BAYES-EQUIVALENT PROMPTS REVEAL HIDDEN NUISANCE DEPENDENCE IN IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

In-context learners can closely match Bayesian predictions on synthetic tasks. But do they also learn which aspects of the input a Bayesian predictor knows to ignore? We introduce Bayes-equivalent prompts: inputs that differ in raw form but induce exactly the same posterior and posterior predictive distribution, a direct test of whether learned predictors acquire the Bayesian solution's invariances. In multilayer perceptrons (MLPs) and Transformers, sensitivity to Bayes-irrelevant variation falls during training and is attenuated more in later layers, yet the identity of the nuisance perturbation remains linearly decodable from final representations. In a separate twelve-seed Transformer study trained to 800k updates, fit keeps improving and responses to task-relevant changes stay close to the Bayesian reference, while sensitivity to the same physical change when it is Bayes-irrelevant declines but stays above a pre-specified, study-specific benchmark. A consistency penalty (one weight, short continuation) sharply reduces disagreement on the targeted perturbation but raises held-out Kullback–Leibler (KL) divergence; in the sampled-target branch, this gain does not clearly carry over to a second Bayes-equivalent transformation. A public pretrained regression Transformer also shows small but measurable sensitivity to an exactly posterior-preserving recoding. Predictive fit therefore does not determine invariance, and invariance gained on one transformation need not transfer: evaluating whether a model has captured a Bayesian procedure requires testing its invariances, not only its predictions.

1 INTRODUCTION

In-context learning (ICL) allows a model to adapt its predictions from examples provided in the prompt, without updating its parameters. On controlled synthetic tasks, in-context learners can closely approximate task-specific estimators and exactly computable Bayesian predictors (Garg et al., 2022; Panwar et al., 2024). These comparisons typically summarize predictive agreement across inputs: a model can match the Bayesian predictor closely overall while still answering differently on two inputs that pose the same inference problem. When a learned predictor closely approximates Bayesian predictions, has it also learned which aspects of the input a Bayesian predictor knows to ignore?

A Bayesian model determines not only the prediction, but also which changes to the input leave that prediction unchanged. From the analytically known data-generating process, we construct Bayes-equivalent prompts: pairs that differ in raw form but induce the same posterior over the latent task and the same posterior predictive distribution (Figure 1). The equivalence is derived from the probabilistic model, not from a judgment that two prompts mean the same thing, so any disagreement within a pair is direct evidence of dependence on variation the Bayesian predictor ignores, which we call nuisance sensitivity. This criterion is stronger than equal predictive probabilities, since distinct posteriors can yield the same one.

Bayesian and latent-task accounts interpret ICL as approximate posterior inference, meta-learning, or Bayesian model averaging (Mikulik et al., 2020; Xie et al., 2022; Müller et al., 2022; Zhang et al., 2025), while algorithmic accounts emphasize learned estimation and optimization (Akyürek et al., 2023; von Oswald et al., 2023); whether such behavior emerges depends on pretraining conditions and can evolve non-monotonically (Raventos et al., 2023; Singh et al., 2023; Wang et al., 2025; Wurgaft et al., 2025). Structural audits test Bayesian coherence (Falck et al., 2024; Jesson et al., 2025)

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

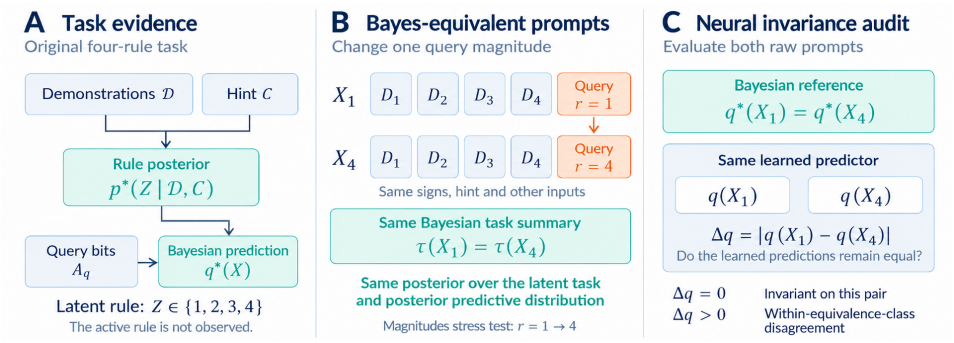

Figure 1: Exact Bayes-equivalence auditing in the original four-rule task. (A) Demonstrations and hint determine the rule posterior; query bits complete the Bayesian prediction. (B) Only one query-coordinate magnitude changes, $r = 1$ to 4: four demonstrations, hint, query signs and other inputs stay fixed, preserving this task's τ, posterior and posterior predictive. (C) The same frozen predictor's absolute probability difference measures disagreement on this pair.

and separate structural Bayesian properties from predictive competitiveness using task-preserving order interventions (Chlon et al., 2026); session-handover theory likewise shows that mutually recoverable records have equal ideal risks yet can incur different risks under the same procedure (Kato & Kato, 2026, Section 6.2). ICL can also be insensitive to demonstration labels (Min et al., 2022) yet sensitive to prompt order and format (Lu et al., 2022; Xiang et al., 2024; Sclar et al., 2024), which permutation-invariant methods address architecturally (Fang et al., 2025). We extend these audits beyond prompt order to Bayes-equivalent perturbations defined by the inference problem itself.

Consistency regularization encourages agreement between predictions on original and transformed inputs (Xie et al., 2020; Hendrycks et al., 2020; Zhang et al., 2019), though inherent robustness–accuracy conflicts (Tsipras et al., 2019), estimator costs despite compatible optima (Raghunathan et al., 2020) and train–test transformation mismatch (Mintun et al., 2021) can limit its gains. We apply it as a controlled ICL intervention on prompt pairs whose Bayesian posterior and prediction are analytically unchanged.

