When Do Structured Prompts Help In-Context Learning? A Relevance–Redundancy Theory of In-Context Prediction

Anonymous authors
Paper under double-blind review

Abstract

In-context learning is commonly analyzed under independent demonstrations, yet real prompts are often structured: demonstrations can be correlated with one another and with the query. When does this structure improve in-context prediction, and when does it merely make the context redundant? We develop a relevance–redundancy account in the solvable setting of in-context linear regression, allowing an arbitrary PSD unit-diagonal sample-similarity kernel over the demonstrations and query. For a linear-attention predictor trained on independent prompts and deployed on a structured one, we derive an exact finite-dimensional excess-risk law. Structure decomposes into a redundancy cost among demonstrations, a relevance gain between demonstrations and the query, and a higher-order mediated interaction. In the weak-structure regime, prediction improves precisely when $Q > \alpha R/(1+\rho)$, where $\alpha = \ell/d$ is the context-to-dimension ratio and ρ is the label-noise variance. We then characterize which relevance–redundancy combinations arise from coherent structure. Under graph-induced covariance, the two terms become internal context connectivity and query connectivity. In a sparse stochastic block model, a single latent context-composition variable changes relevance linearly but redundancy quadratically, yielding explicit homophily- and heterophily-dependent help–hurt phase transitions. Finally, trained full linear attention closely follows the exact law; nonlinear softmax attention often depends on the same coordinates, but with architecture-dependent coefficients. Thus relevance and redundancy describe a shared prompt geometry, while their exchange rate is model-specific.

1 Introduction

Most theoretical analyses of in-context learning (ICL) assume independent demonstrations, especially in the now-standard linear-regression testbed (Brown et al., 2020; Akyürek et al., 2023; Von Oswald et al., 2023). Real prompts rarely look that clean: retrieved examples can be mutually similar (Liu et al., 2022; Rubin et al., 2022), graph neighbors share latent structure, and examples selected for their similarity to a query are usually correlated with the query itself. The key issue for in-context prediction is therefore not only how many demonstrations a prompt contains, but where the dependence lies inside it.

Structure has two opposing effects. Correlation among demonstrations creates *redundancy*: several examples may repeat the same information about the latent task. Correlation between demonstrations and the query creates *relevance*: context labels can become directly informative about the query target. Thus "more correlation" is not intrinsically good or bad; two prompts with comparable overall dependence can have different risks depending on how that dependence is split between context–context and query–context interactions.

This motivates the question behind our title:

> *When does structure help in-context learning, rather than merely make the context redundant?*

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

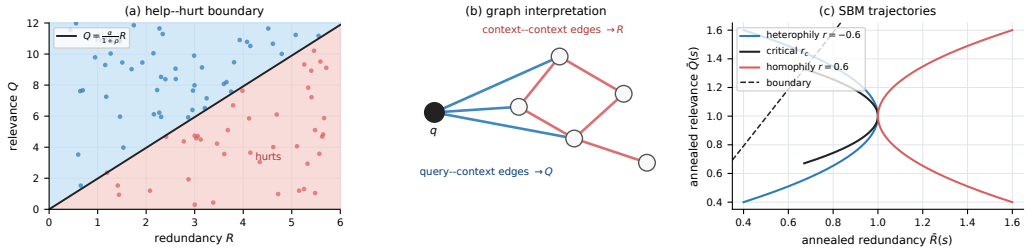


Figure 1: **Relevance–redundancy geometry.** (a) The linear-attention law separates structured prompts by the boundary $Q = \alpha R/(1+\rho)$. (b) On graphs, R is induced by context–context connectivity and Q by query–context connectivity. (c) A coherent latent graph does not move freely in the (R, Q) plane: sparse-SBM contexts trace homophilic or heterophilic trajectories through the same boundary.

We answer this question exactly in a solvable model: in-context linear regression under linear attention. The model captures one-step gradient-descent ICL and has served as a bridge from theory to nonlinear Transformer experiments (Zhang et al., 2024; Wu et al., 2024; Lu et al., 2025). A prompt contains ℓ labeled demonstrations and one query, with features jointly Gaussian under an arbitrary unit-diagonal similarity kernel $S = I + T$. Its geometry enters through redundancy $\mathcal{R}(C) = \ell^{-1}\|T_{CC}\|_F^2$, query relevance $\mathcal{Q}(q, C) = \|T_{Cq}\|_2^2$, and a higher-order mediated term $\mathcal{M}(q, C) = T_{Cq}^\top T_{CC} T_{Cq}$. For a linear-attention predictor trained on independent prompts and deployed, without retraining, on any structured prompt, the large-dimensional form of our exact finite-d law is

$$\mathcal{L}(S) - \mathcal{L}_{\text{iid}} = \frac{\alpha\mathcal{R}(C) - (1+\rho)\mathcal{Q}(q, C) + \mathcal{M}(q, C)}{(\alpha + 1 + \rho)^2}, \qquad \alpha := \frac{\ell}{d}, \tag{1}$$

with exact coefficients stated in Theorem 1. The law makes the sign transparent: redundancy is costly, direct query relevance is helpful, and the mediated term matters at higher order. In the weak-structure regime, the prompt improves over an independent prompt of the same length precisely when

$$\boxed{\mathcal{Q}(q, C) > \frac{\alpha}{1+\rho}\mathcal{R}(C)}. \tag{2}$$

The ratio $\alpha/(1+\rho)$ is the exchange rate between relevance and redundancy for one-step linear attention. It rises with the number of demonstrations per feature dimension and falls with label noise. Figure 1 summarizes the resulting geometry.

This also clarifies effective-context-length interpretations of correlated ICL (Letey et al., 2026). When the query is independent of the demonstrations, $\mathcal{Q} = \mathcal{M} = 0$ and dependence only discounts the amount of useful context. Once the query is correlated with the context, that one-dimensional picture breaks: even a redundant context can help if it is sufficiently relevant to the query. Effective context length is therefore an independent-query slice of a broader relevance–redundancy geometry. This is where arbitrary covariance changes the problem rather than only the notation: sequential kernels trace a one-dimensional ray in the $(\mathcal{R}, \mathcal{Q}, \mathcal{M})$ space, while general structured prompts can move relevance and redundancy separately before a coherent generator, such as a graph, couples them again.

The exact law treats $(\mathcal{R}, \mathcal{Q})$ as coordinates of an arbitrary covariance matrix. A coherent source of structure, however, need not vary these coordinates freely. We study graph-induced covariance as such a source. For a normalized graph heat kernel, $S_G(\tau) = I + \tau A + O_G(\tau^2)$, so $\mathcal{R}(C) = \tau^2 2|E(C)|/\ell + O_G(\tau^3)$ and $\mathcal{Q}(q, C) = \tau^2 d_C(q) + O_G(\tau^3)$. The abstract boundary becomes a comparison between internal context connectivity and query-to-context connectivity.

Graph structure also couples the two quantities. In a sparse two-community SBM, where homophily and heterophily refer to community-dependent feature covariance rather than graph-smooth labels, let s be the signed composition of the context relative to the query's community and write $r = b/a$ for assortativity. Conditional on s, expected query connectivity is linear, $a + bs$, while expected context redundancy is quadratic, $a(1 - 1/\ell) + b(s^2 - 1/\ell)$. Combining this trajectory with the help–hurt boundary gives the large-ℓ phase function $\Psi(s; t, r) = t(1 + rs^2) - (1 + rs)$ with $t = \alpha/(1+\rho)$. Homophily and heterophily therefore trace different paths through the same boundary: homophily can increase relevance but eventually pays a quadratic redundancy cost, whereas heterophily can reduce internal redundancy while maintaining query relevance.

Finally, we ask whether these coordinates remain predictive beyond the solvable reduced model. A full-parameter one-layer linear-attention network trained on independent prompts closely follows the exact law when deployed on structured prompts. Softmax attention shows a related low-dimensional dependence on the same statistics, but with a different exchange rate and larger architecture-level variation. These results separate shared prompt geometry from algorithm-specific pricing: relevance and redundancy remain predictive coordinates, but their exchange rate depends on the inference algorithm.

Contributions. Our results form a progression from exact theory to coherent structure and trained attention models:

1. **Exact relevance–redundancy law for in-context prediction.** For arbitrary unit-diagonal sample-similarity kernel and finite d, ℓ, we derive the exact population risk of reduced linear attention and an exact linear excess-risk law for a predictor trained at independence and deployed on a structured prompt. The leading terms yield the model-specified help–hurt boundary $\mathcal{Q} > \alpha\mathcal{R}/(1+\rho)$.
2. **Coherent structure constrains the trade-off.** For graph-induced covariance, relevance and redundancy reduce to query connectivity and internal context connectivity. In a sparse SBM, one latent context-composition variable controls them linearly and quadratically, producing explicit homophily- and heterophily-dependent phase transitions.
3. **Transfer beyond the reduced model.** Trained full linear attention closely follows the analytic law under a fixed-model deployment protocol, while softmax attention exhibits a related low-dimensional dependence with architecture-dependent coefficients.

A consequence of the first result is a demonstration-selection criterion: at fixed context budget, the leading linear-attention theory favors sets maximizing $(1+\rho)\mathcal{Q} - \alpha\mathcal{R}$. We treat this as an implication rather than a separate contribution; its links to submodular selection, MMR, and determinantal diversity (Carbonell and Goldstein, 1998; Kulesza and Taskar, 2012) are developed in the appendix.

The paper proceeds as follows. Section 2 discusses related work; Section 3 defines structured in-context regression and linear attention; Section 4 proves the exact law; Section 5 derives the graph and SBM phase transitions; Section 6 tests trained linear and softmax attention.

2 Related Work

ICL theory. Linear regression has become a canonical testbed for understanding the algorithms implemented by in-context learners. Transformers can realize least-squares, gradient-based, and closed-form regression procedures inside a single forward pass (Akyürek et al., 2023; Von Oswald et al., 2023), and subsequent work characterizes when trained attention implements one-step gradient descent, preconditioned updates, higher-order methods, or competing shortcut algorithms (Ahn et al., 2023; Mahankali et al., 2024; Fu et al., 2024; Vladymyrov et al., 2024; Bai et al., 2023; Kawata et al., 2025). A complementary line studies population training, task diversity, sample complexity, and the high-dimensional risk of in-context regression (Zhang et al., 2024; Wu et al., 2024; Atanasov et al., 2025; Moniri and Hassani, 2025). Our work uses this solvable linear-regression setting differently: rather than identifying which algorithm a Transformer learns, we fix the deployed predictor and ask how structured prompt covariance changes its risk. Methodologically, we follow theory-to-architecture studies that derive exact laws for linear attention and then test their transfer to nonlinear Transformers (Lu et al., 2025). The closest technical point of contact is Letey et al. (2026), who study sequentially correlated prompts and separate context-correlation from query-correlation effects. Our theorem recovers their weak-correlation exponential-kernel expressions on the sequential ray (Remark 1) but removes the restriction that one correlation-length parameter controls all prompt dependence. Away from that ray, two prompts can have comparable average correlation and opposite excess-risk signs because query–context relevance and context–context redundancy vary independently. The graph and SBM results exploit precisely this extra freedom: a coherent kernel then ties the coordinates together in a structural, rather than sequential, way.

Structured prompting. Many practical prompting pipelines deliberately introduce dependence: retrieved examples are chosen for query similarity (Liu et al., 2022; Rubin et al., 2022), selection rules often add coverage or diversity (Su et al., 2023; Levy et al., 2023), and MMR or determinantal

objectives make the similarity–diversity trade-off explicit (Carbonell and Goldstein, 1998; Kulesza and Taskar, 2012; Ye et al., 2023; Yang et al., 2023). Graph-based prompting and graph foundation models similarly exploit relational structure in graph reasoning, molecular prediction, and graph-language systems (Mao et al., 2024; Wang et al., 2023; Lin et al., 2025; Fu et al., 2026; Al-Lawati et al., 2025; Liu et al., 2024; 2025). Our use of graphs is not to propose a graph architecture or a new selector, but to make dependence coherent: a single graph kernel simultaneously determines context–context redundancy and query–context relevance. This lets us convert the abstract relevance–redundancy boundary into degree-level conditions and, for sparse SBMs, into an explicit homophily–heterophily phase diagram. The selection objective in the appendix is therefore a consequence of the same risk law: at fixed budget, the leading theory favors maximizing $(1+\rho)\mathcal{Q}(q,C) - \alpha\mathcal{R}(C)$, providing a model-specified statistical explanation for why similarity-only retrieval can fail when the selected demonstrations are mutually redundant.

3 Setup

A prompt is a context $C = \{1, \dots, \ell\}$ of labelled demonstrations and a query q. Demonstration i carries $x_i \in \mathbb{R}^d$ and $y_i = x_i^\top w + \epsilon_i$; the query carries x_q and the target $y_q = x_q^\top w + \epsilon_q$, with $w \sim \mathcal{N}(0, I_d)$ and $\epsilon_i, \epsilon_q \overset{\text{iid}}{\sim} \mathcal{N}(0, \rho)$. The learner sees $\{(x_i, y_i)\}_{i\in C}$ and x_q and predicts y_q; $\alpha := \ell/d$.

Structured prompts. Structure enters through a similarity kernel: a PSD matrix S with unit diagonal indexed by $C \cup \{q\}$, such that features are jointly Gaussian with $\mathbb{E}[x_u x_v^\top] = S_{uv} I_d/d$, i.e. the full prompt feature matrix satisfies $\mathsf{X} = d^{-1/2} S^{1/2} Z$ for i.i.d. Gaussian Z. We reserve $X_C \in \mathbb{R}^{\ell\times d}$ and $y_C \in \mathbb{R}^\ell$ for the context design and labels used by the predictor. Thus $S = I$ is independent demonstrations and S_{uv} the feature correlation of u and v. We write $S = I + T$ with T symmetric and zero-diagonal, and $K := S[C, C] = I + T_{CC}$, $k_q := S[C, q] = T_{Cq}$. They are two blocks of one matrix: the same structure fixes how much the context repeats itself and how much it says about the query. The task is not itself structured by S: w is isotropic and independent of the kernel and label noise is i.i.d., so homophily creates feature dependence rather than a separate smooth-label signal.

