statistician" of sorts (Edwards & Storkey, 2017), compressing the context dataset into a single sufficient statistic (e.g., a mean vector or energy scalar) that determines the downstream decision rule.

2 Problem Setup: Dynamic Statistical Discrimination

We study ICL in the setting of binary hypothesis testing with task parameters that vary across episodes (i.e., independent task instances consisting of a context set and a query drawn from a shared latent task). Let Φ denote a family of binary classification tasks, where each task $\phi \in \Phi$ specifies two class-conditional distributions $(p_\phi(x \mid H_0), p_\phi(x \mid H_1))$ and an associated label space $y \in \{0, 1\}$. In each episode, we sample task parameters $\phi \sim p(\Phi)$ and generate a context dataset $C = \{(x_i, y_i)\}_{i=1}^N$ where $y_i \sim \text{Bernoulli}(1/2)$ and $x_i \sim p_\phi(x \mid H_{y_i})$. A query point (x_q, y_q) is drawn from the same task distribution. A Transformer model f_θ is trained to predict the label (source distribution) y_q given (x_q, C) by minimizing the binary cross-entropy (BCE) loss:

$$\mathcal{L} = -\mathbb{E}_{\phi \sim p(\Phi)} \mathbb{E}_{C, x_q} \left[y_q \log f_\theta(x_q, C) + (1 - y_q) \log(1 - f_\theta(x_q, C)) \right]. \tag{1}$$

Minimizing BCE is equivalent to estimating the posterior probability $p(y_q = 1 \mid x_q, C)$. The logit of the Bayes-optimal predictor satisfies

$$\log \frac{p(y_q = 1 \mid x_q, C)}{p(y_q = 0 \mid x_q, C)} = \text{LLR}(x_q; \phi) + \log \frac{\pi_1}{\pi_0}, \tag{2}$$

where π_1, π_0 denote the class priors. Thus, under BCE training, the Bayes-optimal internal decision statistic is identifiable up to an affine transformation of the LLR.

Conditioned on the context dataset C, each episode induces a simple binary hypothesis testing problem between H_0 and H_1. By the Neyman-Pearson lemma, the likelihood-ratio test

$$\frac{p(x_q \mid H_1, C)}{p(x_q \mid H_0, C)}$$

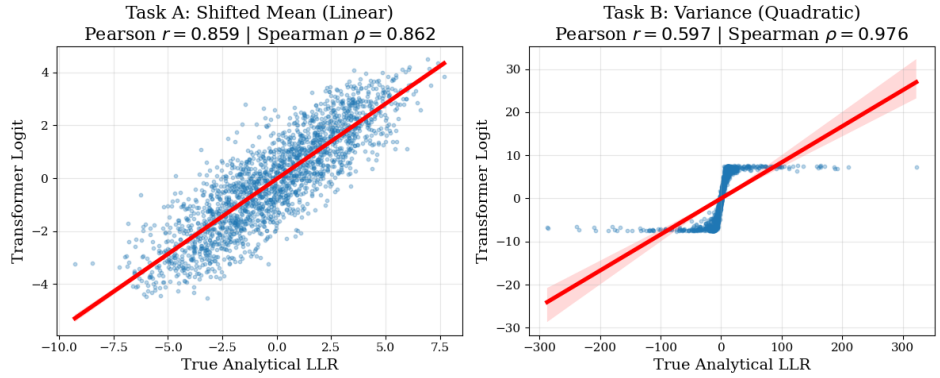

Figure 1: **Approximation of the LLR.** Regression of the Transformer's output logits against the true analytical LLR for validation episodes. **(Left) Task A:** The model exhibits a strong linear correlation ($r = 0.859$), indicating it approximates the affine sufficient statistic $\mu^\top(x - k)$. **(Right) Task B:** The model achieves near-perfect rank correlation ($\rho = 0.976$), effectively recovering the quadratic sufficient statistic $\|x\|^2$ up to a monotone transform. The sigmoidal shape suggests the model has learned a calibrated probability mapping, saturating for high-confidence inputs while preserving the optimal decision ordering.

is the uniformly most powerful decision rule, and any Bayes-optimal classifier must implement a statistic that is monotone in the corresponding log-likelihood ratio. Consequently, recovery of the LLR up to an affine transformation is both necessary and sufficient for optimal in-context prediction under BCE training.

