INFORMATIVITY THRESHOLDS FOR IN-CONTEXT LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

When does in-context learning (ICL) succeed, and when does it fail? We give a quantitative answer by building an informativity theory of ICL on Willems' fundamental lemma: the prompt is recorded input–output data, the lemma specifies how much excitation that data must contain before the continuation is determined, and the conditioning of the data sets how error grows as information thins. Measured on this scale, a transformer trained on the task class nearly matches the analytic Bayes predictor and beats the model-free data-enabled predictor by 3–$11\times$ where only identification is possible, collapsing near the point where the prompt first determines the answer. The theory yields one universal law (a prompt that leaves the answer undetermined within the task class defeats every architecture) and one model-relative law: identification provably needs less context information than model-free interpolation, so how much is enough depends on the predictor's prior. Pretrained open-weight LLMs obey the same two laws in two regimes. On latent-system identification probes (2.7B–120B, five families) their error is flat in the prompt's information content, 11–$18\times$ above least squares regardless of scale or instruction tuning and 47–$75\times$ with a 3072-token reasoning budget: the bottleneck is the model's extraction, not the prompt. On short-memory sources, which their prior covers, error falls as context information grows and, where emissions are ambiguous, ends below both fixed-order n-gram baselines (four models, 2.7B–14B). Swapping only the training prior on one fixed transformer moves the same task between the regimes. Prompt engineering can supply information; whether a model can use it is fixed at training time.

1 INTRODUCTION

Large language models predict surprisingly well from a few in-context examples (Brown et al., 2020), and just as often fail on prompts that look equally informative. What separates the two? This paper gives a quantitative answer: ICL succeeds when the prompt carries enough information for the task, and "enough" can be computed once one asks which of two strategies the model uses: *interpolate* the prompt as raw data, recombining past pieces to match the query, or *identify* the latent process behind the examples and simulate it. The two strategies need provably different amounts of context information, which is what makes the question measurable.

The instrument we borrow is Willems' fundamental lemma (Willems et al., 2005), the foundation of data-enabled predictive control (DeePC; Coulson et al., 2019) and a starting point of the theory of data informativity (van Waarde et al., 2020). The lemma states that if the input of a single trajectory of a controllable system is persistently exciting (PE) of an order that grows with the prediction horizon, then every trajectory of that length lies in the column span of a Hankel matrix built from that data: one can predict without ever identifying the system, at the price of needing rich data. Classical system identification is the alternative: estimate the parameters, which needs excitation scaling only with the state dimension, and then simulate forward. Throughout, the lemma's linearity is a condition on the data-generating class, never on the predictor: we use it to compute what a prompt from a known class determines, while the predictor under test may be arbitrary, which is what makes the resulting lower bound learner-agnostic.

The analogy is exact: the ICL prompt plays the role of the offline data, and the requested continuation plays the role of a data-enabled prediction (§3). As the context's informativity varies, does

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.

the error collapse at the behavioral threshold, which grows with the horizon, or at the identification threshold, which does not? A k-tone multisine is PE of order exactly $2k$, so k places both thresholds at known, separated locations; what gates success is the context's informativity (its persistency of excitation and, robustly, its conditioning $\sigma_{\min}$), not the number of examples. A single softmax head can already implement kernel interpolation (Prop. 5); identification is what training adds when the prior makes it Bayes-optimal (Thm. 2), so the training prior, not the architecture, sets the strategy (§5.6). Below the identification threshold the failure is in the data itself: a learner-agnostic lower bound (Thm. 1) caps every architecture. The same analysis resolves pretrained-LLM ICL into two regimes with a measured boundary.

Contributions.

- **An informativity theory of ICL** (§4): a two-threshold dichotomy separating the data requirements of interpolation ($k_{\mathrm{beh}} = \Theta(L)$) and identification ($k_{\mathrm{id}} = \Theta(n)$) (Props. 1–2), a learner-agnostic lower bound under which a prompt that leaves the continuation undetermined within the class defeats every architecture (Thm. 1), a proof that the population-optimal (Bayes) ICL map succeeds already at the smaller threshold (Thm. 2), and a discrete counterpart with a $\mathrm{poly}(r)$ rate and a worst-case exponential separation (Thm. 3, Cor. 2).
- **The predicted phase transition, measured** (§5.1): transformers trained on random linear systems collapse at the horizon-free identification knee ($k \approx 2$–3) and beat model-free DeePC 3–$11\times$ throughout the gap; the early knee holds across order, input distribution, and context length.
- **Conditioning organizes the error past identifiability** (§5.2): an exact negative control shows the rank condition holds throughout the gap, and under noise $\sigma_{\min}/\sigma$ organizes the median dose–response, turning where $\sigma_{\min}$ crosses the noise level (Prop. 3).
- **Bayes confirmation and a discrete bridge** (§5.4): on exact-posterior families the trained transformer nearly matches the analytic Bayes predictor and hedges to the posterior mean; on HMM/WFA sources it identifies the latent predictive state functionally and representationally.
- **Two measured regimes in pretrained LLMs, 2.7B–120B** (§5.5): on identification probes error is flat in prompt information (11–$18\times$ above least squares across scale and tuning, 47–$75\times$ with a 3072-token reasoning budget); on short-memory sources four of them are information-gated (error falls with context in ≥ 22 of 24 settings each) and, on ambiguous emissions, end below both n-gram baselines; the boundary is latent memory depth (head-to-head with Dai et al., 2025).
- **The cause is the training prior** (§5.6): swapping only the training distribution on one fixed transformer flips the mechanism, consistent with the pretrained-LLM pattern.

2 Related work

Mechanisms of in-context learning. ICL has been framed as implicit Bayesian inference over a latent concept/document prior (Xie et al., 2022), as the in-context emulation of learning algorithms such as least squares and gradient descent (Garg et al., 2022; Akyürek et al., 2023; von Oswald et al., 2023; Li et al., 2023), and as a function of pretraining task diversity, with a memorizing-to-generalizing transition as diversity grows (Raventós et al., 2023). Our Theorem 2 specializes the prior-data-fitted-network view (Müller et al., 2022) to a system class. Our account is quantitative about the context itself, extending the $n{=}d$ collapse of Garg et al. (2022) to dynamical systems: we locate the persistency-of-excitation threshold that separates success from failure and show that the trained model sits at the smaller, identification threshold. Earlier work ties the training distribution to whether ICL emerges and which predictor it implements (Chan et al., 2022; Raventós et al., 2023; Wurgaft et al., 2025; Yadlowsky et al., 2023), and transformers can select among in-context algorithms (Bai et al., 2023); we add the identification-versus-kernel-interpolation axis and isolate it with a same-architecture prior swap (§5.6). A complementary mechanistic line attributes much of ICL to induction heads (Olsson et al., 2022) and to function and task vectors (Todd et al., 2024; Hendel et al., 2023); our probes (§5.1, 5.3) tie the decoded object to an identification threshold. The Bayesian reading is developed further by Panwar et al. (2024); transformers simulate fixed automata through shallow shortcuts (Liu et al., 2023), and in-context learning of formal languages is studied by Akyürek et al. (2024).

Data-driven control and informativity. Willems' fundamental lemma (Willems et al., 2005) and its behavioral framing (Markovsky & Dörfler, 2021) underpin DeePC (Coulson et al., 2019); the

generic persistency bound is sharpened in Markovsky et al. (2023), behavioral identifiability is characterized in Markovsky & Dörfler (2023), and data informativity relative to a given analysis or control task is developed in van Waarde et al. (2020); we use these tools to characterize ICL.

ICL for dynamical systems and identification. In-context system identification and state estimation have been studied empirically (Forgione et al., 2023; Piga et al., 2024; Rufolo et al., 2025; Akram & Vikalo, 2026; Du et al., 2023), a transformer can represent the Kalman filter (Goel & Bartlett, 2024), the expressive power of ICL for dynamical systems has been bounded (Cole et al., 2025), and pretrained LLMs forecast dynamical systems better with longer contexts (Liu et al., 2024). This line establishes that a transformer can identify an LTI class in context; we add the two-threshold informativity framing, the conditioning ($\sigma_{\min}$) account of the empirical knee, the training-prior flip, and a lower bound that, unlike expressivity results, is an obstruction in the data (Thm. 1). The discrete version rests on spectral learning of HMMs/WFAs (Hsu et al., 2012; Balle et al., 2014) and the Hankel-rank theorem (Fliess, 1974).

3 Setup: in-context learning as data-enabled prediction

System class. Fix $n \in \mathbb{N}$ and consider single-input single-output (SISO) discrete-time linear time-invariant (LTI) systems

$$x_{t+1} = Ax_t + Bu_t, \qquad y_t = Cx_t + e_t, \tag{1}$$

with state dimension $n_x \leq n$, (A, B) controllable, (A, C) observable, and e_t observation noise. Let $\mathcal{S}_n$ denote this set and π a prior on it (in experiments, n=4, n_x uniform on $\{2, 3, 4\}$, and A, B, C Gaussian with A rescaled to spectral radius $U[0.5, 0.95]$, conditioned on controllability and observability). Write $\ell(S)$ for the lag (observability index).

Prompt and query. A length-T *context* is the interleaved sequence $z_{1:T} = (u_1, y_1, \ldots, u_T, y_T) \in \mathbb{R}^{2T}$. Fix a horizon N and a lag window T_{ini}, and set $L = T_{\text{ini}} + N$. A *query* is a future input $u_f = u_{T+1:T+N}$, and an *ICL predictor* is a map $f : (z_{1:T}, u_f) \mapsto \hat{y}_f \in \mathbb{R}^N$ predicting $y_f = y_{T+1:T+N}$. The trained transformer is one such f, obtained by minimizing the expected next-token loss over π.

Hankel matrices and the informativity knob. For a signal $w_{1:T} \in (\mathbb{R}^q)^T$ and depth L, $\mathscr{H}_L(w) \in \mathbb{R}^{qL \times (T-L+1)}$ stacks the length-L windows of w as columns. An input is *persistently exciting (PE) of order* r if $\operatorname{rank} \mathscr{H}_r(u) = r$; a k-tone multisine with distinct frequencies in $(0, \pi)$ and nonzero amplitudes is PE of order exactly $2k$ once $T \geq 4k - 1$, so k controls context informativity.

Behavioral predictor (DeePC). Partitioning $\mathscr{H}_L(z) = [U_p; Y_p; U_f; Y_f]$ (rows reordered) into past (T_{ini}) and future (N) blocks, the data-enabled predictor is

$$\hat{y}_f^{\text{beh}} = Y_f\, g^\star, \qquad g^\star = M^+ [u_{\text{ini}}; y_{\text{ini}}; u_f], \qquad M := [U_p; Y_p; U_f], \tag{2}$$

the minimum-norm solution of $Mg = [u_{\text{ini}}; y_{\text{ini}}; u_f]$ (least squares if inconsistent), with $(u_{\text{ini}}, y_{\text{ini}})$ the most recent T_{ini}-window of the context. Equation (2) predicts from the context without identifying (A, B, C); the competing strategy fits S by ARX least squares and simulates forward.

4 Theory

We first treat the noiseless case ($e_t \equiv 0$). Let $\mathcal{B}_L(S) \subseteq \mathbb{R}^{2L}$ be the length-L behavior of S; for n_x states and $L \geq \ell(S)$, $\dim \mathcal{B}_L(S) = L + n_x$. Full proofs are in Appendix B. The ingredients are standard, specialized to ICL: Willems' lemma and ARX identifiability (Ljung, 1999) (Props. 1–2), an indistinguishability lower bound (Thm. 1), and the squared-loss Bayes act (Thm. 2).

4.1 Two thresholds: behavioral vs. identification

Proposition 1 (Fundamental lemma, ICL form)**.** *If the context input $u_{1:T}$ is PE of order $L + n_x$ then $\operatorname{col} \mathscr{H}_L(z_{1:T}) = \mathcal{B}_L(S)$, and whenever $T_{\text{ini}} \geq \ell$ the behavioral predictor is exact: $\hat{y}_f^{\text{beh}} = y_f(S)$.*

For a k-tone context the classical sufficient condition is $k \geq \lceil (L+n_x)/2 \rceil$; PE of order L is necessary and, for almost every initial state, sufficient (Markovsky et al., 2023), so the generic requirement is $k \geq k_{\mathrm{beh}} := \lceil L/2 \rceil$, and this is where the collapse occurs in our experiments. The point is that $k_{\mathrm{beh}} = \Theta(L)$ grows with the horizon.

Proposition 2 (Identification sufficiency)**.** *Let $S \in \mathcal{S}_n$ with known order bound n, $T \geq 4n$, and $k_{\mathrm{id}} := \lceil n/2 \rceil$. For almost every initial state, $k \geq k_{\mathrm{id}}$ (PE of order n) makes the length-n windows of $z_{1:T}$ span $\mathcal{B}_n(S)$ (Markovsky et al., 2023); $k \geq n$ (PE of order $2n$) suffices for every initial state. Every least-squares ARX(n) fit (unique iff $n_x = n$) then rolls out the true continuation: $\hat{y}_f^{\mathrm{id}} = y_f(S)$ for every query and horizon, and $k_{\mathrm{id}} = \Theta(n)$ is independent of L.*

Theorem 1 (Separation and a learner-agnostic lower bound)**.** *Let $T \geq 4n$ and $L \geq n+2$, so that $k_{\mathrm{id}} < k_{\mathrm{beh}}$. For every $k \in [k_{\mathrm{id}}, k_{\mathrm{beh}})$, almost every initial state, and a generic system and query: (i) the context determines S ($\hat{y}_f^{\mathrm{id}} = y_f(S)$); yet (ii) $\operatorname{col} \mathscr{H}_L(z) \subsetneq \mathcal{B}_L(S)$, the system $Mg = [u_{\mathrm{ini}}; y_{\mathrm{ini}}; u_f]$ is inconsistent, and the behavioral predictor errs. Whenever two systems in $\mathcal{S}_n$ are consistent with $z_{1:T}$ but disagree on the continuation of u_f (generically the case for $k < k_{\mathrm{id}}$ and, for almost every initial state, never for $k \geq k_{\mathrm{id}}$), every predictor f incurs*

$$\max_j \left\| f(z_{1:T}, u_f) - y_f(S_j) \right\|_2 \;\geq\; \tfrac{1}{2} \| y_f(S_1) - y_f(S_2) \|_2 > 0. \tag{3}$$

4.2 The training objective selects identification

Theorem 2 (Bayes-optimal ICL is implicit identification)**.** *The minimizer of the population risk $\mathcal{R}(f) = \mathbb{E}_{S\sim\pi} \mathbb{E}_{x_0,u,e} \| f(z_{1:T}, u_f) - y_f \|_2^2$ is the posterior mean $f^\star(z_{1:T}, u_f) = \mathbb{E}[y_f \mid z_{1:T}, u_f] = \int_{\mathcal{S}_n} \mathbb{E}[y_f \mid S, \cdot]\, d\pi(S \mid z_{1:T})$. In the noiseless regime, if $T \geq 4n$ and $k \geq k_{\mathrm{id}}$ the posterior concentrates (for almost every initial state) on the identified $\hat{S}$ and $f^\star = y_f(\hat{S}) = \hat{y}_f^{\mathrm{id}}$. Hence the optimal error vanishes for every $k \geq k_{\mathrm{id}} = \Theta(n)$, independently of L, whereas the behavioral predictor needs $k \geq k_{\mathrm{beh}} = \Theta(L)$.*

A transformer trained to the Bayes act, with generic queries in its training law, therefore dominates model-free DeePC throughout the gap $[k_{\mathrm{id}}, k_{\mathrm{beh}})$, where the Bayes predictor is exact (Cor. 1).