Three summaries. The risk of the predictor below depends on (K, k_q) only through $\kappa_{\mathrm{c}} := \mathrm{tr}(K^2)/\ell$, $\kappa_{\mathrm{q}} := \|k_q\|^2$ and $\kappa_{\mathrm{m}} := k_q^\top K k_q$, the summaries of Letey et al. (2026). In terms of similarities they are elementary.

Lemma 1 (Prompt statistics)**.** *Let* $\mathcal{R}(C) := \frac{1}{\ell}\|T_{CC}\|_F^2$, $\mathcal{Q}(q,C) := \|T_{Cq}\|^2$ *and* $\mathcal{M}(q,C) := T_{Cq}^\top T_{CC} T_{Cq}$. *Then, exactly,* $\kappa_{\mathrm{c}} = 1 + \mathcal{R}$, $\kappa_{\mathrm{q}} = \mathcal{Q}$ *and* $\kappa_{\mathrm{m}} = \mathcal{Q} + \mathcal{M}$.

The proof is one line: $K^2 = I + 2T_{CC} + T_{CC}^2$ and $\mathrm{tr}\, T_{CC} = 0$. $\mathcal{R}$ is the mean squared similarity of a demonstration to the rest of the context, $\mathcal{Q}$ the squared similarity of the query to the context, and $\mathcal{M}$ query similarity routed through pairs of demonstrations that are themselves similar — third order in T where $\mathcal{R}$, $\mathcal{Q}$ are second order.

In-context learning with linear attention. The prompt is presented as $\ell+1$ tokens $z_i = (x_i, y_i) \in \mathbb{R}^{d+1}$ for the demonstrations and $z_q = (x_q, 0)$ for the query, whose label slot is masked; there are no positional encodings, so the prompt is a set. A single linear-attention layer with merged query–key and value–output matrices $W_{QK}, W_{VO} \in \mathbb{R}^{(d+1)\times(d+1)}$ reads the prediction off the label coordinate of the query's output. Equivalently, after absorbing the value readout and the conventional factor d into a reduced parameter $\Gamma \in \mathbb{R}^{d\times(d+1)}$, this class can be written as

$$\widehat{y}_q = x_q^\top \Gamma H_C e_y, \qquad H_C := \frac{d}{\ell}\sum_{i\in C} z_i z_i^\top, \qquad e_y := e_{d+1}. \tag{3}$$

This is the standard population reduction of Zhang et al. (2024); Wu et al. (2024); Lu et al. (2025): the query feature x_q acts on the context moment H_C, and $H_C e_y = (d/\ell)\sum_i z_i y_i$ is the label column of that moment. We make the usual assumptions of this line of work: the task is isotropic ($w \sim \mathcal{N}(0, I_d)$), the population risk is the squared error at the query, and the model is trained to its population optimum on a given prompt distribution. Since $x_v \mapsto Ox_v$, $w \mapsto Ow$ and $z_v \mapsto \widetilde{O} z_v$,

with $\widetilde{O} = \mathrm{diag}(O, 1)$, preserve the model law for every $O \in O(d)$ and every S, and the risk is a convex quadratic in Γ invariant under $\Gamma \mapsto O\Gamma\widetilde{O}^\top$, the population optimum is rotation-invariant: $\Gamma^\star = [\gamma^\star I_d \; 0]$. Hence the predictor collapses to the scalar family

$$\widehat{y}_q = \gamma \frac{d}{\ell} x_q^\top X_C^\top y_C, \qquad \mathcal{L}(\gamma; S) = 1 + \rho - 2a\gamma + b\gamma^2, \tag{4}$$

which is one step of gradient descent on the context from a zero task estimate, with step γ (Von Oswald et al., 2023; Mahankali et al., 2024). We write $\mathcal{L}^\star(S) = \min_\gamma \mathcal{L}(\gamma; S)$ and $\gamma_0 := \arg\min_\gamma \mathcal{L}(\gamma; I) = \alpha/(\alpha+\rho+1+1/d)$ for the predictor trained at independence. The reduction is exact for the population optimum; the trained full-parameter model of Section 6 is where it is tested.

The graph instance. For a graph with adjacency A and normalized heat kernel $S_G(\tau) = D_\tau^{-1/2} e^{-\tau L} D_\tau^{-1/2}$, one has $S_G = I + \tau A + O(\tau^2)$ (Appendix D), so for an unweighted graph

$$\mathcal{R} = \tau^2 \bar{d}_{\mathrm{in}}(C) + O_G(\tau^3), \quad \mathcal{Q} = \tau^2 d_C(q) + O_G(\tau^3), \quad \mathcal{M} = O_G(\tau^3), \tag{5}$$

the mean degree of the context-induced subgraph and the degree of the query into it. We use graphs to generate prompts throughout, since they make the boundary a statement about degrees and supply a natural candidate pool. These experiments are therefore tests of structured covariance and topology, not of a separate graph-smooth target model.

4 The Relevance–Redundancy Law

Proposition 1 (Exact population risk). *For any unit-diagonal S, finite d, ℓ and isotropic tasks, $\mathcal{L}^\star = 1 + \rho - a^2/b$ with $\gamma^\star = a/b$ and, writing $e_1 = 1 + \frac{1}{d}$, $e_3 = 1 + \frac{3}{d} + \frac{4}{d^2}$,*

$$a = 1 + \frac{\kappa_{\mathrm{q}} e_1}{\alpha}, \qquad b = 1 + \frac{\rho}{\alpha} + \frac{(\kappa_{\mathrm{c}} + 2\kappa_{\mathrm{q}})e_1}{\alpha} + \frac{\kappa_{\mathrm{m}} e_3 + \rho\kappa_{\mathrm{q}} e_1}{\alpha^2}. \tag{6}$$

The only non-elementary ingredient is a sixth-order Gaussian moment (Appendix E); $d \to \infty$ recovers the weak-correlation expressions of Letey et al. (2026). The statement closes on the three summaries, so their sufficiency is exact. The finite-d re-optimised surface follows directly from the ratio a^2/b. For readability we record the $d \to \infty$ surface, which is still nonperturbative in S and uses no weak-similarity expansion:

$$\mathcal{L}^\star < \mathcal{L}^\star_{\mathrm{iid}} \iff (1 - \kappa_{\mathrm{c}}) + \frac{(2+\rho)\kappa_{\mathrm{q}} - \kappa_{\mathrm{m}}}{\alpha} + \frac{D_\alpha \kappa_{\mathrm{q}}^2}{\alpha^2} > 0, \qquad D_\alpha := \alpha + 1 + \rho. \tag{7}$$

4.1 An exact law for a deployed predictor

A selector does not retrain the model for each prompt it builds; the relevant object is the risk of a predictor trained on one distribution and deployed on the prompt the selector chose.

Theorem 1 (Exact relevance–redundancy law). *Let $\mathcal{L}(S) := \mathcal{L}(\gamma_0; S)$. For any unit-diagonal S and finite d, ℓ, with no remainder,*

$$\mathcal{L}(S) - \mathcal{L}_{\mathrm{iid}} = \frac{\alpha e_1 \mathcal{R} - \big((1+\rho)e_1 - \frac{2}{d^2}\big)\mathcal{Q} + e_3 \mathcal{M}}{(\alpha + \rho + e_1)^2} \xrightarrow{d\to\infty} \frac{\alpha \mathcal{R} - (1+\rho)\mathcal{Q} + \mathcal{M}}{D_\alpha^2}. \tag{8}$$

The re-optimised predictor satisfies $\mathcal{L}^\star(S) - \mathcal{L}^\star_{\mathrm{iid}} = \mathcal{L}(S) - \mathcal{L}_{\mathrm{iid}} - b(\gamma^\star - \gamma_0)^2$ with $\gamma^\star - \gamma_0 = O(\|T\|^2)$, so Equation (8) holds for it to fourth order.

The proof is short (Appendix F): a and b are affine in the summaries, so $\mathcal{L}(\gamma; \cdot)$ is affine in them for any fixed γ, Lemma 1 converts summaries to similarities, and the statement about $\gamma^\star$ is the envelope theorem. The same computation gives the law for any fixed γ, which we will need.

Corollary 1 (Any fixed parameter). *For every fixed γ, $\mathcal{L}(\gamma; S) - \mathcal{L}(\gamma; I) = c_R(\gamma)\mathcal{R} - c_Q(\gamma)\mathcal{Q} + c_M(\gamma)\mathcal{M}$ with $c_R = \gamma^2 e_1/\alpha$, $c_Q = 2\gamma e_1/\alpha - \gamma^2\big(2e_1/\alpha + (e_3 + \rho e_1)/\alpha^2\big)$, $c_M = \gamma^2 e_3/\alpha^2$. Near γ_0 the rate $\beta(\gamma) := c_R/c_Q$ is increasing in γ: a predictor with a smaller step pays less for redundancy and gains more from relevance.*

Reading the law. At γ_0 and $d \to \infty$ the local derivatives are $\partial_{\kappa_c}\mathcal{L} = \alpha/D_\alpha^2$, $\partial_{\kappa_q}\mathcal{L} = -(2+\rho)/D_\alpha^2$, $\partial_{\kappa_m}\mathcal{L} = 1/D_\alpha^2$: redundancy is harmful, direct query coupling beneficial, and coupling mediated through the context harmful, since a demonstration similar both to the query and to other demonstrations mostly repeats them. Up to the third-order term the prompt beats an independent one of the same length iff

$$\mathcal{Q}(q, C) > \frac{\alpha}{1+\rho}\,\mathcal{R}(C). \tag{9}$$

The ratio $\alpha/(1+\rho)$ is the exchange rate between the two criteria every selector balances. It rises with the number of demonstrations per dimension because the one-step task estimate is then already good and repeated demonstrations mostly add variance to it, and falls with label noise because the direct route from y_i to y_q is then worth relatively more. Setting $\mathcal{Q} = \mathcal{M} = 0$ recovers the effective context length of Letey et al. (2026), $\ell_{\text{eff}} = \ell(1 - \mathcal{R}/(1+\rho)) + O(\mathcal{R}^2)$: correlation as a discount is the independent-query slice of the law. With a correlated query that picture fails at every length — a context with no internal similarity whose query is similar to most of it has $\ell_{\text{eff}} = \ell$ and a much lower risk — because in-context prediction is transductive and similarity opens a second route in which the y_i inform y_q directly.

Corollary 2 (Graphs)**.** *Under the heat kernel, as $\tau \to 0$ at fixed G, C, q,*

$$\mathcal{L}_G(\tau) - \mathcal{L}_{\text{iid}} = \frac{\tau^2}{(\alpha+\rho+e_1)^2}\Big[\alpha e_1 \bar{d}_{\text{in}}(C) - \big((1+\rho)e_1 - \tfrac{2}{d^2}\big)d_C(q)\Big] + O_G(\tau^3).$$

Thus below some $\tau_0(G, C, q)$ the prompt helps iff

$$d_C(q) > \frac{\alpha e_1}{(1+\rho)e_1 - 2/d^2}\,\bar{d}_{\text{in}}(C),$$

whose large-d limit is $\alpha\bar{d}_{\text{in}}(C)/(1+\rho)$.

The query must be attached to the context more strongly than the context is attached to itself, by a factor set by α, ρ, and a small finite-d correction. Across 170 random graphs the sign of $\Delta\mathcal{L}$ follows this boundary on every instance, and the empirical threshold tracks $\alpha/(1+\rho)$ as each parameter is varied separately, without a fitted constant (Appendix J). On four real topologies the boundary holds on every random prompt for three and degrades only on the densest, where the two-hop term the degree expansion drops is large (Appendix M).

Remark 1 (Sequences)**.** *For the exponential kernel of sequential models the summaries lie on a ray, $(\kappa_c - 1, \kappa_q, \kappa_m) = \kappa_q(2(1-\frac{1}{\ell}), 1, 1) + O(e^{-4/\xi})$ (Appendix G), so at leading order a sequential prompt beats an independent one iff $\alpha < \frac{1+\rho}{2}$ in the weak-correlation regime $e^{-1/\xi} \ll 1$. At the $\alpha = 4$ of Letey et al. (2026) every weakly correlated sequential prompt is harmful, consistent with their curves; the law predicts the leading sign flips once $\ell < (1+\rho)d/2$. On a set the ray opens into a cone and the boundary (9) lies in its interior: allocation becomes a question only once the geometry can allocate.*

4.2 Beyond weak similarity

Because Proposition 1 is exact, the ways the two-statistic reading fails can be named (Appendix C). The mediated term $\mathcal{M}$ is an independent direction, exactly zero only when demonstrations similar to the query are not similar to each other, and it is what a selector must add on a dense context (Appendix A). The re-optimised predictor carries a positive second-order term $D_\alpha\kappa_q^2/\alpha^2$, so its boundary bends towards tolerating more redundancy as similarity strengthens. And on graphs the binding error is the kernel expansion itself, which drops two-hop paths, so the condition is $\tau \cdot (\text{second-order neighbourhood}) \ll 1$ rather than $\tau \ll 1$; heavy-tailed degrees turn out not to be the problem, density is.

5 Coherent Graph Coupling and Phase Transitions

The degree form in Corollary 2 makes graph structure a coherent generator of relevance and redundancy. We now ask which pairs $(\mathcal{R}, \mathcal{Q})$ such a generator can realize. The answer is already visible in the simplest latent-community model: one composition variable changes the query degree linearly but the internal context degree quadratically, forcing an explicit phase transition.

