Bayes-Optimal Coexistence via Fact Localizability in Trainable-Feature Decoder-Only Transformers

Manoj Saravanan [1]

Abstract

We study when exact memorization of sparse facts can coexist with Bayes-optimal rule generalization in a trainable-feature decoder-only transformer. For a causal rules-and-facts model, we define the Bayes-coexistence gap $\Delta_{F_m}(\mathcal{T})$, the excess population rule risk forced by exact interpolation of a finite fact set, and the fact-localizability functional $\Lambda_{F_m}(\mathcal{T}; P_{\mathrm{rule}})$, the smallest $L^2(P_X)$-mass of a Bayes-centered fact correction. Under squared loss we prove the exact identity $\Delta_{F_m} = \Lambda_{F_m}$. We then show that a minimal trainable-feature causal decoder realizes a rule–residual factorization $A_{\bar{\Theta}}(X) = (S^\star(X), Z^\perp(X), 0)$, and that sparse ReLU tent functions in the independent residual block exactly memorize arbitrary bounded facts with exponentially small excess rule risk. A lazy/tangent version of the same decoder obeys an effective-dimension lower bound, separating feature-learning coexistence from fixed-representation interpolation. The construction also gives exact structural deletability: residual fact units can be zeroed while preserving the Bayes rule.

1. Introduction

Foundation models must often learn reusable rules while also storing rare, idiosyncratic facts. This coexistence is useful when the facts are legitimate knowledge, but it becomes a trustworthiness problem when the facts are private, copyrighted, or otherwise sensitive. A useful architecture should therefore separate the reusable rule from the removable factual residue: exact memorization should not automatically imply a population penalty on the underlying rule, and deleting a fact should not destroy the rule computation.

Existing theory captures important parts of this tension. Rules-and-Facts models formalize simultaneous generalization and memorization (Farné et al., 2026); attention-indexed models give tractable high-dimensional theories of attention learning (Boncoraglio et al., 2025); and recent transformer analyses study factual acquisition, extraction, and storage (Xu & Chen, 2025; Nichani et al., 2024; Dong et al., 2025). What remains less explicit is an architectural separation statement: can a trainable-feature causal decoder exactly interpolate arbitrary sparse facts while remaining Bayes-optimal on the rule distribution, and is this ability absent from the corresponding lazy/tangent class unless the tangent kernel has enough effective dimension?

This paper gives a theorem-level answer in a controlled decoder-only model. The key invariant is *fact localizability*: the minimum population mass of a perturbation that implants the desired facts on top of the Bayes rule. We prove that this invariant is exactly equal to the excess rule risk forced by memorization. We then construct a trainable decoder whose attention representation splits into a rule block and an independent high-dimensional residual block. Sparse residual-only MLP units memorize the facts with exponentially small $L^2(P_X)$-mass, whereas a bounded lazy/tangent model must pay an effective-dimension cost. The construction is also structurally deletable: fact units can be removed by zeroing their output coefficients.

2. Setup and Objectives

Let $X = (x_0, \ldots, x_{L-1}) \in \mathcal{X}_d$ denote a causal token sequence. The rule distribution is generated by a low-dimensional teacher statistic

$$S^\star(X) \in \mathbb{R}^r, \tag{1}$$

and a squared-loss Bayes rule

$$f^\star(X) = \mathbb{E}[Y \mid X] = \varphi^\star(S^\star(X)). \tag{2}$$

A finite fact set is $F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$. Its residual vector is

$$\delta(F_m) = (\delta_1, \ldots, \delta_m), \qquad \delta_\mu := \xi^\mu - f^\star(X^\mu). \tag{3}$$

[1]Department of Electrical and Computer Engineering, Virginia Polytechnic Institute and State University, Blacksburg, VA, USA. Correspondence to: Manoj Saravanan <manoj663@vt.edu>.

The Second Workshop on the Impact of Memorization on Trustworthy Foundation Models at ICML 2026, Seoul, South Korea. Non-archival camera-ready version, 2026. Copyright 2026 by the author(s).

For a predictor f, define

$$R_{\text{rule}}(f) = \mathbb{E}_{(X,Y)\sim P_{\text{rule}}}\big[(f(X)-Y)^2\big], \quad R_{\text{Bayes}} = R_{\text{rule}}(f^\star). \tag{4}$$

Definition 2.1 (Bayes-coexistence gap)**.** For $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$,

$$\Delta_{F_m}(\mathcal{T}) := \inf_{f\in\mathcal{T}\,:\,f(X^\mu)=\xi^\mu\ \forall\mu} \left(R_{\text{rule}}(f) - R_{\text{Bayes}}\right), \tag{5}$$

with value $+\infty$ if the interpolation constraints are infeasible.

Definition 2.2 (Fact localizability)**.** For $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$,

$$\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}) := \inf_{\substack{g\in\mathcal{M}_2(P_X)\,:\,f^\star+g\in\mathcal{T}\\ g(X^\mu)=\delta_\mu\ \forall\mu}} \|g\|^2_{L^2(P_X)}. \tag{6}$$

The trainable-feature decoder class used below has attention representation

$$A_\Theta(X) = (A_1(X), \dots, A_H(X)) \tag{7}$$

followed by a bounded-width ReLU MLP output

$$f_\Theta(X) = \sum_{j=1}^{W_f} c_j\, \sigma(\langle b_j, A_\Theta(X)\rangle - \tau_j) + c_0. \tag{8}$$

The appendices give the full Gaussian teacher assumptions, causal-head construction, norm budgets, and high-probability events.

3. Main Results

Theorem 3.1 (Coexistence–localizability identity)**.** *For any predictor class $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ and finite fact set F_m, squared loss gives*

$$\Delta_{F_m}(\mathcal{T}) = \Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}). \tag{9}$$

Equivalently,

$$\Delta_{F_m}(\mathcal{T}) = \inf_{\substack{g:\ f^\star+g\in\mathcal{T}\\ g(X^\mu)=\xi^\mu - f^\star(X^\mu)\ \forall\mu}} \|g\|^2_{L^2(P_X)}. \tag{10}$$

The proof is the Bayes Pythagorean identity $R_{\text{rule}}(f) - R_{\text{Bayes}} = \|f - f^\star\|^2_{L^2(P_X)}$, followed by translating the interpolation constraints from f to $g = f - f^\star$; see Section C. Thus the price of exact fact memorization is not an informal generalization penalty but the minimum rule-distribution visibility of the smallest fact correction.

Theorem 3.2 (Bayes-optimal coexistence for trainable-feature decoders)**.** *Assume the standing Gaussian teacher assumptions of Section 2 and Sections B and D. Fix residual dimension $d_\perp \le d - p_\star$. If $H \ge r + d_\perp$, $W_f \ge 3m$, and F_m is a bounded generic fact set, then for every budget $B \ge B_{\text{upper}}(d_\perp, \|\delta(F_m)\|_\infty)$, with probability at least $1 - 2me^{-d_\perp/128} - 2m(m-1)e^{-d_\perp/128}$, writing $\mathcal{D} = \mathcal{T}_{\text{dec}}(H, W_f, B)$,*

$$\Delta_{F_m}(\mathcal{D}) \le \|\delta(F_m)\|^2_\infty \Big[me^{-25d_\perp/128} + m(m-1)e^{-75d_\perp/224}\Big]. \tag{11}$$

Equivalently, there exists $f_{\Theta_F} \in \mathcal{D}$ interpolating all facts and satisfying the same bound on $R_{\text{rule}}(f_{\Theta_F}) - R_{\text{Bayes}}$.

The construction learns a split representation $A_{\bar\Theta}(X) = (S^\star(X), Z^\perp(X), 0)$, where the first block carries the Bayes rule and $Z^\perp(X) \sim \mathcal{N}(0, \gamma_\psi^2 I_{d_\perp})$ is independent of the teacher sigma-field. Three-ReLU tent bumps centered at high-threshold residual projections interpolate the residuals δ_μ while having exponentially small mass. Combining these bumps with Theorem 3.1 proves the bound; see Sections D to F.

Corollary 3.3 (Vanishing coexistence gap)**.** *Under Theorem 3.2, if $\log m = o(d_\perp)$, $\|\delta(F_m)\|_\infty = O(1)$, and $me^{-25d_\perp/128} \to 0$, then $\Delta_{F_m}(\mathcal{T}_{\text{dec}}(H, W_f, B)) \to 0$ with high probability.*

Theorem 3.4 (Lazy/kernel coexistence lower bound)**.** *Let $K = K_{\Theta_0}$ be the tangent kernel, $\mathcal{H}_K$ its RKHS, and $\mathcal{N}_K(\zeta) = \operatorname{tr}(T_K(T_K + \zeta I)^{-1})$. Under the trace-class assumptions of Section G, suppose $f_{\Theta_0} - f^\star \in \mathcal{H}_K$ with norm R_0. For the affine lazy class $\mathcal{T}_{\text{lazy}}(R; \Theta_0)$ and signed facts $\xi^\mu = f^\star(X^\mu) + \tau\omega_\mu$, with ω_μ i.i.d. Rademacher, for any $\zeta > 0$ and $\eta \in (0,1)$, with probability at least $1-\eta$,*

$$\Delta_{F_m}(\mathcal{T}_{\text{lazy}}(R;\Theta_0)) \ge \left(\eta^2 \frac{\tau^2 m}{\mathcal{N}_K(\zeta)} - \zeta(R+R_0)^2\right)_+. \tag{12}$$

Corollary 3.5 (Effective-dimension bottleneck)**.** *If $\mathcal{N}_K(\zeta_d) \le C_N \eta_d^2 m$, $\zeta_d(R+R_0)^2 \le \tau^2/(2C_N)$, and $\eta_d \to 0$, then*

$$\Delta_{F_m}(\mathcal{T}_{\text{lazy}}(R;\Theta_0)) \ge \frac{\tau^2}{2C_N}$$

with probability at least $1 - \eta_d$.