Mechanistic studies examine what ICL models retain internally: task and function representations (Hendel et al., 2023; Todd et al., 2024), concept subspaces (Tang et al., 2026), task-oriented information removal (Cho et al., 2026), and depth-dependent statistical computation (Chaudhry & Gadkari, 2026), which does not by itself imply invariance to posterior-preserving changes. Theories of deep representation suggest that training can progressively discard nuisance information (Achille & Soatto, 2018). When output sensitivity decreases, has the nuisance information therefore disappeared, or has its influence merely been attenuated? Because probe recoverability need not imply relevance or causal use (Ravichander et al., 2021; Elazar et al., 2021; Opiełka et al., 2026), we test whether the nuisance perturbation remains linearly decodable from late representations.

We audit trained MLPs and Transformers on two synthetic tasks and one public pretrained model, ask whether the residual sensitivity can be reduced by a consistency penalty, and test whether the sampled-target consistency gain transfers to a second equivalence (Section 4 gives the experiments and results).

2 Controlled Task and Bayes-Equivalent Inputs

2.1 Task and Bayesian reference

In the four-rule task, a latent rule Z is uniform on $\{1, 2, 3, 4\}$. Each of four demonstrations has a four-bit pattern A_i (uniform over the six patterns with two ones) and a label S_i equal to A_{iZ} with probability 0.75 and otherwise flipped. The hint satisfies $C \mid Z = j \sim \mathcal{N}(\sqrt{2}e_j, I_4)$, independent of the demonstrations given Z. The query A_q has independent fair bits and target $Y = A_{qZ}$. The model observes the demonstrations, hint and query, but not Z or Y.

Up to terms constant in j, the rule-dependent log-likelihoods are $\ell_{Dj} = \sum_{i=1}^{4} \big[\mathbf{1}\{S_i = A_{ij}\} \log 0.75 + \mathbf{1}\{S_i \neq A_{ij}\} \log 0.25\big]$ and $\ell_{Cj} = \sqrt{2}\, C_j$. The answer-feature magnitudes of

Section 2.2 are independent of Z given the encoded bits, so they add no evidence among the rules. With the uniform prior,

$$p^*(X) = \mathrm{softmax}(\ell_D + \ell_C), \qquad q^*(X) = \sum_{j=1}^{4} A_{qj}\, p_j^*(X), \tag{1}$$

where $q^*(X)$ is the posterior probability that the answer is one. Softmax is unchanged by a common constant, so with $\mathrm{center}(\ell) = \ell - \mathrm{mean}(\ell)\mathbf{1}$ we define the Bayesian task summary and audit equivalence by

$$\tau(X) = \big(\mathrm{center}(\ell_D),\, \mathrm{center}(\ell_C),\, A_q\big), \qquad X \sim_B X' \iff \tau(X) = \tau(X'). \tag{2}$$

Equivalent inputs thus share relative demonstration evidence, relative hint evidence and query answer mapping, hence the same posterior over Z and posterior predictive. The Bayesian predictor is constant within each class: invariance follows from the inference problem rather than being imposed. The construction is task-specific and neither unique nor minimal; later audits use the structure of their own latent-task models.

2.2 Equivalence-preserving pairs

Each binary answer feature $a \in \{0, 1\}$, in the demonstrations and the query, is encoded as $x = (2a - 1)r$ with magnitude $r = \max(|g|, 10^{-6})$, $g \sim \mathcal{N}(0, 1)$, independent of the task evidence: the sign of x fixes the bit a, while the magnitude carries no information about Z.

Our primary Bayes-equivalent pair changes the magnitude of one query coordinate j from $r = 1$ to $r = 4$, holding its sign, all other query coordinates, the demonstrations and the hint fixed. Then A_{qj}, ℓ_D and ℓ_C are unchanged, so $\tau(X_1) = \tau(X_4)$ and the two inputs have the same posterior over Z and the same posterior predictive. Any prediction difference therefore reflects sensitivity to a Bayes-irrelevant magnitude change, not to a change in task evidence.

Both magnitudes lie in the support of the generating distribution, but $r = 4$ is in a low-density tail, so we treat $1 \to 4$ as a deliberately large stress perturbation, not a typical nuisance change. As a check we also use a $Q_{25} \to Q_{75}$ perturbation (the 25th and 75th percentiles of r), with adjacent-quartile comparisons as secondary checks. Encoding details, likelihood calculations and equivalence validation are in Supplement S1–S3.

3 Experimental Design and Measures

3.1 Original models and longitudinal measures

We ask how sensitivity to a Bayes-irrelevant magnitude change evolves over training and across representation depth, following the same models and prompt pairs across checkpoints (a longitudinal comparison within models, not across independently trained models at different budgets).

Models and pairs. Each architecture has five independently trained models (seeds): two-hidden-layer MLPs and two-layer, width-64 Transformers of comparable size (about 68k parameters), each with a scalar logit head. The Transformers must first pass a pre-specified task-learning criterion on Bayes regret (excess log loss over the Bayesian predictor): final mean at most 0.05 bits, with every seed improving and below 0.10 (Supplement S4). We evaluate six checkpoints from 5k to 200k updates on 4,096 (MLP) or 8,192 (Transformer) prompt pairs shared across models, each pair differing only in one query-coordinate magnitude (1 or 4). We read out both post-GELU MLP layers and, for the Transformer, the projected query token (with positional embedding), both block outputs, and the final LayerNorm, all at the query token (Supplement S2, S4 and S5).

Measures. For each pair we record two responses. The normalized hidden sensitivity M compares the hidden-state change to the scale of that representation, keeping it comparable as hidden norms grow during training. The behavioral disagreement P is the root-mean-square change in predicted probability q. With h_r and q_r the hidden state and output probability at magnitude $r \in \{1, 4\}$, and E the average over prompt pairs,

$$M = \sqrt{\mathbb{E}\|h_4 - h_1\|^2}/\sigma_h, \quad \sigma_h^2 = \mathbb{E}\|h - \mathbb{E}h\|^2, \qquad P = \sqrt{\mathbb{E}(q_4 - q_1)^2}, \tag{3}$$

Here σ_h pools both endpoints with equal weight, and M and P are computed per model and checkpoint (and hidden location for M) before averaging over models. Because σ_h changes during training, a falling M does not imply a falling raw displacement (Supplement S5).