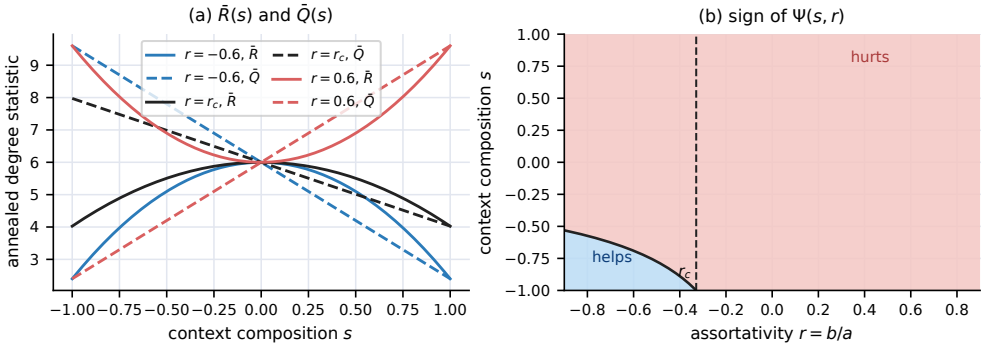


Figure 2: **Sparse-SBM structural trajectories.** Left: the annealed redundancy statistic $\bar{R}(s)$ is quadratic in the context composition s, while the query statistic $\bar{Q}(s)$ is linear. Right: substituting this coupled trajectory into the help–hurt law yields homophily- and heterophily-dependent phase transitions; $\Psi < 0$ is the beneficial region.

Consider a sparse two-block SBM in which context node i has label $z_i \in \{\pm 1\}$, the query has $z_q = +1$, and an edge between two nodes of labels z, z' is present with probability $(a + bzz')/\ell$; write $r = b/a$ for the assortativity and $s = \ell^{-1} \sum_{i \in C} z_i$ for the signed community composition of the context. Throughout this section "homophily" means assortative feature covariance induced by the graph kernel, not a target label that is smooth on the graph.

Proposition 2 (SBM relevance–redundancy coupling)**.** *Conditional on s,*

$$\mathbb{E}[d_C(q) \mid s] = a + bs, \qquad \mathbb{E}[\bar{d}_{\text{in}}(C) \mid s] = a\Big(1 - \frac{1}{\ell}\Big) + b\Big(s^2 - \frac{1}{\ell}\Big). \tag{10}$$

Proof. $\mathbb{E}[A_{iq} \mid z_i] = (a + bz_i)/\ell$ and $\sum_i z_i = \ell s$ give the first identity. For the second, $\bar{d}_{\text{in}}(C) = \ell^{-1} \sum_{i \neq j} A_{ij}$ and $\mathbb{E}[A_{ij} \mid z] = (a + bz_i z_j)/\ell$, so $\mathbb{E}[\bar{d}_{\text{in}} \mid z_C] = \ell^{-2} \sum_{i \neq j} (a + bz_i z_j)$; the constant term contributes $a(1 - 1/\ell)$ and $\sum_{i \neq j} z_i z_j = (\sum_i z_i)^2 - \sum_i z_i^2 = \ell^2 s^2 - \ell$. □

Query relevance varies linearly with s and context redundancy quadratically, so the family traces the one-dimensional trajectory $s \mapsto (a + bs^2,\ a + bs)$ in the $(\mathcal{R}, \mathcal{Q})$ plane at large ℓ, rather than filling it.

Corollary 3 (Annealed SBM phase law)**.** *Let $t = \alpha/(1 + \rho)$. In the large-d coefficient limit, for fixed ℓ and weak diffusion, $\mathbb{E}_G[\mathcal{L}_G(\tau) - \mathcal{L}_{\text{iid}} \mid s] = \tau^2 \Phi_\ell(s)/(\alpha + 1 + \rho)^2 + O_G(\tau^3)$ with $\Phi_\ell(s) = \alpha[a(1 - \frac{1}{\ell}) + b(s^2 - \frac{1}{\ell})] - (1 + \rho)(a + bs)$. Ignoring the $O(\ell^{-1})$ correction the sign is governed by $\Psi(s; t, r) = t(1 + rs^2) - (1 + rs)$.*

The law is annealed because $d_C(q)$ is an $O(1)$ query degree that does not concentrate around $a + bs$ as $\ell \to \infty$, whereas $\bar{d}_{\text{in}}$ aggregates $O(\ell)$ edges; individual prompts form a noisy cloud around the trajectory while conditional means follow the phase law.

Corollary 4 (Homophily and heterophily)**.** *For $r > 0$, Ψ is convex in s with minimiser $s^\star = 1/(2t)$ and minimum $t - 1 - r/(4t)$, so for $t \geq 1$ a homophilic beneficial regime exists only if $r > 4t(t - 1)$; since $r \leq 1$, no homophilic SBM helps once $t \geq (1 + \sqrt{2})/2 \approx 1.207$. For $r < 0$, $\Psi(-1; t, r) = t(1+r) - (1-r)$, so heterophily enters the helpful phase at $s = -1$ exactly when $r < -(t-1)/(t+1)$. For the experimental setting $\alpha = 2$, $\rho = 0.01$, $t \approx 1.980$ and $r_c \approx -0.329$.*

Both statements are elementary algebra on Ψ. The key point is that the same graph cannot choose relevance and redundancy independently: once the context composition s is fixed, query connectivity moves linearly while internal context density moves quadratically. Thus homophily and heterophily trace different paths through the same help–hurt boundary.

To check that the annealed law describes finite sampled prompts, we sample SBM graphs at $\ell = 80$, $a = 6$, $\alpha = 2$, and $\rho = 0.01$, condition on the rounded composition s, and compare the sampled leading numerator $\alpha \bar{d}_{\text{in}}(C) - (1 + \rho) d_C(q)$ with $(1 + \rho) a\, \Psi(s; t, r)$. Figure 3 shows that the conditional means follow the predicted curves, while the query degree produces the expected finite-prompt scatter. The heterophilic curve crosses into the beneficial phase near the analytic threshold $r_c = -(t - 1)/(t + 1)$; the homophilic curve stays harmful in this high-α setting.

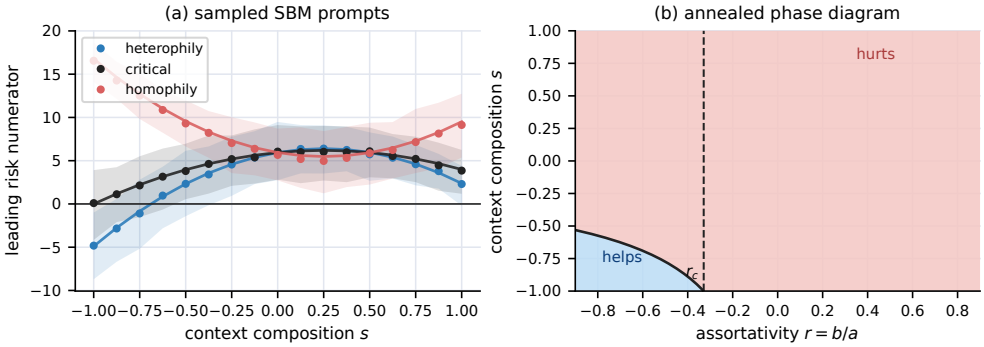


Figure 3: **Finite-sample SBM phase validation.** (a) Sampled SBM prompts follow the annealed phase curves as context composition varies; bands show the central 80% of prompt-level variation. Negative values predict improvement over independence. (b) The annealed help–hurt region in the (r, s) plane for $\alpha = 2$, $\rho = 0.01$; the dashed line marks the heterophilic threshold $r_c = -(t-1)/(t+1)$.

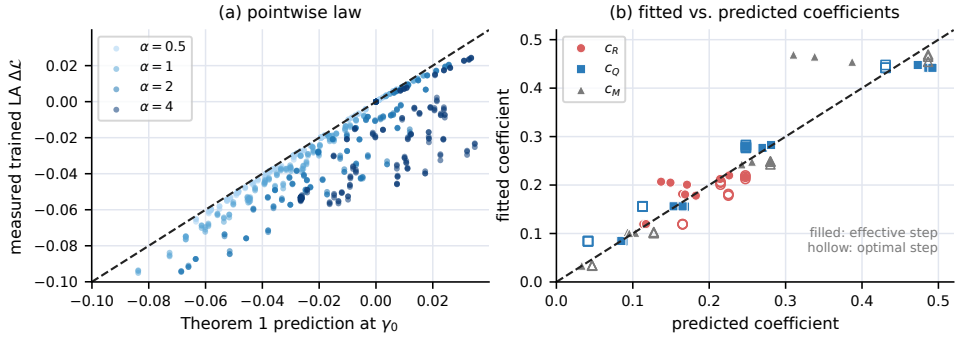


Figure 4: **The law holds for a trained model, and its residual is a single scalar.** (a) Measured $\Delta\mathcal{L}$ of a one-layer linear-attention model trained on independent prompts, against Equation (8) at γ_0, on every prompt of the bank. (b) The three fitted coefficients of the trained model against Corollary 1 evaluated at the γ read off the model's own independence risk (filled) and at γ_0 (hollow); all α and seeds.

6 Trained Linear Attention and Softmax Transfer

Theorem 1 describes a model trained on one distribution and deployed on another, and that is how we test it. We train a one-layer linear-attention model with full parameter matrices $W_{QK}, W_{VO} \in \mathbb{R}^{(d+1)\times(d+1)}$ — not the scalar γ — on independent prompts ($d = 16$, AdamW, 3000 steps), then evaluate it, unchanged, on a bank of 60 structured prompts spanning the $(\mathcal{R}, \mathcal{Q})$ plane at $S = I + 0.15A$, for $\alpha \in \{0.5, 1, 2, 4\}$ and three seeds. Each $\Delta\mathcal{L}$ is a paired Monte-Carlo estimate with the reduced predictor's exact paired difference as a control variate, resolving effects of 10^{-3} (Appendix K). For any predictor, trained or not, we define its coefficients (c_R, c_Q, c_M) as the least-squares fit through the origin of $\Delta\mathcal{L} = c_R\mathcal{R} - c_Q\mathcal{Q} + c_M\mathcal{M}$ over the bank and its exchange rate as $\beta := c_R/c_Q$; for the reduced predictor these are the quantities of Corollary 1, and the fit is exact. No intercept is fitted, and R^2 is reported as $1 - \|\text{residual}\|^2/\|\Delta\mathcal{L}\|^2$.

Why deployment matters. The fixed-model protocol isolates how structure changes the information in a prompt, rather than how a learner adapts after seeing that structure throughout training. This is the closer analogue of a pre-trained model used with retrieved, graph-neighbour, or otherwise correlated demonstrations at test time. All architectures see the same prompt bank, paired random draws, and statistics $(\mathcal{R}, \mathcal{Q}, \mathcal{M})$; only the learned inference rule changes. A match for full linear attention tests whether training recovers the analytic one-step learner, while softmax deviations measure how that architecture prices the same geometry.

At $\alpha = 0.5$ the measured $\Delta\mathcal{L}$ coincides with the prediction on all 60 prompts (correlation 1.000); at $\alpha = 1$ the correlation is 0.996 (Figure 4a). At every α the three-statistic linear fit through the origin has $R^2 \geq 0.999$, so $(\mathcal{R}, \mathcal{Q}, \mathcal{M})$ are sufficient for the trained model as for the reduced one; this is the isotropy argument at work, since the population optimum of the full parameterisation at independence

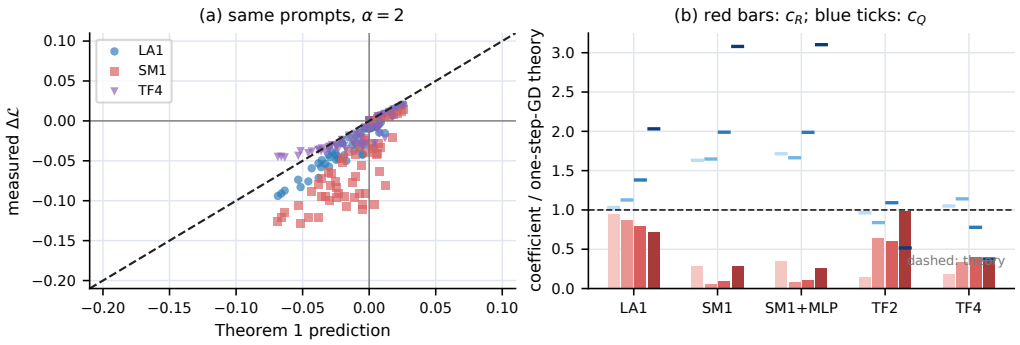


Figure 5: **Architectures on the same prompts.** (a) At $\alpha = 2$, trained architectures respond differently on the same structured prompts; softmax attention is often helped where the one-step law predicts harm. (b) Redundancy prices c_R (bars) and relevance gains c_Q (ticks), normalized by the one-step-GD coefficients.

is $[\gamma_0 I\ 0]$. At $\alpha = 2$ and 4 the points move above the diagonal: the trained model pays less for redundancy and gains more from relevance than γ_0 would. Corollary 1 says what this is. The model's independence risk exceeds the closed form (0.241 against 0.221 at $\alpha = 4$), which fixes an effective $\gamma < \gamma_0$ (0.65 against 0.79); the corollary at that γ reproduces all three fitted coefficients to within 5–10%, without fitting structured-prompt coefficients (Figure 4b). A model that has not reached the optimal step — the loss curvature $b_0 = 1 + (\rho + e_1)/\alpha$ flattens with α, so weight decay and finite training leave a larger shrinkage — is a reduced predictor with a smaller step, and the law prices it accordingly. The deployed rate of a linear-attention model can thus be read off its independence risk alone.