4.3 Conditioning, noise, attention, and the discrete version

Four further results (Apps. A–B) complete the picture.

First, past identifiability the error is governed by conditioning (Prop. 3): the true-order ARX regressor is full rank for every $k \geq k_{\mathrm{id}}$ (an exact negative control, §5.2), and under a small perturbation δ the least-squares error is $O(\delta/\sigma_{\min}(\Phi_k))$, guaranteed small once $\sigma_{\min}(k)$ rises well above δ.

Second, identification is robust under noise (Prop. 4): for $\sigma \ll \sigma_{\min}(\Phi)$ its excess error is $O(\sigma/\sigma_{\min}(\Phi))$, whereas DeePC errs in the gap already at $\sigma = 0$ and just above k_{beh} amplifies noise as $O(\sigma/\sigma_{\min}(M))$. Both are one $\sigma_{\min}$ bound at two perturbation scales (Rem. 2), so under noise the dose–response is organized by $\sigma_{\min}/\sigma$.

Third, kernel interpolation is within attention's reach (Prop. 5). A single softmax head with past-window keys and values computes a Nadaraya–Watson smoother (Goel & Bartlett, 2024), a local-averaging counterpart of the kernel-regression form of the behavioral predictor (Molodchyk & Faulwasser, 2024). This is the interpolation side of the dichotomy, which the Bayes predictor provably exceeds by identifying (Thm. 2) and where LLMs stay on identification probes (§5.5).

Fourth, the dichotomy has a discrete version (Thm. 3, Cor. 2). For rank-r WFA/HMMs the Hankel rank equals the predictive (PSR) rank r (Fliess, 1974), and given a margin $\sigma_r \geq \sigma_0$, Bayes-optimal ICL reaches π-averaged excess KL $\leq \varepsilon$ at a context length polynomial in r, $1/\sigma_0$ and $1/\varepsilon$, with a factor in $|\Sigma|$ and the mixing time but no horizon or $|\Sigma|^d$ factor. A count-based order-d predictor instead needs $T \gtrsim |\Sigma|^{\tau(\varepsilon)}$ in the worst case, with $\tau(\varepsilon)$ the source's ε-memory depth: exponentially more context on long-memory sources.

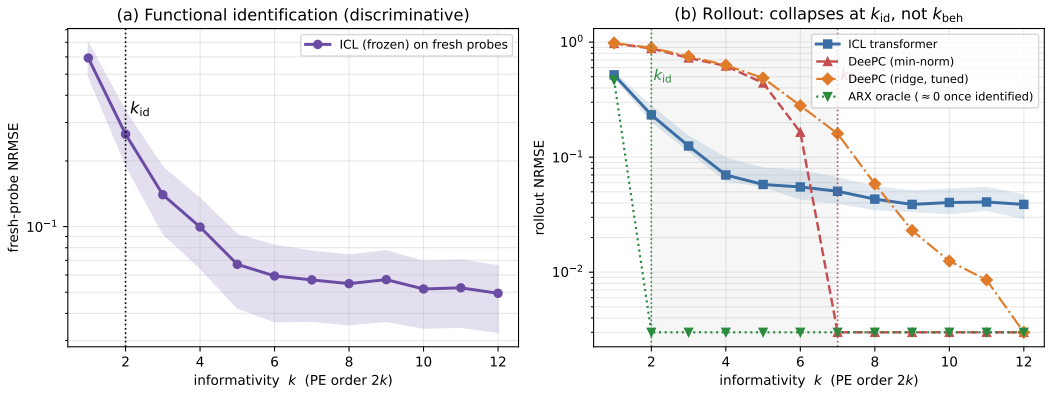


Figure 1: **ICL functionally identifies, collapsing at the identification knee.** (a) Fresh-probe functional identification: from one context the frozen model predicts $M{=}20$ independent probe responses, a task that requires a reusable system model; NRMSE collapses $0.59 \to 0.05$ with a knee at $k \approx 2$–3 (one training seed; band: within-context spread). (b) Rollout NRMSE under a white-noise query (log axis; 5-seed median, band = range): the transformer drops $\sim 7\times$ by $k{=}4$ and matches the ARX oracle's knee at its ~ 0.04 capacity floor, whereas min-norm DeePC collapses only at its rank knee $k_{\mathrm{beh}} = 7$ and ridge-tuned DeePC improves gradually, overtaking the transformer past k_{beh}.

5 Experiments

The experiments follow one line of argument: trained transformers identify (§5.1–5.4), pretrained LLMs split into a flat and an information-gated regime (§5.5), and a same-architecture flip shows that the cause is the prior rather than the architecture (§5.6). We train decoder-only transformers on interleaved (u, y) tokens (continuous inputs linearly embedded; the discrete model uses a symbol embedding and softmax head), predicting the next output autoregressively. The main LTI model is a 21M-parameter GPT ($d{=}384$, 12 layers); a 2.7M model and matched oracles (DeePC, ARX(n_x), Kalman observer, exact posterior) are used throughout. Full details in Appendix D; all trained-model LTI rollout claims are medians over ≥ 256 held-out systems, replicated across seeds. Table 4 lists each experiment's models, sample and seed counts, metric, and scope.

5.1 The phase transition: ICL tracks identification, not DeePC

Trained transformers identify: their error collapses at the horizon-free knee, and in rollout they beat model-free DeePC throughout the gap. The discriminating test is functional identification on fresh probes (Fig. 1a). From a single length-T context ($T{=}256$), the frozen model must predict the same system's response to $M{=}20$ independent probe inputs that are unrelated to the context input. Succeeding here requires a reusable model of the system: a behavioral predictor can only recombine the trajectory it was handed. Fresh-probe NRMSE collapses $0.59 \to 0.27 \to 0.14 \to 0.05$ with a knee at $k \approx 2$–3 (near the identification threshold $k_{\mathrm{id}} {=} \lceil n/2 \rceil {=} 2$; §5.2), and the within-context spread is small ($0.10 \to 0.017$): the model has identified the system.

Rollout under a fresh white-noise query input (Fig. 1b) then tests the threshold locations directly. On a log axis the transformer's error drops $\sim 7\times$ by $k{=}4$ and floors at ~ 0.04, with its knee at the conditioning ($\sigma_{\min}$) floor at $k \approx 2$, far before the behavioral rank knee ($k_{\mathrm{beh}} = 7$, $L = 13$) where min-norm DeePC collapses. In the gap $k \in \{2, \ldots, 6\}$ it matches the ARX oracle's identification knee (at its own ~ 0.04 capacity floor; the oracle is exact) and beats DeePC 3–11$\times$ (e.g. $n_x{=}3$, $k{=}4$: 0.07 vs. 0.62). It beats ridge-regularized DeePC by a similar margin (3–9$\times$): the obstruction in the gap is that the data underdetermine the behavior, so ridge regularization cannot supply the missing behavioral rank (Rem. 1 extends this to subspace predictive control and regularized DeePC variants (Dörfler et al., 2023; Breschi et al., 2023)). A one-step readout, by contrast, is nearly flat in k: informativity is relative to the task. The effect is robust across $n_x \in \{2, 3, 4\}$ and across multisine vs. white training inputs. This is Cor. 1's prediction, met out of support (training never pairs a multisine context with a white query).

5.2 IDENTIFIABILITY VS. CONDITIONING

Context length moves the DeePC knee. Retraining at short contexts ($T_{\mathrm{ctx}} \in \{16, 32\}$, $L{=}T_{\mathrm{ini}}{+}N{=}12$) separates the two thresholds sharply (Fig. 4, App. C): the transformer's knee stays pinned at $k \approx 2$ regardless of context length, while DeePC never collapses at $T_{\mathrm{ctx}}{=}16$ (too few Hankel columns) and collapses at its rank knee $k{\approx}6$ at $T_{\mathrm{ctx}}{=}32$, as the generic thresholds predict. A horizon$\times$order sweep confirms the scaling (Fig. 15, App. D): DeePC's knee rises with L and its collapse sits at $\lceil L/2 \rceil$ in every cell ($k_{\mathrm{beh}} = \Theta(L)$), while the transformer's knee is flat in L, tracks the ARX oracle, and on the order axis stays at $k \approx$ 2–3, within one tone of $\lceil n_x/2 \rceil$, rather than climbing with n_x.

Conditioning, not identifiability. An exact negative control (App. D) shows that the system is identifiable from $k{=}k_{\mathrm{id}}{=}2$ at long context: the knee at $k{\approx}2$–3 is far before the behavioral threshold k_{beh}, and past k_{id} the error keeps falling as the regressor's conditioning improves. Under observation noise conditioning organizes the median dose–response: error rises as the identification $\sigma_{\min}$ falls, and rescaling by the perturbation aligns the noisy medians onto one curve in $\sigma_{\min}/\sigma$ (Fig. 14, App. C; matched-SNR medians $\approx$0.3 $\to$ 0.03–0.07, transition at $\sigma_{\min}/\sigma \approx 0.3$; the alignment is of medians, with wide per-system spread quantified in the caption). Without noise the per-system error is flat in $\sigma_{\min}$ (median $\approx$ 0.06 over full-rank contexts); only the medians across k track the conditioning, down to the $\sim$0.04 capacity floor.

5.3 ROBUSTNESS AND REPRESENTATION

Identification degrades gracefully under noise exactly where DeePC is brittle, as Prop. 4 predicts (Fig. 5, App. C): unregularized DeePC spikes at its rank knee (NRMSE 3.4 at noise 0.3), ARX degrades as $O(\sigma)$, and noise-matched transformers keep the early knee with a nearly noise-independent floor, beating model-free DeePC 7.5–10$\times$ at $k{=}4$ across all noise levels. On the representation side, a linear probe of the frozen hidden state decodes the gauge-invariant impulse-response shape, and this decodability is learned rather than generic (best-layer R^2 0.20 $\to$ 0.41 with k, vs. $\leq$0.06 at random init).

5.4 DIRECT BAYESIAN CONFIRMATION AND A DISCRETE BRIDGE

Exact-posterior LTI families. To test the Bayesian predictions (Thms. 1, 2) directly, we use one- and two-parameter families with analytic posteriors: $x_{t+1} = a\,x_t + u_t$, $y_t = x_t + e_t$, with unknown pole $a \sim U[0.4, 0.9]$ (AR(1)) or two unknown poles $p_1, p_2 \sim U[0.2, 0.9]$ (AR(2)), and with the input amplitude as the informativity knob. On both families the trained transformer is near-Bayes-optimal: its error equals the analytic Bayes-predictor error to within the posterior-spread floor (error/floor $\approx$ 0.8–1.2 up to amplitude $\approx$2 in 1-D and $\approx$1 in 2-D; §6). It also hedges: the gap to the posterior mean stays small (0.005–0.09 in 1-D, 0.01–0.22 in 2-D) while the gap to the prior mean grows with informativity (to 0.82 and 2.5), and at low SNR the model falls back to the prior mean. The exact-posterior results are in Fig. 6 (App. C); sampling the high-dimensional ARX null space as a posterior is biased (App. D).

A discrete bridge: HMM/WFA identification. We repeat the analysis on discrete symbols from long-memory cyclic HMMs (rank $r \in \{2, 3, 4\}$, $|\Sigma|{=}3$), the canonical language-like source. An exact NumPy computation on random rank-r HMMs confirms the discrete dichotomy: the Hankel rank saturates at r, spectral identification reaches machine precision at basis rank r, and no finite-order Markov model does. Trained 21M-parameter symbol transformers (Fig. 7, App. C) drive next-symbol KL toward zero, beat a fair empirical back-off n-gram at every context length (3.5–4.5$\times$ at $T_{\mathrm{ctx}}{=}128$), and drop below the order-1 Markov oracle at long context. They capture latent rank-r structure rather than surface n-grams, the context they need grows with r, consistent with the $\mathrm{poly}(r)$ rate (Thm. 3), and for $T_{\mathrm{ctx}} \geq 8$ a linear probe decodes the predictive state at $R^2 \approx 0.83$–0.98, above raw context (0.56–0.87) and far above random init (0.25–0.57) (Fig. 7).

5.5 BEHAVIORAL VS. IDENTIFICATION IN REAL PRETRAINED LLMS

Pretrained LLMs need probes that separate mechanism from arithmetic and tokenization competence, so we pair numeric tests with a confound-free symbolic test on frozen off-the-shelf models (OPT 6.7B–66B, Qwen2.5 7B–32B base and 7B/14B INSTRUCT, Mistral-7B, phi-2, gpt-oss-20b/120b, QwQ-32B, R1-Distill-Qwen-32B; inference only).