Depth contrast. For the post-confirmatory depth analysis, we fix an early and late location: MLP hidden layers 1 and 2, and Transformer block 1 and final LayerNorm (not assumed computationally equivalent across architectures). With $\bar{M}$ the model-averaged M,

$$A_t = \bar{M}_{\text{late},t}/\bar{M}_{\text{early},t}, \qquad T_A = A_{200\text{k}}/A_{5\text{k}}. \tag{4}$$

$A_t < 1$ means lower normalized sensitivity at the late location (for example, 0.3 means 30% of the early value), and $T_A < 1$ means this contrast strengthens during training.

Inference and status. H1 (late attenuation at 200k) requires the upper 95% bound of $A_{200\text{k}}$ below one and the same direction in at least four of five seeds; H2 (strengthening) applies the same rule to T_A. Intervals are nominal per-quantity bootstrap intervals over seeds and prompt pairs (both versions of a pair kept together), without familywise control; intermediate checkpoints are descriptive. The MLP depth analysis revisits previously inspected results, whereas the Transformer final LayerNorm was fixed as the late location before the depth evaluation because it feeds the output head. The H1/H2 depth analysis and the central-quartile extension (Supplement S6) were nevertheless designed after earlier related results were known and are therefore treated as post-confirmatory.

3.2 Posterior-relevance study and extended training

The nuisance-versus-signal comparison in the original task changed both relevance and the form of the perturbation, since the signal change crossed a sign boundary (Supplement S9). To isolate relevance, we build a separate task in which one physical change is held fixed and only its relevance to the Bayesian prediction varies.

Task. A latent coordinate $Z \in \{1, 2, 3, 4\}$ generates Bernoulli labels with probability $\sigma(x_Z)$, where σ is the logistic function. The model sees an explicit prior π, six labeled demonstrations and a query. With ℓ the vector of demonstration log-likelihoods, the posterior is $p^* = \text{softmax}(\log \pi + \ell)$ and the exact query prediction is $q^*(x) = \sum_j p^*_j \sigma(x_j)$. We call q^* the oracle.

Relevance control. The relevance ρ of a query coordinate is its posterior probability. From each base prompt we construct priors that set $\rho \in \{0, .10, .25, .50, .75\}$, with the other coordinates sharing mass $1 - \rho$ in fixed proportions. For each fixed ρ, the paired inputs differ only in that coordinate's magnitude, from its 25th percentile $Q_{25} \simeq .3186$ to its 75th percentile $Q_{75} \simeq 1.1503$; its sign and all other inputs are held fixed. Across values of ρ, we adjust the prior to control the coordinate's Bayesian relevance. The sign-aligned Bayesian change is $s(q^*_{75} - q^*_{25}) = \rho\Delta_\sigma$, where $s \in \{-1, +1\}$ is the sign of the perturbed query coordinate and $\Delta_\sigma = \sigma(Q_{75}) - \sigma(Q_{25}) \simeq .1805821$. It is exactly zero at $\rho = 0$ (a Bayes-equivalent pair) and grows with ρ. ρ is an external control, not an internal neural variable. In training, 20% of episodes have a prior with one zero-probability coordinate, so the $\rho = 0$ audit uses a prior type the models have seen.

Measures. We measure three aspects of behavior. Nuisance sensitivity $N_{0,\text{ABS}}$ is the mean of $|q_{75} - q_{25}|/\Delta_\sigma$ at $\rho = 0$, averaged within each coordinate-by-sign group (four coordinates $\times$ two signs) and then equally over groups and models; $\epsilon_0 = .01$ is a pre-specified, study-specific benchmark for practical invariance, not an application-harm threshold. Relevant response β_{POS} measures how the response grows with relevance: with the input sign s defined above, $\xi = s(q_{75} - q_{25})/\Delta_\sigma$ equals ρ for the oracle. For each model and base prompt we fit ξ on the four positive relevance values (recomputed from the encoded inputs; equal weights, with an intercept; $\rho = 0$ excluded) and average the slopes within the eight coordinate-by-sign groups, then over groups and models. The oracle slope is one, and a slope near one does not by itself establish perfect fidelity. Predictive fit K is the mean Bernoulli $\text{KL}(q^* \| q)$ in bits on held-out episodes from the task distribution (lower is better); unlike the paired measures, it scores individual episodes.

Models and status. We train a separate cohort of 12 six-layer, width-128 Transformers, unrelated to the original models. The original confirmatory analysis evaluated them at 200k updates with 97.5% intervals. After seeing those results, we continued the same models to 400k and 800k, reevaluating on the unchanged prompt pairs and held-out episodes without replacing seeds, and summarize the continuation with paired 95% pointwise descriptive intervals. This allows paired comparisons but is

neither an independent replication nor familywise coverage, and does not replace the 200k analysis (Supplement S10–S14).

Paired-error decomposition. To separate broad predictive improvement from targeted reduction in disagreement, we split the error on the zero-relevance paired panel. With $\bar{q} = (q_{25} + q_{75})/2$,

$$E_{\mathrm{PAIR}} = \tfrac{1}{2}\mathbb{E}[(q_{25} - q^*)^2 + (q_{75} - q^*)^2] = E_{\mathrm{SHARED}} + E_{\mathrm{WITHIN}},$$
$$E_{\mathrm{SHARED}} = \mathbb{E}[(\bar{q} - q^*)^2], \quad E_{\mathrm{WITHIN}} = \tfrac{1}{4}\mathbb{E}[(q_{75} - q_{25})^2], \quad F_{\mathrm{WITHIN}} = E_{\mathrm{WITHIN}}/E_{\mathrm{PAIR}}. \tag{5}$$

E_{SHARED} is the error of the pair's mean prediction, and E_{WITHIN} reflects endpoint disagreement. F_{WITHIN}, a ratio of cohort mean errors, is the algebraic share of paired error due to disagreement, not a causal attribution. We apply this identity to the zero-relevance paired panel; it does not decompose held-out K, which uses a different loss and distribution.