Softmax transfer. Under the same protocol we train one-layer softmax attention, softmax+MLP, and two- and four-layer transformers (Appendix K). These nonlinear runs are transfer evidence, not an exact law. The fitted exchange rate $\beta_{\mathcal{A}} = c_R/c_Q$ generally rises with α (Figure 7c), though c_R itself need not be monotone; much of the change comes from decreasing query-relevance gain c_Q. One-layer softmax behaves like the smoother, with rates 0.03–0.11 for $\alpha \leq 2$ and much smaller redundancy prices than one-step GD. Deeper transformers sit between but are less stable: at $\alpha = 2$ their median rates are locally close to the Bayes-optimal rate, while Appendix K reports seed ranges, near-zero relevance coefficients, and failed two-layer fits. We therefore use the nonlinear runs to show architecture-dependent pricing, not a systematic Bayes-optimal trajectory. On the same prompts, softmax attention is helped where the one-step law says a linear-attention model is harmed (Figure 5).

7 Discussion

In the model we can solve, structure helps when its relevance to the query outweighs its redundancy within the context. For one-step linear attention this finite-d excess-risk law is exact and includes the mediated term $\mathcal{M}$ and finite-dimensional coefficients. Its weak-structure reading reduces to the two-statistic criterion $(1 + \rho)\mathcal{Q}(q, C) - \alpha\mathcal{R}(C)$, with the weight set by the shot-count-to-dimension ratio rather than tuned. This gives a statistical explanation for why similarity retrieval can fail at larger budgets, why MMR and determinantal diversity appear as first-order approximations, and why "effective context length" is only the independent-query slice of a broader geometry. The coordinates $(\mathcal{R}, \mathcal{Q}, \mathcal{M})$ are structural, but their prices are algorithm-dependent: smoothers, Bayes-optimal prediction, one-step attention, trained linear attention, and softmax Transformers can place different boundaries on the same structured prompt.

Scope. The theory is exact for isotropic linear regression and linear attention, while softmax transfer is measured experimentally rather than derived. The second-order reading drops $\mathcal{M}$, which matters for denser structures, and the graph degree law is local in the kernel strength τ. Because the target in our model is global, graph homophily supplies redundancy rather than a separate smooth-signal channel; the harmful homophily regime in Corollary 4 should therefore not be read as a claim about domains where target labels are intrinsically graph-smooth. Adding tasks that vary over the same similarity structure is the natural extension. Practically, retrieval should balance relevance and diversity according to task dimension, shot count, and the model's inference algorithm.

AI USE STATEMENT

Generative AI tools were used to assist with manuscript organization, editing for clarity, LaTeX maintenance, bibliography checking, figure and experiment-code preparation, and feedback on the presentation of mathematical arguments and experimental protocols. The authors reviewed all AI-assisted text, code, claims, proofs, and experimental outputs, and take responsibility for the final content of the submission.

ETHICS STATEMENT

This work studies theoretical and synthetic models of in-context learning and uses no human subjects, personal data, privacy-sensitive datasets, or deployed decision systems. We do not anticipate direct ethical risks beyond the general need for accurate reporting and reproducible scientific claims.

REPRODUCIBILITY STATEMENT

The main assumptions, finite-dimensional risk law, graph coupling, and phase-transition derivations are stated in the main text, with complete proof details in the appendix. Experimental protocols, hyperparameters, and additional validation results are reported in Appendices K, M, and N. The supplementary material includes anonymized code for regenerating the synthetic experiments and figures.

REFERENCES

Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

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

Ali Al-Lawati, Jason Lucas, Zhiwei Zhang, Prasenjit Mitra, and Suhang Wang. Graph-based molecular in-context learning grounded on Morgan fingerprints. *arXiv preprint arXiv:2502.05414*, 2025.

Alexander Atanasov, Jacob A. Zavatone-Veth, and Cengiz Pehlevan. Risk and cross validation in ridge regression with correlated samples. In *Proceedings of the 42nd International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=GMwKpJ9TiR.

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*, volume 36, 2023. doi: 10.52202/075280-2498.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901, 2020.

Niv Buchbinder, Moran Feldman, Joseph (Seffi) Naor, and Roy Schwartz. A tight linear time (1/2)-approximation for unconstrained submodular maximization. *SIAM Journal on Computing*, 44(5): 1384–1402, 2015.

Jaime Carbonell and Jade Goldstein. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, 1998. doi: 10.1145/290941.291025.

Deqing Fu, Tian-Qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. In *Advances in Neural Information Processing Systems*, volume 37, 2024. arXiv:2310.17086.

Jiale Fu, Yaqing Wang, Simeng Han, Jiaming Fan, and Xu Yang. GraphIC: A graph-based in-context example retrieval model for multi-step reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, 2026.

Ryotaro Kawata, Yujin Song, Alberto Bietti, Naoki Nishikawa, Taiji Suzuki, Samuel Vaiter, and Denny Wu. From shortcut to induction head: How data diversity shapes algorithm selection in transformers. In *Advances in Neural Information Processing Systems*, volume 38, 2025. arXiv:2512.18634.

Alex Kulesza and Ben Taskar. Determinantal point processes for machine learning. *Foundations and Trends in Machine Learning*, 5(2–3):123–286, 2012.

Jon Lee, Vahab S. Mirrokni, Viswanath Nagarajan, and Maxim Sviridenko. Non-monotone submodular maximization under matroid and knapsack constraints. In *Proceedings of the 41st Annual ACM Symposium on Theory of Computing (STOC)*, pages 323–332, 2009.

Mary Letey, Yue M. Lu, Cengiz Pehlevan, and Jacob A. Zavatone-Veth. Sequential correlations change in-context learning: Effective context length and architectural mismatch. *arXiv preprint arXiv:2607.03660*, 2026.

Itay Levy, Ben Bogin, and Jonathan Berant. Diverse demonstrations improve in-context compositional generalization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, 2023. arXiv:2212.06800.

Yukang Lin, Bingchen Zhong, Shuxin Jiang, Joanna Siebert, and Qingcai Chen. Reasoning graph enhanced exemplars retrieval for in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*, pages 9737–9759, 2025. arXiv:2409.11147.

Hao Liu, Jiarui Feng, Lecheng Kong, Ningyue Liang, Dacheng Tao, Yixin Chen, and Muhan Zhang. One for all: Towards training one graph model for all classification tasks. In *International Conference on Learning Representations*, 2024.

Jiachang Liu, Dinghan Shen, Yizhe Zhang, Bill Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for GPT-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.deelio-1.10.

Yixin Liu, Shiyuan Li, Yu Zheng, Qingfeng Chen, Chengqi Zhang, and Shirui Pan. ARC: A generalist graph anomaly detector with in-context learning. In *Advances in Neural Information Processing Systems*, 2025.

Yue M. Lu, Mary Letey, Jacob A. Zavatone-Veth, Anindita Maiti, and Cengiz Pehlevan. Asymptotic theory of in-context learning by linear attention. *Proceedings of the National Academy of Sciences*, 122(28):e2502599122, 2025. doi: 10.1073/pnas.2502599122.

Arvind Mahankali, Tatsunori B. Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. In *International Conference on Learning Representations*, 2024. arXiv:2307.03576.

Haitao Mao, Zhikai Chen, Wenzhuo Tang, Jianan Zhao, Yao Ma, Tong Zhao, Neil Shah, Michael Galkin, and Jiliang Tang. Position: Graph foundation models are already here. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Behrad Moniri and Hamed Hassani. Asymptotics of linear regression with linearly dependent data. In *Proceedings of the 7th Annual Learning for Dynamics and Control Conference*, volume 283 of *Proceedings of Machine Learning Research*, pages 72–85. PMLR, 2025.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, pages 2655–2671. Association for Computational Linguistics, 2022.

Hongjin Su, Jungo Kasai, Chen Henry Wu, Weijia Shi, Tianlu Wang, Jiayi Xin, Rui Zhang, Mari Ostendorf, Luke Zettlemoyer, Noah A. Smith, and Tao Yu. Selective annotation makes language models better few-shot learners. In *International Conference on Learning Representations*, 2023. arXiv:2209.01975.

Max Vladymyrov, Johannes von Oswald, Mark Sandler, and Rong Ge. Linear transformers are versatile in-context learners. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1546.

Johannes Von Oswald, Eyvind Niklasson, Ettore Randazzo, João Sacramento, Alexander Mordvintsev, Andrey Zhmoginov, and Max Vladymyrov. Transformers learn in-context by gradient descent. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 35151–35174. PMLR, 2023.

Heng Wang, Shangbin Feng, Tianxing He, Zhaoxuan Tan, Xiaochuang Han, and Yulia Tsvetkov. Can language models solve graph problems in natural language? In *Advances in Neural Information Processing Systems*, volume 36, pages 30840–30861, 2023.

Jingfeng Wu, Difan Zou, Zixiang Chen, Vladimir Braverman, Quanquan Gu, and Peter Bartlett. How many pretraining tasks are needed for in-context learning of linear regression? In *International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=vSh5ePa0ph.

Zhao Yang, Yuanzhe Zhang, Dianbo Sui, Cao Liu, Jun Zhao, and Kang Liu. Representative demonstration selection for in-context learning with two-stage determinantal point process. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5443–5456, 2023.

Jiacheng Ye, Zhiyong Wu, Jiangtao Feng, Tao Yu, and Lingpeng Kong. Compositional exemplars for in-context learning. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 39818–39833. PMLR, 2023. URL https://proceedings.mlr.press/v202/ye23c.html.

Ruiqi Zhang, Spencer Frei, and Peter L. Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024. URL http://jmlr.org/papers/v25/23-1042.html.

A Demonstration Selection, MMR, and Determinantal Diversity

At fixed budget $|C| = \ell$, Theorem 1 says the deployed predictor's risk is smaller the larger

$$\mathcal{J}(q, C) := (1+\rho)\,\mathcal{Q}(q, C) - \alpha\,\mathcal{R}(C) = (1+\rho)\sum_{i\in C} S_{iq}^2 - \frac{2\alpha}{\ell}\sum_{\{i,j\}\subset C} S_{ij}^2, \tag{11}$$

up to the mediated term $\mathcal{M}$, which can be added on a dense context.

Proposition 3 (Structure of $\mathcal{J}$)**.** *For any unit-diagonal kernel and fixed budget ℓ, $\mathcal{J}$ is submodular in C, the sum of the modular $\mathcal{Q}$ and the negative of a supermodular function; it is not monotone.*

The objective is non-monotone, so deterministic forward greedy is used below as an empirical heuristic rather than as a worst-case approximation algorithm. The cited non-monotone submodular guarantees apply to randomized or local-search procedures (Lee et al., 2009; Buchbinder et al., 2015); our local-search oracle is included for that reason. The greedy marginal gain of adding e is $(1+\rho)S_{eq}^2 - \frac{2}{d}\sum_{i\in C} S_{ei}^2$: each pair of similar demonstrations costs $2/d$ regardless of the budget, and it is the number of pairs that grows with ℓ.

Applied selectors are first-order approximations. MMR scores a candidate by $\lambda\,\mathrm{sim}(e, q) - (1-\lambda)\max_{i\in C}\mathrm{sim}(e, i)$ (Carbonell and Goldstein, 1998); the marginal gain above has the same shape with squared similarity and a sum in place of the maximum. The global set-level exchange rate is $\alpha/(1+\rho)$, while the per-candidate marginal penalty after dividing by the query coefficient is $2/[d(1+\rho)]$. Determinantal selectors score a set by $\log\det$ of the kernel restricted to it (Kulesza and Taskar, 2012; Ye et al., 2023), and this is our redundancy statistic for any unit-diagonal kernel.

Proposition 4 (Determinantal identity)**.** *For any unit-diagonal $S = I + T$, $\log\det K = -\frac{\ell}{2}\,\mathcal{R}(C) + O(\|T_{CC}\|^3)$.*

Selection experiment. From a pool of 60 candidates we select ℓ demonstrations greedily by similarity only, diversity only, $\mathcal{J}$ with the model-specified weight $w = \alpha$, the exact objective including $\mathcal{M}$, tuned MMR, a tuned quality–diversity determinantal selector, and a local-search oracle on the exact risk. Evaluated under the fixed predictor γ_0, $\mathcal{J}$ improves on similarity retrieval by 0.6–1.0% of the risk for $\ell \geq 8$, winning on 93–100% of pools. In this binary graph-kernel pool the tuned determinantal selector chooses the same contexts as $\mathcal{J}$ at every budget, because linear and squared off-diagonal similarities are collinear; continuous similarity spectra need not have this equivalence. Full experimental details and proofs are in Appendix N.

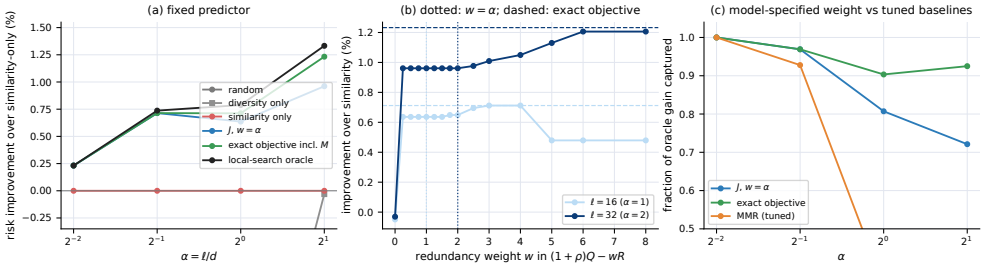


Figure 6: **Selection under the fixed predictor.** (a) Improvement over similarity-only selection. (b) Risk improvement over similarity-only against the redundancy weight at $\ell = 16$ and 32; dotted $w = \alpha$, dashed the exact objective including $\mathcal{M}$. (c) Fraction of the oracle's gain recovered by the model-specified weight, the exact objective, and a tuned MMR.

Prediction for selection outside the model. Since $\alpha = \ell/d$, the theory predicts that the optimal weight on diversity increases with the number of demonstrations relative to the effective task dimension. The MMR parameter is not linearly related to α, and a language model's in-context inference is not linear regression, so the transferable claim is directional: sweeping shot count on a fixed task should shift the best relevance–diversity balance toward diversity when the learner is closer to a one-step estimator, and less so when it is closer to a Bayes-optimal learner.