The converse is oracle-strengthened: the appendix first grants the kernel class the Bayes rule and still proves that bounded RKHS corrections cannot interpolate independent signed facts with small $L^2(P_X)$-mass unless $\mathcal{N}_K(\zeta)$ is large. The affine lazy statement then follows by containment; see Section G. This separates the feature-learning mechanism from generic kernel interpolation.

Corollary 3.6 (Structural localization and deletability)**.** *Under Theorem 3.2, the coexistence predictor decomposes as*

$$f_{\Theta_F}(X) = f^\star(X) + g_F(Z^\perp(X)),$$

where $f^\star$ is implemented by rule coordinates and g_F by at most $3m$ residual-only ReLU units. For any subset $U \subseteq [m]$, zeroing the three output coefficients attached to each $\mu \in U$ removes those residual facts, preserves all nondeleted facts, and maps deleted fact locations back to $f^\star(X^\mu)$. Zeroing all fact units recovers the Bayes rule exactly.

4. Proof Roadmap and Implications

The proof stack has four ingredients. First, Bayes orthogonality converts rule excess risk into squared $L^2(P_X)$-mass, yielding Theorem 3.1. Second, causal attention heads are chosen to expose $S^\star(X)$ and an independent Gaussian residual block. Third, high-dimensional concentration makes the residual codes of the fact locations separable, so sparse ReLU tent functions interpolate bounded residuals with exponentially small population mass. Fourth, a localized Rademacher anti-interpolation inequality lower-bounds any bounded RKHS correction by the inverse effective dimension, giving the lazy converse. The trustworthiness implication is conceptual rather than a deployed unlearning algorithm: in this model, useful rule computation and removable factual residue are represented in separate blocks, which provides a precise target for future auditing and mitigation work.

References

Boncoraglio, F., Troiani, E., Erba, V., and Zdeborová, L. Bayes optimal learning of attention-indexed models, 2025.

Bourtoule, L., Chandrasekaran, V., Choquette-Choo, C. A., Jia, H., Travers, A., Zhang, B., Lie, D., and Papernot, N. Machine unlearning. In *2021 IEEE Symposium on Security and Privacy*, pp. 141–159. IEEE, 2021. doi: 10.1109/SP40001.2021.00019.

Cao, Y. and Yang, J. Towards making systems forget with machine unlearning. In *2015 IEEE Symposium on Security and Privacy*, pp. 463–480. IEEE, 2015. doi: 10.1109/SP.2015.35.

Chizat, L., Oyallon, E., and Bach, F. On lazy training in differentiable programming. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Dong, Y., Noci, L., Khodak, M., and Li, M. Attention retrieves, mlp memorizes: Disentangling trainable components in the transformer, 2025.

Farné, G., Boncoraglio, F., and Zdeborová, L. The rules-and-facts model for simultaneous generalization and memorization in neural networks, 2026.

Huang, Y., Zhu, H., Guo, T., Jiao, J., Sojoudi, S., Jordan, M. I., Russell, S., and Mei, S. Generalization or hallucination? understanding out-of-context reasoning in transformers, 2025.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Nichani, E., Lee, J. D., and Bietti, A. Understanding factual recall in transformers via associative memories, 2024.

Vasudeva, B., Deora, P., Bietti, A., Sharan, V., and Thrampoulidis, C. Understanding contextual recall in transformers: How finetuning enables in-context reasoning over pretraining knowledge, 2026.

Xu, R. and Chen, K. Filtering with self-attention and storing with MLP: One-layer transformers can provably acquire and extract knowledge, 2025.

A. Exact Novelty Boundary and Relation to Prior Theory

This appendix states precisely what is, and is not, proved by the present paper. The purpose is to make the novelty boundary checkable. The paper is closest to recent theoretical work on simultaneous memorization and generalization, attention-indexed teacher–student models, factual recall in transformers, component specialization, lazy/NTK theory, and machine unlearning. Our contribution is not a new empirical memorization metric and not a claim about arbitrary trained large language models. It is a theorem-level separation: a trainable-feature causal decoder class admits Bayes-optimal coexistence of a structured rule and arbitrary sparse facts through a localized residual subspace, whereas a bounded lazy/kernel version of the same architecture has a nonvanishing coexistence gap unless its effective dimension is sufficiently large.

A.1. What This Paper Proves

For clarity, we summarize the theorem package as a sequence of logically independent advances.

1. **A variational invariant for coexistence.** We introduce the Bayes-coexistence gap
$$\Delta_{F_m}(\mathcal{T})$$
and the fact-localizability functional
$$\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}).$$
Theorem C.5 proves the exact identity
$$\Delta_{F_m}(\mathcal{T}) = \Lambda_{F_m}(\mathcal{T}; P_{\text{rule}})$$
under squared loss. Thus the population price of exact fact memorization is exactly the minimum $L^2(P_X)$-mass of a Bayes-centered fact-correcting perturbation.

2. **A trainable-feature rule–residual factorization.** Theorem D.9 constructs a causal decoder representation
$$A_{\bar{\Theta}}(X) = \big(S^\star(X), Z^\perp(X), 0\big),$$
where $S^\star(X)$ is sufficient for the Bayes rule and $Z^\perp(X)$ is a Gaussian residual block independent of the teacher sigma-field.

3. **Sparse fact localization in the residual block.** Theorem E.9 constructs explicit residual-only ReLU tent functions that exactly interpolate arbitrary bounded fact residuals and have exponentially small population mass:
$$\|g_F \circ Z^\perp\|^2_{L^2(P_X)} \lesssim \|\delta(F_m)\|^2_\infty \left[me^{-25d_\perp/128} + m(m-1)e^{-75d_\perp/224}\right].$$

4. **Bayes-optimal coexistence for trainable-feature decoders.** Theorem F.4 combines the previous ingredients to show that trainable-feature decoders can exactly memorize all facts while paying vanishing excess rule risk whenever the residual dimension dominates the number of facts in the stated exponential regime.

5. **A lazy/kernel converse.** Theorems G.9 and G.12 prove that a bounded lazy/tangent class cannot generally realize the same localizability phenomenon unless its kernel effective dimension $\mathcal{N}_K(\zeta)$ is large relative to m.

6. **Structural deletability.** Theorems H.6 and H.7 show that in the coexistence construction the fact residuals are represented by sparse residual-only MLP units. Zeroing the corresponding output coefficients deletes chosen fact residuals exactly while preserving the Bayes rule block.

A.2. Relation to Rules-and-Facts Theory

The closest conceptual predecessor is the Rules-and-Facts (RAF) model of Farné et al. (2026). RAF studies a mixture in which a structured teacher rule generates most labels while a subset of labels are unstructured facts. It gives a solvable framework for simultaneous generalization and memorization and analyzes capacity allocation between rule recovery and factual storage.

Our paper differs from RAF in four precise ways.

First, RAF is primarily a rules-and-facts statistical-physics model for classical learner families such as perceptrons, random features, and kernels. Here the learner is a causal decoder class with trainable attention features and an MLP correction layer.

Second, RAF characterizes when rule learning and fact memorization can coexist, whereas our central object is the exact variational identity

$$\Delta_{F_m} = \Lambda_{F_m}.$$

This identity reduces coexistence to a geometric localization problem inside an architecture class.

Third, the main upper theorem here is not only a capacity statement. It proves a constructive representation theorem:

$$f_{\Theta_F} = f^\star + g_F \circ Z^\perp,$$

where $f^\star$ lives in a low-dimensional rule block and g_F lives in a high-dimensional residual block.

Fourth, our lazy/kernel lower bound is architecture-relative: it compares the trainable-feature decoder to the tangent/lazy version of the same parameterization. This isolates the role of feature learning in creating the localized residual geometry.

Thus the present paper should be read as a decoder-theoretic extension and separation theorem for the rules-and-facts question, not as another analysis of the RAF model itself.

A.3. Relation to Bayes-Optimal Attention Theory

Boncoraglio et al. (2025) introduce attention-indexed models and derive Bayes-optimal predictions and phase transitions for learning in deep attention layers. Their framework is highly relevant because it shows that tractable attention models can support sharp high-dimensional theory. Our teacher rule is deliberately attention-indexed for this reason.

The difference is that AIM studies structured attention learning, whereas our problem contains an additional adversarial finite fact set

$$F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$$

whose values are arbitrary and need not follow the teacher. The central question is therefore not only whether the attention model learns the Bayes rule, but whether it can simultaneously implant arbitrary sparse exceptions with vanishing population cost. This requires the fact-localizability identity, the residual Gaussian block, and the sparse ReLU localization construction. None of these objects appears in AIM.