To test whether Transformers rely on simple heuristics or perform optimal, context-dependent statistical inference, we design two Gaussian discrimination tasks with differing optimal statistics.

Task A: Shifted Mean Discrimination (Linear Regime). We sample a discriminative direction $\mu \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and a shift $k \sim \mathcal{N}(0, \sigma_k^2 I)$. The class-conditional distributions are

$$H_0 : x \sim \mathcal{N}(-\mu + k, I), \quad H_1 : x \sim \mathcal{N}(\mu + k, I). \tag{3}$$

The optimal decision boundary is linear but not centered at the origin. The sufficient statistic is the shifted projection $S(x) = \mu^\top(x - k)$, requiring the model to infer both μ and k from the context. This task probes whether the model can dynamically estimate local centroids and perform linear discrimination. Static models that assume fixed centering fail on this task.

Task B: Variance Discrimination (Nonlinear Regime). We sample two variances $\sigma_0, \sigma_1 \sim \mathrm{Unif}[0.5, 3.0]$ and fix the mean at zero. The distributions are

$$H_0 : x \sim \mathcal{N}(0, \sigma_0^2 I), \quad H_1 : x \sim \mathcal{N}(0, \sigma_1^2 I). \tag{4}$$

Since the class means coincide, dot-product similarity is uninformative. The optimal decision statistic depends on the quadratic energy $\|x\|^2$, with the sign determined by the relative ordering of (σ_0, σ_1). This task tests whether the model can adapt its internal geometry from linear projections to norm-based estimation.

3 Approximation of the LLR

3.1 Recovery of Optimality

To quantify the model's ability to recover the sufficient statistic, we compare its in-context accuracy against a theoretical Bayes-optimal classifier. The oracle computes the exact log-likelihood ratio using the ground-truth task parameters (μ, k, σ), representing the theoretical performance ceiling.

In the nonlinear variance task (Task B), the model achieves an accuracy of $83.0 \pm 0.5\%$, effectively matching the oracle performance of $84.0 \pm 1.0\%$. While the model's raw logits do not linearly

track the analytical LLR (Pearson $r = 0.60$), they achieve near-perfect rank alignment (Spearman $\rho = 0.98$). This indicates that the model has successfully recovered the ordering induced by the quadratic sufficient statistic $\|x\|^2$, but maps it through a nonlinear calibration function (Figure 1).

In the linear shifted mean task (Task A), the model achieves $78.3 \pm 0.3\%$. While discriminative, it remains below the oracle accuracy of $84.6 \pm 1.0\%$, leaving an optimality gap of approximately 6.3%. This discrepancy is reflected in the regression analysis, which shows a noisy linear approximation ($r = 0.86$) rather than the clean functional relationship observed in Task B. This suggests that instead of performing exact symbolic inference, the model implements some approximation. We verify this hypothesis in Appendix C.1 by evaluating the model on OOD tasks with significantly larger nuisance shifts ($\sigma_k = 9.0$). Under these conditions, the correlation with the true LLR degrades to $r = 0.567$, demonstrating that the learned decision rule is a local approximation calibrated to the training support rather than an exact symbolic recovery. Nonetheless, the model does eventually begin to generalize OOD, exhibiting a delayed rise in validation accuracy characteristic of partial grokking.

3.2 Ablations and Failure Modes

We isolate the necessary components for in-context learning by modifying the architecture and data structure, as detailed in Table 1. Comprehensive results for all experimental conditions are provided in Appendix C.2.

Table 1: **Key Ablations (Task A).** We test the necessity of specific architectural features. **1) Permutation Invariance:** Removing positional encodings (`NoPos`) has negligible impact, confirming the model treats the context as a set rather than a sequence. **2) Learned Metric:** Freezing attention weights (`FrozenQK`) destroys performance, indicating the model must learn a task-specific similarity metric. **3) Supervision:** Shuffling labels (`ShuffledLabels`) causes collapse to random chance, ruling out unsupervised clustering heuristics.

Model Variant	Validation Accuracy	Implication
`Regular (Baseline)`	$78.3 \pm 0.3\%$	—
`NoPos`	$78.2 \pm 0.5\%$	Permutation Invariant
`ShuffledLabels`	$49.6 \pm 1.2\%$	Requires $x \to y$ mapping
`FrozenQK`	$49.6 \pm 1.3\%$	Requires Learned Metric

4 Mechanistic Evidence

We now investigate how the model implements these statistical decision boundaries. Our analysis reveals that the model does not use a universal algorithm, but adapts its circuit depth to the task geometry.