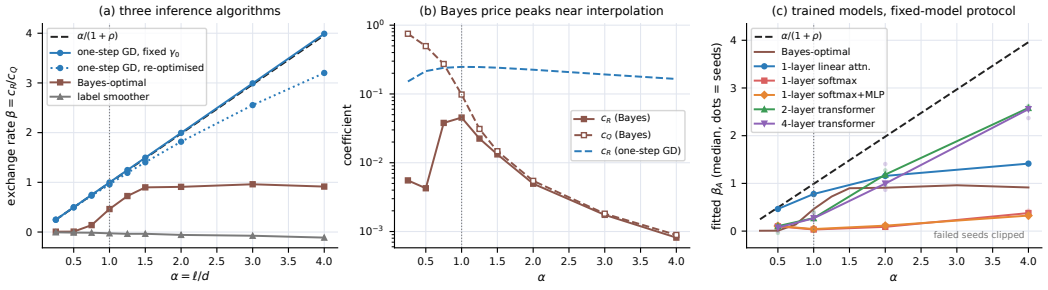


Figure 7: **The price of redundancy is set by the algorithm.** Rates are c_R/c_Q from the regression $\Delta\mathcal{L} = c_R\mathcal{R} - c_Q\mathcal{Q} + c_M\mathcal{M}$ over one 60-prompt bank ($d = 16$, $\rho = 0.01$). (a) Exact or Monte-Carlo rates for reduced linear attention, the label smoother, and the Bayes-optimal predictor. (b) Bayes-optimal redundancy and relevance coefficients, with the redundancy price peaking near interpolation. (c) Fitted rates of trained architectures under the fixed-model protocol.

B Algorithm-Dependent Exchange Rates

Other inference algorithms price redundancy differently. This appendix keeps the architectural spectrum separate from the main linear-attention law: the constant $\alpha/(1+\rho)$ is the price a one-step learner pays, not a property of the data alone.

Proposition 5 (Smoother rate)**.** *For $\widehat{y}_q = \omega^\top y_C$ with $\omega \in \mathbb{R}^\ell$ chosen for the prompt, $\mathcal{L}(\omega) = 1 + \rho - 2\omega^\top k_q + \omega^\top(K + \rho I)\omega$, the optimum is $\omega^\star = (K + \rho I)^{-1}k_q$, and*

$$\mathcal{L}^\star_{\text{smooth}}(S) = 1 + \rho - \frac{\mathcal{Q}}{1+\rho} + \frac{\mathcal{M}}{(1+\rho)^2} + O(\|T\|^4)\,: \tag{12}$$

$\mathcal{R}$ does not enter and the rate is zero. At independence $\mathcal{L}(\omega) = (1+\rho)(1+\|\omega\|^2) \geq (1+\rho)(1+1/\ell)$ on the simplex, so no convex combination of context labels beats predicting zero.

The smoother never estimates the task; it exploits the direct route from y_i to y_q, and redundancy costs it nothing because its weights whiten the context. The predictor that uses both routes optimally is the posterior mean $\widehat{y}_q = x_q^\top X_C^\top (X_C X_C^\top + \rho I)^{-1} y_C$, ridge regression on the context.

Proposition 6 (Bayes-optimal rate)**.** *Let $\kappa^\star := k_q^\top K^{-1} k_q$ and $\Sigma := I - X_C^\top (X_C X_C^\top + \rho I)^{-1} X_C$. For any unit-diagonal S with $K \succ 0$,*

$$\mathcal{L}_{\text{B}}(S) = \rho + (1-\kappa^\star)\,\tfrac{1}{d}\,\mathbb{E}\,\text{tr}\,\Sigma + \mathbb{E}\big[k_q^\top K^{-1} X_C\,\Sigma\,X_C^\top K^{-1} k_q\big]. \tag{13}$$

(i) If $\ell < d$ and $\rho = 0$, then exactly $\mathcal{L}_{\text{B}}(S) = (1-\alpha)(1-\kappa^\star)$, so $\mathcal{L}_{\text{B}}(S) - \mathcal{L}_{\text{B}}(I) = -(1-\alpha)(\mathcal{Q} - \mathcal{M}) + O(\|T\|^4)$: $\mathcal{R}$ does not enter at any order and the rate is zero; for $\rho > 0$ its contribution is $O(\rho)$. (ii) If $\ell > d + 1$, then $\mathcal{L}_{\text{B}}(S) = \rho + \rho\,\mathbb{E}[x_q^\top (X^\top X + \rho I)^{-1} x_q]$: both the redundancy penalty and the relevance gain are $O(\rho)$, and the rate is their $O(1)$ ratio, set by how K conditions $X^\top X$.

Figure 7 measures rates on one prompt bank for $\alpha \in [0.25, 4]$. The one-step learner follows $\alpha/(1+\rho)$, the smoother stays near zero, and the Bayes-optimal redundancy price peaks around the interpolation threshold before saturating in the noisy overdetermined regime. The trained architectures in Table 1 lie between these analytic learners: one-layer softmax behaves much like the smoother, while deeper transformers are locally closer to the Bayes-optimal rate near $\alpha = 2$. They do not reproduce the full non-monotone ridge curve, so this comparison is evidence for algorithm-dependent pricing rather than an architectural identification.

C Failure Modes of the Two-Statistic Description

The mediated direction. In Equations (7) and (8) the coefficients of κ_{q} and κ_{m} have opposite signs, so their difference $\kappa_{\text{m}} - \kappa_{\text{q}} = k_q^\top (K - I) k_q = \mathcal{M}(q, C)$ is an independent direction. It measures query relevance *mediated* through the context: similarity between q and a demonstration i that is itself similar to the rest of the context. The reduction of Theorem 1 to $(\mathcal{R}, \mathcal{Q})$ holds exactly when $\mathcal{M} = 0$, and to leading order whenever $\mathcal{M}$ is negligible against $\mathcal{R}, \mathcal{Q}$, which is what a sparse similarity structure or a weak-diffusion expansion delivers.

Curvature of the population-optimal boundary. For the re-optimised predictor the term $D_\alpha \kappa_q^2/\alpha^2$ in Equation (7) is positive and second order in the summary κ_q (fourth order in the raw perturbation T), so as similarity strengthens the boundary bends towards tolerating more redundancy: query relevance becomes superlinearly more valuable. The overshoot of the empirically optimal weight $w^\star$ above α at larger budgets (Appendix A) is in this direction.

On graphs the binding condition is local density. The dominant error at moderate τ is not the Taylor expansion of the risk but the kernel expansion $S_G(\tau) = I + \tau A + O(\tau^2)$, which discards two-hop contributions $\frac{\tau^2}{2}A^2$. Numerically, the relative error of κ_q against $\tau^2 d_C(q)$ is negligible on a graph of mean degree 0.91 but reaches 48% at $\tau = 0.2$ on a graph of mean degree 7.39. The correct condition is therefore $\tau \cdot (\text{second-order neighbourhood size}) \ll 1$. Replacing $d_C(q)$ by $\|a_{Cq} + \frac{\tau}{2}(A^2)_{Cq}\|_2^2$ restores agreement at larger τ, at the cost of the elementary degree interpretation; the similarity form $\mathcal{Q} = \|T_{Cq}\|^2$ of Theorem 1 needs no such repair, since it is exact.

D Kernel Expansions

Lemma (normalized heat kernel). *As $\tau \to 0$, $S_G(\tau) = I + \tau A + O(\tau^2)$.* Since $L = D - A$ and $A_{ii} = 0$, $e^{-\tau L} = I - \tau D + \tau A + O(\tau^2)$ and $D_\tau = \mathrm{diag}(e^{-\tau L}) = I - \tau D + O(\tau^2)$. Using $(1-z)^{-1/2} = 1 + \frac{z}{2} + O(z^2)$ entrywise, $D_\tau^{-1/2} = I + \frac{\tau}{2}D + O(\tau^2)$, whence

$$S_G(\tau) = \Big(I + \tfrac{\tau}{2}D\Big)(I - \tau D + \tau A)\Big(I + \tfrac{\tau}{2}D\Big) + O(\tau^2) = I + \tau A + O(\tau^2), \tag{14}$$

the two half-degree terms cancelling $-\tau D$ exactly. □

Degree form of the statistics. *Equation (5) holds.* Restricting Equation (14) to blocks gives $T_{CC} = \tau A_{CC} + \tau^2 B + O_G(\tau^3)$ and $T_{Cq} = \tau a_{Cq} + O(\tau^2)$, where B has zero diagonal because S_G has unit diagonal for every τ. Then $\|T_{CC}\|_F^2 = \tau^2 \|A_{CC}\|_F^2 + O_G(\tau^3)$ and $\|T_{Cq}\|^2 = \tau^2 \|a_{Cq}\|^2 + O_G(\tau^3)$; for an unweighted graph $\|A_{CC}\|_F^2/\ell = 2|E(C)|/\ell = \bar{d}_{\mathrm{in}}(C)$ and $\|a_{Cq}\|^2 = d_C(q)$. The mediated term $\mathcal{M} = T_{Cq}^\top T_{CC} T_{Cq}$ is $O_G(\tau^3)$. □

Other graph kernels. The heat kernel is not essential. For the regularized Laplacian $(I + \tau L)^{-1} = I - \tau L + O(\tau^2)$, diagonal normalization again cancels the degree term at first order and gives $I + \tau A + O(\tau^2)$, so Equation (5) holds verbatim. For kernels built on the normalized adjacency, such as personalised PageRank, the first-order term is $D^{-1/2} A D^{-1/2}$ and the statistics become those of the weighted graph with weights $A_{ij}/\sqrt{d_i d_j}$. Higher-order behaviour depends on the kernel.

Weighted graphs. For a symmetric nonnegative A with zero diagonal the expansion is unchanged and $\mathcal{R}(C) = \frac{\tau^2}{\ell}\sum_{i,j\in C} A_{ij}^2$, $\mathcal{Q}(q, C) = \tau^2 \sum_{i\in C} A_{iq}^2$: edge counts become squared edge strengths. This is a special case of Lemma 1.

E Proof of Proposition 1

Isotropy. For every $O \in O(d)$ the map $x_v \mapsto O x_v$, $w \mapsto O w$ preserves the joint law, because $\mathbb{E}[x_u x_v^\top] = \frac{1}{d} S_{uv} I_d$ is rotation covariant and $w \sim \mathcal{N}(0, I_d)$. The risk is a convex quadratic in Γ and is invariant under $\Gamma \mapsto O \Gamma \widetilde{O}^\top$, where $\widetilde{O} = \mathrm{diag}(O, 1)$; its unique minimiser is therefore self-invariant, giving $\Gamma^\star = [\gamma^\star I_d \; 0]$. No expansion is used.

Reduction to a scalar quadratic. With $\widehat{y}_q = \gamma \frac{d}{\ell} x_q^\top X_C^\top y_C$ and $y_C = X_C w + \epsilon_C$, independence of w, ϵ from the features and $\mathbb{E}[w w^\top] = I_d$ give $\mathcal{L}(\gamma) = 1 + \rho - 2a\gamma + b\gamma^2$ with

$$a = \frac{d}{\ell}\sum_i \mathbb{E}\big[(x_i^\top x_q)^2\big], \qquad b = \Big(\frac{d}{\ell}\Big)^2 \sum_{i,j} \mathbb{E}\big[(x_i^\top x_q)(x_j^\top x_q)(x_i^\top x_j)\big] + \rho\Big(\frac{d}{\ell}\Big)^2 \sum_i \mathbb{E}\big[(x_i^\top x_q)^2\big].$$

Hence $\gamma^\star = a/b$ and $\mathcal{L}^\star = 1 + \rho - a^2/b$.

The two moments. Writing $k_i := S_{iq}$, Isserlis gives

$$\mathbb{E}\big[(x_i^\top x_q)^2\big] = k_i^2\Big(1+\tfrac{1}{d}\Big) + \tfrac{1}{d}, \tag{15}$$

and, collecting all fifteen pairings of the sixth-order moment,

$$\mathbb{E}\big[(x_i^\top x_j)(x_i^\top x_q)(x_j^\top x_q)\big] = e_3\, S_{ij}k_ik_j + \Big(\tfrac{1}{d}+\tfrac{1}{d^2}\Big)\big(S_{ij}^2 + k_i^2 + k_j^2\big) + \tfrac{1}{d^2}, \tag{16}$$

with $e_3 = 1 + \frac{3}{d} + \frac{4}{d^2}$. (Check: at $j = i$, $k_i = 0$, Equation (16) gives $\frac{1}{d} + \frac{2}{d^2}$, matching $\mathbb{E}\|x_i\|^4/d = (1+\frac{2}{d})/d$.)

Closure on $(\kappa_c, \kappa_q, \kappa_m)$. Summing Equation (16) over $i, j \in C$ and using $\sum_{i,j} S_{ij}k_ik_j = \kappa_m$, $\sum_{i,j} S_{ij}^2 = \ell\kappa_c$, $\sum_{i,j}(k_i^2 + k_j^2) = 2\ell\kappa_q$, together with $\sum_i k_i^2 = \kappa_q$ in Equation (15), gives Equation (6) directly. No fourth invariant appears, so the three summaries are exactly sufficient for this predictor. Letting $d \to \infty$ sets $e_1, e_3 \to 1$ and recovers the weak-correlation population expressions of Letey et al. (2026). □

Numerical check. Against 2×10^5 Monte-Carlo prompts at $d = 32$, $\ell = 64$, the exact formula attains an absolute error of 1.9×10^{-4} ($\tau = 0.15$) and 4.2×10^{-5} ($\tau = 0.3$, dense graph), against 7.1×10^{-3} and 1.15×10^{-2} for the $d \to \infty$ form, i.e. a 20–100× reduction, at the level of the Monte-Carlo noise.