A.4. Relation to Factual Recall and Knowledge Extraction in Transformers

Recent work has made substantial progress on theorem-level models of factual knowledge in transformers. Nichani et al. (2024) analyze factual recall through associative memories and prove near-optimal storage behavior in shallow transformer abstractions. Xu & Chen (2025) study a one-layer self-attention-plus-MLP model and prove conditions under which the model can acquire and extract knowledge. Vasudeva et al. (2026) analyze how fine-tuning can enable in-context access to pretrained knowledge.

These works address storage, recall, or extraction of factual knowledge. Our theorem package addresses a different question: whether arbitrary memorized facts can be added on top of a Bayes-optimal structured rule without degrading the population rule risk. The distinction is formal. In our notation, factual storage alone asks whether there exists f such that

$$f(X^\mu) = \xi^\mu \quad \forall \mu.$$

Bayes-optimal coexistence asks whether this can be done while also making

$$R_{\text{rule}}(f) - R_{\text{Bayes}} \to 0.$$

By Theorem C.5, this is equivalent to constructing fact corrections whose $L^2(P_X)$-mass vanishes. The sparse fact-bump theorem is precisely this stronger statement.

A.5. Relation to Generalization–Hallucination Mechanisms

Work on out-of-context reasoning and hallucination, such as Huang et al. (2025), gives a theory for how one-layer attention-only transformers may generalize or hallucinate through associative mechanisms. That line of work is complementary. It

studies when a model extrapolates factual implications, including spurious implications. Our paper instead fixes a Bayes rule $f^\star$, injects a finite arbitrary fact residual vector $\delta(F_m)$, and studies the exact population cost of forcing those facts to be memorized.

The technical objects are different. OCR-style analyses concern the mechanism by which a model associates facts or implications. Our analysis concerns a variational localizability functional and an explicit decomposition

$$\text{rule block} \oplus \text{residual fact block}.$$

The deletability theorem then shows that the residual facts in our construction are structurally removable. This is not a hallucination mitigation theorem for arbitrary models; it is an exact structural theorem for a controlled coexistence representation.

A.6. Relation to Component-Specialization Work

Dong et al. (2025) study how different transformer components contribute to memorization, retrieval, and algorithmic tasks; their findings suggest that MLPs and attention mechanisms play different roles. Related one-layer theories also distinguish between attention-mediated selection and MLP-mediated storage (Xu & Chen, 2025).

Our result is not an empirical component ablation and not a qualitative claim that "attention retrieves" while "MLPs memorize." Instead, Theorems H.3 and H.13 prove a structural separation for the constructed coexistence predictor:

$$f_{\Theta_F}(X) = \underbrace{f^\star(X)}_{\text{rule block}} + \underbrace{g_F(Z^\perp(X))}_{\text{residual fact block}}.$$

The rule block is supported on the first r attention coordinates, while the fact block is supported on at most $3m$ residual-only MLP units. Moreover, the coefficient-space cross Gram between rule and fact dictionaries vanishes in the coexistence regime. This is a theorem-level localization statement rather than a component-level empirical observation.

A.7. Relation to Lazy Training and Kernel Interpolation

The lazy/tangent regime of neural networks is classically modeled by the neural tangent kernel (Jacot et al., 2018) and by lazy-training analyses (Chizat et al., 2019). Kernel methods can certainly interpolate data, and our paper does not claim otherwise. The question is sharper: can a bounded kernel correction interpolate arbitrary signed fact residuals while remaining $L^2(P_X)$-invisible?

Theorem G.9 proves an effective-dimension bottleneck. Even if the kernel class is granted the Bayes rule $f^\star$ for free, a bounded RKHS correction cannot interpolate independent signed facts with small population mass unless

$$\mathcal{N}_K(\zeta)$$

is large relative to the number of facts. The affine lazy result then follows by containment of the tangent class in an oracle-centered RKHS ball. This converse is deliberately not a universal impossibility theorem for all kernels; it is a separation between learned residual geometry and bounded fixed-representation interpolation.

A.8. Relation to Machine Unlearning

Machine unlearning asks for mechanisms by which trained systems can remove the effect of selected data (Cao & Yang, 2015; Bourtoule et al., 2021). The present paper does not propose a general-purpose unlearning algorithm for arbitrary trained foundation models. Instead, it proves an exact deletability corollary inside the coexistence construction.

The distinction is important. General unlearning is algorithmic and distributional: after training has occurred, one asks for a procedure whose output resembles retraining without the deleted data. Our result is structural: for the constructed predictor, the memorized residuals are represented by identifiable residual-only MLP units. Therefore, deletion is implemented by a deterministic parameter projection:

$$\Theta_F \mapsto \Theta_F^{(-U)},$$

which zeroes the output coefficients of the three units attached to each deleted fact $\mu \in U$. This gives exact identities

$$f_{\Theta_F^{(-U)}}(X^\nu) = f^\star(X^\nu) \quad (\nu \in U), \qquad f_{\Theta_F^{(-U)}}(X^\nu) = \xi^\nu \quad (\nu \notin U),$$

and full deletion recovers

$$f_{\Theta_F^{(-\mathrm{all})}} = f^\star$$

on the entire input space. Thus the result is narrower than general unlearning, but stronger within its theorem-level model.

A.9. Summary Table

Table 1. Novelty boundary relative to the closest theoretical lines.

Prior line	What it establishes	What this paper adds
Rules-and-facts theory (Farné et al., 2026)	Solvable framework for simultaneous rule learning and fact memorization.	Causal decoder theorem with trainable attention features, fact localizability, lazy/kernel converse, and exact deletability.
Bayes-optimal attention theory (Boncoraglio et al., 2025)	Bayes-optimal learning and phase transitions in attention-indexed models.	Adds arbitrary exception facts and proves coexistence via a residual Gaussian block and sparse fact localization.
Factual recall/storage theory (Nichani et al., 2024; Xu & Chen, 2025)	Mechanisms and capacity for storing or extracting factual knowledge.	Proves exact fact memorization with vanishing excess Bayes rule risk.
Component-specialization studies (Dong et al., 2025)	Evidence and theory that transformer components contribute differently to retrieval, memorization, and sequence modeling.	Proves a sparse-plus-low-rank functional decomposition and vanishing rule–fact Gram coupling for the constructed coexistence predictor.
Lazy/NTK theory (Jacot et al., 2018; Chizat et al., 2019)	Fixed-representation kernel description of lazy neural networks.	Shows an effective-dimension bottleneck for fact localizability in bounded lazy decoder classes.
Machine unlearning (Cao & Yang, 2015; Bourtoule et al., 2021)	Algorithmic goals and protocols for removing data effects from trained models.	Proves exact structural deletion of memorized residual facts in the constructed coexistence representation.

A.10. Claims Deliberately Not Made

We close by listing claims that are outside the scope of the theorem package.

First, we do not claim that arbitrary trained large language models learn the specific split representation constructed in Theorem D.9. The result is a theorem about an explicit causal decoder class.

Second, we do not claim an optimization theorem for stochastic gradient descent. The core results are representation, localization, and lower-bound statements. Any optimization result would require a separate analysis and should not be used as a hidden premise.

Third, we do not claim that kernels cannot memorize. The lazy converse is an effective-dimension lower bound: sufficiently rich kernels may interpolate many facts, but bounded low-effective-dimension tangent classes cannot do so with vanishing population mass.

Fourth, we do not claim a general-purpose machine-unlearning algorithm. The deletability theorem is exact for the constructed sparse residual representation and should be understood as a structural possibility result.

These restrictions are not weaknesses of the theorem package; they are the conditions that make the separation mathematically precise.

B. Formal Setup, Assumptions, and Notation

This appendix fixes the theorem-level model used throughout the paper. Our setup combines the structured/fact decomposition of the Rules-and-Facts (RAF) model (Farné et al., 2026) with a causal decoder architecture closer to recent tractable attention models and one-layer transformer abstractions (Boncoraglio et al., 2025; Xu & Chen, 2025; Nichani et al., 2024; Vasudeva et al., 2026; Dong et al., 2025). The role of this appendix is purely foundational: it specifies the ambient probability space, the teacher rule, the finite fact set, the decoder function classes, the lazy/tangent class used in the converse, and the exact risk functionals optimized in the subsequent appendices. Unless explicitly stated otherwise, Appendices C–H are *representation-first*: they establish existence and separation results for function classes and do not assume that a particular

optimization algorithm finds the relevant predictors.

B.1. Ambient Space, Tokens, and Asymptotic Regime

Fix an integer $L \geq 2$, interpreted as the observed prefix length of a decoder-only next-token prediction problem. For every ambient dimension $d \in \mathbb{N}$, define the raw content-token space

$$\mathcal{X}_d := (\mathbb{R}^d)^L.$$

An input sequence is denoted by

$$X = (x_1, \ldots, x_L) \in \mathcal{X}_d,$$

where x_L is the final observed token and plays the role of the *readout/query token*. We write $[m] := \{1, \ldots, m\}$ for every $m \in \mathbb{N}$.

To retain exact positional information without burdening the analysis with learned positional embeddings, we work with position-augmented tokens. Let $e_1, \ldots, e_L$ denote the canonical basis of $\mathbb{R}^L$, and define

$$z_t(X) := (x_t, e_t) \in \mathbb{R}^{d+L}, \qquad t \in [L].$$

Thus the model has access to both content and position, but only through a fixed, deterministic augmentation.