First, a common hypothesis is that ICL performs nearest-neighbor smoothing (Han et al., 2025). To test this, we compared the model's logits against a Nadaraya-Watson kernel regression estimator. The correlation is weak, confirming that the model is not merely averaging labels based on similarity, but computing a context-dependent sufficient statistic (e.g., centering by k). More details are provided in Appendix C.3.

4.1 Decision Latency and Logit Lens

Using the Logit Lens technique (nostalgebraist, 2020), we project intermediate residual states into the vocabulary space. As shown in Figure 2 (Left), Task A exhibits an early decoding pattern: the representation at Layer 1 shows a partial but decisive correlation with the final target. This suggests that the model is performing a form of preprocessing or summary statistic calculation early in the network which is then refined into a decision. In contrast, nonlinear tasks (Task B) show near-zero correlation until the final layer, indicating a need for deeper composition to estimate energy terms ($\|x\|^2$). See Appendix C.4.

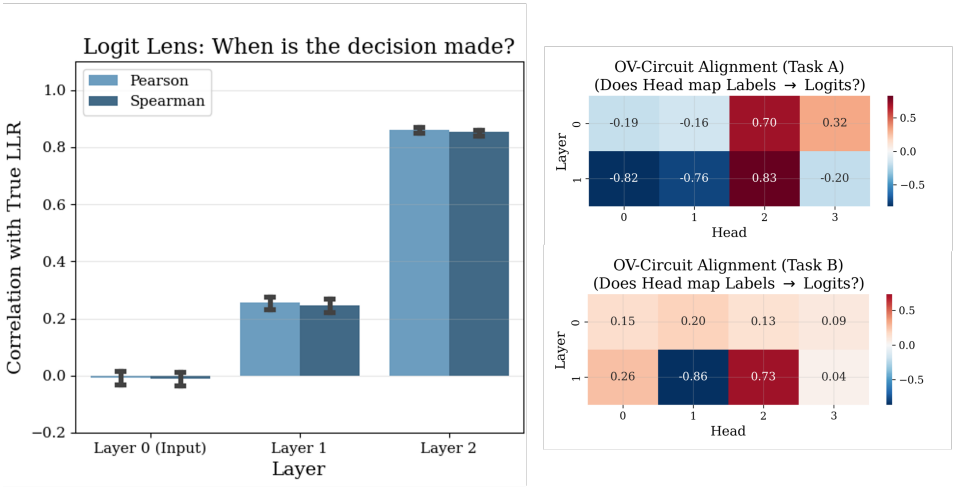

Figure 2: **Mechanistic Adaptivity. (Left) Logit Lens (Task A):** The correlation with the true LLR rises significantly in Layer 1, suggesting early linear decoding or aggregation. **(Right) OV Circuit Alignment:** In Task A (top), Layer 0 heads (e.g., Head 2) show strong positive alignment (> 0.7) with the logit direction, acting as voting ensemble. In Task B (bottom), Layer 0 heads are effectively silent (< 0.26), implying that the model suppresses early voting to perform deeper sequential processing in Layer 1. Both OV circuits are taken from representative seeds; qualitatively similar behavior persisted across seeds.

4.2 Hypothesis: Adaptive Circuit Depth and Voting Ensembles

We find that this decision latency manifests as distinct circuit architectures (Figure 2, Right). To interpret the role of individual attention heads, we analyze their Output-Value (OV) circuits (Elhage et al., 2021). The OV matrix $W_{OV} = W_V W_O$ determines how the features read by a head are projected into the residual stream and, subsequently, the output logits.

In Task A, Layer 0 heads exhibit strong positive alignment ($|\cos\theta| > 0.7$) with the final decision direction. We hypothesize that in this linear regime, the model utilizes a greedy voting ensemble, where heads independently compute partial summary statistics (via forwarding and suppression) that are linearly aggregated to form the decision boundary immediately.

On the other hand, in Task B, Layer 0 heads are effectively silent regarding the decision ($|\cos\theta|$ small). Significant alignment only emerges in Layer 1. This suggests a sequential algorithm where Layer 0 is suppressed or repurposed to compute intermediate features (e.g., squared norms) rather than voting directly.