F Proof of Theorem 1 and Corollary 2

The fixed predictor. At independence $\kappa_c = 1$, $\kappa_q = \kappa_m = 0$, so $a_0 = 1$ and $b_0 = 1 + (\rho + e_1)/\alpha = \widetilde{D}/\alpha$ with $\widetilde{D} := \alpha + \rho + e_1$, and $\gamma_0 = a_0/b_0 = \alpha/\widetilde{D}$. For a fixed γ, $\mathcal{L}(\gamma; S) = 1 + \rho - 2a\gamma + b\gamma^2$ is affine in (a, b), hence affine in $(\kappa_c, \kappa_q, \kappa_m)$ by Equation (6). Writing $\delta a = a - a_0 = e_1\kappa_q/\alpha$ and $\delta b = b - b_0 = e_1(\kappa_c - 1 + 2\kappa_q)/\alpha + (e_3\kappa_m + \rho e_1\kappa_q)/\alpha^2$,

$$\begin{aligned}\mathcal{L}(\gamma_0; S) - \mathcal{L}(\gamma_0; I) &= -2\gamma_0\,\delta a + \gamma_0^2\,\delta b \\ &= \frac{1}{\widetilde{D}^2}\Big[\alpha e_1(\kappa_c - 1) + \big(-2e_1\widetilde{D} + 2\alpha e_1 + \rho e_1 + e_3\big)\kappa_q \\ &\qquad + e_3(\kappa_m - \kappa_q)\Big],\end{aligned}$$

after multiplying through by $\alpha^2/\widetilde{D}^2$ and collecting terms. Since $-2e_1\widetilde{D} + 2\alpha e_1 + \rho e_1 = -\rho e_1 - 2e_1^2$ and $2e_1^2 - e_3 = e_1 - 2/d^2$, the coefficient of κ_q is $-\big((1+\rho)e_1 - 2/d^2\big)$. Substituting $\kappa_c - 1 = \mathcal{R}$, $\kappa_q = \mathcal{Q}$ and $\kappa_m - \kappa_q = \mathcal{M}$ from Lemma 1,

$$\mathcal{L}(S) - \mathcal{L}_{\text{iid}} = \frac{\alpha e_1\,\mathcal{R}(C) - \big((1+\rho)e_1 - \frac{2}{d^2}\big)\mathcal{Q}(q, C) + e_3\,\mathcal{M}(q, C)}{(\alpha + \rho + e_1)^2}, \tag{17}$$

with no remainder, which is Equation (8); letting $d \to \infty$ gives its limiting form.

The re-optimised predictor. $\mathcal{L}(\gamma; S)$ is a quadratic in γ with minimiser $\gamma^\star = a/b$ and curvature b, so $\mathcal{L}(\gamma_0; S) = \mathcal{L}^\star(S) + b(\gamma_0 - \gamma^\star)^2$, and subtracting the same identity at $S = I$ (where $\gamma_0 = \gamma^\star$) gives $\mathcal{L}^\star(S) - \mathcal{L}^\star_{\text{iid}} = \mathcal{L}(S) - \mathcal{L}_{\text{iid}} - b(\gamma^\star - \gamma_0)^2$. Now $\gamma^\star - \gamma_0 = (b_0\delta a - a_0\delta b)/(bb_0)$ and $\delta a, \delta b$ are linear in $\mathcal{R}, \mathcal{Q}, \mathcal{M}$, each of which is $O(\|T\|^2)$ in operator norm ($\|T_{CC}\|_F^2 \le \ell\|T\|^2$, $\|T_{Cq}\|^2 \le \|T\|^2$, $|\mathcal{M}| \le \|T\|^3$), so $\gamma^\star - \gamma_0 = O(\|T\|^2)$ and the discrepancy is $O(\|T\|^4)$. □

Local derivatives. Differentiating $\mathcal{L}^\star = 1 + \rho - a^2/b$ at the independent point in the $d \to \infty$ form, with $D_\alpha = \alpha + 1 + \rho$,

$$\partial_{\kappa_c}\mathcal{L}^\star\big|_{\text{iid}} = \frac{\alpha}{D_\alpha^2}, \quad \partial_{\kappa_q}\mathcal{L}^\star\big|_{\text{iid}} = -\frac{2+\rho}{D_\alpha^2}, \quad \partial_{\kappa_m}\mathcal{L}^\star\big|_{\text{iid}} = \frac{1}{D_\alpha^2}, \tag{18}$$

which are the coefficients of Equation (8) with $\kappa_m = \kappa_q + \mathcal{M}$; this is the envelope theorem in coordinates. Along a similarity-induced direction with $\kappa_m = \kappa_q$ to leading order the combined derivative is $-(1+\rho)/D_\alpha^2 < 0$.

Corollary 2. Substitute Equation (5) into Equation (8): $\mathcal{R} = \tau^2 \bar{d}_{\text{in}} + O_G(\tau^3)$, $\mathcal{Q} = \tau^2 d_C(q) + O_G(\tau^3)$, and $\mathcal{M} = O_G(\tau^3)$. The finite-$d$ coefficient of the τ^2 term is therefore

$$\frac{\alpha e_1 \bar{d}_{\text{in}}(C) - \big((1+\rho)e_1 - \frac{2}{d^2}\big) d_C(q)}{(\alpha + \rho + e_1)^2}.$$

If its numerator is nonzero it is the leading term, so its sign is the sign of the difference for all sufficiently small $\tau > 0$. The displayed large-d boundary follows by letting $d \to \infty$. □

G The Exponential-Kernel Ray

For $K_{ab} = e^{-|a-b|/\xi}$ and $(k_{\text{test}})_a = e^{-(\ell+1-a)/\xi}$, write $u = e^{-1/\xi}$. Then

$$\kappa_{\text{q}} = \sum_{a=1}^{\ell} u^{2(\ell+1-a)} = \frac{u^2(1-u^{2\ell})}{1-u^2}, \qquad \kappa_{\text{c}} - 1 = \frac{1}{\ell}\sum_{a \neq b} u^{2|a-b|} = \frac{2}{\ell}\sum_{m=1}^{\ell-1} (\ell - m) u^{2m}.$$

For small u the leading terms are $\kappa_{\text{q}} = u^2 + O(u^4)$ and $\kappa_{\text{c}} - 1 = 2(1 - \frac{1}{\ell})u^2 + O(u^4)$, and $\kappa_{\text{m}} = \kappa_{\text{q}} + O(u^4)$ because the off-diagonal contributions to $k^\top K k$ are of order $u^2 \cdot u^2$. Hence $(\kappa_{\text{c}} - 1, \kappa_{\text{q}}, \kappa_{\text{m}}) = \kappa_{\text{q}}\,(2(1 - \frac{1}{\ell}), 1, 1) + O(u^4)$. Numerically the ratios remain close to $(1.984, 1)$ beyond the small-u regime: at $\ell = 128$ they are $(1.9844, 1.0000)$ at $\xi = 0.1$ and $(1.9753, 2.164)$ by $\xi = 2$, the drift in $\kappa_{\text{m}}/\kappa_{\text{q}}$ reflecting the $O(u^4)$ term. Substituting the ray into Equation (8) gives $\Delta\mathcal{L} \propto \kappa_{\text{q}}[2\alpha(1 - \frac{1}{\ell}) - (1 + \rho)]$, so the sign is fixed by α alone at this order. The statement is a leading-order one: at large ξ, κ_{q} is no longer small, the $D_\alpha \kappa_{\text{q}}^2/\alpha^2$ term of Equation (7) enters, and the ray itself drifts.

H Proof of Proposition 5

For $\widehat{y}_q = \omega^\top y_C$, $\mathbb{E}[y_q^2] = 1 + \rho$, $\mathbb{E}[y_i y_q] = S_{iq} = k_{q,i}$ and $\mathbb{E}[y_i y_j] = S_{ij} + \rho\delta_{ij}$, so $\mathcal{L}(\omega) = 1 + \rho - 2\omega^\top k_q + \omega^\top (K + \rho I)\omega$. The minimiser is $\omega^\star = (K + \rho I)^{-1} k_q$ with $\mathcal{L}^\star_{\text{smooth}} = 1 + \rho - k_q^\top (K + \rho I)^{-1} k_q$. Writing $K + \rho I = (1 + \rho)I + T_{CC}$,

$$(K + \rho I)^{-1} = \frac{1}{1+\rho}\sum_{n \geq 0}\Big(-\frac{T_{CC}}{1+\rho}\Big)^n,$$

$$k_q^\top (K + \rho I)^{-1} k_q = \frac{\mathcal{Q}}{1+\rho} - \frac{\mathcal{M}}{(1+\rho)^2} + \frac{T_{Cq}^\top T_{CC}^2 T_{Cq}}{(1+\rho)^3} - \cdots,$$

and the third term is $O(\|T\|^4)$. This is Equation (12); $\mathcal{R}$ does not appear because the smoother's weights invert the context kernel. At independence, $\mathcal{L}(\omega) = 1 + \rho + (1 + \rho)\|\omega\|^2$, and $\|\omega\|^2 \geq 1/\ell$ on the simplex gives the bound. □

I Proof of Proposition 6

Posterior and risk. For compactness in this proof only, write $X = X_C$ and $y = y_C$ for the context design and labels. Conditional on X and y, $w \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = X^\top (XX^\top + \rho I)^{-1} y$ and $\Sigma = (I + X^\top X/\rho)^{-1} = I - X^\top (XX^\top + \rho I)^{-1} X$ by Woodbury. The Bayes predictor is $\widehat{y}_q = x_q^\top \mu$, and since $y_q = x_q^\top w + \epsilon_q$ with ϵ_q independent, $\mathbb{E}[(y_q - \widehat{y}_q)^2 \mid X, x_q] = \rho + x_q^\top \Sigma x_q$.

Conditional law of the query feature. For each coordinate j, $(X_{\cdot j}, x_{qj}) \sim \mathcal{N}(0, S/d)$ independently across j. Hence $x_q \mid X \sim \mathcal{N}\big(X^\top m, \frac{1-\kappa^\star}{d} I_d\big)$ with $m = K^{-1} k_q$ and $\kappa^\star = k_q^\top K^{-1} k_q$. Taking the conditional expectation of $x_q^\top \Sigma x_q$,

$$\mathbb{E}[x_q^\top \Sigma x_q \mid X] = m^\top X \Sigma X^\top m + \frac{1 - \kappa^\star}{d} \operatorname{tr} \Sigma,$$

which is Equation (13) after averaging over X.

(i) Below the threshold. If $\ell < d$ and $\rho = 0$, X has full row rank almost surely and $\Sigma = P_\perp := I - X^\top(XX^\top)^{-1}X$ is the orthogonal projection onto the complement of the row space of X. Then $XP_\perp = 0$, so the first term vanishes, and $\operatorname{tr} P_\perp = d - \ell$ for every X, so $\mathcal{L}_{\mathrm{B}}(S) = (1-\kappa^\star)(d-\ell)/d = (1-\alpha)(1-\kappa^\star)$ with no expectation left to take. At $S = I$, $\kappa^\star = 0$. Expanding $K^{-1} = (I + T_{CC})^{-1}$ gives $\kappa^\star = \mathcal{Q} - \mathcal{M} + T_{Cq}^\top T_{CC}^2 T_{Cq} - \cdots$, whose third term is $O(\|T\|^4)$. For $\rho \geq 0$, $\Sigma = P_\perp + \rho X^\top(XX^\top)^{-1}(XX^\top + \rho I)^{-1}X$, so $\operatorname{tr}\Sigma = d - \ell + \rho \operatorname{tr}(XX^\top + \rho I)^{-1}$ and $m^\top X\Sigma X^\top m = \rho\, m^\top(XX^\top + \rho I)^{-1}XX^\top m$; the only dependence on the conditioning of $XX^\top$, and hence on $\mathcal{R}$, sits in these $O(\rho)$ terms.

(ii) Above the threshold. If $\ell > d$, $X^\top X$ is invertible almost surely and $\Sigma = \rho(X^\top X + \rho I)^{-1}$, so $\mathcal{L}_{\mathrm{B}} = \rho + \rho\,\mathbb{E}[x_q^\top (X^\top X + \rho I)^{-1} x_q]$. The expectation is finite as $\rho \to 0$ when $\ell > d + 1$ (inverse Wishart), so every structural effect is $O(\rho)$. Substituting the conditional law of x_q, $\mathbb{E}[x_q^\top (X^\top X)^{-1} x_q] = \mathbb{E}[m^\top P_X m] + \frac{1-\kappa^\star}{d}\mathbb{E}\operatorname{tr}(X^\top X)^{-1}$ with $P_X = X(X^\top X)^{-1}X^\top$ the rank-d projection onto the column space of X. The second term carries the redundancy penalty: $X^\top X = \frac{1}{d} Z^\top K Z$, the first-order change of $\mathbb{E}\operatorname{tr}(X^\top X)^{-1}$ in T_{CC} vanishes because $\operatorname{tr} T_{CC} = 0$, and the second-order change is positive and proportional to $\|T_{CC}\|_F^2 = \ell\mathcal{R}$. The first term and the factor $(1-\kappa^\star)$ carry the relevance gain, of order $\mathcal{Q}$. Both are multiplied by ρ, which gives the claim; the $O(1)$ ratio is the exchange rate, whose value in the proportional limit requires the deterministic equivalents for correlated-sample ridge regression of Atanasov et al. (2025) and is left to numerics here. □

J Additional Figures for the Population Law

These diagnostics isolate two finite-sample consequences of the degree-level boundary. Figure 8 checks the sign transition while varying the problem scale and noise; Figure 9 checks the weak-diffusion expansion and where its local approximation begins to degrade.