Assumption B.1 (Input law)**.** For each d, the random vectors $x_1, \ldots, x_L$ are independent and identically distributed according to a distribution $\mathcal{D}_d$ on $\mathbb{R}^d$ satisfying

$$\mathbb{E}[x_t] = 0, \qquad \mathbb{E}[x_t x_t^\top] = I_d,$$

and

$$\sup_{\|u\|_2=1} \|\langle u, x_t \rangle\|_{\psi_2} \leq \kappa$$

for some finite constant $\kappa > 0$ independent of d and t. Here $\|\cdot\|_{\psi_2}$ denotes the usual subgaussian Orlicz norm.

Remark B.2 (Why this stylized input model). Theorems about coexistence are simplest when the input law is isotropic and high-dimensional, as in teacher–student and modern attention-indexed models (Farné et al., 2026; Boncoraglio et al., 2025). The position augmentation above is a deterministic device that makes the causal ordering explicit while preserving the concentration geometry of the content coordinates.

Assumption B.3 (Asymptotic regime)**.** All statements are indexed by $d \to \infty$. The sequence length L, the teacher rank r introduced below, the subgaussian parameter κ, and all structural constants appearing in the standing assumptions are fixed with respect to d. The quantities

$$n = n(d), \qquad m = m(d), \qquad H = H(d), \qquad W_f = W_f(d), \qquad B = B(d)$$

may vary with d. When a residual dimension parameter $d_\perp = d_\perp(d)$ appears later, it is understood to satisfy $1 \leq d_\perp \leq H - r$. Unless explicitly stated otherwise, the budget $B(d)$ is allowed to grow at most polynomially in d.

Throughout, $c, C, c_0, C_0, \ldots$ denote positive constants whose values may change from line to line; they may depend on fixed structural parameters such as L, r, κ, and the teacher/activation constants, but never on d, n, m, H, W_f unless explicitly indicated. We write $a_d \lesssim b_d$ if $a_d \leq C b_d$ for all sufficiently large d, and $a_d \asymp b_d$ if both $a_d \lesssim b_d$ and $b_d \lesssim a_d$. The phrase *with high probability* means probability tending to one as $d \to \infty$.

B.2. Teacher Rule and the Rule Distribution

The structured part of the data is generated by a finite-rank causal attention-type teacher. This matches the score nonlinearity used by the student architecture and isolates the coexistence question from approximation error. Such matched teacher–student abstractions are standard in solvable theory for memorization/generalization and attention-based models (Farné et al., 2026; Boncoraglio et al., 2025).

Assumption B.4 (Score nonlinearity and MLP activation)**.** The scalar attention score map $\psi : \mathbb{R} \to \mathbb{R}$ is fixed, twice continuously differentiable, satisfies $\psi(0) = 0$, and obeys

$$\sup_{u \in \mathbb{R}} \big(|\psi'(u)| + |\psi''(u)|\big) \leq B_\psi, \qquad |\psi(u)| \leq B_\psi(1 + |u|)$$

for some finite constant $B_\psi > 0$. The hidden activation in the final MLP is the ReLU

$$\sigma(u) := u_+ := \max\{u, 0\}.$$

Assumption B.5 (Rank-r causal teacher)**.** Fix $r \in \mathbb{N}$. There exist parameters

$$\{q_a^\star, k_a^\star, v_a^\star, \beta_a^\star\}_{a=1}^r \subset \mathbb{R}^{d+L} \times \mathbb{R}^{d+L} \times \mathbb{R}^{d+L} \times \mathbb{R}$$

such that

$$\max_{a\in[r]} \Big\{ \|q_a^\star\|_2,\ \|k_a^\star\|_2,\ \|v_a^\star\|_2,\ |\beta_a^\star| \Big\} \le B_\star$$

for a constant $B_\star > 0$ independent of d, and the teacher rule is

$$f^\star(X) = \sum_{a=1}^r \beta_a^\star S_a^\star(X), \qquad S_a^\star(X) := \sum_{t=1}^{L-1} \psi\big(\langle q_a^\star, z_L(X)\rangle \langle k_a^\star, z_t(X)\rangle\big)\, \langle v_a^\star, z_t(X)\rangle. \tag{13}$$

The observed response is

$$Y = f^\star(X) + \varepsilon, \tag{14}$$

where ε is independent of X, centered, and subgaussian with

$$\mathbb{E}[\varepsilon^2] = \sigma_\varepsilon^2 < \infty.$$

Let P_X denote the law of X, and P_{rule} the joint law of (X, Y) induced by Theorems B.1 and B.5. By construction,

$$f^\star(X) = \mathbb{E}[Y \mid X],$$

so $f^\star$ is the unique $L^2(P_X)$-Bayes regressor.

Definition B.6 (Square-integrable predictor space)**.** We write $\mathcal{M}_2(P_X)$ for the set of all measurable functions $f : \mathcal{X}_d \to \mathbb{R}$ such that

$$\|f\|_{L^2(P_X)}^2 := \mathbb{E}_{X\sim P_X}[f(X)^2] < \infty.$$

Whenever $g \in \mathcal{M}_2(P_X)$, we also write

$$\|g\|_{L^2(P_{\text{rule}})}^2 := \mathbb{E}_{(X,Y)\sim P_{\text{rule}}}[g(X)^2] = \|g\|_{L^2(P_X)}^2.$$

B.3. Regular Samples, Fact Sets, and the Representation-First Viewpoint

For a regular sample size n, let

$$S_n := \{(X_i, Y_i)\}_{i=1}^n$$

be i.i.d. draws from P_{rule}. In addition to these structured examples, we consider a finite set of *facts* (or exceptions)

$$F_m := \{(X^\mu, \xi^\mu)\}_{\mu=1}^m,$$

where the locations X^μ are sampled from the same input marginal P_X, but the values ξ^μ are *arbitrary* responses that need not follow the teacher rule. This models sparse exceptions or memorized facts on top of the structured rule, in the spirit of RAF (Farné et al., 2026), but in a causal decoder setting.

Definition B.7 (Fact residual vector)**.** Given a fact set F_m, define its residual vector

$$\delta(F_m) := (\delta_1, \dots, \delta_m)^\top \in \mathbb{R}^m, \qquad \delta_\mu := \xi^\mu - f^\star(X^\mu).$$

Definition B.8 (Admissible fact set)**.** Fix $B_\delta > 0$. A finite fact set F_m is called B_δ-*admissible* if

$$\|\delta(F_m)\|_\infty = \max_{\mu\in[m]} |\xi^\mu - f^\star(X^\mu)| \le B_\delta.$$

Definition B.9 (Generic fact set)**.** A random admissible fact set F_m is called *generic* if its locations $X^1, \dots, X^m$ are i.i.d. draws from P_X, independent of the regular sample S_n and of any random initialization introduced later, while the residual vector $\delta(F_m)$ may be arbitrary (possibly deterministic), subject only to the admissibility constraint $\|\delta(F_m)\|_\infty \le B_\delta$.

Whenever a theorem is stated for generic fact sets, the probability is over the draw of the locations $X^1, \dots, X^m$, and the conclusion is uniform over all admissible residual vectors $\delta(F_m)$.

Definition B.10 (Evaluation operator on a fact set)**.** For a fixed fact set F_m, define the evaluation operator

$$E_{F_m} : \mathcal{M}_2(P_X) \to \mathbb{R}^m, \qquad E_{F_m}(g) := \big(g(X^1), \dots, g(X^m)\big)^\top.$$

The central theorems in this paper are *class-level* statements. Thus, rather than postulating a specific training algorithm, we study the existence or nonexistence of predictors in a function class that simultaneously achieve low rule risk and exact interpolation of the fact set. For intuition only, one may view this as the constrained empirical problem

$$\min_\Theta \ \frac{1}{n}\sum_{i=1}^{n}\big(f_\Theta(X_i) - Y_i\big)^2 \quad \text{subject to} \quad f_\Theta(X^\mu) = \xi^\mu \ \ \forall \mu \in [m], \tag{15}$$

but Appendices C–H do *not* assume that any specific algorithm solves Equation (15). Optimization is deferred to a separate, optional appendix.

B.4. Trainable-Feature Decoder Class

We now define the decoder-only architecture that underlies the upper theorem. It is intentionally minimal: a single causal self-attention block, evaluated at the final readout token, followed by a one-hidden-layer ReLU MLP. This class is rich enough to contain the teacher rule and, later, a sparse fact-correction layer, while still remaining explicit enough for exact analysis.

Fix integers $H \ge 1$ and $W_f \ge 0$, interpreted as the number of scalar attention heads and the width of the final MLP, respectively.