4 Results

We organize the results around three questions. **Part I** asks whether trained models disagree on Bayes-equivalent inputs, how this changes with training and depth, and whether it also appears in an independent model (§4.1–4.3). **Part II** asks whether reduced output sensitivity means the nuisance information has been removed (§4.4). **Part III** asks whether the residual can be reduced by changing the training target or adding a consistency penalty, and, in the sampled-target branch, whether the consistency gain transfers to a different Bayes-equivalent transformation (§4.5–4.6). The setups for §4.3–4.6 open their subsections; those for §4.1–4.2 are in Section 3.

In the original four-rule models, all primary pairs have identical Bayesian task summaries (Section 2.1), so prediction differences isolate sensitivity at fixed Bayesian evidence. Additional panels test demonstration-order and equal-likelihood variation as breadth checks with different cohorts and estimands (Supplement S7). Unless stated otherwise, intervals are nominal 95% bootstrap intervals over seeds and prompts, without familywise coverage guarantees. Post-confirmatory intervals are descriptive. The original 200k confirmatory results of the relevance study use 97.5% intervals.

4.1 Output disagreement falls with training; hidden attenuation is late

In both architectures, output disagreement under the $1 \to 4$ magnitude change decreases over training, while normalized hidden sensitivity is lower at the selected late than early representations at 200k updates (Figure 2, Table 1). At 200k the late-to-early ratio is well below one ($A_{200\mathrm{k}} = 0.30$ for the MLP, 0.41 for the Transformer) and strengthens from 5k ($T_A = 0.30, 0.51$); all five seeds per architecture agree, so both architectures meet the stated H1 and H2 criteria. For the MLP this assessment revisits previously inspected results, whereas the Transformer final LayerNorm was fixed before the depth evaluation as the layer feeding the output head. Output disagreement falls by about 60% ($P_{200\mathrm{k}}/P_{5\mathrm{k}} \approx 0.40$) but stays nonzero. The Transformer meets its task-learning criterion (final mean Bayes regret 0.00793 bits).

Three qualifications apply. The pattern is not only a final-LayerNorm effect, since attenuation is already present before it; the profile is not monotonic at every location and checkpoint; and because M is normalized, its decline can coexist with growing raw displacement (Supplement S5), so the result is attenuation relative to scale, not raw contraction.

The post-confirmatory central pair ($Q_{25} \to Q_{75}$) gives the same directions in all five seeds per architecture, with adjacent-quartile checks agreeing (Supplement S6). Final responses are smaller than for the $1 \to 4$ stress pair but nonzero. Because perturbation size and input density change together, this extends the evidence to central changes, not to arbitrary magnitudes.

4.2 Fit improves with longer training; residual nuisance sensitivity persists

Continuing the twelve relevance-controlled models from 200k to 800k improves held-out fit (K falls from 0.0019 to 0.0011 bits) while leaving measurable sensitivity to Bayes-irrelevant changes (Figure 3). Nuisance sensitivity $N_{0,\mathrm{ABS}}$ declines from 0.049 to 0.042 to 0.038 (a 22% reduction from 200k), but the 800k value (descriptive 95% interval $[0.034, 0.042]$) remains 3.8 times the benchmark

Table 1: Longitudinal results for the four-rule models (five seeds per architecture; 95% bootstrap intervals). A_{200k} is the late-to-early ratio at 200k, $T_A = A_{200k}/A_{5k}$, and P_{200k}/P_{5k} is the change in output disagreement. Dashes indicate values not reported for the post-confirmatory central-pair check.

	Pair	A_{200k}	T_A	P_{200k}/P_{5k}
MLP	$1 \to 4$	$0.299\ [0.282, 0.316]$	$0.297\ [0.276, 0.318]$	$0.399\ [0.367, 0.433]$
	$Q_{25} \to Q_{75}$	$0.255\ [0.245, 0.267]$	$0.250\ [0.239, 0.261]$	—
Transformer	$1 \to 4$	$0.412\ [0.364, 0.450]$	$0.508\ [0.467, 0.537]$	$0.402\ [0.333, 0.463]$
	$Q_{25} \to Q_{75}$	$0.300\ [0.270, 0.338]$	$0.383\ [0.345, 0.422]$	—

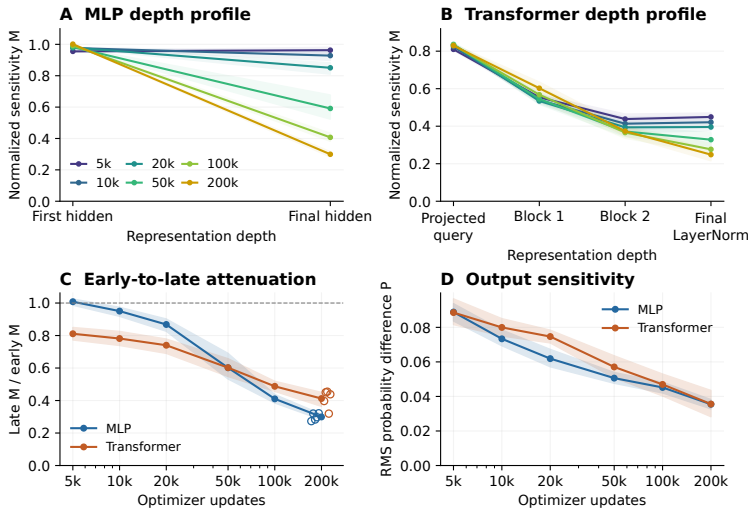

Figure 2: Longitudinal magnitude sensitivity: normalized hidden M in (A) MLP and (B) Transformer; (C) late-to-early ratio A_t; (D) RMS output disagreement P. Bands: 95% intervals from paired seed/base resampling. Open circles: individual 200k seed ratios, offset for visibility. (C)–(D) use log update axes. Intermediate checkpoints are descriptive; locations are architecture-specific. Lower normalized M need not mean smaller raw displacement.