Numeric and symbolic-stream tasks: LLMs predict behaviorally. On the canonical regression testbed (Garg et al., 2022), where the prompt holds n pairs from $y = w^\top x$, ordinary least squares (OLS) is the identification predictor and Nadaraya–Watson (NW) and 1-nearest-neighbor (1-NN) are the behavioral ones, every model predicts behaviorally: within $1.8\times$ of 1-NN, 11–$18\times$ off OLS, with a flat curve through the identifiability threshold $n{=}d$. The same pattern holds across scale, RLHF, and six prompt formats, and it survives explicit reasoning: chain-of-thought matches or increases the direct error on the INSTRUCT models, and the reasoning-trained QwQ-32B and R1-Distill-Qwen-32B, given a 3072-token thinking budget (100% parse), remain 47–$75\times$ off OLS and 2–$3\times$ above the 1-NN error, the trace worsening the point prediction (Table 3; Fig. 3c). A PCA probe of Qwen2.5-7B finds the NW prediction far more decodable than the OLS answer (best layer: $R^2 \approx 0.7$–0.88 vs. 0.11–0.43; Fig. 12, App. C), consistent with attention as kernel smoothing (Prop. 5; Han et al., 2025). At fixed n we sweep the design conditioning $\kappa = \sigma_{\max}/\sigma_{\min}(X)$: OLS follows it closely, near-exact at $\kappa{=}1$ and an order of magnitude worse by $\kappa{=}50$, while every LLM stays within $\sim 2\times$ of 1-NN across all κ (Fig. 10, App. C). LLM regression error is therefore flat in both example count and conditioning, matching the interpolation side of the dichotomy (§4.3). On long-memory cyclic HMM streams the split reappears under the discriminating metric: every LLM reaches near-optimal accuracy, yet in KL it sits near or above the fair same-context n-gram (0.8–$3.4\times$ it; Qwen2.5 dips just below, OPT 6.7B–66B is flat in scale), and above the $\sim 4\times$ lower KL (ratio 0.22) that a class-trained transformer attains (Fig. 13, App. C).

Head-to-head with Dai et al. (2025): the boundary is latent memory depth. Dai et al. (2025) report that pretrained LLMs predict HMMs in-context near-optimally. We sweep random HMMs over two of their axes, mixing $|\lambda_2| \in \{0.5, 0.75, 0.95, 0.99\}$ and emission entropy, and score with their accuracy metric against the Bayes optimum (they also report Hellinger distance). We reproduce their finding (Fig. 8, App. C): at $T_{\text{ctx}}{=}128$ the Qwen models reach 0.9–$1.0\times$ the optimal accuracy over most of the grid, including the slow-mixing column, and with $M{=}8$ states dip to $\approx 0.75\times$ only at near-maximal emission entropy, where the Bayes optimum itself is low, and beat the same-context n-gram. The reconciliation is latent structure: their random HMMs are short-memory (the sticky state is recoverable locally, and the LLMs recover it), while our cyclic class hides a period-r phase that only whole-context aggregation reveals. Within HMMs the split is thus set by latent memory depth: pretrained LLMs track latent state on short-memory sources, while long-memory automaton identification is reached only by the class-trained transformer (§5.4).

Inside their working regime, LLM error is information-gated. On the short-memory family, where the head-to-head shows the LLMs recover latent state, we measure how error scales with context. We sweep the context length $T \in \{8, \ldots, 256\}$ across the full mixing$\times$entropy grid (24 settings, 150 trials each) and score next-symbol KL to the true filter (Fig. 2). Error falls with T in every setting for Qwen2.5-7B, Qwen2.5-14B, and Mistral-7B (24/24 each; pooled log-log slopes -0.56 to -0.60; phi-2: 22/24, -0.34): inside this regime, LLM in-context prediction is information-gated. The gain over surface statistics appears exactly where latent inference is required. On ambiguous-emission settings the LLMs cross below the same-context back-off n-gram between $T{=}16$ and 32; Qwen and Mistral end 3.6–$4.3\times$ below it and phi-2 $1.9\times$ (Qwen2.5-7B: KL 0.025 vs. 0.109 at order 2, 0.141 at order 3), while both n-gram orders stay between 0.109 and 0.144 at every context length; on sharp-emission settings, where the surface suffices, Qwen and Mistral match it (≈ 0.03). A spectral-WFA identifier given the same single trajectory stays far above the LLMs (slow-mixing median KL 2.2 at $T{=}8$, still 0.24 at $T{=}256$): the LLM extracts more from one context than the textbook estimator. Together with the flat curves, this completes the informativity account: the prompt decides the outcome inside the regime the training prior covers, and the model's extraction decides it outside (protocol and pre-registered criteria in App. D; the combined surface-separation criterion missed its n-gram-flatness gate).

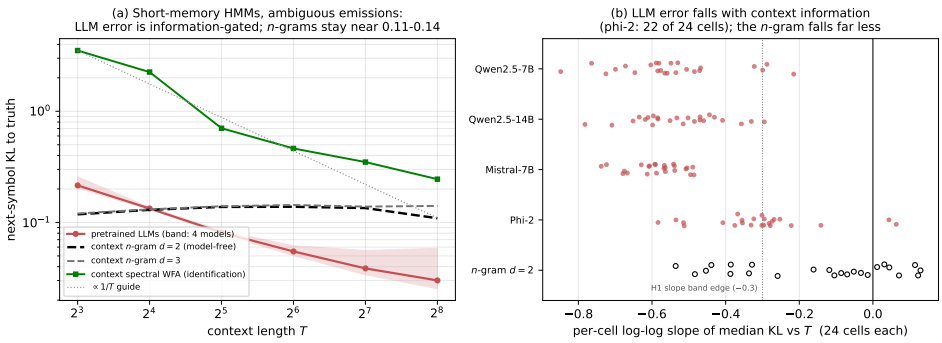


Figure 2: **Inside their working regime, pretrained LLMs are information-gated** (§5.5; short-memory HMMs, four models). (a) Ambiguous-emission settings: the LLM band is the only curve that is both low and falling; its median crosses below the order-2 and order-3 context n-grams between $T{=}16$ and 32 and ends $3.6\times$ and $4.7\times$ below them, while a single-trajectory spectral-WFA identifier stays far above. (b) Per-setting log-log slopes of median KL vs. T: Qwen2.5-7B/14B and Mistral-7B fall in every setting (near -0.55), phi-2 in 22 of 24; the n-gram straddles zero.

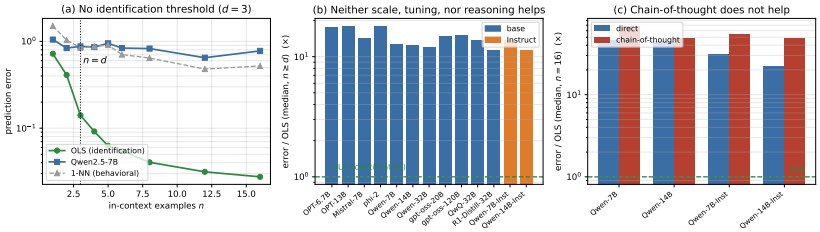


Figure 3: **Pretrained LLMs do behavioral in-context regression, base and INSTRUCT alike.** (a) OLS collapses at the identifiability threshold $n = d$ while the LLM (and 1-NN) stay flat. (b) Error/OLS at $n \geq d$: every base and RLHF-INSTRUCT model is $\sim$11–18$\times$ off OLS, with no trend in scale, tuning, or reasoning training (OPT-30B omitted: 0% parseable). (c) Chain-of-thought matches or increases the direct error at $n{=}16$ for base and INSTRUCT Qwen 7B/14B.

The theory's core object, and a confound-free ladder. On continuous LTI/DeePC trajectories a frozen LLM shown an interleaved (u, y) trajectory predicts far from the ARX identifier (NRMSE $\approx$0.7 vs. $\approx$0, flat in k) and closer to the NW kernel for every model (per-model medians; $68/88$ cells; Fig. 9, App. C; parse $\geq 98\%$). To remove the numeric confound entirely we use symbolic, arithmetic-free rules read from logits (parse 100%), with a fresh query whose nearest seen exemplars unanimously disagree with the rule answer, so that matching the rule requires identification (Fig. 11, App. C; identification index 0 chance, 1 perfect, 4 seeds). On copy (out=key$[c^\star]$), every model identifies (index 0.45–0.79, OPT included). On the abstract map (a hidden bijection) there is a spectrum: Qwen2.5/Mistral reach 0.21–0.37, OPT stays at chance. On conjunction (a two-feature pattern), all tested models are at chance. Re-dressed as CVCV pseudowords under a vowel-harmony rule, the partial identifiers revert to copying the nearest word (mean rule accuracy 0.11, nearest-word match 0.83). In-context identification in pretrained LLMs is graded by distance from the copy/induction primitive (Olsson et al., 2022) (template-stable, App. D) and easily overridden by surface similarity; the latent-system identification our theory concerns sits at the hard end, reached by the class-trained transformer and no pretrained LLM.

5.6 THE MECHANISM IS SET BY THE TRAINING PRIOR: A SAME-ARCHITECTURE FLIP

Changing only the training prior on one fixed architecture (same GPT, optimizer, next-output objective) flips the mechanism: trained on the LTI class, where identification is Bayes-optimal, the model tracks the ARX oracle; trained on a non-identifiable Gaussian-process map $y_t = f(u_t)$, where kernel smoothing is Bayes-optimal, it tracks the GP posterior (Table 1). The mechanism follows the training prior rather than the test data: on perfectly identifiable LTI data, the GP-trained

Table 1: **Mechanism is set by the training prior, not the architecture.** One 21M GPT architecture, two training and two test priors. Mean ± std over 5 seeds of per-seed median gaps to the ARX and GP-posterior oracles; the lower (bold) gap is the implemented mechanism; the decisive cell is GP→LTI.

Train prior	Test prior	Gap to ARX identifier	Gap to GP smoother	Mechanism
LTI	LTI	$\mathbf{0.009 \pm 0.001}$	0.794 ± 0.002	identification
LTI	GP	$\mathbf{0.336 \pm 0.022}$	0.843 ± 0.012	identification
GP	LTI	0.802 ± 0.017	$\mathbf{0.223 \pm 0.010}$	interpolation
GP	GP	0.854 ± 0.006	$\mathbf{0.009 \pm 0.002}$	interpolation

network remains much closer to the interpolation oracle than to the identifier (gap 0.80 ± 0.02 to ARX, vs. 0.009 for the LTI-trained net; 5 seeds), matching the pretrained-LLM pattern: the training distribution selects the mechanism because the Bayes act is relative to π (Thm. 2). This refines the in-context least-squares and gradient-descent account (Akyürek et al., 2023; von Oswald et al., 2023): it holds when training makes identification Bayes-optimal, while general pretraining yields the partial identification of §5.5.

6 Discussion and limitations

Our account makes ICL on dynamical systems measurable: under noise, the "knee" practitioners observe tracks the conditioning $\sigma_{\min}$ of the in-context data. For pretrained LLMs the analysis has a practical reading: inside the regime the training prior covers, richer context pays off; outside it, added informativity is wasted, and neither prompt engineering, scale, nor elicited reasoning substitutes for a covering prior.

Several limitations bound the internal results. On the exact-posterior families the transformer is near-Bayes-optimal over most of the informativity range (error/floor ≈ 0.8–1.2) but drifts off the floor at high amplitude, reaching 1.9 in 1-D and 2.2 in 2-D; a widening control (6.0 amplitude range, 3 seeds) reduces the drift only partially ($1.6/2.0$), pointing to a precision limit as the posterior sharpens. The LTI representational probe gives partial linear decodability ($R^2 \approx 0.4$), with the functional fresh-probe evidence decisive and the discrete predictive-state probe strong ($R^2 \approx 0.9$).

The controlled systems are SISO with synthetic priors; MIMO systems, nonlinear Koopman-embeddable classes, and natural-language priors would extend the same informativity question to richer settings. The LLM conclusions cover open-weight models from 2.7B to 120B (five families, including MoE and reasoning-trained models) on synthetic probes; the information-gated regime was measured on four of them (2.7B–14B). Frontier closed models and some larger open ones, which can rival ensemble regressors and beat k-NN on low-dimensional regression (Vacareanu et al., 2024), still need evaluation under the same informativity-controlled probes. Finally, the learned identifier extrapolates across system order: a NARROW model trained only on $n_x{=}3$ matches or beats the WIDE model (gap $\leq +0.1\sigma$) at every $n_x \in \{2, \ldots, 6\}$ (App. D); the prior sets the mechanism, not the identifier's reach.

7 Conclusion

Treating the ICL prompt as the recorded data of a behavioral system makes the success of ICL provable and measurable: interpolation needs context information that grows with the horizon, identification needs a small, horizon-free amount gated by the conditioning $\sigma_{\min}$, and the training prior sets which one a model runs. Pretrained LLMs follow the same laws in two measured regimes, flat outside the prior's coverage and falling with context information inside it. The prompt decides what a model can know; the training prior decides what it can use.

AI USE STATEMENT

We used generative AI throughout this project: Claude (Anthropic), mainly through the Claude Code agent. The pretrained language models evaluated in §5.5 are objects of study, not research assistants.

Tasks with required disclosure. We used generative AI tools to help us develop the conceptual framework, propose and refine hypotheses, design experiments and improve the research methodology, implement the method and the experiments (including generating the synthetic data), format the synthetic data and examples into prompts for evaluating the language models, and interpret results. We also used generative AI tools to provide critical ingredients for proving the mathematical claims and to assist in writing the proofs; we formulated the claims ourselves, without generative AI tools. We did not use generative AI tools to translate literature or to perform qualitative or thematic analysis.

Tasks with recommended disclosure. We also used generative AI tools to help us suggest experimental parameters, write and edit code (including training, evaluation, analysis, plotting, table-generation, and cluster job scripts), generate scientific figures through scripts, revise the text of the paper for readability, search for information, organize references and check their formatting, and suggest titles for the paper.

Review. The authors reviewed all AI-assisted content item by item, including text, mathematics, code, analyses, figures, and references. Before submission we recomputed the paper's numbers from the logged results and corrected those that did not match. All values in tables and figures were computed by analysis scripts from the logged experimental results. The short-memory informativity sweep was pre-registered: we fixed the hypotheses and pass criteria before the experiments began and used a random-predictor control to verify that the criteria detect failure. Verdicts were computed automatically by script, and App. D lists which criteria were met. Before testing the pretrained LLMs at scale, we compared alternative inference engines with the single-GPU Hugging Face inference path on pre-registered acceptance criteria and adopted only that path. All figures were generated by plotting scripts from logged results; we did not use image-generation models to produce any image, and no figure was retouched. We take responsibility for the final content of this work, including all text, arguments, and conclusions produced with the aid of generative AI.

ETHICS STATEMENT

This work analyzes the in-context prediction mechanism of small transformers on synthetic dynamical-system and automaton sources, and evaluates open pretrained language models in an inference-only setting. It uses no human subjects, personal data, or sensitive content.