5 Conclusion, Limitations, and Future Work

Importantly, binary hypothesis testing provides a setting where mechanistic interpretability techniques can be compared to a known ground truth. We have demonstrated that toy Transformers trained on dynamic hypothesis testing tasks can approximate the Neyman-Pearson optimal decision rule in-context. By adapting their internal circuit depth (e.g., employing greedy heuristics for linear tasks and sequential processing for nonlinear boundaries) the models recover a sufficient statistic that is highly monotonically correlated with the LLR, matching the performance of a Bayes-optimal oracle in the quadratic regime.

Limitations. While our controlled synthetic environment allows for exact analytical baselines, it relies on a small two-layer Transformer and relatively low-dimensional Gaussian data. Consequently, it remains an open question to what extent these specific mechanistic behaviors—such as the discrete shift from early voting ensembles to deeper sequential processing—scale to more general statistical tasks or even large language models operating on complex, real-world distributions. Furthermore, our mechanistic interpretability results, including the Logit Lens and OV circuit align-

ment, establish strong correlational evidence rather than strict causal proofs of the model's internal algorithms. Future work incorporating causal interventions could further substantiate these structural hypotheses.

Future work. Firstly, conditioning on the in-context dataset reduces each episode to a simple binary hypothesis test, for which the optimal decision rule is characterized by the likelihood-ratio test. A natural extension is to consider composite hypotheses, where class-conditional distributions depend on latent parameters that cannot be eliminated by conditioning alone. In such settings, optimal decision-making requires either marginalization over nuisance parameters or plug-in estimation. Studying ICL in this regime would help distinguish whether models behave more like Bayesian model averaging or approximate maximum-likelihood estimators.

Secondly, our experiments assume balanced class priors and symmetric loss, leading to decision thresholds centered at zero log-likelihood ratio. Extending the framework to asymmetric priors or cost-sensitive objectives would test whether ICL adapts not only the sufficient statistic but also the optimal decision threshold, as prescribed by statistical decision theory.

Finally, binary hypothesis testing provides a minimal setting with sharp optimality guarantees. Extending the analysis to multi-class or sequential testing problems, such as multi-way likelihood-ratio tests or Wald-style sequential procedures, would probe whether ICL can recover more complex decision rules under uncertainty while retaining decision-theoretic interpretability.

REFERENCES

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models, 2023. URL https://arxiv.org/abs/2211.15661.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection, 2023. URL https://arxiv.org/abs/2306.04637.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL https://arxiv.org/abs/2005.14165.

Damai Dai, Yutao Sun, Li Dong, Yaru Hao, Shuming Ma, Zhifang Sui, and Furu Wei. Why can GPT learn in-context? language models implicitly perform gradient descent as meta-optimizers, 2023. URL https://arxiv.org/abs/2212.10559.

Harrison Edwards and Amos Storkey. Towards a neural statistician, 2017. URL https://arxiv.org/abs/1606.02185.

Nelson Elhage et al. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. URL https://transformer-circuits.pub/2021/framework/index.html.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can Transformers learn in-context? a case study of simple function classes, 2023. URL https://arxiv.org/abs/2208.01066.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. Understanding emergent in-context learning from a kernel regression perspective, 2025. URL https://arxiv.org/abs/2305.12766.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors, 2023. URL https://arxiv.org/abs/2310.15916.

Erich L Lehmann and Joseph P Romano. *Testing Statistical Hypotheses*. Springer, 2005.

Yingcong Li, M. Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning, 2023. URL https://arxiv.org/abs/2301.07067.

E. A. Nadaraya. On estimating regression. *Theory of Probability & Its Applications*, 9(1):141–142, 1964. doi: 10.1137/1109020. URL https://doi.org/10.1137/1109020.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. Progress measures for grokking via mechanistic interpretability, 2023. URL https://arxiv.org/abs/2301.05217.

nostalgebraist. Interpreting GPT: the logit lens. https://www.lesswrong.com/posts/AcKRB8wDpdaN6v6ru/interpreting-gpt-the-logit-lens, Aug 31 2020. LessWrong blog post.

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, 2022. URL https://arxiv.org/abs/2209.11895.