$\epsilon_0 = 0.01$. On the raw probability scale, this corresponds to a mean absolute disagreement of about 0.69 percentage points, compared with 0.18 percentage points for the benchmark (Supplement S12). The residual is therefore small on the raw probability scale; its significance here is structural rather than evidence that the discrepancy is necessarily practically consequential. The benchmark is an operational criterion for this study, not an application-harm threshold. Responses to task-relevant changes stay close to, but below, the oracle ($\beta_{\text{POS}} = 0.95 \to 0.97$), and their change intervals include zero, so we do not infer improved tuning. Nuisance sensitivity was still falling over the last interval (400k to 800k change -0.0039, interval $[-0.0069, -0.0009]$), but trajectories are heterogeneous: 9 of 12 seeds decline between 400k and 800k (Supplement S13). The finite continuation does not determine an asymptotic limit.

4.3 A PUBLIC PRETRAINED TRANSFORMER SHOWS THE SAME SENSITIVITY

Setup. We audit one public pretrained linear-regression Transformer of Garg et al. (2022) (no fine-tuning) on dense noiseless Gaussian regression, $w, x \sim \mathcal{N}(0, I_d)$ with $d = 20$ and $y = x^\top w$, using 4,096 episodes at each of $n = 10$ (primary) and $n = 15$ (secondary) demonstrations, both below d. The *sign recoding* replaces one preassigned demonstration (x_i, y_i) with $(-x_i, -y_i)$ ($X' = JX$, $y' = Jy$, J flipping one row's sign); since $Xw = y \iff JXw = Jy$, the posterior and posterior predictive are preserved. The Bayesian predictive mean is $m_B = x_q^\top X^\dagger y$ (the minimum-norm least-squares prediction, treating the recorded numbers as exact); the network outputs a point prediction, so we assess its mean, not its variance. Task MSE is the squared error against the generated query label, and R_{APPROX} is the mean squared difference between the network prediction and m_B on original prompts. Recoding disagreement is the RMS prediction difference and the $E_{\text{WITHIN}}/F_{\text{WITHIN}}$

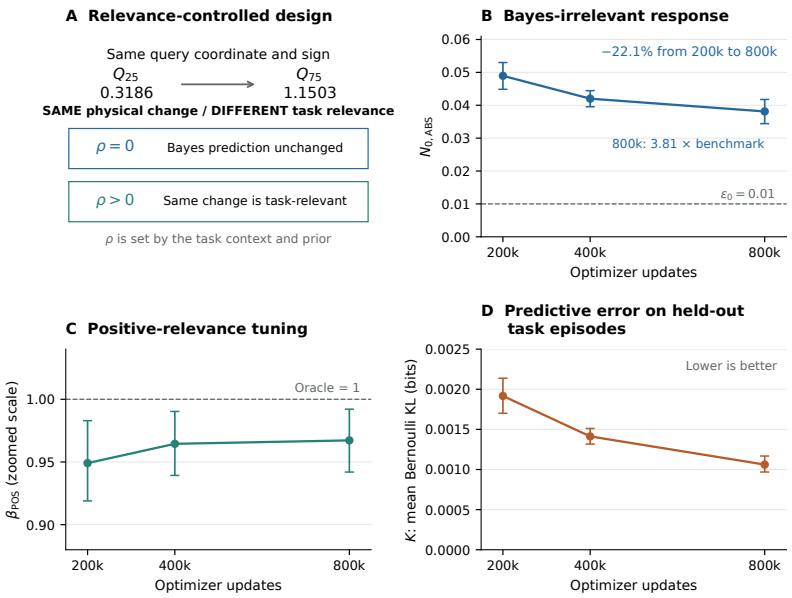

Figure 3: Relevance-controlled continuation of twelve Transformer seeds. (A) A fixed-sign Q_{25}-to-Q_{75} change is irrelevant at $\rho = 0$, relevant at positive ρ. (B) $N_{0,\mathrm{ABS}}$ remains above the pre-specified operational benchmark $\epsilon_0 = 0.01$. (C) β_{POS} stays close to, but below, one (zoomed scale). (D) Held-out Bayesian KL decreases. Error bars: paired 95% descriptive intervals from 20,000 draws, not replacements for original 200k confirmatory 97.5% intervals. Lines join observed checkpoints, without asymptotic fitting.

split of Eq. (5) with m_B in place of q^*. Intervals are 2,000 bootstrap resamples of prompts stratified by recoded position; they describe one checkpoint, and the two context lengths are not replications (Supplement S24).

The pretrained model performs close to the minimum-norm baseline on the task (n=10: MSE 9.79 vs. 9.74; n=15: 5.08 vs. 5.02). Despite this, sign recoding changes its prediction while preserving the Bayesian posterior and posterior predictive: the RMS change is 0.076 $[0.074, 0.079]$ at n=10 and 0.078 $[0.075, 0.081]$ at n=15 (Bayes-mean approximation error on original prompts $R_{\mathrm{APPROX}} = 0.039$ and 0.053), and the within-pair component accounts for 3.7% $[3.4, 4.0]$ and 2.8% $[2.4, 3.2]$ of the paired Bayes-mean approximation error E_{PAIR}. Most paired error is therefore shared with the unrecoded prompt, and the share is a fraction of approximation error, not of task MSE. This extends the phenomenon to another model, task and provenance; it does not test transfer of the consistency intervention (Supplement S24).

4.4 Lower output sensitivity does not mean the nuisance is gone

In the original four-rule models we test whether magnitude (1 vs. 4) remains linearly decodable from the final pre-head state after removing the single direction that drives the output head (a logistic-regression probe on the high-posterior subset G with posterior mass above 0.5 on the perturbed coordinate, called the strong subset in the supplement; post-confirmatory; Supplement S15–S17). At 200k it does: held-out AUC is 0.740 from the full state and 0.745 from its head-null component for the MLP, and 0.971 and 0.967 for the Transformer, with every seed above chance. Decodability was already near ceiling at 5k, so these results do not show that training created it; for the MLP it has fallen substantially (head-null AUC $0.999 \to 0.745$) but stays well above chance, while the Transformer's barely moves. The full perturbation still changes the logit (RMS 0.873 MLP, 0.732 Transformer), so recoverability coexists with output-relevant sensitivity. Because the nullspace has all but one dimension, this shows retention, not preferential storage or causal use.