REPRODUCIBILITY STATEMENT

The main experiments are specified for reproduction. Model and training hyperparameters, oracles, baselines, input/PE construction, exact-posterior grids, and the empirical-knee definition are in Appendix D; each trained-model LTI sweep point is a median over ≥ 256 held-out systems, and Table 4 gives the sample and seed counts of the main experiments together with their models and scope. All data are generated on the fly from the stated priors. Appendix D also gives the complete pretrained-LLM protocol: checkpoints, dtype, single-GPU requirement, decoding, prompt templates, and parse rules. Compute is a Slurm GPU cluster (H100/H200). Training, evaluation, and figure-generation code is provided as anonymized supplementary material and will be publicly released upon acceptance.

REFERENCES

Usman Akram and Haris Vikalo. Transformers as implicit state estimators: In-context learning in dynamical systems. *Transactions on Machine Learning Research*, 2026.

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 (ICLR)*, 2023.

Ekin Akyürek, Bailin Wang, Yoon Kim, and Jacob Andreas. In-context language learning: Architectures and algorithms. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 787–812, 2024.

Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 57125–57211, 2023. doi: 10.52202/075280-2498.

Borja Balle and Odalric-Ambrym Maillard. Spectral learning from a single trajectory under finite-state policies. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pp. 361–370, 2017.

Borja Balle, Xavier Carreras, Franco M. Luque, and Ariadna Quattoni. Spectral learning of weighted automata: A forward-backward perspective. *Machine Learning*, 96(1–2):33–63, 2014. doi: 10.1007/s10994-013-5416-x.

Valentina Breschi, Alessandro Chiuso, and Simone Formentin. Data-driven predictive control in a stochastic setting: A unified framework. *Automatica*, 152:110961, 2023. doi: 10.1016/j.automatica.2023.110961.

Tom B. Brown et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pp. 1877–1901, 2020.

Stephanie C. Y. Chan, Adam Santoro, Andrew K. Lampinen, Jane X. Wang, Aaditya K. Singh, Pierre H. Richemond, James L. McClelland, and Felix Hill. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 35, pp. 18878–18891, 2022. doi: 10.52202/068431-1371.

Frank Cole, Yuxuan Zhao, Yulong Lu, and Tianhao Zhang. In-context learning of linear dynamical systems with transformers: Approximation bounds and depth-separation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 187641–187679, 2025. doi: 10.52202/085713-5641.

Jeremy Coulson, John Lygeros, and Florian Dörfler. Data-enabled predictive control: In the shallows of the DeePC. In *18th European Control Conference (ECC)*, pp. 307–312, 2019. doi: 10.23919/ECC.2019.8795639.

Yijia Dai, Zhaolin Gao, Yahya Sattar, Sarah Dean, and Jennifer J. Sun. Pre-trained large language models learn to predict hidden Markov models in-context. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 38, pp. 32682–32728, 2025. doi: 10.52202/085713-0970.

Florian Dörfler, Jeremy Coulson, and Ivan Markovsky. Bridging direct and indirect data-driven control formulations via regularizations and relaxations. *IEEE Transactions on Automatic Control*, 68(2):883–897, 2023. doi: 10.1109/TAC.2022.3148374.

Zhe Du, Haldun Balim, Samet Oymak, and Necmiye Ozay. Can transformers learn optimal filtering for unknown systems? *IEEE Control Systems Letters*, 7:3525–3530, 2023. doi: 10.1109/LCSYS.2023.3335318.

Michel Fliess. Matrices de Hankel. *Journal de Mathématiques Pures et Appliquées*, 53:197–222, 1974.

Marco Forgione, Filippo Pura, and Dario Piga. From system models to class models: An in-context learning paradigm. *IEEE Control Systems Letters*, 7:3513–3518, 2023. doi: 10.1109/LCSYS.2023.3335036.

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 (NeurIPS)*, volume 35, pp. 30583–30598, 2022. doi: 10.52202/068431-2217.

Gautam Goel and Peter Bartlett. Can a transformer represent a Kalman filter? In *Proceedings of the 6th Annual Learning for Dynamics & Control Conference (L4DC)*, volume 242 of *Proceedings of Machine Learning Research*, pp. 1502–1512, 2024.

Chi Han, Ziqi Wang, Han Zhao, and Heng Ji. Understanding emergent in-context learning from a kernel regression perspective. *Transactions on Machine Learning Research*, 2025.

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.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. A spectral algorithm for learning hidden Markov models. *Journal of Computer and System Sciences*, 78(5):1460–1480, 2012. doi: 10.1016/j.jcss.2011.12.025.

Yingcong Li, Muhammed Emrullah Ildiz, Dimitris Papailiopoulos, and Samet Oymak. Transformers as algorithms: Generalization and stability in in-context learning. In *Proceedings of the 40th International Conference on Machine Learning (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19565–19594, 2023.

Bingbin Liu, Jordan T. Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. In *International Conference on Learning Representations (ICLR)*, 2023.

Toni J. B. Liu, Nicolas Boullé, Raphaël Sarfati, and Christopher Earls. LLMs learn governing principles of dynamical systems, revealing an in-context neural scaling law. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 15097–15117, 2024. doi: 10.18653/v1/2024.emnlp-main.842.

Lennart Ljung. *System Identification: Theory for the User*. Prentice Hall PTR, 2nd edition, 1999.

Ivan Markovsky and Florian Dörfler. Behavioral systems theory in data-driven analysis, signal processing, and control. *Annual Reviews in Control*, 52:42–64, 2021. doi: 10.1016/j.arcontrol.2021.09.005.

Ivan Markovsky and Florian Dörfler. Identifiability in the behavioral setting. *IEEE Transactions on Automatic Control*, 68(3):1667–1677, 2023. doi: 10.1109/TAC.2022.3209954.

Ivan Markovsky, Eduardo Prieto-Araujo, and Florian Dörfler. On the persistency of excitation. *Automatica*, 147:110657, 2023. doi: 10.1016/j.automatica.2022.110657.

Oleksii Molodchyk and Timm Faulwasser. Exploring the links between the fundamental lemma and kernel regression. *IEEE Control Systems Letters*, 8:2045–2050, 2024. doi: 10.1109/LCSYS.2024.3406053.

Samuel Müller, Noah Hollmann, Sebastian Pineda Arango, Josif Grabocka, and Frank Hutter. Transformers can do Bayesian inference. In *International Conference on Learning Representations (ICLR)*, 2022.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, et al. In-context learning and induction heads. *Transformer Circuits Thread*, 2022. arXiv:2209.11895.

Madhur Panwar, Kabir Ahuja, and Navin Goyal. In-context learning through the Bayesian prism. In *International Conference on Learning Representations (ICLR)*, 2024.

Dario Piga, Filippo Pura, and Marco Forgione. On the adaptation of in-context learners for system identification. *IFAC-PapersOnLine*, 58(15):277–282, 2024. doi: 10.1016/j.ifacol.2024.08.541.

Allan Raventós, Mansheej Paul, Feng Chen, and Surya Ganguli. Pretraining task diversity and the emergence of non-Bayesian in-context learning for regression. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 14228–14246, 2023. doi: 10.52202/075280-0626.

Matteo Rufolo, Dario Piga, Gabriele Maroni, and Marco Forgione. Enhanced transformer architecture for in-context learning of dynamical systems. In *European Control Conference (ECC)*, pp. 819–824, 2025. doi: 10.23919/ECC65951.2025.11186918.

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 (ICLR)*, 2024.

Robert Vacareanu, Vlad-Andrei Negru, Vasile Suciu, and Mihai Surdeanu. From words to numbers: Your large language model is secretly a capable regressor when given in-context examples. In *Conference on Language Modeling (COLM)*, 2024.

Henk J. van Waarde, Jaap Eising, Harry L. Trentelman, and M. Kanat Camlibel. Data informativity: A new perspective on data-driven analysis and control. *IEEE Transactions on Automatic Control*, 65(11):4753–4768, 2020. doi: 10.1109/TAC.2020.2966717.

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 (ICML)*, volume 202 of *Proceedings of Machine Learning Research*, pp. 35151–35174, 2023.

Jan C. Willems, Paolo Rapisarda, Ivan Markovsky, and Bart L. M. De Moor. A note on persistency of excitation. *Systems & Control Letters*, 54(4):325–329, 2005. doi: 10.1016/j.sysconle.2004.09.003.

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 (NeurIPS)*, volume 38, pp. 131263–131323, 2025. doi: 10.52202/085713-3950.

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 (ICLR)*, 2022.

Steve Yadlowsky, Lyric Doshi, and Nilesh Tripuraneni. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv preprint arXiv:2311.00871*, 2023.

A Additional theoretical results

These are the results summarized in §4.3; proofs are in Appendix B.

Corollary 1 (Domination in the gap)**.** *For $k \in [k_{\mathrm{id}}, k_{\mathrm{beh}})$ the Bayes-optimal ICL predictor is exact while model-free DeePC errs on generic queries (it is exact when $Mg = [u_{\mathrm{ini}}; y_{\mathrm{ini}}; u_f]$ is consistent, e.g. for the continuation of the context multisine, and generically only then); a transformer trained to the Bayes act under an input law in which u_f has a density given the context therefore dominates DeePC on generic queries throughout the gap. Our models are trained with the context input continued into the query, so the margin measured in §5.1 is out-of-support generalization.*

Identifiability versus conditioning. An exact negative control (App. D, the null space of the true-order ARX regressor of the context) shows that with the long contexts used in practice the regressor is full rank for every $k \geq k_{\mathrm{id}}$: the system is identified throughout. Past that point the error is governed by the conditioning of that identification, cleanly so under noise (Fig. 14).

Proposition 3 (Past identifiability, conditioning bounds the error)**.** *Let the true-order ARX regressor Φ_k be full column rank for all $k \geq k_{\mathrm{id}}$ and write $\sigma_{\min}(k) = \sigma_{\min}(\Phi_k)$. If the outputs are perturbed at scale $\delta \leq \sigma_{\min}(k)/(2\sqrt{n})$ (finite precision, observation noise, or finite-capacity residual), the least-squares estimate obeys $\|\hat{\theta} - \theta\| = O(\delta/\sigma_{\min}(k))$ and the N-step rollout inherits it. Hence the error is guaranteed small once $\sigma_{\min}(k) \gg \delta$; where $\sigma_{\min}(k)$ increases with k (as it does in median in our experiments), the knee under noise tracks the crossover $\sigma_{\min}(k) \approx \delta$ rather than k_{id}.*

Observation noise: robust persistency of excitation. With $e_t \sim \mathcal{N}(0, \sigma^2)$, the optimal predictor degrades gracefully while the exact lemma is brittle.

Proposition 4 (Graceful vs. brittle)**.** *For $\sigma \ll \sigma_{\min}(\Phi)$ the ARX plug-in predictor's excess error is $O(\sigma/\sigma_{\min}(\Phi))$ (Proposition 3 with $\delta = \sigma$), hence $O(\sigma)$ when the regressor is well conditioned; the posterior-mean predictor, being Bayes-optimal, has expected excess squared error no larger than this plug-in's. The behavioral solution* (2) *inverts the noisy data matrix M: with $\sigma_{\min}(M) := \sigma_{L+n_x}(M)$, the smallest singular value that is nonzero for $k \geq k_{\mathrm{beh}}$, its first-order error for $k \geq k_{\mathrm{beh}}$ is $O(\sigma/\sigma_{\min}(M))$, largest just above k_{beh} where $\sigma_{\min}(M)$ is smallest, and*

ridge regularization bounds it at the cost of bias; for $k < k_{\mathrm{beh}}$, $\operatorname{rank} M < L + n_x$ and DeePC errs already at $\sigma = 0$ (proof of Theorem 1(ii)). Identification is thus graceful wherever Φ is well conditioned, while DeePC is wrong in the gap and noise-amplifying just above k_{beh}.

Remark 1 (The SPC family sits inside the dichotomy)**.** Unregularized subspace predictive control (SPC) coincides exactly with min-norm DeePC (both return $Y_f M^+[u_{\mathrm{ini}}; y_{\mathrm{ini}}; u_f]$), so it inherits k_{beh}. The regularizers of Dörfler et al. (2023) encode an implicit identification of the non-parametric predictor and those of Breschi et al. (2023) counteract noise; neither supplies the noiseless rank deficit, so both stay behavioral; in our runs ridge-regularized DeePC, the only regularizer evaluated, remains far above the identifier throughout the gap.

Remark 2 (One $\sigma_{\min}$ behind both phases)**.** Propositions 3 and 4 are one statement at two perturbation scales: the predictor's excess error is $O(\delta/\sigma_{\min})$ at perturbation δ ($\delta \to 0$ noiseless, $\delta = \sigma$ noisy). Under noise this predicts that the dose–response is organized by $\sigma_{\min}/\sigma$ (empirically the noiseless per-system error is instead flat in $\sigma_{\min}$). This is the continuous counterpart of the discrete spectral margin in Thm. 3.

Attention reach. Kernel interpolation is in the expressive reach of attention, linking the statistical characterization to the architecture.

Proposition 5 (Attention as windowed kernel regression)**.** *If the representation at position t contains the past window $\psi_t = (y_{t-T_{\mathrm{ini}}:t-1}, u_{t-T_{\mathrm{ini}}:t})$ (e.g. assembled by earlier layers) and the output y_t, a single softmax attention head with keys and queries built from ψ_t and values y_t computes the Nadaraya–Watson estimator $\hat{y} = \sum_t \kappa(\psi_\star, \psi_t)\, y_t / \sum_t \kappa(\psi_\star, \psi_t)$ over context windows (cf. Goel & Bartlett, 2024). This is a windowed kernel smoother of the context (behavioral local weighting): the local-averaging counterpart of, not identical to, the kernel-regression form of the behavioral predictor (Molodchyk & Faulwasser, 2024), which is least squares with a linear kernel. That ICL itself admits a kernel-regression reading is known (Han et al., 2025); in our dichotomy this is the interpolation side, which the Bayes-optimal predictor, the target of class training, provably exceeds by identifying (Thm. 2) and where LLMs stay on identification probes (§5.5).*

Discrete sequences: the WFA/PSR version. Replacing the LTI class by rank-r weighted finite automata (which include r-state HMMs) and the trajectory Hankel by the Hankel of a string function, the same dichotomy holds and becomes quantitative. By Fliess's theorem (Fliess, 1974) the Hankel rank equals the rank r of a minimal WFA (the predictive-state, PSR, rank), the discrete analogue of n_x; the identifiability margin is $\sigma_r(\mathbf{H}_A)$ over a fixed complete basis. We fix a prior π under which this margin is bounded below, $\sigma_r(\mathbf{H}_A) \geq \sigma_0 > 0$ π-a.s.; the rate below is stated for this σ_0, and without such a margin bound the spectral constant can be arbitrarily large (some HMMs have exponentially small σ_r).