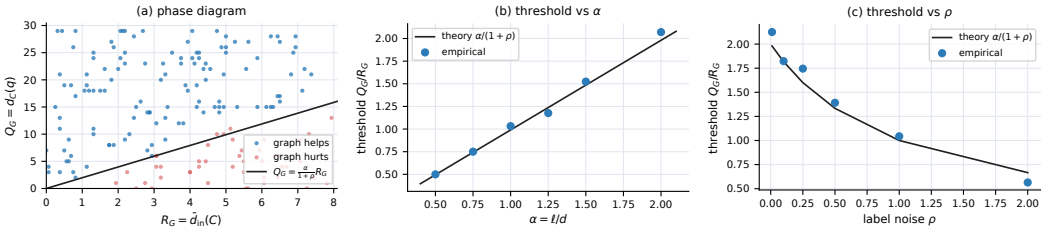


Figure 8: **A model-specified phase boundary in degrees.** (a) Across 170 graphs at $\tau = 0.06$ the sign of $\Delta\mathcal{L}$ agrees with the predicted boundary on every instance. (b,c) The empirical zero-crossing of $d_C(q)/\bar{d}_{\mathrm{in}}(C)$ tracks $\alpha/(1+\rho)$ as each parameter varies; varying both is what pins the functional form, since a fit along α alone is consistent with α/c for any c.

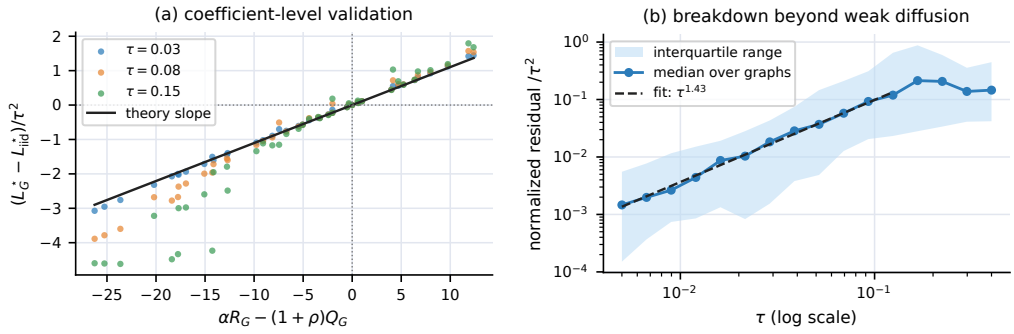


Figure 9: **Quantitative validation of the degree law.** (a) Weak-diffusion graph instances approach the model-specified relation of Corollary 2. (b) The normalized residual grows with τ as a power close to the predicted one, consistent with the higher-order remainder neglected by the degree expansion.

K TRAINED-MODEL EXPERIMENTS

Table 1: **Trained models under the fixed-model protocol.** Each model is trained once on independent prompts ($d = 16$) and evaluated on the 60-prompt bank at $S = I + 0.15A$. Entries are medians over three seeds with the seed range in brackets for the rate; c_R, c_Q are the fitted redundancy price and relevance gain, R^2 that of the three-statistic linear fit through the origin, "corr" the correlation of the measured $\Delta\mathcal{L}$ with the prediction of Theorem 1. Theory rows give the one-step-GD values.

α	architecture	$\mathcal{L}_{\text{iid}}$	$\beta_{\mathcal{A}}$	c_R	c_Q	R^2	corr
0.5	1-layer linear attention	0.701	0.46 [0.45, 0.47]	0.205	0.442	1.000	1.000
	1-layer softmax	0.863	0.10 [0.05, 0.11]	0.071	0.712	0.986	0.982
	1-layer softmax + MLP	0.861	0.11 [0.08, 0.11]	0.077	0.725	0.987	0.984
	2-layer transformer	0.779	0.10 [−0.04, 0.10]	0.050	0.480	0.934	0.971
	4-layer transformer	0.788	0.08 [0.00, 0.15]	0.022	0.415	0.925	0.966
	theory, one-step GD	0.692	0.50	0.215	0.431		
1	1-layer linear attention	0.528	0.78 [0.75, 0.79]	0.217	0.279	1.000	0.996
	1-layer softmax	0.789	0.03 [0.03, 0.04]	0.013	0.410	0.975	0.941
	1-layer softmax + MLP	0.793	0.04 [0.03, 0.08]	0.018	0.414	0.978	0.944
	2-layer transformer	0.597	0.27 [−4.25, 0.40]	0.205	0.166	0.733	0.871
	4-layer transformer	0.486	0.27 [0.24, 0.36]	0.075	0.279	0.927	0.946
	theory, one-step GD	0.526	0.99	0.247	0.248		
2	1-layer linear attention	0.371	1.16 [1.14, 1.17]	0.181	0.156	1.000	0.947
	1-layer softmax	0.705	0.09 [0.07, 0.13]	0.020	0.224	0.980	0.777
	1-layer softmax + MLP	0.702	0.11 [0.08, 0.14]	0.026	0.225	0.979	0.778
	2-layer transformer	0.376	1.18 [0.95, 1.27]	0.137	0.144	0.900	0.840
	4-layer transformer	0.198	1.00 [0.86, 1.41]	0.092	0.092	0.898	0.874
	theory, one-step GD	0.359	1.98	0.225	0.113		
4	1-layer linear attention	0.241	1.41 [1.41, 1.42]	0.119	0.084	0.999	0.752
	1-layer softmax	0.619	0.38 [0.31, 0.43]	0.049	0.127	0.986	0.421
	1-layer softmax + MLP	0.610	0.33 [0.28, 0.38]	0.043	0.129	0.982	0.440
	2-layer transformer	0.296	2.58 [−22.7, 2.62]	0.123	0.029	0.884	0.737
	4-layer transformer	0.082	2.56 [2.37, 8.13]	0.061	0.011	0.963	0.763
	theory, one-step GD	0.221	3.96	0.165	0.041		

Uncertainty of the fitted rates. We also bootstrap prompts within each seed when fitting the exchange rate $\beta_{\mathcal{A}} = c_R/c_Q$. Full linear attention is tight across the whole range: median seed-level 95% intervals are $[0.46, 0.47]$, $[0.77, 0.79]$, $[1.15, 1.16]$, and $[1.39, 1.43]$ for $\alpha = 0.5, 1, 2, 4$. Nonlinear models have wider intervals because their redundancy coefficients are smaller and the linear-fit residuals are larger. For one-layer softmax, representative intervals are $[-0.16, 0.17]$ at $\alpha = 1$, $[-0.09, 0.23]$ at $\alpha = 2$, and $[0.26, 0.47]$ at $\alpha = 4$. Deep transformers are wider still at small α, and two seeds of the two-layer model give negative or near-zero relevance coefficients. This is why the main text treats softmax transfer as qualitative and architecture-dependent rather than as an exact extension of Theorem 1.

Within-bank held-out fits. As a quick guard against purely in-sample linear fits, we repeatedly fit c_R, c_Q, c_M on 40 randomly chosen prompts and evaluate on the remaining 20, using 100 splits per seed. Full linear attention remains essentially exact, with median held-out $R^2 \geq 0.999$ at every α. One-layer softmax and softmax+MLP also remain stable on this bank, with median held-out R^2 between 0.971 and 0.984 and median sign accuracy 1.0. Deep transformers are less stable: median held-out R^2 ranges from 0.728 to 0.918 for the two-layer model and 0.888 to 0.954 for the four-layer model, with the two-layer sign accuracy dropping to 0.90 at $\alpha = 4$. This check does not replace a separately generated prompt-bank validation, but it shows that the reported three-statistic structure is not only the result of fitting all 60 prompts at once.

Architectures. One-layer linear attention: $\widehat{y}_q = [\ell^{-1} \sum_i (z_q^\top W_{QK} z_i)\, W_{VO} z_i]_{d+1}$ with W_{QK} initialised at the identity and W_{VO} near zero. One-layer softmax attention: the same with weights $\mathrm{softmax}_i(z_q^\top W_{QK} z_i)$, W_{QK} initialised at $d\,I$, a linear readout initialised to read the label coordinate, and optionally a GELU MLP of width 64 added to the attention output. Two- and four-layer transformers: pre-LayerNorm blocks of width 32 with two heads and a width-64 MLP, full attention over the $\ell + 1$ tokens with the query label zeroed, and a linear head on the query token. Tokens are $(x_i, y_i) \in \mathbb{R}^{d+1}$.

Training and evaluation. Training data are independent prompts ($S = I$) with $w \sim \mathcal{N}(0, I_d)$, $\rho = 0.01$; AdamW with learning rate 10^{-3}, 10% linear warm-up, weight decay 10^{-4}, gradient

Table 2: **Context selection at fixed budget under the fixed predictor.** Exact risk of the selected context for $\widehat{y}_q = \gamma_0(d/\ell)x_q^\top X_C^\top y_C$ ($d = 16$, $S = I + 0.1A$, $\rho = 0.01$, pool $N = 60$; means over 30 pools; lower is better). MMR and the quality–diversity DPP report their best tuned weight. "oracle" is local search on the exact risk. Summary rows report the paired improvement of $\mathcal{J}$ ($w = \alpha$) over similarity-only, its win fraction, and the fraction of oracle gain captured.

ℓ	α	random	similarity	diversity	MMR*	DPP*	$\mathcal{J}$, $w{=}\alpha$	exact ($+\mathcal{M}$)	oracle
4	0.25	0.8142	0.7985	0.8135	0.7966	0.7966	0.7966	0.7966	0.7966
8	0.5	0.6839	0.6627	0.6795	0.6582	0.6580	0.6580	0.6580	0.6578
16	1	0.5216	0.5016	0.5164	0.5002	0.4984	0.4984	0.4980	0.4976
32	2	0.3576	0.3523	0.3524	0.3512	0.3490	0.3490	0.3480	0.3476
paired gain of $\mathcal{J}$ over similarity (win)		$+0.0019$ (63%), $+0.0047$ (100%), $+0.0032$ (93%), $+0.0034$ (100%)							
fraction of oracle gain: $\mathcal{J}$ / exact / MMR*		1.00/1.00/1.00, 0.97/0.97/0.93, 0.81/0.90/0.35, 0.72/0.93/0.24							

clipping at 1, batch 64, 3000 steps. Evaluation draws, for each prompt kernel S in the bank, 32,768 samples of (Z, w, ϵ) and evaluates the model on both $X = d^{-1/2}S^{1/2}Z$ and $X = d^{-1/2}Z$ with the same draws; the paired difference of squared errors is averaged after subtracting a control variate, the reduced predictor's paired difference on the same samples, whose mean is known exactly from Theorem 1. The resulting standard errors are 10^{-4} for linear attention and 1–4 $\times 10^{-3}$ for the other architectures, against median $|\Delta\mathcal{L}|$ of 2–7 $\times 10^{-2}$. The bank is the one of Figure 7 (context edge probabilities 0–0.15, query edge probabilities 0–0.8, three graphs per cell, $S = I + 0.15A$); $\mathcal{R}$, $\mathcal{Q}$, $\mathcal{M}$ are computed exactly from S. Rates are least-squares fits of $\Delta\mathcal{L} = c_R\mathcal{R} - c_Q\mathcal{Q} + c_M\mathcal{M}$ over the 60 prompts.

The effective step of the trained linear-attention model. For each seed and α we solve $\mathcal{L}_{\text{iid}}^{\text{trained}} = 1 + \rho - 2\gamma + b_0\gamma^2$ for the smaller root γ and evaluate Corollary 1 there. At $\alpha = 1, 2, 4$ the predicted (c_R, c_Q, c_M) are $(0.226, 0.270, 0.256)$, $(0.164, 0.168, 0.093)$, $(0.114, 0.088, 0.032)$ against measured $(0.220, 0.277, 0.248)$, $(0.182, 0.155, 0.103)$, $(0.119, 0.084, 0.034)$ for seed 0, with the other seeds within the same margins; the effective steps are 0.46, 0.56, 0.65 against $\gamma_0 = 0.48$, 0.65, 0.79. At $\alpha = 0.5$, the inverted scalar diagnostic is nearly γ_0, so the γ_0 prediction is the better summary. We do not attribute the remaining small independence excess to a particular off-diagonal parameter block without direct parameter-level measurements.

Reading the table. Two seeds of the two-layer model ($\alpha = 1$, seed 1; $\alpha = 4$, seed 0) fit a negative or near-zero relevance gain and give uninterpretable rates; they are optimization failures rather than structure, and medians are used throughout. The correlation with the one-step prediction falls at $\alpha = 4$ for every architecture, as expected once an architecture's own boundary differs from the one-step one; the linear-fit R^2 is the informative measure of whether the three statistics suffice, and it stays above 0.98 for all one-layer models.

Reproducing the figures. `scripts/make_additional_figures.py` builds Figures 1, 2 and 3. `exp3_trained.py` trains and evaluates one architecture at one (ℓ, seed); `scripts/make_trained_figures_v2.py` builds Figures 4 and 5 from the fixed-model trained runs. `scripts/make_appendix_figures.py` builds Figures 7, 6, 8, 9, and 10 from the saved JSON outputs, and `FIGURE_SOURCES.md` records which files come from the `Graph_ICL` experiment directory.

L Heavy-Tailed Hosts and Scale

The graphs in the main text are Erdős–Rényi or SBM, both with Poisson degrees. Real networks have hubs, and since Section 4.2 shows that the binding condition involves the second-order neighbourhood, heavy tails are the natural place to look for a breakdown. We repeat the phase-boundary test on Barabási–Albert hosts under two sampling regimes: a uniformly random induced subgraph, and the two-hop ball around the query, which is what retrieval-style applications do and which produces much denser induced subgraphs.