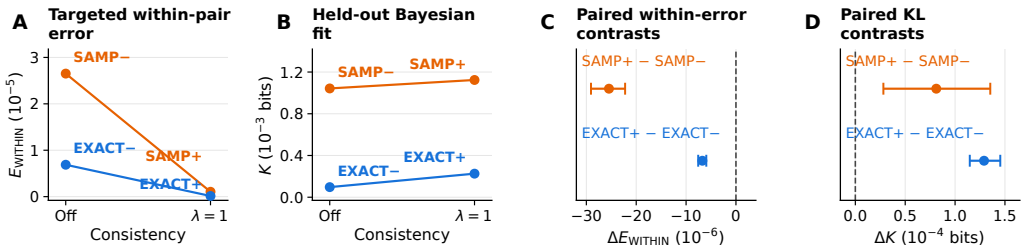

Figure 4: Post-confirmatory 2×2 diagnostic, twelve paired training trajectories and +10k updates. A = SAMP$^-$, C = EXACT$^-$, D = SAMP$^+$, E = EXACT$^+$ (sampled vs exact targets; consistency off vs $\lambda = 1$). (A)–(B): targeted E_{WITHIN} and held-out K (bits); E's within-error is nonzero (1.40×10^{-7}). (C)–(D): paired descriptive 95% intervals using the same resampling draws across all four conditions (Supplement S22). Lines join discrete conditions. Extra D/E partners prevent matched-FLOP interpretation.

4.5 CONSISTENCY CUTS DISAGREEMENT AT A FIT COST

Setup. Nuisance sensitivity remains at 800k updates; can it be reduced? From each of the twelve 800k models (one per seed, with optimizer state) we continue training for 10k updates (1.25% more) in a 2×2 design: the query loss uses either a sampled binary outcome or the exact Bayesian probability $q^*(X)$, with (+) or without (−) a consistency penalty, giving SAMP$^-$, EXACT$^-$, SAMP$^+$ and EXACT$^+$; an auxiliary SAMP$^-$ (low LR) uses learning rate 3×10^{-5} instead of 3×10^{-4} (Supplement S20). The penalty adds the mean absolute probability disagreement between an eligible *anchor* (a sampled training prompt with a zero-prior coordinate) and its *partner* (a Bayes-equivalent prompt with that coordinate's query magnitude resampled) to the binary cross-entropy (BCE) loss on all sampled anchors, with $\lambda = 1$ untuned (Supplement S21–S22). Partners enter only the penalty, so the "+" conditions process extra examples and equal update counts do not imply equal compute. EXACT$^+$ starts from the 800k state, not from EXACT$^-$. These are paired continuations of the same trajectories, not independent cohorts, and measure the short-term response. All contrasts were designed after earlier results were known and use paired descriptive 95% intervals with the twelve seeds as independent units; EXACT$^+$ vs EXACT$^-$ on E_{WITHIN} and K is primary (Supplement S22).

Figure 4 reports the 2×2 comparison (condition means for all metrics are in Supplement S22). Adding consistency under exact targets (EXACT$^+$ vs. EXACT$^-$, the primary comparison) lowers targeted within-pair error ($\Delta E_{\text{WITHIN}} = -6.7 \times 10^{-6}$, 95% interval $[-7.6, -5.9] \times 10^{-6}$) but raises held-out K ($+1.3 \times 10^{-4}$ bits, $[1.1, 1.5] \times 10^{-4}$), in all twelve paired trajectories; SAMP$^+$ vs. SAMP$^-$ shows the same pattern. The paired panel evaluates zero-relevance prompts, whereas K averages over the full task distribution, which assigns 80% probability to all-positive priors. Such prompts are not directly included in the consistency term, although updates to the shared model parameters can affect their predictions.

Consistency drives $N_{0,\text{ABS}}$ below the benchmark $\epsilon_0 = 0.01$ (0.0059 and 0.0023, against 0.036 and 0.016 without the penalty) and cuts the within-pair share F_{WITHIN} to 0.24% and 0.14%. Exact targets alone instead improve both fit and raw within-pair error E_{WITHIN} while raising the share (5.1% $\rightarrow$ 8.2%), so broad approximation improvement and targeted suppression are different changes. Lowering the learning rate likewise improves disagreement and fit ($N_{0,\text{ABS}}$ 0.036 $\rightarrow$ 0.032, K $10.4 \times 10^{-4} \rightarrow 4.0 \times 10^{-4}$ bits) without preferentially reducing the share (5.1% $\rightarrow$ 7.5%; Supplement S20). Consistency also pushes positive-relevance slopes above the oracle value (1.068, 1.035), which does not indicate better fidelity. These results characterize one penalty weight and a short +10k continuation, not a universal fit–invariance trade-off.

4.6 SAMPLED-TARGET CONSISTENCY GAINS DO NOT CLEARLY TRANSFER

Setup. We recode one demonstration row (x_i, y_i) as $(-x_i, 1 - y_i)$ (the *complement recoding* T_{COMP}), leaving the prior, query and other demonstrations unchanged. The signed evidence for every candidate rule is unchanged, $(2(1-y_i)-1)(-x_{ij}) = (2y_i-1)x_{ij}$, so the likelihood vector, posterior and posterior predictive are preserved (zero numerical discrepancy), and given the prior type the recoded prompt has the same marginal distribution. We evaluate two separately sampled panels of

8,192 prompts, the zero-prior panel and the all-positive panel (recoded position balanced), and report them separately, not averaged. Disagreement is $E_{\text{WITHIN,COMP}} = \frac{1}{4}\mathbb{E}[(q(T_{\text{COMP}}X) - q(X))^2]$, the counterpart of the within-pair term in Eq. (5). A label-only recoding T_{LABEL} (complementing the label without negating features) changes the Bayesian evidence, so regressing the model's prediction change on the Bayesian change (with intercept; oracle slope one) tests whether low recoding disagreement reflects reduced responsiveness to demonstrations. This post-confirmatory audit evaluates existing checkpoints at 810k for SAMP$^-$, SAMP$^-$ (low LR), EXACT$^-$ and SAMP$^+$; EXACT$^+$ was not evaluated (Supplement S23). Thus the transfer comparison specifically tests the sampled-target consistency intervention, not the primary EXACT$^+$ versus EXACT$^-$ contrast of Section 4.5.