Theorem 3 (Spectral-rate upper bound on Bayes-optimal next-symbol KL, under a margin)**.** *For log loss the optimal next-symbol predictor is the posterior predictive $q^\star$. If a context-measurable spectral estimator $\hat{q}$ satisfies $\mathbb{E}_{x_{1:T}\sim A} \operatorname{KL}(f_A(\cdot \mid x_{1:T}) \,\|\, \hat{q}(\cdot \mid x_{1:T})) \leq \rho_T$ for π-a.e. A, then $\mathbb{E}_{A^\star\sim\pi}\, \mathbb{E}_{x_{1:T}\sim A^\star} \operatorname{KL}(f_{A^\star}(\cdot \mid x_{1:T}) \,\|\, q^\star(\cdot \mid x_{1:T})) \leq \rho_T$. Under the margin, a mixing time τ bounded on the support of π, and the regularity conditions of these analyses, spectral learning gives $\rho_T \leq \varepsilon$ at a context length polynomial in r, $1/\sigma_0$ and $1/\varepsilon$, times a factor depending on $|\Sigma|$ and τ (Hsu et al., 2012; Balle & Maillard, 2017): no horizon or $|\Sigma|^d$ factor.*

Corollary 2 (Exponential separation on long-memory sources)**.** *Let $\tau(\varepsilon) := \min\{d : \beta_d(A^\star) \leq \varepsilon\}$ be the ε-memory depth, with β_d the order-d truncation bias defined in the proof. A predictor that depends on the context only through its last d symbols has excess KL $\geq \beta_d(A^\star)$, so it needs $d \geq \tau(\varepsilon)$; in the worst case over sources whose length-d windows all have probability $\Omega(|\Sigma|^{-d})$, estimating it from one context by counts needs $T \gtrsim |\Sigma|^{\tau(\varepsilon)}$, exponentially more context than the $\operatorname{poly}(r)$ of identification whenever $\tau(\varepsilon) \gg \log_{|\Sigma|} \operatorname{poly}(r)$.*

B PROOFS

Throughout, $\mathcal{B}_L(S) \subseteq \mathbb{R}^{2L}$ is the length-L behavior of S, of dimension $L + n_x$ for $L \geq \ell$, and $z_{1:T}$ is a trajectory of $S \in \mathcal{S}_n$ with state dimension n_x and lag ℓ.

Proof of Proposition 1. Every column of $\mathscr{H}_L(z_{1:T})$ is a length-L window of a trajectory of S, hence a point of $\mathcal{B}_L(S)$; therefore $\operatorname{col} \mathscr{H}_L(z) \subseteq \mathcal{B}_L(S)$. The two subspaces have equal dimension:

$\dim \mathcal{B}_L(S) = L + n_x$, while persistency of excitation of order $L + n_x$ together with controllability of (A, B) forces $\operatorname{rank} \mathscr{H}_L(z) = L + n_x$ (Willems et al., 2005). A subspace inclusion with equal finite dimensions is an equality, so $\operatorname{col} \mathscr{H}_L(z) = \mathcal{B}_L(S)$. Now take $T_{\mathrm{ini}} \geq \ell$. The past block $(u_{\mathrm{ini}}, y_{\mathrm{ini}})$ of a length-L window then fixes the internal state at the start of the future block (observability index $\leq T_{\mathrm{ini}}$), so the affine constraints $[U_p; Y_p; U_f]g = [u_{\mathrm{ini}}; y_{\mathrm{ini}}; u_f]$ select exactly those trajectories of S with the prescribed initial condition and future input. All such trajectories share the same future output, whence $\hat{y}_f^{\mathrm{beh}} = Y_f g^\star = y_f(S)$. □

Proof of Proposition 2. A SISO system in $\mathcal{S}_n$ admits an ARX representation $y_t = \sum_{i=1}^n a_i y_{t-i} + \sum_{i=1}^n b_i u_{t-i}$ with parameter $\theta \in \mathbb{R}^{2n}$ and regressors $\varphi_t = (y_{t-1}, \ldots, y_{t-n}, u_{t-1}, \ldots, u_{t-n})$, the columns of the depth-n Hankel matrix $\mathscr{H}_n(z)$ (rows reordered). For $k \geq n$ the input is PE of order $2n \geq n + n_x$, so $\operatorname{col} \mathscr{H}_n(z) = \mathcal{B}_n(S)$ by the fundamental lemma (Willems et al., 2005); for $k \geq k_{\mathrm{id}}$ the input is PE of order n, which gives the same equality for almost every initial state (Markovsky et al., 2023). The regressor therefore has rank $n + n_x$, full column rank iff $n_x = n$; otherwise the exact least-squares fits form an affine set. Every fit θ' satisfies $\varphi^\top \theta' = \varphi^\top \theta$ for all $\varphi \in \mathcal{B}_n(S)$, since each such φ is a linear combination of context regressors whose targets both parameters reproduce. Rolling the recursion forward from the last n context samples under u_f keeps every regressor in $\mathcal{B}_n(S)$, so by induction over the horizon every fit returns $y_f(S)$, for every query and horizon, and $k_{\mathrm{id}} = \Theta(n)$ does not depend on L. (For the steady-state initial state, whose free response is absent, the requirement rises to $2k \geq n + n_x$.) □

Proof of Theorem 1. ***(i)*** is Proposition 2. ***(ii)***: $k < k_{\mathrm{beh}}$ gives $2k < L$, so the input rows of $\mathscr{H}_L(z)$, which form $\mathscr{H}_L(u)$, have rank $2k < L$; since the input of S is free, $\mathcal{B}_L(S)$ projects onto $\mathbb{R}^L$ in the input coordinates, and $\operatorname{col} \mathscr{H}_L(z) \subsetneq \mathcal{B}_L(S)$ for every initial state (PE of order L is necessary; Markovsky et al., 2023). The u_f with $[u_{\mathrm{ini}}; u_f] \in \operatorname{col} \mathscr{H}_L(u)$ form a proper affine subspace of $\mathbb{R}^N$ (elements of $\operatorname{col} \mathscr{H}_L(u)$ are sums of the k context sinusoids, i.e. of $2k \leq L - 1$ complex exponentials, and no nonzero such sum vanishes at $L - 1$ consecutive samples), so for a generic query the system $Mg = [u_{\mathrm{ini}}; y_{\mathrm{ini}}; u_f]$ is inconsistent and (2) returns the output block of the least-squares fit $\mathscr{H}_L(z)\, g^\star$, which differs from $y_f(S) = \hat{y}_f^{\mathrm{id}}$ for a generic system and query (the two affine maps $u_f \mapsto Y_f g^\star$ and $u_f \mapsto y_f(S)$ do not coincide generically). (When the system is consistent and $T_{\mathrm{ini}} \geq \ell$, (2) is exact: every solution g gives a window of $\mathcal{B}_L(S)$ with the true past and input.) For the lower bound, suppose S_1, S_2 both produce $z_{1:T}$. Then $f(z_{1:T}, u_f) = \hat{y}$ is a single vector independent of j, and the triangle inequality gives $\|\hat{y} - y_f(S_1)\| + \|\hat{y} - y_f(S_2)\| \geq \|y_f(S_1) - y_f(S_2)\|$, so the larger summand is $\geq \frac{1}{2}\|y_f(S_1) - y_f(S_2)\|$. The hypothesis is not implied by $\operatorname{rank} \mathscr{H}_L(z) < L + n_x$: for $k \geq k_{\mathrm{id}}$, part (i) shows that all systems in $\mathcal{S}_n$ consistent with $z_{1:T}$ share the continuation, so the bound is vacuous in the gap. It is strictly positive whenever $2k < n$, i.e. $k < k_{\mathrm{id}}$: the length-n windows of z then span at most $2k + n_x < n + n_x = \dim \mathcal{B}_n(S)$ dimensions, the exact least-squares ARX(n) fits of the context form an affine family that does not determine the behavior of S, and a generic member defines a system in $\mathcal{S}_n$ that reproduces $z_{1:T}$ but continues a generic u_f differently. □

Proof of Theorem 2. For squared loss the Bayes-optimal act is the conditional mean, so the risk minimizer is $f^\star(z_{1:T}, u_f) = \mathbb{E}[y_f \mid z_{1:T}, u_f]$; the tower property, conditioning on S, yields the integral form against the posterior $\pi(\cdot \mid z_{1:T})$. In the noiseless regime with $k \geq k_{\mathrm{id}}$ and almost every initial state, Proposition 2 shows that every system in $\mathcal{S}_n$ consistent with the context shares the identified continuation, so $\pi(\cdot \mid z_{1:T})$ is supported on systems with the input–output behavior of $\hat{S}$ and $f^\star = \mathbb{E}[y_f \mid \hat{S}, \cdot] = y_f(\hat{S}) = \hat{y}_f^{\mathrm{id}}$. The minimizer's error therefore vanishes for every $k \geq k_{\mathrm{id}} = \Theta(n)$, independently of L. The transformer is trained on the next-token loss; in this regime its one-step Bayes act $\mathbb{E}[y_t \mid z_{1:t-1}, u_t]$ is exact at every step on the support of the training input law, so there its autoregressive rollout coincides with $f^\star$. □

Proof of Corollary 1. For $k \in [k_{\mathrm{id}}, k_{\mathrm{beh}})$, Theorem 1(i) and Theorem 2 give $f^\star = y_f(S)$ (exact), while Theorem 1(ii) gives $\hat{y}_f^{\mathrm{beh}} \neq y_f(S)$ for generic queries. A transformer approaching $f^\star$ thus dominates the model-free predictor on generic queries throughout the gap. □

Proof of Proposition 3. With Φ_k of full column rank, perturbing the outputs by δ_y ($\|\delta_y\| \leq \delta$) perturbs both the target and the lagged-output columns of the regressor, $\Phi_k \mapsto \Phi_k + \Delta$ with

$\|\Delta\| \le \sqrt{n}\,\delta$. For $\sqrt{n}\,\delta \le \sigma_{\min}(\Phi_k)/2$ the perturbed regressor keeps full column rank with $\sigma_{\min}(\Phi_k + \Delta) \ge \sigma_{\min}(\Phi_k)/2$, and $\hat\theta - \theta = (\Phi_k + \Delta)^+(\delta_y - \Delta\theta)$, so $\|\hat\theta - \theta\| \le 2(1 + \sqrt{n}\|\theta\|)\,\delta/\sigma_{\min}(\Phi_k)$. On a compact set of stable parameters the N-step ARX rollout is Lipschitz in θ, so the prediction error is $O(\delta/\sigma_{\min}(k))$. The error is therefore guaranteed small once $\sigma_{\min}(k) \gg \delta$; where $\sigma_{\min}(k)$ increases with k (observed in median, not guaranteed per system), this crossover, not the identifiability boundary, locates the empirical knee under noise. □

Proof of Proposition 4. Identification. Under output noise the ARX equation error $e_t - \sum_i a_i e_{t-i}$ is correlated with the lagged outputs in φ_t, so least squares is the errors-in-variables estimate of Proposition 3 with $\delta = \sigma$ (per-sample normalization): its error, including an $O(\sigma^2/\sigma_{\min}(\Phi)^2)$ bias, is $O(\sigma/\sigma_{\min}(\Phi))$ for $\sigma \ll \sigma_{\min}(\Phi)$, hence $O(\sigma)$ where the regressor is well conditioned. The posterior mean minimizes the Bayes risk, so its expected excess squared error is at most that of this plug-in predictor. *Behavioral.* For $k \ge k_{\mathrm{beh}}$ the rows of the noiseless Y_f lie in the row space of M ($T_{\mathrm{ini}} \ge \ell$), $Y_f = ZM$ with Z the exact prediction map. For noisy data $\tilde M \tilde g = \tilde b$ (noise makes $\tilde M$ full row rank) and $Mg^\star = b$ give $M(\tilde g - g^\star) = (\tilde b - b) - (\tilde M - M)\tilde g$, so $Y_f(\tilde g - g^\star) = Z[(\tilde b - b) - (\tilde M - M)\tilde g] = O(\sigma\|Z\|\|\tilde g\|)$; the noise in Y_f adds $O(\sigma\|\tilde g\|)$, and $\|\tilde g\| = O(\|b\|/\sigma_{\min}(M))$ to first order, so $\|\Delta\hat y_f^{\mathrm{beh}}\| = O(\sigma/\sigma_{\min}(M))$, largest just above k_{beh}. For $k < k_{\mathrm{beh}}$, $\operatorname{rank} M < L + n_x$ and the predictor errs already at $\sigma = 0$ (proof of Theorem 1(ii)). Replacing $g^\star$ by the ridge solution bounds $\|g\|$ and hence the variance, at the price of an irreducible bias. □

Proof of Proposition 5. With the stated keys, queries, and values a single softmax head outputs the convex combination $\hat y = \sum_t w_t y_t$ with $w_t = \kappa(\psi_\star, \psi_t)/\sum_s \kappa(\psi_\star, \psi_s)$ and $\kappa(\psi_\star, \psi_t) = \exp\langle Q\psi_\star, K\psi_t\rangle$; this is precisely the Nadaraya–Watson estimator with kernel κ. Its weights are positive and sum to one, so $\hat y$ lies in the convex hull of the context outputs; when the logits include $-\frac{1}{2}\|K\psi_t\|^2$ (a Gaussian kernel, cf. Goel & Bartlett, 2024) or the keys have equal norm, it tends to the nearest-window output as the bandwidth sharpens. The behavioral predictor (2) is instead least-squares kernel regression with the linear kernel (Molodchyk & Faulwasser, 2024), linear in the query with signed weights; in that limit the two coincide when the query window equals a context window, and generically only then. □

Proof of Theorem 3. The posterior predictive $q^\star$ is the Bayes act for log loss, so its π-averaged expected excess KL is at most that of *any* context-measurable predictor, in particular $\hat q$; averaging the assumed bound over $A^\star \sim \pi$ gives the claim. Bayes optimality is a statement about the π-average and does not bound the excess KL at an individual $A^\star$. For the rate, the empirical Hankel moments of an ergodic source concentrate from a single trajectory at a mixing-dependent rate (Balle & Maillard, 2017), and the perturbation analysis of the spectral algorithm (Hsu et al., 2012, Thm. 7), under its regularity conditions, converts this into a conditional-KL guarantee polynomial in r, $1/\sigma_r \le 1/\sigma_0$ and $1/\varepsilon$ (see also Balle et al., 2014); mixing $\hat q$ with the uniform distribution keeps its log loss bounded on the failure event. □

Proof of Corollary 2. An order-d predictor's conditional ignores history older than d, so by the chain rule its excess KL is at least the order-d truncation bias $\beta_d(A^\star) = \mathbb{E}\,\mathrm{KL}(f_{A^\star}(\cdot \mid x_{1:T}) \,\|\, f_{A^\star}(\cdot \mid x_{T-d+1:T}))$; excess KL $\le \varepsilon$ therefore needs $d \ge \tau(\varepsilon)$. A count-based estimate of the order-d statistics from a single context requires populating the length-d windows, which in the worst case over sources whose windows all have probability $\Omega(|\Sigma|^{-d})$ needs $T \gtrsim |\Sigma|^d$. Together, $T \gtrsim |\Sigma|^{\tau(\varepsilon)}$, exponentially more context than the $\mathrm{poly}(r)$ of Theorem 3 on long-memory sources, where $\tau(\varepsilon)$ is large. □

C Additional figures

Figures supporting the experiments (§5); each caption names the section it supports.