Table 3 shows that the concern is mostly controlled in this regime. Degree coefficients of variation of 0.86–0.91 and maximum degrees up to 108 against a mean of 15.8 leave sign accuracy at 97–100%; BA $m{=}8$ is the most accurate configuration under neighbourhood sampling. The single weak cell is the sparse Poisson host under neighbourhood sampling at $\tau = 0.2$ (91.2%), where the two-hop ball

Table 3: **Sign accuracy of the predicted boundary on Poisson- and power-law-degree hosts.** 400 prompts per cell, $n = 400$ nodes, $d = 16$, $\ell = 32$, $\rho = 0.01$. Accuracy is the fraction of prompts on which $\mathrm{sign}(\Delta\mathcal{L})$ agrees with $\mathrm{sign}(\alpha\bar{d}_{\mathrm{in}} - (1+\rho)d_C(q))$.

			random subgraph			two-hop ball		
host	deg. CV	max deg	τ=0.03	τ=0.1	τ=0.2	τ=0.03	τ=0.1	τ=0.2
ER p=0.02	0.36	19	100.0	99.8	99.5	98.9	96.2	91.2
ER p=0.05	0.23	33	99.2	99.5	98.2	99.5	99.2	98.0
BA m=4	0.91	57	100.0	99.2	99.5	98.7	98.2	97.4
BA m=8	0.86	108	99.5	99.2	98.0	99.0	99.2	99.8

Table 4: **Context selection at larger scale** (re-optimised analytical predictor, heat kernel at $\tau = 0.1$, $\rho = 0.01$, 40 pools per row, candidate pool $N = \max(60, 2\ell)$). "gain" is the paired improvement of $\mathcal{J}$ over similarity-only selection and "win" the fraction of pools on which it is positive.

d	ℓ	α	similarity	$\mathcal{J}$	gain	relative	win
16	8	0.5	0.6539	0.6498	+0.0040	0.62%	98%
32	16	0.5	0.6197	0.6171	+0.0025	0.41%	88%
32	32	1.0	0.4778	0.4729	+0.0049	1.03%	100%
64	32	0.5	0.6072	0.6017	+0.0055	0.90%	100%
64	64	1.0	0.4185	0.4022	+0.0163	3.90%	100%

is barely larger than the budget, so the context is close to the entire ball, and the statistics are small in absolute terms ($\bar{d}_{\mathrm{in}} \approx 1.47$) so the $O_G(\tau^3)$ remainder is relatively larger. Degradation tracks τ and the absolute size of the statistics, not the degree tail.

For selection the tail helps rather than hurts. At matched mean degree (7.9) the power-law host yields roughly three times the improvement of $\mathcal{J}$ over similarity-only selection (+0.0019 at 60% win rate versus +0.0007 at 45%), and the densest power-law host gives +0.0027 at 85%. Hubs create exactly the concentrated redundancy that $\bar{d}_{\mathrm{in}}$ penalises and pure similarity ignores.

Scale. Repeating the selection experiment at larger dimension and budget shows the effect strengthening in the direction the theory predicts: since $\alpha = \ell/d$ weights redundancy, the improvement should grow with α at fixed d. It does, by $1.96\times$ at $d = 32$ and $2.98\times$ at $d = 64$ when α moves from 0.5 to 1 (Table 4).

M Real Graph Topologies

Here we repeat the two central claims on four real network topologies — two homophilic citation graphs (Cora, CiteSeer) and two heterophilic ones (Actor, Squirrel) — with the protocol of Appendix L. This is a topology stress test rather than a real-feature ICL benchmark: node features are drawn from the heat kernel on the real adjacency, so the Gaussian and unit-diagonal assumptions hold exactly and all losses come from Proposition 1.

Table 5: **Phase-boundary sign accuracy on real graphs** (%). Fraction of prompts on which $\mathrm{sign}(\Delta\mathcal{L})$ agrees with $\mathrm{sign}(\alpha\bar{d}_{\mathrm{in}} - (1+\rho)d_C(q))$; $d = 16$, $\ell = 32$, $\rho = 0.01$, up to 400 prompts per cell.

				random subgraph			two-hop ball		
host	mean deg	CV	max deg	0.03	0.1	0.2	0.03	0.1	0.2
CiteSeer (hom.)	2.7	1.24	99	100	100	100	100	97.2	97.3
Cora (hom.)	3.9	1.34	168	100	100	100	100	96.8	95.0
Actor (het.)	7.0	2.53	1303	100	100	100	98.4	95.9	94.4
Squirrel (het.)	76.3	2.12	1903	99.5	96.5	95.8	91.4	90.9	78.6

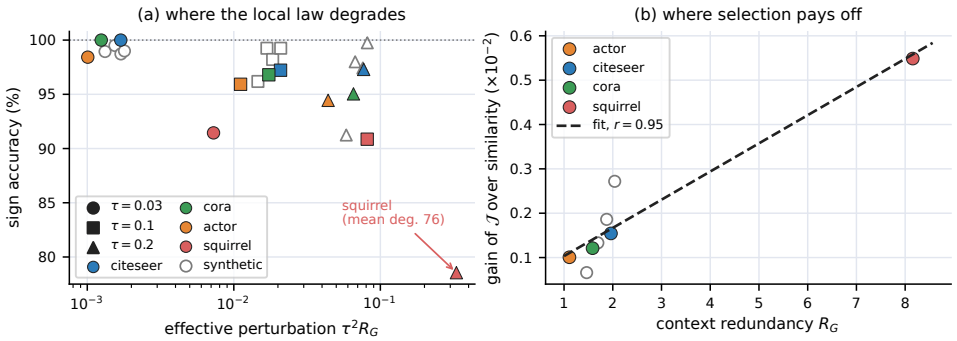


Figure 10: **Real graph topologies.** Filled markers are real graphs, open grey markers the synthetic hosts of Appendix L. (a) Sign accuracy against the effective perturbation $\tau^2\mathcal{R}$, under two-hop-ball sampling. Degradation tracks this quantity, not the degree tail: Actor has maximum degree 1303 against a mean of 7 and stays above 94%, while Squirrel, whose mean degree is 76, leaves the perturbative regime entirely. (b) The improvement of $\mathcal{J}$ over similarity-only selection is close to linear in $\mathcal{R}$ across real and synthetic hosts alike.

The boundary holds, and fails where predicted. Under random subgraph sampling the predicted sign is correct on every prompt for three of the four graphs, including Actor, whose maximum degree of 1303 against a mean of 7.0 is far more extreme than anything in Appendix L. Squirrel is the exception and is instructive: it is by far the densest host (198 353 edges on 5 201 nodes, mean degree 76.3), so its two-hop ball is essentially the whole graph and the induced context is dense, $\bar{d}_{\text{in}} \approx 8.1$ against 1.1–2.1 elsewhere. At $\tau = 0.2$ this gives $\tau^2\mathcal{R} \approx 0.33$, whereas the other hosts sit at 0.04–0.08, and accuracy falls to 78.6%. This is the regime Section 4.2 identifies. Across all 25 real and synthetic configurations, $\text{corr}(\log \tau^2\mathcal{R},\ \text{accuracy}) = -0.58$ (Figure 10a).

Selection gains scale with redundancy. Across the four real graphs and the four synthetic hosts, the improvement of $\mathcal{J}$ over similarity-only selection is nearly linear in $\mathcal{R}$, with correlation 0.95 (Figure 10b), markedly tighter than against mean degree (0.80) or the degree coefficient of variation (0.38). Squirrel, the most redundant host, gives the largest improvement ($+0.0055$, 0.81% of the similarity-only risk, positive on 72% of pools). One thing the data do not establish: Squirrel is both heterophilic and by far the densest host, so its gain cannot be attributed to heterophily — Actor, also heterophilic but sparse, gives the smallest gain of the four. Density and assortativity are confounded in these datasets.

N Proofs and Details for Selection

N.1 Proof of Proposition 4

For any unit-diagonal $S = I + T$, $K = I + T_{CC}$ with $\operatorname{tr} T_{CC} = 0$. Hence

$$\begin{aligned}\log\det K &= \operatorname{tr}\log(I + T_{CC}) \\ &= \operatorname{tr} T_{CC} - \frac{1}{2}\operatorname{tr}(T_{CC}^2) + O(\|T_{CC}\|^3) \\ &= -\frac{1}{2}\|T_{CC}\|_F^2 + O(\|T_{CC}\|^3) \\ &= -\frac{\ell}{2}\mathcal{R}(C) + O(\|T_{CC}\|^3).\end{aligned}$$

On a graph $T_{CC} = \tau A_{CC} + O(\tau^2)$ with a zero-diagonal correction, and $\operatorname{tr} A_{CC} = 0$, giving the degree form. Numerically, on Erdős–Rényi graphs the ratio of $\log\det K_G(\tau)$ to $-\tau^2\ell\bar{d}_{\text{in}}/2$ is 1.0015 at $\tau = 0.02$ and 1.0131 at $\tau = 0.2$ for $\ell = 32$, $p = 0.08$; at $p = 0.15$ the ratios are 1.0050 and 1.0654, consistent with the local-density dependence of the remainder. □

N.2 Proof of Proposition 3

Fix the budget ℓ and write $\mathcal{J}(C) = (1+\rho)f(C) - \frac{2\alpha}{\ell}g(C)$ with $f(C) = \sum_{i\in C} S_{iq}^2$ and $g(C) = \sum_{\{i,j\}\subset C} S_{ij}^2$. f is modular: $f(C\cup\{e\}) - f(C) = S_{eq}^2$ independently of C. g is supermodular: for $A \subseteq B$ and $e \notin B$, $g(A\cup\{e\}) - g(A) = \sum_{i\in A} S_{ie}^2 \le \sum_{i\in B} S_{ie}^2 = g(B\cup\{e\}) - g(B)$. Hence $-g$ is submodular and $\mathcal{J}$, a nonnegative combination of a modular and a submodular function with ℓ held fixed, is submodular. It is not monotone: adding e changes $\mathcal{J}$ by $(1+\rho)S_{eq}^2 - \frac{2\alpha}{\ell}\sum_{i\in C} S_{ie}^2$, which is negative whenever e is dissimilar to q but similar to enough of C; on random graphs with $n = 12$, $p = 0.35$, $\ell = 5$, $\alpha = 2$, a random single-element addition decreases $\mathcal{J}$ in $119/200$ instances. The normalisation matters: with $\bar{d}_{\text{in}}(C) = 2|E(C)|/|C|$ and $|C|$ varying, over 12,000 random triples $(A \subseteq B, e)$ we found 5,500 violations of diminishing returns, against 0 for the fixed-budget form. Since selection is posed at a fixed shot budget, we use the latter. □

N.3 Experimental details

Pool construction. For each replicate we draw $N = 60$ candidate nodes with context–context edges from an Erdős–Rényi model at $p = 0.12$, and attach the query to each candidate independently with probability 0.30. The kernel is $S = I + 0.1A$ (shrunk if needed to keep $\lambda_{\min}(S) \ge 0.2$), so $\mathcal{R}, \mathcal{Q}, \mathcal{M}$ are exact, and all risks are computed with Proposition 1 at $d = 16$, $\rho = 0.01$ for the fixed predictor γ_0; no weak-similarity approximation enters the evaluation. Table 4 uses the earlier heat-kernel pools with $N = \max(60, 2\ell)$ candidates and the re-optimised analytical predictor; it is a scale stress test of Proposition 1, not a trained checkpoint or a replacement for the fixed-deployment benchmark in Table 2.

Selection rules. All greedy rules add one element at a time until the required budget is filled, maximising the marginal gain of, respectively, $\sum_i S_{iq}^2$ (similarity), $-\sum_{\{i,j\}} S_{ij}^2$ (diversity), $\mathcal{J}$ of Equation (11), and the exact objective $-[\alpha e_1\mathcal{R} - ((1+\rho)e_1 - 2/d^2)\mathcal{Q} + e_3\mathcal{M}]$ of Theorem 1 evaluated on the partial context with the budget ℓ in the normalisation. If all remaining marginal gains are negative, the rule chooses the least negative candidate so that all methods are compared at the same shot budget. The random baseline draws ℓ candidates uniformly without replacement. Table 2 reports means over 30 independently drawn pools under the fixed predictor γ_0; the paired gain compares $\mathcal{J}$ against similarity-only on the same pool.

Weight sweep. To test the predicted dependence on α we sweep w in the two-statistic objective $(1+\rho)\mathcal{Q} - w\,\mathcal{R}$ over $w \in \{0, 0.25, \dots, 8\}$, selecting greedily and evaluating the exact fixed-predictor risk on the same 30 pools (Figure 6b). The risk is flat over $w \in [0.25, 2]$ at every budget: greedy selection changes its choices only at a few thresholds in w. The best two-statistic weight is 0.25 at $\ell = 4$ and 8, 3 at $\ell = 16$ and 6 at $\ell = 32$, but its advantage over $w = \alpha$ is 0.08% and 0.25% of the risk at $\ell = 16$ and 32, and the exact objective including $\mathcal{M}$ at $w = \alpha$ reaches the same level (0.4980 against 0.4980; 0.3480 against 0.3481). The overshoot of a tuned two-statistic weight above α is therefore the mediated term $\mathcal{M}$, which grows with the query degree, and not a discrepancy in the exchange rate. Under the re-optimised predictor $\gamma^\star$ the same sweep gives 0.25, 0.25, 2.5, 5.

Baselines. MMR scores a candidate by $\lambda S_{eq} - (1-\lambda)\max_{i\in C} S_{ei}$ over $\lambda \in \{0.2, 0.35, 0.5, 0.65, 0.8, 0.95\}$. The quality–diversity determinantal selector scores by $\theta S_{eq} + \log\det S_{C\cup\{e\}}$ over $\theta \in \{0, 2, 5, 10, 20, 50\}$. Both report the best weight per budget. The oracle runs single-swap local search on the exact fixed-predictor risk from the exact-objective greedy solution until no swap improves. For every $\theta \ge 2$ the determinantal selector returns the same context as $\mathcal{J}$, at every budget and pool in this unweighted graph-kernel experiment. This equality is not asserted for continuous similarities: here $S_{ij} \in \{0, \tau\}$ off the diagonal, so S_{ij} and S_{ij}^2 induce identical rankings, while continuous kernels can decouple the DPP score from the quadratic relevance term in $\mathcal{J}$.