Under sampled targets, consistency strongly reduces disagreement for query-magnitude changes but has a small, uncertain effect on demonstration recoding: against the SAMP$^-$ baseline $E_{\text{WITHIN,COMP}}$ of 18.8 (zero-prior panel) and 19.3 (all-positive panel) in units of 10^{-6}, SAMP$^+$ changes it by -0.74 $[-2.09, 0.48]$ and -0.57 $[-1.98, 0.65]$. The intervals span zero in both panels, which supports unclear transfer, not its absence. For scale, the point-estimate reduction in within-pair error is about 96% for the magnitude family (Section 4.5), about 4% for recoding in the zero-prior panel, and about 3% in the all-positive-prior panel. This comparison is descriptive; no formal cross-family effect-difference test was performed. Exact targets alone (EXACT$^-$ vs. SAMP$^-$) reduce recoding disagreement far more (-15.6 $[-17.7, -13.6]$ and -16.3 $[-18.6, -14.1]$), consistent with broad approximation improvement rather than an equivalence-specific mechanism. The label-only control shows that SAMP$^+$ still responds to demonstrations (slopes $\simeq 0.93$, modestly below SAMP$^-$'s $\simeq 0.95$), so it does not simply ignore them, but reduced responsiveness remains a competing explanation for its small recoding gain (Supplement S23). The two panels are separate prior conditions, not averages over the full task distribution.

5 DISCUSSION

Fit and invariance are separable. Agreement with a Bayesian predictor does not show how a model treats inputs that pose the same inference problem; Bayes-equivalent pairs expose this by fixing the reference posterior and prediction, complementing structural audits (Chlon et al., 2026; Kato & Kato, 2026). The two are not mathematically independent on the paired distribution (within-pair disagreement is part of squared error), but held-out K uses a different loss and distribution. Sensitivity persists as fit improves (Section 4.2) and in a public model with near-baseline task performance (Section 4.3), unlike our internal regression extension, which failed its competence gate (Supplement S18); one checkpoint does not establish a shared mechanism.

Attenuation is not removal. Lower normalized hidden sensitivity need not mean smaller raw displacement, and lower output disagreement does not mean the nuisance information is absent: the magnitude endpoint remains linearly decodable after removing the readout direction (Section 4.4), which shows retention, not causal use (Ravichander et al., 2021; Elazar et al., 2021; Todd et al., 2024; Opiełka et al., 2026).

Suppression is not better fit. Exact targets improve fit but raise the within-pair share; consistency lowers it at a held-out KL cost (Section 4.5). This need not mean the ideal objectives conflict: an unrestricted Bayes predictor minimizes population BCE with zero penalty. Our result concerns one model, $\lambda = 1$ and $+10$k updates, leaving other recipes and invariant architectures (Fang et al., 2025) open; nor do the sampled-target consistency gains clearly transfer to a second equivalence (Section 4.6), which shows neither zero transfer nor a mechanism.

Scope and implications. These conclusions concern controlled synthetic tasks, specified transformations and finite horizons, not foundation models. The cohorts differ, paired interventions reuse trajectories (extra partners break compute matching), declines fix no limit, normalization is coordinate-dependent, central-quartile checks cover only tested magnitudes, and follow-ups were designed after earlier results (an earlier first-layer MLP analysis missed its criteria; Supplement S8). Within this scope, auditing Bayes-equivalence turns a known invariance of the inference problem into a direct test of the learned predictor: judging whether a model has captured a Bayesian procedure requires testing its invariances, not only its predictions. Where an exact equivalence is known, such as controlled pretraining studies or tasks with a provable symmetry, invariance can be reported alongside accuracy and used as an evaluation or training signal; whether this extends to natural-language prompts, where equivalence is not exactly defined, is open.

AI USE STATEMENT

AI tools assisted implementation of training, evaluation, and synthetic-data-generation code; protocol and methodological drafting; literature review, analysis support, and assistance with interpreting and checking results; figure preparation; and manuscript drafting and editorial restructuring. All scientific decisions, protocol approvals, interpretations, and responsibility for the final manuscript remain with the authors. All AI-assisted work was reviewed by the authors.

REPRODUCIBILITY STATEMENT

The anonymous reviewer artifact at `https://github.com/AuditTesting/bayes-equivalent-icl-review-artifact` provides code and saved inputs for reconstructing principal numerical summaries and selected uncertainty estimates for Figures 2–4, the recoding contrasts in Section 4.6 (Supplement S23), and the mapped external-regression audit metrics in Section 4.3 (Supplement S24). Within the release artifact, two archived tables summarize the recoding contrasts and external-regression metrics, respectively. Prior verification of the complete release archive within that mapped scope included independent download, hash verification, and checks: 150 mapped claims and 414 numerical comparisons passed. See its README for commands and scope. It does not support running the saved models on new inputs or repeating the full training procedure. Additional historical records and checkpoints remain retained by the authors.

REFERENCES

Alessandro Achille and Stefano Soatto. Emergence of invariance and disentanglement in deep representations. *Journal of Machine Learning Research*, 19(50):1–34, 2018. URL `http://jmlr.org/papers/v19/17-646.html`.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. In *International Conference on Learning Representations*, 2023. URL `https://openreview.net/forum?id=0g0X4H8yN4I`.

Faris Chaudhry and Siddhant Gadkari. Implicit statistical inference in transformers: Approximating likelihood-ratio tests in-context. In *Latent and Implicit Thinking Workshop at ICLR*, 2026. URL `https://arxiv.org/abs/2603.10573v1`. arXiv:2603.10573v1.

Leon Chlon, Zein Khamis, Fatima Sheaib, Maggie Chlon, Mahdi El Zein, and MarcAntonio M. Awada. LLMs are bayesian in expectation, not realization. *arXiv preprint arXiv:2507.11768v3*, 2026. URL `https://arxiv.org/abs/2507.11768v3`. Version 3, June 23, 2026.