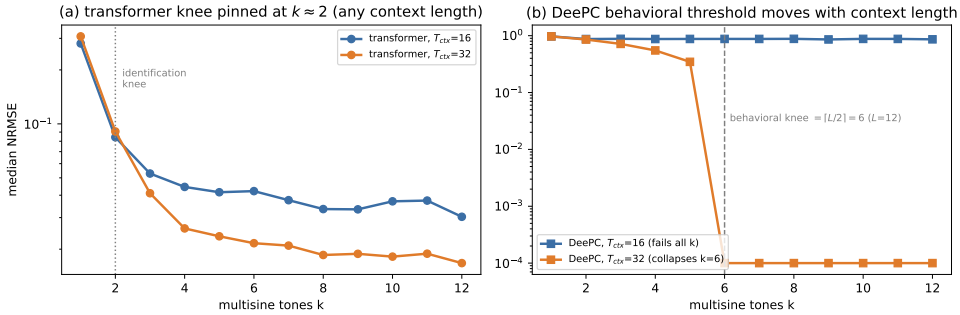


Figure 4: **Short-context retrain** (§5.2). (a) The transformer's identification knee is pinned at $k \approx 2$ independent of context length. (b) DeePC's behavioral threshold is context-dependent: it never collapses at T_{ctx}=16 and collapses at its rank knee $k \approx 6$ (L=12) at T_{ctx}=32.

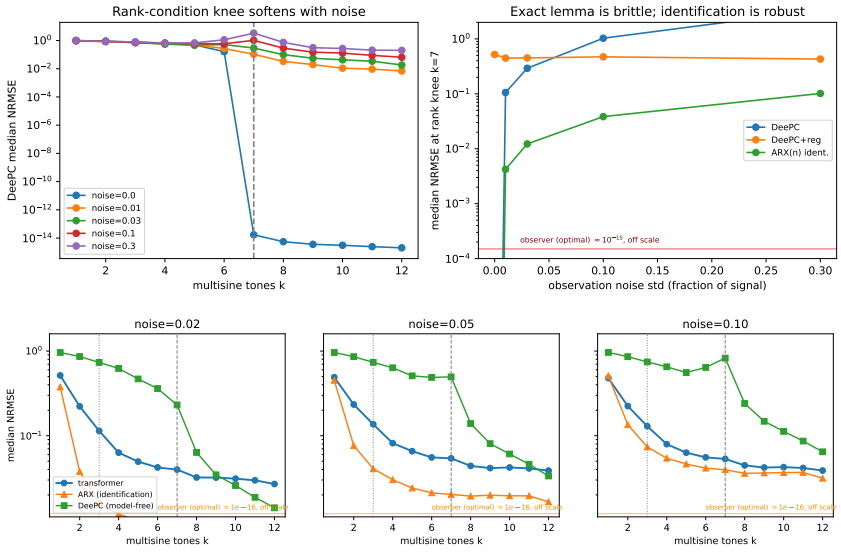


Figure 5: **Robust PE** (§5.3). Top: oracle dose–response; DeePC is brittle at the rank knee, ARX and the observer are robust. Bottom: noise-matched transformers keep the identification knee and beat DeePC 7.5–10× at k=4 across noise levels.

D Experimental Details

Models and training. Decoder-only transformer over interleaved (u, y) tokens; the output read at each u-token predicts the corresponding y. Main LTI model: d=384, 12 layers, 6 heads ($\sim$21M), 80k steps, AdamW, cosine schedule, bf16; a 2.7M model (d=192, 6 layers) is used for ablations. Discrete model: symbol embedding + softmax head, same depth. Data are generated on the fly (no dataset files); outputs are normalized to unit std (the context std at evaluation) and clipped. Each training sequence uses one input type throughout (white noise or a random k-multisine, "mix"), so the multisine-context/white-query pairing of §5.1 is out of support.

Oracles. DeePC (2) (min-norm and ridge-regularized), ARX(n_x) least squares, Kalman-style observer (known A, B, C), and, for the discrete, AR(1), and AR(2) settings, exact posteriors (forward algorithm / grid marginalization).

Negative control. The consistent set is the affine null space of the true-order (n_x=3) ARX regressor of the context; at high k it is a point (identified), at low k a sampleable continuum (underdetermined). Sampling this continuum with a fixed Gaussian scale does not follow the prior re-

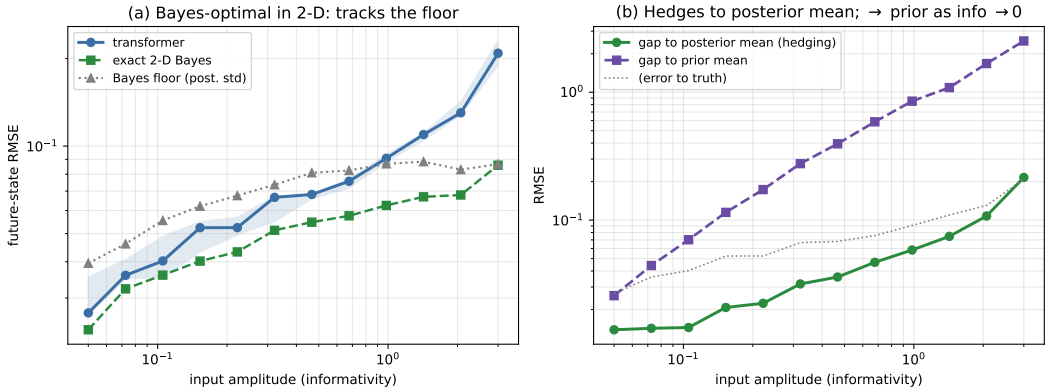
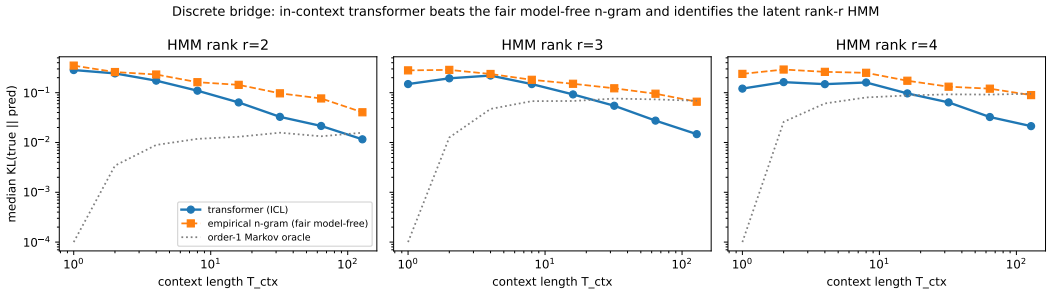
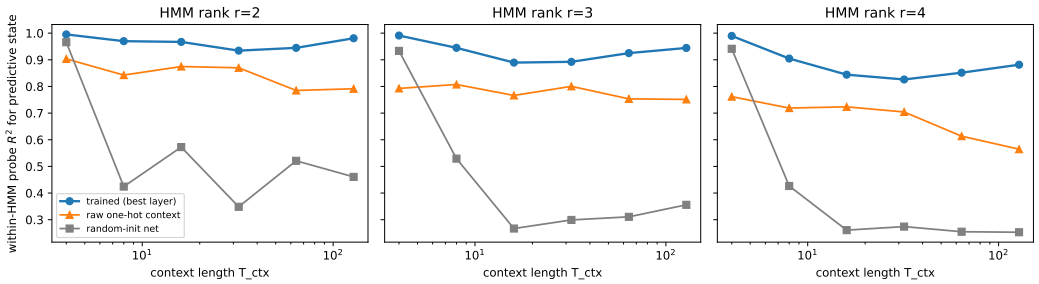


Figure 6: **Exact-posterior confirmation** (§5.4; AR(2): two unknown poles, exact 2-D posterior, 3 seeds). (a) The transformer tracks the exact 2-D Bayes error and sits at the information floor (error/floor $\approx$ 0.8–1.1 up to amplitude $\approx$1). (b) It hedges to the posterior mean (small green gap) and falls back to the prior as informativity $\to$ 0 (the purple gap to the prior mean shrinks). The 1-D AR(1) family behaves identically, with posterior-mean gap 0.005–0.09.



(a) Functional: next-symbol KL falls toward zero, beats the fair empirical n-gram, and drops below the order-1 Markov oracle for $r \in \{3, 4\}$.



(b) Representational: the predictive state is linearly decodable from hidden states (R^2 up to 0.9+), above raw context and, for $T_{\text{ctx}} \geq 8$, far above a random-init net.

Figure 7: **Discrete bridge and mechanism** (§5.4). Trained 21M symbol transformers identify the latent rank-r HMM both (a) functionally (model-free n-grams never reach the transformer, and the context needed grows with r) and (b) representationally.

stricted to the consistent set, so as a posterior it is biased in the high-dimensional class; the rank test (relative tolerance 10^{-3}; median null dimension 0 for $k \geq 2$) is unaffected, and the AR(1) and AR(2) families of §5.4 provide exact posteriors.

Compute. Training runs on a Slurm GPU cluster (H100/H200); each 21M/2.7M transformer trains in a few GPU-hours on a single H100, and all pretrained-LLM results are inference-only (no train-

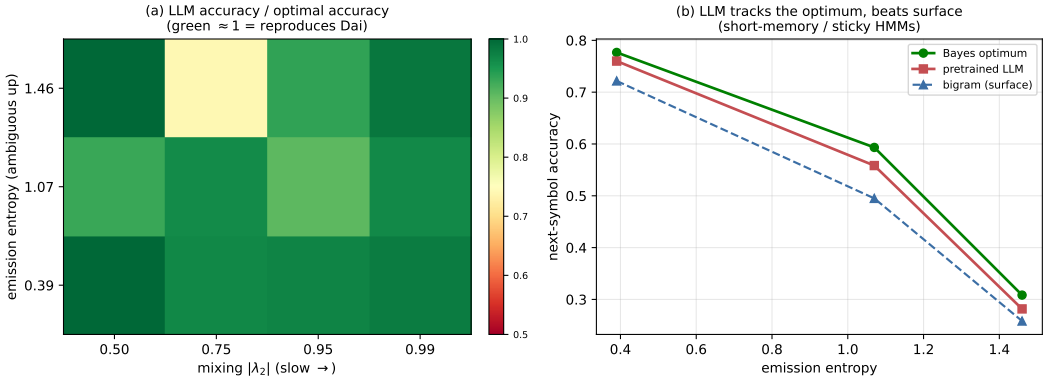


Figure 8: **Head-to-head reproduction of Dai et al. (2025)** (§5.5). Random HMMs on two of their axes, mixing×entropy (Qwen2.5-14B, $M{=}8$ states, $T_{\text{ctx}}{=}128$). (a) LLM accuracy / Bayes-optimal accuracy is $\approx$1 (green) over most of the grid and drops only at near-maximal emission entropy; we reproduce their near-optimal result, slow-mixing column included. (b) The LLM (red) tracks the Bayes optimum (green) and beats the surface bigram (blue): on these short-memory HMMs it does genuine in-context state estimation, unlike on our long-memory cyclic class (Fig. 13), where it stays at the n-gram.

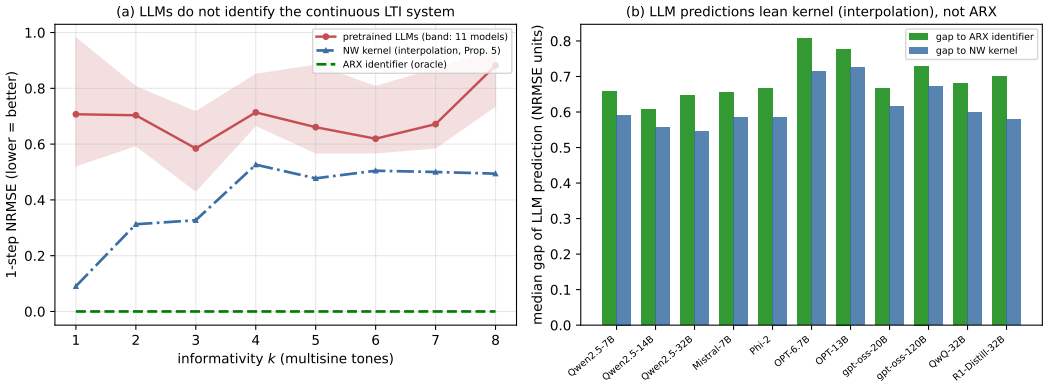


Figure 9: **On continuous LTI, the theory's core object, pretrained LLMs remain kernel-leaning interpolators** (§5.5). 1-step NRMSE on (u, y) trajectories vs. informativity k (single GPU, parse $\geq 98\%$). (a) The LLM band (11 models, 2.7B–120B) sits far above the ARX identifier ($\approx$0) and is flat in k. (b) Median gap of the LLM prediction to the ARX identifier vs. to the NW kernel (Prop. 5): the prediction matches neither, but for every model it is relatively closer to the kernel, the continuous analogue of "closer to 1-NN than OLS".

ing). Per-run wall-clock and the aggregate GPU-hour budget are recorded in the released run logs. Analysis and oracles run on a single GPU or CPU.