Lee Sharkey, Bilal Chughtai, Joshua Batson, Jack Lindsey, Jeff Wu, Lucius Bushnaq, Nicholas Goldowsky-Dill, Stefan Heimersheim, Alejandro Ortega, Joseph Bloom, Stella Biderman, Adria Garriga-Alonso, Arthur Conmy, Neel Nanda, Jessica Rumbelow, Martin Wattenberg, Nandi Schoots, Joseph Miller, Eric J. Michaud, Stephen Casper, Max Tegmark, William Saunders, David Bau, Eric Todd, Atticus Geiger, Mor Geva, Jesse Hoogland, Daniel Murfet, and Tom McGrath. Open problems in mechanistic interpretability, 2025. URL https://arxiv.org/abs/2501.16496.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models, 2024. URL https://arxiv.org/abs/2310.15213.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. URL https://arxiv.org/abs/1706.03762.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent, 2023. URL https://arxiv.org/abs/2212.07677.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference, 2022. URL https://arxiv.org/abs/2111.02080.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context, 2023. URL https://arxiv.org/abs/2306.09927.

A Derivation of Optimal Test Statistics

For completeness, we derive the analytical LLR for both tasks. Although the marginal problem involves latent task parameters ϕ, conditioning on the context C renders the hypotheses H_0 and H_1 simple for each episode. Classical Neyman-Pearson optimality therefore applies at the episode level, and the optimal decision statistic is given by the likelihood ratio conditioned on C. The following derivations make this dependence explicit for the two task families considered.

A.1 Task A: Shifted Mean Discrimination

Let $\phi = \{\mu, k\}$. The class-conditional distributions are isotropic Gaussians with means $\mu_1 = \mu + k$ and $\mu_0 = -\mu + k$, and covariance $\Sigma = I$. For $x \sim \mathcal{N}(m, I)$,

$$\log p(x \mid m) = -\frac{d}{2}\log(2\pi) - \frac{1}{2}\|x - m\|^2. \tag{5}$$

The LLR is

$$\Lambda(x) = -\frac{1}{2}\|x - (\mu + k)\|^2 + \frac{1}{2}\|x - (-\mu + k)\|^2 \tag{6}$$

$$= 2\mu^\top x - 2\mu^\top k. \tag{7}$$

Thus, the optimal statistic is affine in $\mu^\top(x-k)$; correct classification requires centering with respect to the context-dependent shift.

A.2 Task B: Variance Discrimination

For centered Gaussians with variances σ_1^2 and σ_0^2,

$$\log p(x \mid \sigma) = -\frac{d}{2}\log(2\pi\sigma^2) - \frac{\|x\|^2}{2\sigma^2}. \tag{8}$$

The LLR is

$$\Lambda(x) = \frac{d}{2}\log\frac{\sigma_0^2}{\sigma_1^2} + \frac{\|x\|^2}{2}\left(\frac{1}{\sigma_0^2} - \frac{1}{\sigma_1^2}\right). \tag{9}$$

The first term is a constant bias, while the data-dependent term is proportional to the energy $\|x\|^2$. Hence, the optimal statistic is purely quadratic.

B Experimental Details

Code, results, and figures are available on GitHub.

B.1 Model Architecture

We use a toy Transformer architecture designed for set-to-scalar tasks, which we refer to as `ICLTransformer`.

- **Type:** Bidirectional Transformer Encoder (PyTorch `nn.TransformerEncoder`).
- **Layers:** 2
- **Attention Heads:** 4
- **Embedding Dimension (d_{model}):** 128
- **Feedforward Dimension (d_{ff}):** 512
- **Activation:** GELU
- **Normalization:** Post-LayerNorm (`norm_first=False`)
- **Input Processing:** The input $x \in \mathbb{R}^{16}$ is linearly projected to d_{model}. The binary label $y \in \{0, 1\}$ is projected via a separate learnable linear layer. These two projections are combined via element-wise addition to form the final context token embedding, effectively binding the label information to the input features via superposition.
- **Positional Encodings:** Standard learned absolute positional embeddings are added to the sequence.

B.2 TASK SPECIFICATIONS

Data is generated on-the-fly during training. Each batch consists of $B = 64$ independent episodes.

Task A: Shifted Mean (Linear).