Hakaze Cho, Haolin Yang, Gouki Minegishi, and Naoya Inoue. Mechanism of task-oriented information removal in in-context learning. In *The Fourteenth International Conference on Learning Representations*, 2026. URL `https://openreview.net/forum?id=VAv1rrPR1A`.

Yanai Elazar, Shauli Ravfogel, Alon Jacovi, and Yoav Goldberg. Amnesic probing: Behavioral explanation with amnesic counterfactuals. *Transactions of the Association for Computational Linguistics*, 9:160–175, 2021. doi: 10.1162/tacl_a_00359. URL `https://aclanthology.org/2021.tacl-1.10/`.

Fabian Falck, Ziyu Wang, and Christopher C. Holmes. Is in-context learning in large language models bayesian? A martingale perspective. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 12784–12805. PMLR, 2024. URL `https://proceedings.mlr.press/v235/falck24a.html`.

Lizhe Fang, Yifei Wang, Khashayar Gatmiry, Lei Fang, and Yisen Wang. Rethinking invariance in in-context learning. In *International Conference on Learning Representations*, 2025. URL `https://proceedings.iclr.cc/paper_files/paper/2025/hash/01e8ecf628f8d9f62a1fd433d44d34ab-Abstract-Conference.html`.

Shivam Garg, Dimitris Tsipras, Percy Liang, and Gregory Valiant. What can transformers learn in-context? a case study of simple function classes. In *Advances in Neural Information Processing Systems*, volume 35, 2022. URL https://papers.nips.cc/paper_files/paper/2022/hash/c529dba08a146ea8d6cf715ae8930cbe-Abstract-Conference.html.

Roee Hendel, Mor Geva, and Amir Globerson. In-context learning creates task vectors. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 9318–9333, 2023. doi: 10.18653/v1/2023.findings-emnlp.624. URL https://aclanthology.org/2023.findings-emnlp.624/.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. In *International Conference on Learning Representations*, 2020. URL https://arxiv.org/abs/1912.02781.

Andrew Jesson, Nicolas Beltran-Velez, and David Blei. Can generative AI solve your in-context learning problem? A martingale perspective. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=bcynT7s2du.

Masahiro Kato and Taka Kato. Handover of in-context learning state across session boundaries. *arXiv preprint arXiv:2608.14528v1*, 2026. URL https://arxiv.org/abs/2608.14528v1. Version 1.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8086–8098, 2022. doi: 10.18653/v1/2022.acl-long.556. URL https://aclanthology.org/2022.acl-long.556/.

Vladimir Mikulik, Grégoire Delétang, Tom McGrath, Tim Genewein, Miljan Martic, Shane Legg, and Pedro A. Ortega. Meta-trained agents implement bayes-optimal agents. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/d902c3ce47124c66ce615d5ad9ba304f-Abstract.html.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 11048–11064. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.759. URL https://aclanthology.org/2022.emnlp-main.759/.

Eric Mintun, Alexander Kirillov, and Saining Xie. On interaction between augmentations and corruptions in natural corruption robustness. In *Advances in Neural Information Processing Systems*, volume 34, 2021. URL https://proceedings.neurips.cc/paper/2021/hash/1d49780520898fe37f0cd6b41c5311bf-Abstract.html.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=KSugKcbNf9.

Gustaw Opiełka, Hannes Rosenbusch, and Claire E. Stevenson. Causality $\neq$ invariance: Function and concept vectors in LLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/86b3697c4eb7792c951831636bfdacd5-Abstract-Conference.html.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the bayesian prism. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d81cd83e7f6748af351485d73f305483-Abstract-Conference.html.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 7909–7919, 2020. URL https://proceedings.mlr.press/v119/raghunathan20a.html.

Allan Raventos, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-bayesian in-context learning for regression. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=BtAz4a5xDg.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pp. 3363–3377. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.eacl-main.295. URL https://aclanthology.org/2021.eacl-main.295/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models'sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/6c0e99d736da621403018ca7b32b1a4d-Paper-Conference.pdf.

Aaditya K. Singh, Stephanie C.Y. Chan, Ted Moskovitz, Erin Grant, Andrew M. Saxe, and Felix Hill. The transient nature of emergent in-context learning in transformers. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA, 2023. Curran Associates Inc.

Wei Tang, Xinyan Jiang, Fakhri Karray, and Lijie Hu. In-context learning operates as concept subspace learning. *arXiv preprint arXiv:2605.18830*, 2026. URL https://arxiv.org/abs/2605.18830.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. Function vectors in large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/4ae163cb8788970e53b4fd9578141139-Abstract-Conference.html.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1805.12152.

Johannes von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 35151–35174, 2023. URL https://proceedings.mlr.press/v202/von-oswald23a.html.

Xiaolei Wang, Xinyu Tang, Junyi Li, Wayne Xin Zhao, and Ji-Rong Wen. Investigating the pre-training dynamics of in-context learning: Task recognition vs. task learning. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/83d349b6eb8125588b5f091e2d47525c-Abstract-Conference.html.

Daniel Wurgaft, Ekdeep Singh Lubana, Core Francisco Park, Hidenori Tanaka, Gautam Reddy, and Noah D. Goodman. In-context learning strategies emerge rationally. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://papers.nips.cc/paper_files/paper/2025/hash/abb22b1e4b4ab2988862be9d7ed083f9-Abstract-Conference.html.

Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. Addressing order sensitivity of in-context demonstration examples in causal language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6467–6481, 2024. doi: 10.18653/v1/2024.findings-acl.386. URL https://aclanthology.org/2024.findings-acl.386/.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc V. Le. Unsupervised data augmentation for consistency training. In *Advances in Neural Information Processing Systems*, volume 33, 2020. URL https://proceedings.neurips.cc/paper/2020/hash/44feb0096faa8326192570788b38c1d1-Abstract.html.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=RdJVFCHjUMI.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482, 2019. URL https://proceedings.mlr.press/v97/zhang19p.html.

Yufeng Zhang, Fengzhuo Zhang, Zhuoran Yang, and Zhaoran Wang. What and how does in-context learning learn? Bayesian model averaging, parameterization, and generalization. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 1684–1692. PMLR, 2025. URL https://proceedings.mlr.press/v258/zhang25d.html.