For each head $h \in [H]$, let

$$q_h,\ k_h,\ v_h \in \mathbb{R}^{d+L}.$$

Define the h-th scalar causal head by

$$A_h(X;\theta_h) := \sum_{t=1}^{L-1} \psi\big(\langle q_h, z_L(X)\rangle \langle k_h, z_t(X)\rangle\big)\, \langle v_h, z_t(X)\rangle, \qquad \theta_h := (q_h, k_h, v_h). \tag{16}$$

The full attention representation is

$$A_\Theta(X) := \big(A_1(X;\theta_1), \dots, A_H(X;\theta_H)\big)^\top \in \mathbb{R}^H. \tag{17}$$

The final scalar predictor is

$$f_\Theta(X) = c + \langle w, A_\Theta(X)\rangle + \sum_{j=1}^{W_f} \alpha_j\, \sigma\big(\langle b_j, A_\Theta(X)\rangle - \tau_j\big), \tag{18}$$

where

$$w, b_j \in \mathbb{R}^H, \qquad c, \alpha_j, \tau_j \in \mathbb{R}.$$

The full parameter vector is

$$\Theta = \Big((\theta_h)_{h=1}^H,\ w,\ c,\ (\alpha_j, b_j, \tau_j)_{j=1}^{W_f}\Big).$$

Definition B.11 (Trainable-feature decoder class)**.** For integers H, W_f and a budget $B > 0$, define $\mathcal{T}_{\mathrm{dec}}(H, W_f, B)$ as the set of all predictors $f_\Theta : \mathcal{X}_d \to \mathbb{R}$ of the form Equations (16) to (18) such that

$$\max_{h\in[H]} \big\{\|q_h\|_2,\ \|k_h\|_2,\ \|v_h\|_2\big\} \le B, \qquad \|w\|_2 \le B,$$

and

$$\max_{j\in[W_f]} \left\{ \|b_j\|_2,\ |\alpha_j|,\ |\tau_j| \right\} \le B, \qquad |c| \le B.$$

When the dependence on H, W_f, B is clear from context, we abbreviate this class by $\mathcal{T}_{\mathrm{dec}}$.

Remark B.12 (Matched teacher inclusion). If $H \ge r$, then the teacher rule $f^\star$ belongs to $\mathcal{T}_{\mathrm{dec}}(H, W_f, B)$ for all sufficiently large B: one may set the first r heads equal to the teacher heads $\{q_a^\star, k_a^\star, v_a^\star\}_{a=1}^r$, switch off the remaining heads, take $W_f = 0$, and choose the linear readout w to equal $(\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0)$. This observation is purely representational and will be used later to eliminate approximation error from the coexistence analysis.

B.5. Lazy/Tangent Class and the Associated Kernel

The lower-bound appendix compares the fully trainable decoder class above to the lazy (or tangent) regime of the *same* architecture. We therefore record the corresponding tangent kernel and affine tangent class here.

Definition B.13 (Tangent kernel and lazy class)**.** Fix a base point Θ_0 such that $X \mapsto \nabla_\Theta f_{\Theta_0}(X)$ is square-integrable under P_X. The associated tangent kernel is

$$K_{\Theta_0}(X, X') := \left\langle \nabla_\Theta f_{\Theta_0}(X), \nabla_\Theta f_{\Theta_0}(X') \right\rangle. \tag{19}$$

For $R > 0$, define the affine lazy class

$$\mathcal{T}_{\mathrm{lazy}}(R; \Theta_0) := \left\{ X \mapsto f_{\Theta_0}(X) + \left\langle \nabla_\Theta f_{\Theta_0}(X), u \right\rangle \ :\ \|u\|_2 \le R \right\}. \tag{20}$$

We write $\mathcal{H}_{K_{\Theta_0}}$ for the reproducing kernel Hilbert space (RKHS) associated with K_{Θ_0}.

Remark B.14 (Why the lazy class is the correct converse baseline). The lower-bound theorem will compare a fully trainable feature-learning decoder to the tangent class of the same parameterization. This isolates the gain from feature learning itself, rather than from changing the architecture class.

B.6. Risks, Exact Memorization, and the Two Central Functionals

We now define the functionals that quantify coexistence between rule generalization and exact fact memorization.

Definition B.15 (Rule risk and Bayes risk)**.** For any $f \in \mathcal{M}_2(P_X)$, define the squared rule risk

$$R_{\mathrm{rule}}(f) := \mathbb{E}_{(X,Y)\sim P_{\mathrm{rule}}} \left[(f(X) - Y)^2 \right]. \tag{21}$$

The Bayes risk is

$$R_{\mathrm{Bayes}} := R_{\mathrm{rule}}(f^\star) = \mathbb{E}[\varepsilon^2]. \tag{22}$$

Definition B.16 (Exact memorization of a fact set)**.** A predictor $f \in \mathcal{M}_2(P_X)$ is said to *exactly memorize* the fact set $F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$ if

$$f(X^\mu) = \xi^\mu \qquad \text{for every } \mu \in [m].$$

Equivalently, f exactly memorizes F_m if and only if

$$E_{F_m}(f) = (\xi^1, \dots, \xi^m)^\top.$$

Definition B.17 (Bayes-coexistence gap)**.** Let $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ be any predictor class and let F_m be any fact set. The *Bayes-coexistence gap* of $\mathcal{T}$ with respect to F_m is

$$\Delta_{F_m}(\mathcal{T}) := \inf_{\substack{f\in\mathcal{T} \\ f(X^\mu)=\xi^\mu,\ \forall \mu\in[m]}} \Big(R_{\mathrm{rule}}(f) - R_{\mathrm{Bayes}} \Big). \tag{23}$$

By convention, $\Delta_{F_m}(\mathcal{T}) := +\infty$ if the feasibility set is empty.

Definition B.18 (Fact localizability)**.** Let $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ be any predictor class and let F_m be any fact set. The *fact localizability* of F_m in $\mathcal{T}$ under the rule law is

$$\Lambda_{F_m}(\mathcal{T}; P_{\mathrm{rule}}) := \inf_{\substack{g\in\mathcal{M}_2(P_X) \\ f^\star+g\in\mathcal{T} \\ g(X^\mu)=\delta_\mu,\ \forall \mu\in[m]}} \|g\|_{L^2(P_X)}^2. \tag{24}$$

Again, $\Lambda_{F_m}(\mathcal{T}; P_{\mathrm{rule}}) := +\infty$ if the feasible set is empty.

Remark B.19 (Interpretation of the two functionals). The quantity $\Delta_{F_m}(\mathcal{T})$ measures the unavoidable excess rule risk that the class $\mathcal{T}$ must pay if it insists on exact memorization of all facts in F_m. The quantity $\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}})$ measures the minimum $L^2(P_X)$-mass of a perturbation that implants those facts on top of the Bayes rule $f^\star$. Appendix C will show that, under squared loss, these two quantities coincide exactly whenever $f^\star \in \mathcal{T}$.

B.7. Standing Conventions for Later Appendices

For later use, we record the following conventions.

Remark B.20 (Uniformity conventions). Whenever a theorem is stated for generic fact sets, the probability is over the random locations $X^1, \ldots, X^m$, and the conclusion is uniform over all admissible residual vectors $\delta(F_m)$. Whenever a theorem involves the lazy class, any probability over random initialization is stated explicitly in that theorem.

Remark B.21 (Representation-first scope). Appendices C–H are concerned with *existence*, *lower bounds*, and *structural separation* inside the classes $\mathcal{T}_{\text{dec}}$ and $\mathcal{T}_{\text{lazy}}$. No claim about gradient descent, stochastic gradient descent, or any other training dynamics is made in those appendices. If an optimization theorem is added, it should appear separately and should not be used as a hidden premise for any of the core representation theorems.

Remark B.22 (Loss model). The exact identity between coexistence cost and localizability is cleanest under squared loss, which is why the core theory is developed in that setting. This choice is standard in theorem-level analyses of synthetic transformer tasks and teacher–student models. A local extension to Bernoulli/cross-entropy losses around the Bayes logit can be stated separately without altering the setup above.

C. Proof of the Coexistence–Localizability Identity

This appendix proves the exact identity that underlies the entire paper. At the level of function classes, the coexistence problem considered here is the problem of simultaneously achieving Bayes-optimal rule prediction under the structured rule distribution P_{rule} while exactly interpolating a finite fact set F_m. The key point is that, under squared loss, this problem is *exactly equivalent* to a constrained minimum-seminorm perturbation problem around the Bayes regressor $f^\star$. This reduction is what converts the rules-and-facts coexistence perspective of Farné et al. (2026) into a tractable geometric question for causal decoder classes: all subsequent upper and lower bounds reduce to quantifying the minimum $L^2(P_X)$-mass of a fact-implanting perturbation.

Throughout this appendix, we adopt all notation and assumptions from Section B. In particular, $f^\star$ is the explicit teacher rule defined in Equation (13), P_{rule} is the corresponding joint law of (X, Y), $\Delta_{F_m}(\mathcal{T})$ is the Bayes-coexistence gap from Theorem B.17, and $\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}})$ is the fact localizability functional from Theorem B.18.

C.1. Feasible Sets and a Semiinorm Convention

Because exact memorization is imposed by pointwise constraints at the realized fact locations $X^1, \ldots, X^m$, we work with actual measurable predictors rather than P_X-almost-sure equivalence classes. Consequently,

$$\langle g, h\rangle_X := \mathbb{E}_{X\sim P_X}[g(X)h(X)], \qquad \|g\|^2_{L^2(P_X)} = \langle g, g\rangle_X,$$

is, strictly speaking, a *semi*-inner-product and a *semi*-norm on $\mathcal{M}_2(P_X)$: nonzero functions supported on P_X-null sets may have zero $L^2(P_X)$-mass. This distinction is essential in exact memorization problems and will be exploited later. None of the proofs in this appendix requires quotienting by P_X-almost-sure equality.