Seeds and repeats. Every trained-model LTI rollout sweep point is a median over ≥ 256 held-out systems. The phase figure's rollout panel uses 5 (mix-input) training seeds and its fresh-probe panel one; the $\sigma_{\min}$ figure uses 8 noiseless plus 6 noise-matched seeds; the mechanism-flip and scaling sweeps 5; the exact-posterior AR(1) and AR(2) hedging demonstrations use 3 seeds each; the discrete HMM sweeps use 400 sampled HMMs per cell; the LLM evaluations use 60–150 contexts per cell (automata) and 50–150 problems per cell (regression). Bands in figures are seed ranges, interquartile ranges, within-context spreads, or across-model ranges intervals, as stated in each caption.

LLM parse rates. Errors in §5.5 are computed only over parseable numeric answers; Table 2 reports the parse rate per model. Every base and INSTRUCT model parses at 100% except OPT-30B

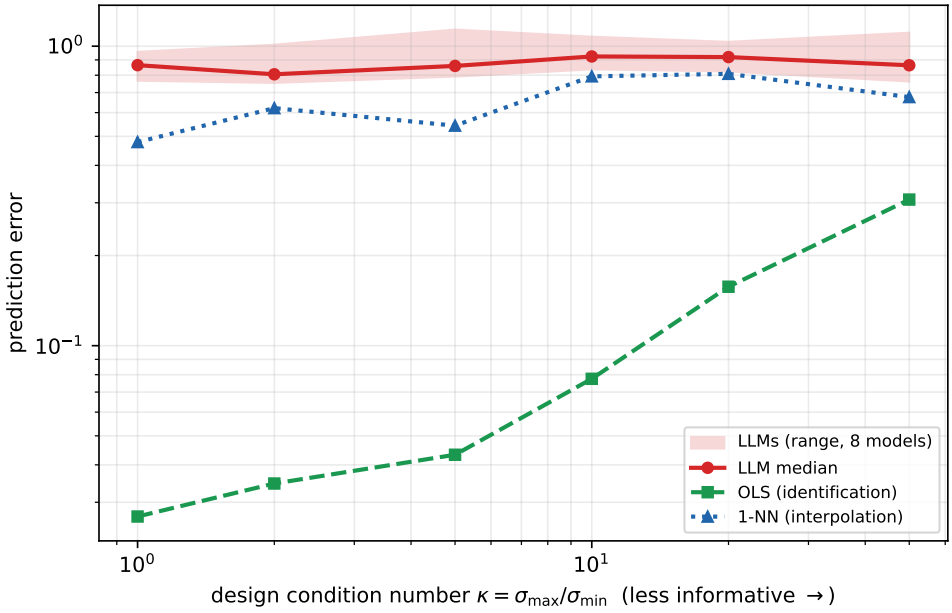


Figure 10: **LLMs remain at the interpolation level across conditioning** (§5.5). At fixed n=12, sweeping the design conditioning $\kappa = \sigma_{\max}/\sigma_{\min}(X)$: OLS (identification) follows $\sigma_{\min}$, near-exact at κ=1 and an order of magnitude worse by κ=50, while every LLM stays within $\sim$2$\times$ of the interpolation (1-NN) level, far above OLS at all κ, including maximal informativity.

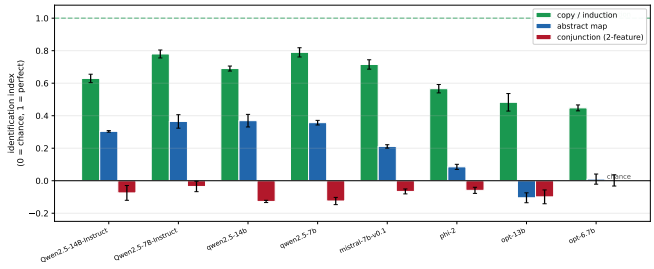


Figure 11: **In-context rule-following in pretrained LLMs degrades with rule complexity** (§5.5). Identification index = (acc−chance)/(1−chance) (0 chance, 1 perfect; 4 seeds; confound-free symbolic tasks with a fresh-query discriminator; easiest setting per task). Copy (out=key[$c^\star$]): all models identify. Abstract map (out=M(key[$c^\star$])): modern models partially, OPT at chance. Conjunction (match a hidden two-feature pattern, balanced accuracy): tested models are at chance. The dynamical-system identification of our theory lies at the hard end, reached by a class-trained transformer and by no pretrained LLM.

(0%, excluded from the bars); on the INSTRUCT models both direct and chain-of-thought parse at 100%, so the CoT degradation is a reasoning effect rather than an extraction artifact.

LLM evaluation protocol. All pretrained-LLM results are inference-only on local snapshots of public checkpoints: `facebook/opt-{6.7b,13b,30b,66b}`, `Qwen/Qwen2.5-{7B,14B,32B}` (7B/14B also `-Instruct`), `mistralai/Mistral-7B-v0.1`, `microsoft/phi-2`, `openai/gpt-oss-{20b,120b}`, `Qwen/QwQ-32B`, and `deepseek-ai/DeepSeek-R1-Distill-Qwen-32B`; exact local revisions are pinned in `configs/llm_models.json` for every model except OPT-66B, which appears only in Fig. 13. Models load in `bfloat16` (regression, continuous LTI) or `float16` (automata, symbolic) on a single GPU (`device_map=cuda:0`). The gpt-oss checkpoints are released in MXFP4; gpt-oss-120b

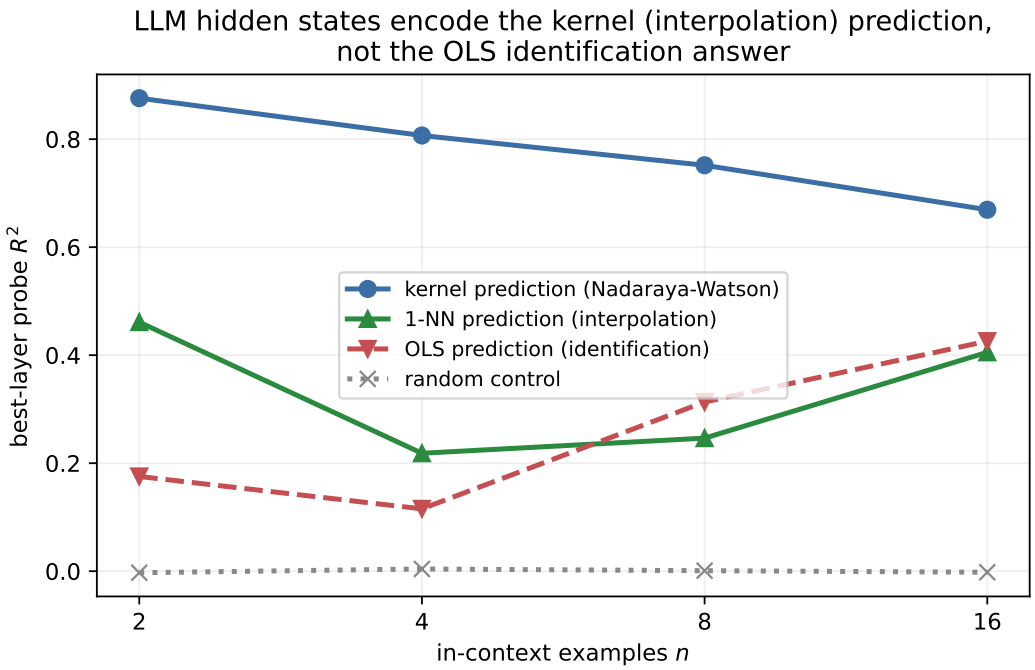


Figure 12: **LLM representations encode kernel smoothing** (§5.5). Best-layer probe R^2 vs. n for Qwen2.5-7B: the Nadaraya–Watson kernel prediction dominates ($\approx$ 0.7–0.88), far above the OLS answer (dashed); the random control (dotted) is at 0. The ordering is not a layer-selection artifact: at the fixed last layer (not plotted; no selection) kernel still dominates OLS (Qwen2.5-7B 0.61–0.82 vs. 0.08–0.41; Qwen2.5-14B 0.61–0.79 vs. 0.06–0.46). Consistent with Proposition 5 (attention as windowed kernel smoothing).

Table 2: Pretrained-LLM parse rate and median error ratio on in-context linear regression ($n \geq d$), and CoT vs. direct on the INSTRUCT models (both at 100% parse). Generated by `src/tab_llm_parse.py`.

Model	parse rate	median LLM/OLS ($n \geq d$)
gpt-oss-120b	100%	15×
gpt-oss-20b	100%	15×
mistral-7b-v0.1	100%	14×
opt-13b	100%	18×
opt-30b (excluded)	0%	–
opt-6.7b	100%	17×
phi-2	100%	18×
qwen2.5-14b	100%	12×
qwen2.5-32b	100%	12×
qwen2.5-7b	100%	13×
qwq-32b	100%	14×
r1-distill-qwen-32b	100%	11×
CoT vs. direct, Instruct (median over $n \geq d$); both at 100% parse		
Qwen2.5-14B-Instruct	100%	direct 11× / CoT 47×
Qwen2.5-7B-Instruct	100%	direct 14× / CoT 32×

loads quantized via `transformers`' `kernels` integration on one H200, and the reported gpt-oss-20b run uses the bf16-dequantized path (a rerun on the MXFP4 kernel path agrees to 1.9% in pooled err/OLS, so the loading path is immaterial). Reasoning-trained models (QwQ, R1-Distill) are evaluated on the same raw-completion protocol; the 8-token cap suppresses any thinking

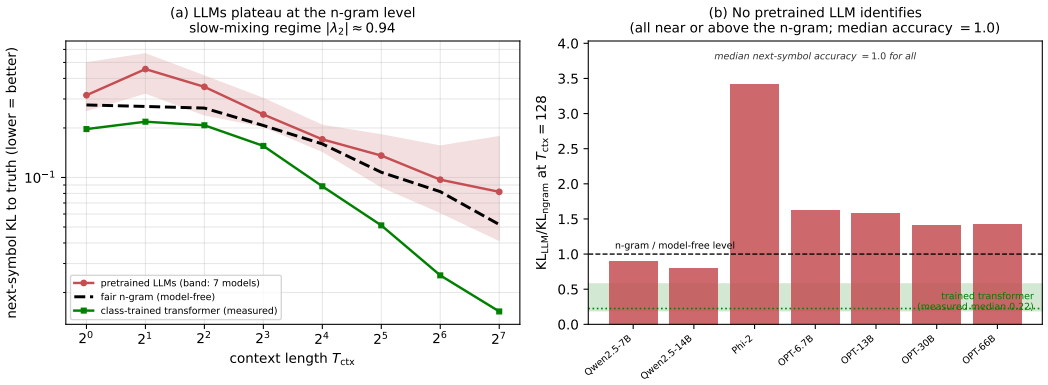


Figure 13: **Pretrained LLMs predict HMMs at the model-free level** (§5.5). Next-symbol KL to the true HMM law, slow-mixing regime ($|\lambda_2| \approx 0.94$), pooled over $r \in \{2, 3, 4\}$ (one GPU per model). (a) Vs. context: the LLM band (7 models) falls with context but plateaus at the fair same-context n-gram (model-free) level, never the measured $\sim 4\times$ lower KL (median ratio 0.22) that the same class-trained transformer reaches on this grid (green curve, §5.4). (b) At T_{ctx}=128: every model's $\text{KL}_{\text{LLM}}/\text{KL}_{\text{ngram}}$ is near or above 1 (0.8–3.4×; OPT 6.7B–66B flat in scale), far from the measured trained-transformer level (ratio ≈0.22, green), while median next-symbol accuracy is 1.0, so accuracy alone (one of the metrics of Dai et al., 2025) would call them optimal; the n-gram-relative KL is what separates interpolation from identification.

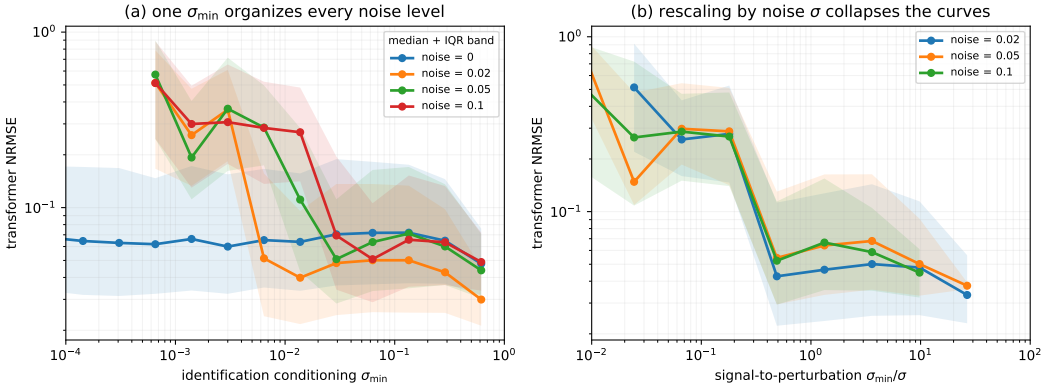


Figure 14: **Under noise, error follows the identification conditioning** $\sigma_{\min}$ (§5.2; binned per-system errors; line = median, band = 25–75%). (a) Error vs. the ARX identification-regressor $\sigma_{\min}$ (full rank from k_{id}, unlike the behavioral Hankel $\sigma_{\min}$, which turns on only at k_{beh}): under noise the error rises as $\sigma_{\min}$ falls (colored curves); the noiseless error is flat in $\sigma_{\min}$ (median ≈ 0.06). (b) Error vs. $\sigma_{\min}/\sigma$: the noisy median curves align on the signal-to-perturbation ratio. The alignment is of medians: binned $\sigma_{\min}/\sigma$ explains only $R^2 \approx 0.2$ of per-system log-error variance (Spearman $\rho \approx -0.2$ pooled; near zero within fixed (σ, k) cells), so per-system spread is genuinely wide and irreducible by $\sigma_{\min}$ alone.

trace, so they act as raw predictors, and the elicited arm (Table 3) removes the cap. Single-GPU execution is mandatory: sharding these models across two GPUs (`device_map=auto`) corrupts the forward pass and produces a spurious flat-high KL; all runners default to one GPU. Decoding is as follows. Numeric tasks (regression, LTI) use greedy decoding (`do_sample=False`, `max_new_tokens=8`) and parse the first float; symbolic and automaton tasks use next-token logits restricted to the candidate tokens and softmax over them (parse 100% by construction). The prompt templates are as follows. Regression: lines `x=`v_1`,` v_2`, ..., y=`v (2-decimal), query ending `x=..., y=`; continuous LTI: lines `u=`v`, y=`v, query ending `u=..., y=`; automata: the symbol stream with each symbol a single token (letters `A`–`H`, digits, or digit+newline), the next-symbol distribution read over the $|\Sigma|$ candidate tokens; symbolic rules: the (key → output) exemplars, the

answer read from logits over the class tokens. The chain-of-thought variant prepends a one-line reasoning instruction. All templates and parse regexes are in the released code.