- **Input Dimension:** $d_x = 16$.
- **Context Size:** $N = 32$.
- **Latent Parameters:**
  - Discriminative direction $\mu \sim \text{Unif}(\mathbb{S}^{d_x-1})$.
  - Nuisance shift $k \sim \mathcal{N}(0, \sigma_k^2 I_{d_x})$.
- **Shift Magnitude:** $\sigma_k = 3.0$ (Training), $\sigma_k = 9.0$ (Out-of-distribution (OOD) evaluation).
- **Data Generation:** $x \mid y \sim \mathcal{N}(k + (2y-1)\mu, I)$.

Task B: Variance (Nonlinear).

- **Input Dimension:** $d_x = 16$.
- **Context Size:** $N = 32$.
- **Latent Parameters:**
  - Class 0 Scale $\sigma_0 \sim \text{Unif}[0.5, 3.0]$.
  - Class 1 Scale $\sigma_1 \sim \text{Unif}[0.5, 3.0]$.
- **Data Generation:** $x \mid y \sim \mathcal{N}(0, \sigma_y^2 I)$.

B.3 TRAINING HYPERPARAMETERS

Models are trained to minimize the Binary Cross Entropy loss on the query label y_q.

- **Optimizer:** AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay $1e-4$).
- **Scheduler:** OneCycleLR.
- **Initial Learning Rate:** 3×10^{-4}.
- **Batch Size:** 64 tasks per step.
- **Training Duration:** 20 epochs.
- **Seeds:** Accuracy results are reported over 3 random seeds.

B.4 ABLATION VARIANTS

To isolate the mechanism of in-context learning, we evaluated several model variants. Each variant tests a specific hypothesis regarding the inductive bias or information flow required for the task.

Positional Encodings (`NoPos`, `FrozenPos`). Standard Transformers use positional encodings to process sequences. However, the statistical tasks (shifted mean, variance) are permutation invariant with respect to the context examples. Namely, the learned methodology should be permutation-invariant like sufficent statistics are.

- `ICLTransformerNoPos`: We completely remove the learned positional embeddings ($P = 0$). This tests whether the model treats the context as a set rather than a sequence.
- `ICLTransformerFrozenPos`: We initialize positional embeddings randomly but freeze them during training. This tests whether the model requires learned positional information or can utilize random absolute position markers.

Attention Mechanism (`FrozenAttention`, `FrozenQK`). We test whether the attention heads must learn a task-specific metric space or if they can function as random associative memories.

- `ICLTransformerFrozenQK`: The Query (W_Q) and Key (W_K) projections are frozen at initialization. Only the Value (W_V) and Output (W_O) matrices are trainable. This enforces a fixed, random similarity kernel.
- `ICLTransformerFrozenAttention`: All attention weights (W_Q, W_K, W_V, W_O) are frozen. Only the feedforward MLPs and embedding projections are trainable.

Tokenization Strategy (`Interleaved`). Our default architecture sums input and label embeddings: $e_i = \text{Proj}(x_i) + \text{Proj}(y_i)$, effectively binding the label to the input in a single token.

- `ICLTransformerInterleavedEmbeddings`: We replace the bound representation with a standard GPT-style interleaved sequence $[x_1, y_1, x_2, y_2, \ldots, x_q]$. This tests whether the additive binding is a necessary inductive bias for efficient learning at this scale ($N = 2$ layers).

Label Dependence (`NoLabels`, `ShuffledLabels`). These ablations verify that the model is performing supervised mapping ($x \to y$) rather than unsupervised clustering ($x \to x$).

- `ICLTransformerNoLabels`: The context consists only of x vectors; y information is zeroed out.
- `ICLTransformerShuffledLabels`: The y labels in the context are randomly permuted within the batch, destroying the specific $x_i \to y_i$ mapping while preserving the marginal distribution of labels.
- `ICLTransformerNoisyLabels`: During training, a fraction p of the context labels are flipped ($0 \leftrightarrow 1$). This tests the model's ability to aggregate evidence robustly despite contradictory data points.

C Supplementary Experimental Results

C.1 Task A OOD Generalization Analysis

To assess whether the model has learned the exact symbolic form of the likelihood ratio or a local approximation, we evaluate it on OOD task where the nuisance shift magnitude σ_k is increased from 3.0 (training) to 9.0 (validation).

Figure 3 presents the learning dynamics and final decision geometry for this OOD setting.