For any predictor class $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ and fact set $F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$, define the feasible predictor and feasible correction sets

$$\mathcal{A}_{F_m}(\mathcal{T}) := \Big\{ f \in \mathcal{T} : f(X^\mu) = \xi^\mu \ \ \forall \mu \in [m] \Big\}, \tag{25}$$

$$\mathcal{G}_{F_m}(\mathcal{T}) := \Big\{ g \in \mathcal{M}_2(P_X) : f^\star + g \in \mathcal{T}, \ \ g(X^\mu) = \delta_\mu \ \ \forall \mu \in [m] \Big\}, \tag{26}$$

where $\delta_\mu = \xi^\mu - f^\star(X^\mu)$ is the fact residual from Theorem B.7. By definition,

$$\Delta_{F_m}(\mathcal{T}) = \inf_{f \in \mathcal{A}_{F_m}(\mathcal{T})} \Big(R_{\text{rule}}(f) - R_{\text{Bayes}} \Big), \tag{27}$$

and

$$\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}) = \inf_{g \in \mathcal{G}_{F_m}(\mathcal{T})} \|g\|^2_{L^2(P_X)}, \tag{28}$$

with the convention that the infimum of the empty set is $+\infty$.

C.2. The Bayes Orthogonality and Pythagorean Identity

The next lemma is the only place where the distributional structure of P_{rule} enters this appendix. It is the precise orthogonality statement that makes the coexistence problem an exact minimum-seminorm problem.

Lemma C.1 (Bayes orthogonality)**.** *For every $g \in \mathcal{M}_2(P_X)$,*

$$\mathbb{E}_{(X,Y)\sim P_{\text{rule}}}\big[g(X)\,(Y - f^\star(X))\big] = 0. \tag{29}$$

Equivalently,

$$\langle g,\, Y - f^\star(X)\rangle_{L^2(P_{\text{rule}})} = 0 \qquad \textit{for every } g(X) \in L^2(P_X).$$

Proof. Fix $g \in \mathcal{M}_2(P_X)$. Since $g(X) \in L^2(P_X)$ and $Y - f^\star(X) = \varepsilon$ has finite second moment by Theorem B.5, the product $g(X)(Y - f^\star(X))$ is integrable by Cauchy–Schwarz. Hence conditional expectation is well-defined, and

$$\begin{aligned}\mathbb{E}\big[g(X)(Y - f^\star(X))\big] &= \mathbb{E}\Big[\mathbb{E}\big[g(X)(Y - f^\star(X)) \mid X\big]\Big] \\ &= \mathbb{E}\Big[g(X)\,\mathbb{E}\big[Y - f^\star(X) \mid X\big]\Big].\end{aligned}$$

By construction of the teacher model,

$$f^\star(X) = \mathbb{E}[Y \mid X]$$

almost surely under P_{rule}. Therefore

$$\mathbb{E}\big[Y - f^\star(X) \mid X\big] = 0 \qquad \text{almost surely,}$$

and the last display is zero. □

Proposition C.2 (Bayes Pythagorean identity)**.** *For every $f \in \mathcal{M}_2(P_X)$,*

$$R_{\text{rule}}(f) - R_{\text{Bayes}} = \|f - f^\star\|^2_{L^2(P_X)}. \tag{30}$$

Equivalently,

$$R_{\text{rule}}(f) = R_{\text{Bayes}} + \|f - f^\star\|^2_{L^2(P_X)}. \tag{31}$$

Proof. Let $f \in \mathcal{M}_2(P_X)$. Since $f - f^\star \in \mathcal{M}_2(P_X)$, all terms below are integrable. Expanding the square gives

$$\begin{aligned}R_{\text{rule}}(f) &= \mathbb{E}\big[(f(X) - Y)^2\big] \\ &= \mathbb{E}\big[(f(X) - f^\star(X) + f^\star(X) - Y)^2\big] \\ &= \mathbb{E}\big[(f(X) - f^\star(X))^2\big] + \mathbb{E}\big[(f^\star(X) - Y)^2\big] \\ &\quad + 2\,\mathbb{E}\big[(f(X) - f^\star(X))(f^\star(X) - Y)\big].\end{aligned}$$

By Theorem C.1 applied to $g = f - f^\star$, the cross term is zero. Since $R_{\text{Bayes}} = R_{\text{rule}}(f^\star) = \mathbb{E}[(f^\star(X) - Y)^2]$, the claim follows. □

Remark C.3 (Loss specificity). Theorem C.2 is exact because squared loss induces a genuine Pythagorean identity around the Bayes regressor. For other losses, one typically obtains only local or inequality-based comparisons around the Bayes predictor, not an exact global identity. This is why the core theorem of the paper is formulated under squared loss.

C.3. Translation of Feasible Sets

The next lemma isolates the purely algebraic part of the argument: exact memorization constraints translate one-to-one between predictors f and Bayes-centered perturbations $g = f - f^\star$.

Lemma C.4 (Translation bijection between predictors and corrections)**.** *Let $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ be any predictor class, and let $F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$ be any fact set. Define*

$$\Phi_{f^\star} : \mathcal{A}_{F_m}(\mathcal{T}) \to \mathcal{G}_{F_m}(\mathcal{T}), \qquad \Phi_{f^\star}(f) := f - f^\star.$$

Then $\Phi_{f^\star}$ is a bijection with inverse

$$\Psi_{f^\star} : \mathcal{G}_{F_m}(\mathcal{T}) \to \mathcal{A}_{F_m}(\mathcal{T}), \qquad \Psi_{f^\star}(g) := f^\star + g.$$

In particular,

$$\mathcal{A}_{F_m}(\mathcal{T}) = \varnothing \quad \Longleftrightarrow \quad \mathcal{G}_{F_m}(\mathcal{T}) = \varnothing. \tag{32}$$

Proof. Take $f \in \mathcal{A}_{F_m}(\mathcal{T})$. By definition, $f \in \mathcal{T}$ and $f(X^\mu) = \xi^\mu$ for all $\mu \in [m]$. Let $g := f - f^\star$. Then $g \in \mathcal{M}_2(P_X)$ because both f and $f^\star$ are square-integrable, and

$$f^\star + g = f \in \mathcal{T}.$$

Moreover, for each $\mu \in [m]$,

$$g(X^\mu) = f(X^\mu) - f^\star(X^\mu) = \xi^\mu - f^\star(X^\mu) = \delta_\mu.$$

Hence $g \in \mathcal{G}_{F_m}(\mathcal{T})$, so $\Phi_{f^\star}$ is well-defined.

Conversely, take $g \in \mathcal{G}_{F_m}(\mathcal{T})$ and define $f := f^\star + g$. Then $f \in \mathcal{T}$ by definition of $\mathcal{G}_{F_m}(\mathcal{T})$, and for each $\mu \in [m]$,

$$f(X^\mu) = f^\star(X^\mu) + g(X^\mu) = f^\star(X^\mu) + \delta_\mu = \xi^\mu.$$

Thus $f \in \mathcal{A}_{F_m}(\mathcal{T})$, so $\Psi_{f^\star}$ is well-defined.

Finally, for every $f \in \mathcal{A}_{F_m}(\mathcal{T})$,

$$\Psi_{f^\star}(\Phi_{f^\star}(f)) = f^\star + (f - f^\star) = f,$$

and for every $g \in \mathcal{G}_{F_m}(\mathcal{T})$,

$$\Phi_{f^\star}(\Psi_{f^\star}(g)) = (f^\star + g) - f^\star = g.$$

Therefore $\Phi_{f^\star}$ and $\Psi_{f^\star}$ are inverse bijections. The equivalence Equation (32) follows immediately. □

C.4. The Core Identity

We can now state and prove the exact identity that grounds the rest of the paper. Importantly, no assumption that $f^\star \in \mathcal{T}$ is needed.

Theorem C.5 (Coexistence–localizability identity)**.** *Let $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ be any predictor class, and let F_m be any fact set. Then*

$$\Delta_{F_m}(\mathcal{T}) = \Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}). \tag{33}$$

Equivalently,

$$\Delta_{F_m}(\mathcal{T}) = \inf_{f \in \mathcal{A}_{F_m}(\mathcal{T})} \|f - f^\star\|^2_{L^2(P_X)} = \inf_{g \in \mathcal{G}_{F_m}(\mathcal{T})} \|g\|^2_{L^2(P_X)}. \tag{34}$$

Proof. If $\mathcal{A}_{F_m}(\mathcal{T}) = \varnothing$, then $\Delta_{F_m}(\mathcal{T}) = +\infty$ by definition. By Theorem C.4, this is equivalent to $\mathcal{G}_{F_m}(\mathcal{T}) = \varnothing$, in which case $\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}) = +\infty$. Hence the identity holds in the infeasible case.