Short-memory informativity sweep (Fig. 2). Dai-style random HMMs (§5.5): $M \in \{4, 8\}$ hidden states, $|\Sigma|$=5, mixing $|\lambda_2| \in \{0.5, 0.75, 0.95, 0.99\}$, emission-entropy knob $h \in \{0.1, 0.4, 0.7\}$, stationary start; 150 HMMs per setting, seed 0, $T \in \{8, 16, 32, 64, 128, 256\}$, one forward pass per sequence with the next-symbol distribution read at every T (digit encoding; median symbol mass 0.97). Baselines see the same single context and nothing else: back-off n-grams of order 2 and 3 (α=0.5) and a spectral-WFA estimator on the length-$\leq$ 2 substring basis with rank set to the number of hidden states M, an upper bound on the Hankel rank (per-step-rescaled forward pass). Hypotheses and pass criteria were fixed before the runs; the pre-registration (supplementary material) discloses earlier exploratory runs on the two Qwen models, for which that hypothesis was not a blind prediction (it was for Mistral-7B and phi-2), and one amendment to the margin definition made before any data of this sweep were seen. A random predictor fails all three criteria. Information gating (median KL falling from T=8 to T=256 in at least 20 of 24 settings, pooled log-log slope in $[-1.2, -0.3]$) passed for all four models. The surface-separation criterion combined three gates on the ambiguous-emission settings: the LLM ends below both n-gram orders at T=256 (passed for all four models); the LLM keeps falling over $T \in \{64, 128, 256\}$ (slope below -0.3; passed for Qwen2.5-7B/14B and Mistral-7B, not phi-2); and the order-2 n-gram is flat over the same range (slope above -0.15). The measured order-2 slope is -0.17, so this gate, and hence the combined criterion, did not pass. A margin-rescaling analysis failed its criterion (the correlation had the opposite sign) and is not claimed. Models: Qwen2.5-7B/14B, Mistral-7B, phi-2.

Elicited reasoning arm (Table 3). QwQ-32B and R1-Distill-Qwen-32B on the regression probe through their chat templates: direct mode (48 new tokens) vs. CoT mode (3072 new tokens; the prompt asks the model to set up and solve the least-squares system; answers parsed after the thinking trace); $d \in \{3, 5\}$, n=16, σ=0.1, 60 trials, greedy decoding, bf16, one GPU. R1-Distill answers only through its trace (direct parse 0–3%), which is why the main-panel protocol uses raw (template-free) completions.

Table 3: Elicited reasoning on in-context regression (n=16; median $|\hat{y} - y|$; OLS 0.024–0.045, 1-NN 0.62–1.20). A full thinking budget parses at 100% and stays 2–3$\times$ above the 1-NN error.

model	mode	d=3	d=5	err/OLS	parse
QwQ-32B	direct	1.36	1.01	23–48$\times$	97–98%
QwQ-32B	CoT	1.76	2.47	54–75$\times$	100%
R1-Distill-32B	direct	unparseable		–	0–3%
R1-Distill-32B	CoT	1.46	2.14	47–62$\times$	100%

Table 4: Per-experiment summary: models, samples and seeds, metric, and the scope of the conclusion each experiment supports. "*tested* LLMs" marks claims bounded to the open-weight panel (2.7B–120B) and the synthetic probes used here.

Experiment (§)	Models	Samples/cell, seeds	Metric	Supports the claim
Phase transition (5.1)	trained 21M	$\geq$256 sys, 1 (probe) / 5 (rollout) seeds	NRMSE	trained transformer identifies
Conditioning (5.2)	trained 21M	$\geq$256 sys, 8+6 seeds	NRMSE vs. ident. $\sigma_{\min}$	identifiable from k_{id}; $\sigma_{\min}/\sigma$ organizes noisy error
Mechanism flip (5.6)	1 arch, 2 priors	$\geq$256 sys, 5 seeds	gap to ARX / GP oracle	training prior sets mechanism
AR(1)/AR(2) hedging (5.4)	trained 2.7M	$\geq$256 sys, 3 seeds	error/floor, gap to means	near-Bayes-optimal; hedges to posterior mean
HMM/WFA bridge (5.4)	trained 21M	400 HMMs, 3 seeds	next-symbol KL	identifies latent rank-r state
LLM regression (5.5)	13 LLMs (2.7–120B)	50–120/cell	error/OLS, /1-NN	*tested* LLMs interpolate
LLM conditioning (5.5)	8 LLMs	150/cell	error vs. κ	*tested* LLMs stay at interpolation level
LLM continuous LTI (5.5)	11 LLMs (2.7–120B)	per-cell	gap to ARX / NW kernel	*tested* LLMs are kernel-leaning
LLM long-memory HMM (5.5)	7 LLMs	60–120/cell	KL / n-gram	at/near n-gram level
Dai head-to-head (5.5)	Qwen 7B/14B	150/cell	acc / optimal	reproduce; boundary = memory depth
Informativity sweep (5.5)	4 LLMs	150/setting, $T \leq 256$	KL vs. T	information-gated in-regime
Elicited reasoning (5.5)	QwQ, R1-Distill	60/cell	err/OLS, parse	thinking does not identify
Symbolic rule ladder (5.5)	8 LLMs	per-cell, 4 seeds	identification index	graded by rule complexity

Prompt-format robustness (control for §5.5). We test template stability by re-running the in-context regression probe under six prompt formats (`x=..,y=..`, `input/output`, arrow (→), `json`, `Q/A`, and a prose sentence) on five models (d=3, 100 trials/cell). Table 5: for the four models that follow alternative templates (Qwen2.5-7B/14B, Mistral-7B, OPT-6.7B), the median $\mathrm{err}_{\mathrm{LLM}}/\mathrm{err}_{\mathrm{OLS}}$ at n=12 is 24–40$\times$ across every format, the prediction is closer to 1-NN than to

OLS in 90 of 96 (format, $n \geq d$) cells, and the $n{=}d$ knee remains absent. phi-2 (2.7B) reliably emits parseable numbers for the baseline template, while alternative templates mostly expose a numeric-output competence limit (parse $< 90\%$).

Table 5: Prompt-format robustness of the regression verdict: median $\mathrm{err}_{\mathrm{LLM}}/\mathrm{err}_{\mathrm{OLS}}$ at $n{=}12$ ($d{=}3$, 100 trials) under six numeric templates. Ratios $\gg 1$ with no $n{=}d$ knee and a 1-NN-closer prediction = behavioral interpolation across formats. †: parse $< 90\%$; "–": parse $< 10\%$ (phi-2 parses $\geq 90\%$ only under the baseline template).

template	Qwen2.5-7B	Qwen2.5-14B	Mistral-7B	OPT-6.7B	phi-2
`x=..,y=..`	30	29	29	40	29
`input/output`	24	24	26	27	22†
arrow (→)	32	28	27	37	–†
`json`	33	31	30	29†	49†
`Q/A`	28	27	30	29	–†
prose	34	29	32	28	–†

Symbolic-template robustness (control for the rule-ladder, §5.5). We test symbolic-template stability by rendering the same trials under five templates (`A B -> X`, `=>`, `:`, comma-separated, and a `key..out..` framing) and re-scoring (Table 6). The per-model verdict is template-stable: copy is identified by every model under all five templates; the abstract map keeps the same ordering (Qwen2.5-14B partially identifies, Qwen2.5-7B/Mistral sit at the boundary, OPT/phi-2 are at or below chance) under every template; and the conjunction stays at or below chance for four of five models under all templates. OPT's marginal above-chance conjunction score is inconsistent and occurs alongside below-chance abstract-map performance, which indicates noise rather than genuine conjunction identification.

Table 6: Symbolic prompt-template robustness of the rule-ladder: balanced-accuracy identification margin (median over $n \geq 8$ and five templates). copy / abstract map: $\mathrm{bacc}_{\mathrm{id}} - \mathrm{bacc}_{\mathrm{exemplar}}$ ($> 0 \Rightarrow$ identifies); conjunction: $\mathrm{bacc}_{\mathrm{id}} - 0.5$ ($> 0 \Rightarrow$ above chance). Copy is positive for every model under all five templates; per-template signs differ only where the pooled margin is near zero. Generated by `src/tab_idrule_ps.py`.

rule	Qwen2.5-7B	Qwen2.5-14B	Mistral-7B	OPT-6.7B	phi-2
copy (induction)	$+0.68$	$+0.58$	$+0.56$	$+0.16$	$+0.36$
abstract map	$+0.03$	$+0.15$	$+0.01$	-0.10	-0.01
conjunction (2-feature)	-0.18	-0.14	-0.23	$+0.07$	-0.06

Hyperparameters. Main LTI transformer: $d_{\mathrm{model}}{=}384$, 12 layers, 6 heads ($\sim$21M params), 80k steps, AdamW (lr$=10^{-4}$, weight decay 0.01), 3000-step warmup then cosine decay to 0.1 lr, gradient clip 1.0, bf16; the symbol (HMM) model uses the same architecture with 40k steps and a 2000-step warmup, and the mechanism-flip models 50k steps at $T_{\mathrm{ctx}}{=}128$. Ablation / exact-posterior models: $d_{\mathrm{model}}{=}192$, 6 layers, 4 heads (2.7M), 24k steps, lr$=3\times10^{-4}$, 300-step warmup. AR families: context $T_{\mathrm{ctx}}{=}24$, horizon $N{=}8$, observation noise $\sigma{=}0.3$, poles drawn $U[0.4, 0.9]$ (AR(1)) or $U[0.2, 0.9]$ (AR(2): two i.i.d. real poles), input amplitude log-uniform on $[0.05, 3.0]$, outputs clipped to ±16; the exact posterior uses a 60×60 grid (AR(2)) or 300-point grid (AR(1)). Phase sweeps: $n_x \in \{2, 3, 4\}$, $T_{\mathrm{ini}}{=}5$, $N{=}8$ (so $L{=}13$), $k \in \{1, \ldots, 12\}$.

Inputs, baselines, and "knee". A k-tone multisine draws k frequencies on $(0.05\pi, 0.95\pi)$ with minimum pairwise gap 0.02, amplitudes $U[0.5, 1.5]$, and uniform random phases, normalized to unit std, giving PE order exactly $2k$ when $T \geq 4k - 1$. DeePC (2): min-norm is the unregularized least-norm g; ridge-DeePC adds $\lambda\|g\|^2$ with a single λ tuned once on a disjoint validation seed (logarithmic grid) and then applied to every context. ARX uses the true order n_x (oracle); the observer uses known (A, B, C). We define the empirical "knee" as the smallest k at which the median error has dropped by half relative to its $k{=}1$ value, applied identically to every predictor. This captures the early collapse before the transformer's slow post-collapse tail and locates the knee

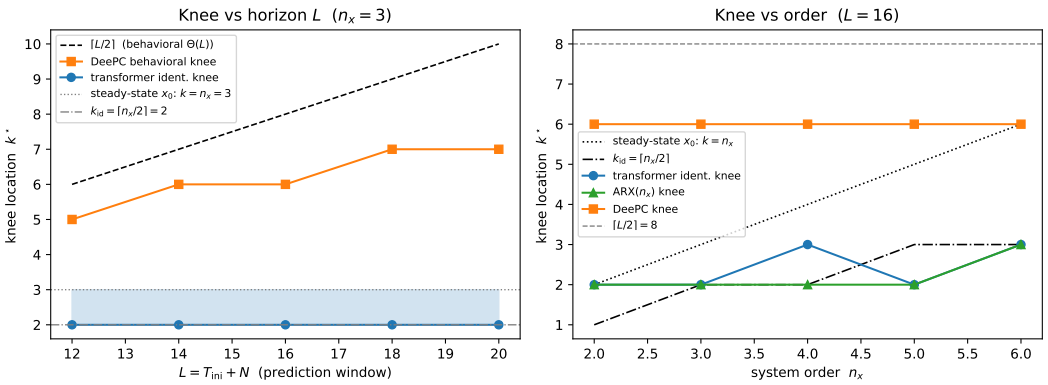


Figure 15: **The two thresholds scale as predicted** (trained 21M model, 5 seeds; knee = 50% drop from k=1). Left: vs. horizon L (n_x=3); DeePC's knee rises with L, tracking $\lceil L/2 \rceil$ from below (its exact collapse sits at $\lceil L/2 \rceil$ in every cell, $k_{\text{beh}} = \Theta(L)$), while the transformer's identification knee is flat in L (horizon-free). Right: vs. order n_x (L=16); the transformer tracks the ARX identifier's knee ($k \approx$ 2–3), within one tone of $\lceil n_x/2 \rceil$ and far below DeePC's behavioral knee. The random initial state supplies the free-response excitation that lowers the requirement from the steady-state $k = n_x$ to $\lceil n_x/2 \rceil$ (Prop. 2).

at $k \approx$ 2–3. For DeePC we also report its rank knee k_{beh}, the exact collapse, which the 50% rule precedes by one to three tones.

Order-extrapolation control (§6). A NARROW 21M model (identical recipe, order prior fixed to n_x=3; 2 seeds) against the WIDE main-sweep models ($n_x \in \{2, 3, 4\}$; 5 seeds), evaluated at $n_x = 2, \ldots, 6$ (T_{ini}=6 for n_x=6, since ARX(n_x) needs $T_{\text{ini}} \geq n_x$). Entries: median rollout-NRMSE plateau over $k \geq n_x$+1 (per-seed median over $\geq$ 256 systems, then across seeds); gap in units of the WIDE seed std.

n_x	2	3	4	5	6
WIDE	0.040	0.043	0.047	0.057	0.063
NARROW	0.032	0.039	0.046	0.053	0.064
gap (σ)	-1.0	-0.6	-0.1	-0.3	$+0.1$

Code. All training, evaluation, and figure-generation scripts will be released.