- **Generalization Gap (Left):** While the training accuracy converges rapidly to $\approx 78\%$ (consistent with the in-distribution baseline), the OOD validation accuracy lags significantly, plateauing at $\approx 64\%$. The delayed rise in validation accuracy suggests a form of partial "grokking," where the model gradually refines its decision rule, but the persistent gap indicates that the learned mechanism does not fully capture the invariant symbolic structure needed for perfect extrapolation.
- **Regression Degradation (Right):** The correlation between the model's logits and the true LLR drops from $r \approx 0.86$ (in-distribution) to $r \approx 0.57$. The increased scatter suggests that the model's internal approximation of the sufficient statistic ($\mu^\top(x - k)$) is calibrated only for the training support and becomes brittle under large shifts.

Taken together, these results support the hypothesis that the Transformer implements an amortized approximate inference algorithm: it constructs a decision boundary that mimics the optimal LLR geometry locally, but relies on heuristics that degrade when the task parameters drift far from the training distribution.

C.2 Full Ablation Results

Table 2: **Full Experimental Results.** We report mean accuracy ± 95% CI over 3 seeds for all experimental conditions. The oracle rows represent the theoretical upper bound (Bayes-Optimal Classifier) computed using the true latent task parameters. The model is close to the oracle on Task B, while Task A ablations demonstrate the necessity of learned attention mechanisms.

Experiment / Condition	Model Variant	Train Acc (%)	Val Acc (%)
Theoretical Oracle			
Task A (Shifted Mean)	`LLR`	—	84.6 ± 1.0
Task B (Variance)	`LLR`	—	84.0 ± 1.0
Main Tasks			
Task A (Shifted Mean)	`ICLTransformer`	77.5 ± 1.1	78.3 ± 0.3
Task B (Variance)	`ICLTransformer`	83.0 ± 0.2	83.0 ± 0.5
Task A OOD ($\sigma_k = 9.0$)	`ICLTransformer`	77.5 ± 1.1	64.7 ± 4.8
Architecture Ablations (Task A)			
No Positional Encodings	`NoPos`	77.5 ± 1.1	78.2 ± 0.5
Frozen Positional Encodings	`FrozenPos`	77.5 ± 1.2	78.1 ± 0.6
Frozen Attention Weights	`FrozenAttention`	49.9 ± 0.2	50.4 ± 0.7
Frozen Q/K Projections	`FrozenQK`	49.7 ± 0.1	49.6 ± 1.3
Interleaved Embeddings (x, y)	`Interleaved`	49.8 ± 0.3	49.4 ± 1.2
Data Structure Ablations (Task A)			
Shuffled Context Pairs	`ShuffledContext`	77.5 ± 1.0	78.0 ± 0.6
Shuffled Labels Only	`ShuffledLabels`	49.8 ± 0.2	49.6 ± 1.3
No Labels	`NoLabels`	50.0 ± 0.1	50.2 ± 1.6
Increased Context Size	`ICLTransformer`	75.4 ± 4.7	75.9 ± 4.3
Label Noise Robustness (Task A)			
Noisy Labels ($p = 0.1$)	`NoisyLabels`	67.7 ± 11.2	70.2 ± 11.6
Noisy Labels ($p = 0.2$)	`NoisyLabels`	52.1 ± 2.9	53.3 ± 5.7
Noisy Labels ($p = 0.4$)	`NoisyLabels`	49.7 ± 0.2	49.7 ± 1.4

C.3 Comparison with Kernel Regression

To verify that the model is performing algorithmic reasoning rather than simple pattern matching, we compare its outputs to a Nadaraya-Watson (Nadaraya, 1964) estimator using a dot-product kernel:

$$\hat{y}_{KR}(x_q) = \sum_{i=1}^{N} \frac{e^{x_q^\top x_i}}{\sum_j e^{x_q^\top x_j}} y_i \tag{10}$$

As shown in Figure 4, the correlation between the Transformer's logits and the Kernel Regression estimator is weak ($\rho \approx 0.33$). This falsifies the hypothesis that the model is merely smoothing labels based on raw input similarity. In Task A, the optimal decision requires computing distances relative to a dynamic shift k, which a simple dot product kernel cannot capture without explicit centering.

C.4 Logit Lens Analysis: Task B

In contrast to the linear regime of Task A, where decision-relevant information emerges early in the residual stream, Task B exhibits a delayed decision profile.