Assume now that $\mathcal{A}_{F_m}(\mathcal{T}) \neq \varnothing$, and equivalently $\mathcal{G}_{F_m}(\mathcal{T}) \neq \varnothing$. Using Equation (27), Theorem C.2, and then Theorem C.4, we obtain

$$\begin{aligned}
\Delta_{F_m}(\mathcal{T}) &= \inf_{f \in \mathcal{A}_{F_m}(\mathcal{T})} \Big(R_{\text{rule}}(f) - R_{\text{Bayes}} \Big) \\
&= \inf_{f \in \mathcal{A}_{F_m}(\mathcal{T})} \|f - f^\star\|^2_{L^2(P_X)} \\
&= \inf_{g \in \mathcal{G}_{F_m}(\mathcal{T})} \|g\|^2_{L^2(P_X)} \\
&= \Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}).
\end{aligned}$$

This proves both Equation (33) and Equation (34). □

Corollary C.6 (Asymptotic coexistence criterion)**.** *Let $\{\mathcal{T}_d\}_{d\geq 1}$ be any sequence of predictor classes and $\{F_{m(d)}\}_{d\geq 1}$ any sequence of fact sets in the asymptotic regime of Theorem B.3. Then*

$$\Delta_{F_{m(d)}}(\mathcal{T}_d) \to 0 \quad \Longleftrightarrow \quad \Lambda_{F_{m(d)}}(\mathcal{T}_d; P_{\mathrm{rule}}) \to 0.$$

Proof. Immediate from Theorem C.5. □

Corollary C.7 (Monotonicity under class enlargement)**.** *If $\mathcal{T}_1 \subseteq \mathcal{T}_2 \subseteq \mathcal{M}_2(P_X)$, then for every fact set F_m,*

$$\Delta_{F_m}(\mathcal{T}_2) \leq \Delta_{F_m}(\mathcal{T}_1), \qquad \Lambda_{F_m}(\mathcal{T}_2; P_{\mathrm{rule}}) \leq \Lambda_{F_m}(\mathcal{T}_1; P_{\mathrm{rule}}).$$

Proof. Both feasible sets enlarge under class inclusion:

$$\mathcal{A}_{F_m}(\mathcal{T}_1) \subseteq \mathcal{A}_{F_m}(\mathcal{T}_2), \qquad \mathcal{G}_{F_m}(\mathcal{T}_1) \subseteq \mathcal{G}_{F_m}(\mathcal{T}_2).$$

Taking infima over larger sets cannot increase the value. □

C.5. Why Structured Function Classes Are Necessary

The next proposition is not used as a technical ingredient later, but it is conceptually important: without architectural or regularity constraints, coexistence is *trivial*. This explains why all nontrivial content in the paper lies in the geometry of structured decoder classes.

Definition C.8 (Closure under finite modifications)**.** A predictor class $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ is said to be *closed under finite modifications* if for every $f \in \mathcal{T}$, every $m \in \mathbb{N}$, every pairwise distinct points $x^1, \ldots, x^m \in \mathcal{X}_d$, and every coefficients $c_1, \ldots, c_m \in \mathbb{R}$, the function

$$x \mapsto f(x) + \sum_{\mu=1}^{m} c_\mu \, \mathbf{1}_{\{x = x^\mu\}} \tag{35}$$

also belongs to $\mathcal{T}$.

Proposition C.9 (Triviality for classes allowing point spikes)**.** *Assume that the input marginal P_X is non-atomic. Let $\mathcal{T} \subseteq \mathcal{M}_2(P_X)$ be a predictor class that contains $f^\star$ and is closed under finite modifications in the sense of Theorem C.8. Then for every finite fact set $F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$ whose locations $X^1, \ldots, X^m$ are pairwise distinct,*

$$\Delta_{F_m}(\mathcal{T}) = \Lambda_{F_m}(\mathcal{T}; P_{\mathrm{rule}}) = 0. \tag{36}$$

In particular, for generic fact sets under a non-atomic P_X, the conclusion holds almost surely.

Proof. Define the spike correction

$$g_F(x) := \sum_{\mu=1}^{m} \delta_\mu \, \mathbf{1}_{\{x = X^\mu\}}, \qquad \delta_\mu = \xi^\mu - f^\star(X^\mu). \tag{37}$$

Since the locations are pairwise distinct and P_X is non-atomic,

$$P_X(\{X^\mu\}) = 0 \qquad \text{for every } \mu \in [m],$$

hence

$$\|g_F\|^2_{L^2(P_X)} = \mathbb{E}\big[g_F(X)^2\big] = \sum_{\mu=1}^{m} \delta_\mu^2 \, P_X(\{X^\mu\}) = 0.$$

Because $f^\star \in \mathcal{T}$ and $\mathcal{T}$ is closed under finite modifications, the function $f^\star + g_F$ belongs to $\mathcal{T}$. By construction,

$$(f^\star + g_F)(X^\mu) = f^\star(X^\mu) + \delta_\mu = \xi^\mu \qquad \text{for every } \mu \in [m],$$

so $f^\star + g_F \in \mathcal{A}_{F_m}(\mathcal{T})$, while $g_F \in \mathcal{G}_{F_m}(\mathcal{T})$. Therefore

$$\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}) \le \|g_F\|^2_{L^2(P_X)} = 0.$$

Since $\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}) \ge 0$ by definition, it follows that $\Lambda_{F_m}(\mathcal{T}; P_{\text{rule}}) = 0$. The identity $\Delta_{F_m}(\mathcal{T}) = 0$ then follows from Theorem C.5.

Finally, if P_X is non-atomic and $X^1, \dots, X^m$ are sampled i.i.d. from P_X, then they are pairwise distinct almost surely, so the last claim follows. □

Remark C.10 (Why later appendices are nontrivial). Theorem C.9 shows that the coexistence problem is vacuous in unrestricted measurable classes: one can implant finitely many facts at zero population cost by modifying the predictor only on P_X-null sets. Thus the entire difficulty of the paper lies in proving that structured decoder classes $\mathcal{T}_{\text{dec}}$ and their lazy limits $\mathcal{T}_{\text{lazy}}$ *do not* admit arbitrary point spikes. The upper theorem later will show that trainable-feature decoders nevertheless admit *approximately localized* fact corrections whose rule mass decays exponentially in a residual dimension parameter; the converse theorem will show that the lazy/kernel regime cannot make this mass vanish.

D. Rule–Residual Factorization of Trainable-Feature Decoders

This appendix proves the structural representation theorem that makes the later fact-localization argument possible. The statement is deliberately stronger than the qualitative module-specialization narratives that have recently appeared in transformer theory (Boncoraglio et al., 2025; Xu & Chen, 2025; Nichani et al., 2024; Dong et al., 2025): we construct an *exact* causal decoder representation in which the first r attention coordinates coincide with the teacher rule statistics and the remaining $d_\perp$ coordinates form a Gaussian residual block independent of the teacher sigma-field. In particular, this appendix isolates a finite rule-relevant subspace of the representation and a complementary residual subspace that is available for later fact implantation.

D.1. Additional Specialization for Exact Factorization

The subgaussian input model of Theorem B.1 is sufficient for the general formalism, but exact independence between rule and residual coordinates is most transparent under a Gaussian specialization. From this appendix onward, whenever exact independence is invoked, we strengthen Theorem B.1 as follows.

Assumption D.1 (Gaussian specialization)**.** In this appendix and all later appendices that explicitly invoke Theorem D.1, the content-token law is

$$\mathcal{D}_d = \mathcal{N}(0, I_d).$$

Equivalently, $x_1, \dots, x_L$ are independent standard Gaussian vectors in $\mathbb{R}^d$.

We also exclude the degenerate case in which the attention score map vanishes identically.

Assumption D.2 (Nontrivial attention gate)**.** The score map ψ from Theorem B.4 is not identically zero. Fix once and for all a scalar $\rho_\psi \in \mathbb{R}$ such that

$$\gamma_\psi := \psi(\rho_\psi) \ne 0. \tag{38}$$

Remark D.3 (Why the Gaussian specialization is the right theorem-level model). The later upper theorem hinges on exact statistical orthogonality between the rule block and a residual block that can host localized fact corrections. For general subgaussian inputs, one can still hope for approximate decorrelation, but exact independence is no longer automatic. The Gaussian specialization captures the essential high-dimensional geometry while keeping the factorization theorem exact.

D.2. Content/Position Projections and the Teacher Content Subspace

For any $u \in \mathbb{R}^{d+L}$, write

$$u = (u^{\text{cnt}}, u^{\text{pos}}) \in \mathbb{R}^d \times \mathbb{R}^L,$$

and let

$$\Pi_{\text{cnt}} u := u^{\text{cnt}} \in \mathbb{R}^d, \qquad \Pi_{\text{pos}} u := u^{\text{pos}} \in \mathbb{R}^L.$$

Recall that the position-augmented token at location t is

$$z_t(X) = (x_t, e_t) \in \mathbb{R}^{d+L},$$

so for any $u = (u^{\mathrm{cnt}}, u^{\mathrm{pos}}) \in \mathbb{R}^{d+L}$,

$$\langle u, z_t(X)\rangle = \langle u^{\mathrm{cnt}}, x_t\rangle + \langle u^{\mathrm{pos}}, e_t\rangle. \tag{39}$$

Definition D.4 (Teacher content subspace)**.** Define the teacher content subspace

$$U_\star := \mathrm{span}\Big(\{\Pi_{\mathrm{cnt}} q_a^\star\}_{a=1}^r \cup \{\Pi_{\mathrm{cnt}} k_a^\star\}_{a=1}^r \cup \{\Pi_{\mathrm{cnt}} v_a^\star\}_{a=1}^r\Big) \subseteq \mathbb{R}^d, \tag{40}$$

and let

$$p_\star := \dim(U_\star) \le 3r.$$

We write $P_\star$ and $P_\star^\perp$ for the orthogonal projectors onto $U_\star$ and $U_\star^\perp$, respectively.