As illustrated in Figure 5, the correlation between the intermediate residual states and the LLR remains negligible (≈ 0) through Layer 0 and Layer 1. A decisive spike in correlation appears only at the final output stage. This latency supports the hypothesis that nonlinear statistical inference requires a deeper, sequential circuit. We posit that the early layers are occupied with computing the necessary sufficient statistics (e.g., the quadratic energy term $\|x\|^2$) which are geometrically orthogonal to the final linear readout until fully assembled.

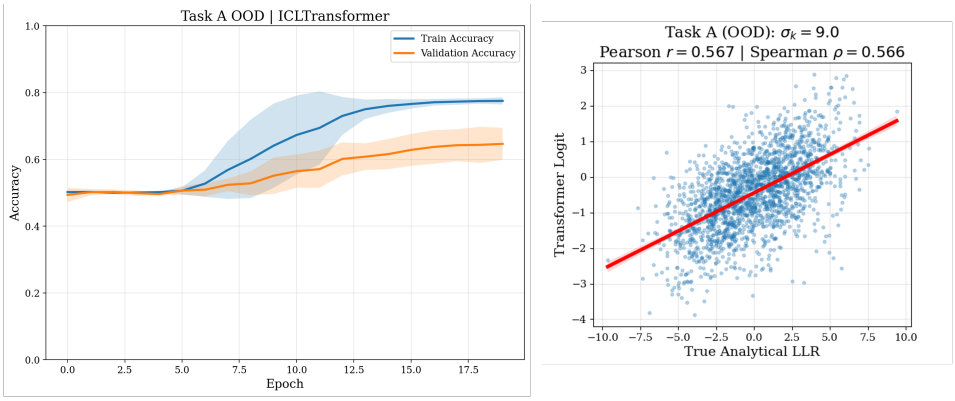

Figure 3: **OOD Generalization Degradation (Task A). (Left)** Learning curves show a significant generalization gap: while the model masters the training distribution (blue), it struggles to extrapolate to large shifts (orange), achieving only partial generalization. **(Right)** The correlation with the true LLR degrades to $r = 0.567$; the learned decision rule is a local approximation rather than the exact symbolic LLR.

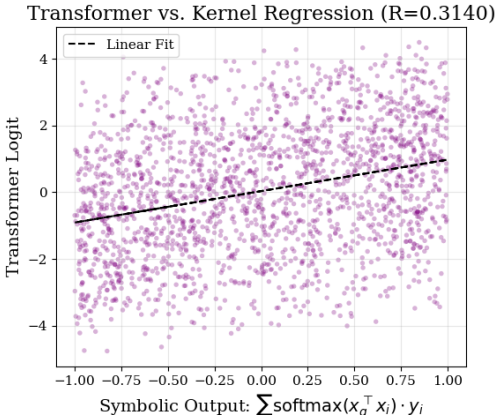

Figure 4: **Transformer vs. Kernel Regression.** The low correlation indicates the model implements a more complex decision rule than similarity-based label smoothing.

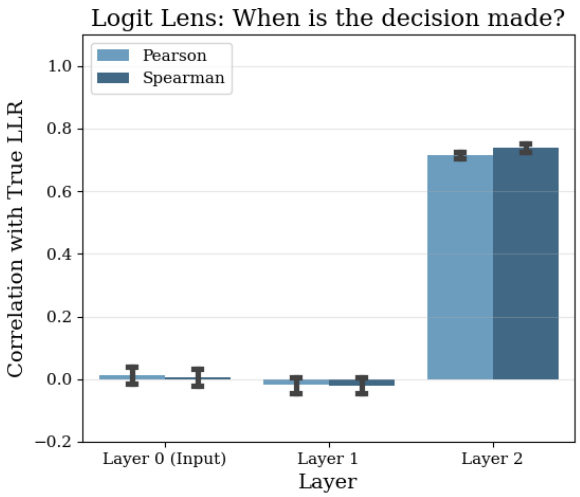

Figure 5: **Logit Lens for Task B.** The Pearson and Spearman correlations with the true LLR are effectively zero for the initial layers, spiking only at the final output. This confirms that the model does not perform a greedy linear approximation early in the network, but relies on the full depth of the Transformer to construct some nonlinear decision boundary.