Definition D.5 (Teacher sigma-field)**.** Define the sigma-field generated by the teacher-relevant content coordinates

$$\mathcal{F}_\star := \sigma\big(P_\star x_t : t \in [L]\big). \tag{41}$$

The key point is that the teacher depends only on the $\mathcal{F}_\star$ part of the input.

Lemma D.6 (Teacher measurability through a low-dimensional content subspace)**.** *Under Theorem B.5, the vector of teacher statistics*

$$S^\star(X) := \big(S_1^\star(X), \dots, S_r^\star(X)\big)$$

is $\mathcal{F}_\star$-measurable. Consequently, the Bayes rule

$$f^\star(X) = \sum_{a=1}^r \beta_a^\star S_a^\star(X)$$

and the response $Y = f^\star(X) + \varepsilon$ are measurable with respect to $\mathcal{F}_\star \vee \sigma(\varepsilon)$.

Proof. Fix $a \in [r]$ and $t \in [L]$. Since $\Pi_{\mathrm{cnt}} q_a^\star, \Pi_{\mathrm{cnt}} k_a^\star, \Pi_{\mathrm{cnt}} v_a^\star \in U_\star$, we have

$$\Pi_{\mathrm{cnt}} q_a^\star = P_\star \Pi_{\mathrm{cnt}} q_a^\star, \qquad \Pi_{\mathrm{cnt}} k_a^\star = P_\star \Pi_{\mathrm{cnt}} k_a^\star, \qquad \Pi_{\mathrm{cnt}} v_a^\star = P_\star \Pi_{\mathrm{cnt}} v_a^\star.$$

Hence, by Equation (39),

$$\begin{aligned}
\langle q_a^\star, z_L(X)\rangle &= \langle \Pi_{\mathrm{cnt}} q_a^\star, x_L\rangle + \langle \Pi_{\mathrm{pos}} q_a^\star, e_L\rangle \\
&= \langle \Pi_{\mathrm{cnt}} q_a^\star, P_\star x_L\rangle + \langle \Pi_{\mathrm{pos}} q_a^\star, e_L\rangle, && (42)\\
\langle k_a^\star, z_t(X)\rangle &= \langle \Pi_{\mathrm{cnt}} k_a^\star, P_\star x_t\rangle + \langle \Pi_{\mathrm{pos}} k_a^\star, e_t\rangle, && (43)\\
\langle v_a^\star, z_t(X)\rangle &= \langle \Pi_{\mathrm{cnt}} v_a^\star, P_\star x_t\rangle + \langle \Pi_{\mathrm{pos}} v_a^\star, e_t\rangle. && (44)
\end{aligned}$$

Each right-hand side is $\mathcal{F}_\star$-measurable. Since ψ is Borel measurable, every summand in the definition

$$S_a^\star(X) = \sum_{t=1}^{L-1} \psi\big(\langle q_a^\star, z_L(X)\rangle \langle k_a^\star, z_t(X)\rangle\big)\, \langle v_a^\star, z_t(X)\rangle$$

is $\mathcal{F}_\star$-measurable, and therefore so is $S_a^\star(X)$. This proves the measurability of $S^\star(X)$. Since $f^\star(X) = \sum_{a=1}^r \beta_a^\star S_a^\star(X)$, the Bayes rule is also $\mathcal{F}_\star$-measurable. Finally, $Y = f^\star(X) + \varepsilon$ is measurable with respect to $\mathcal{F}_\star \vee \sigma(\varepsilon)$. □

D.3. Position-Selective Causal Heads

The next lemma is purely constructive. It shows that, because positions are hard-coded into the augmented tokens and because $\psi(0) = 0$, a single scalar causal head can isolate one chosen position and return an arbitrary linear functional of its content coordinates.

Lemma D.7 (Exact position-selective scalar head)**.** *Fix $s \in [L-1]$ and $u \in \mathbb{R}^d$. Define*

$$q_{\mathrm{gate}}^{(s)} := (0, e_L) \in \mathbb{R}^{d+L}, \qquad k_{\mathrm{gate}}^{(s)} := (0, \rho_\psi e_s) \in \mathbb{R}^{d+L}, \qquad v^{(u)} := (u, 0) \in \mathbb{R}^{d+L}. \tag{45}$$

Then the scalar causal head

$$A^{(s,u)}(X) := \sum_{t=1}^{L-1} \psi\big(\langle q_{\text{gate}}^{(s)}, z_L(X)\rangle \langle k_{\text{gate}}^{(s)}, z_t(X)\rangle\big) \, \langle v^{(u)}, z_t(X)\rangle$$

satisfies, for every $X \in \mathcal{X}_d$,

$$A^{(s,u)}(X) = \gamma_\psi \, \langle u, x_s\rangle. \tag{46}$$

Moreover,

$$\|q_{\text{gate}}^{(s)}\|_2 = 1, \qquad \|k_{\text{gate}}^{(s)}\|_2 = |\rho_\psi|, \qquad \|v^{(u)}\|_2 = \|u\|_2.$$

Proof. For every $X \in \mathcal{X}_d$,

$$\langle q_{\text{gate}}^{(s)}, z_L(X)\rangle = \langle (0, e_L), (x_L, e_L)\rangle = 1.$$

Likewise, for every $t \in [L-1]$,

$$\langle k_{\text{gate}}^{(s)}, z_t(X)\rangle = \langle (0, \rho_\psi e_s), (x_t, e_t)\rangle = \rho_\psi \, \mathbf{1}_{\{t=s\}},$$

and

$$\langle v^{(u)}, z_t(X)\rangle = \langle (u, 0), (x_t, e_t)\rangle = \langle u, x_t\rangle.$$

Therefore,

$$\begin{aligned} A^{(s,u)}(X) &= \sum_{t=1}^{L-1} \psi\big(\rho_\psi \, \mathbf{1}_{\{t=s\}}\big) \, \langle u, x_t\rangle \\ &= \psi(\rho_\psi) \, \langle u, x_s\rangle + \sum_{t \neq s} \psi(0) \, \langle u, x_t\rangle. \end{aligned}$$

By Theorem B.4, $\psi(0) = 0$, and by Theorem D.2, $\psi(\rho_\psi) = \gamma_\psi \neq 0$. Hence

$$A^{(s,u)}(X) = \gamma_\psi \, \langle u, x_s\rangle,$$

which proves Equation (46). The norm identities are immediate from Equation (45). □

D.4. Gaussian Orthogonal Complement Coordinates

The next lemma turns the orthogonal complement of the teacher content subspace into a genuine residual block.

Lemma D.8 (Gaussian residual coordinates are independent of the teacher sigma-field)**.** *Assume Theorem D.1. Let $u_1, \dots, u_{d_\perp} \in U_\star^\perp$ be orthonormal, where $1 \le d_\perp \le d - p_\star$, and define*

$$G^\perp(X) := \big(\langle u_1, x_1\rangle, \dots, \langle u_{d_\perp}, x_1\rangle\big) \in \mathbb{R}^{d_\perp}. \tag{47}$$

Then:

1. *$G^\perp(X) \sim \mathcal{N}(0, I_{d_\perp})$.*
2. *$G^\perp(X)$ is independent of $\mathcal{F}_\star$.*
3. *Consequently, $G^\perp(X)$ is independent of $S^\star(X)$, $f^\star(X)$, and Y.*

Proof. Since $x_1 \sim \mathcal{N}(0, I_d)$ and $u_1, \dots, u_{d_\perp}$ are orthonormal, the random vector $G^\perp(X)$ is centered Gaussian with covariance matrix

$$\mathrm{Cov}\big(G^\perp(X)\big) = \big(\langle u_i, u_j\rangle\big)_{i,j=1}^{d_\perp} = I_{d_\perp},$$

so $G^\perp(X) \sim \mathcal{N}(0, I_{d_\perp})$.

To prove independence from $\mathcal{F}_\star$, decompose each token as

$$x_t = P_\star x_t + P_\star^\perp x_t.$$

Under Theorem D.1, x_t is Gaussian and $P_\star x_t$ is orthogonal to $P_\star^\perp x_t$. Therefore $P_\star x_t$ and $P_\star^\perp x_t$ are independent Gaussian vectors for every t. Since the tokens are also independent across t, the entire family

$$\{P_\star x_t : t \in [L]\} \quad \text{is independent of} \quad \{P_\star^\perp x_t : t \in [L]\}.$$

Because each $u_j \in U_\star^\perp$, we have

$$\langle u_j, x_1 \rangle = \langle u_j, P_\star^\perp x_1 \rangle,$$

so $G^\perp(X)$ is measurable with respect to $\sigma(P_\star^\perp x_1)$, whereas $\mathcal{F}_\star = \sigma(P_\star x_t : t \in [L])$ is measurable with respect to the teacher projection family. Hence $G^\perp(X) \perp\!\!\!\perp \mathcal{F}_\star$.

Finally, Theorem D.6 implies that $S^\star(X)$ and $f^\star(X)$ are $\mathcal{F}_\star$-measurable, so they are independent of $G^\perp(X)$. Moreover, ε is independent of X by Theorem B.5, hence independent of $G^\perp(X)$. Since

$$Y = f^\star(X) + \varepsilon$$

is measurable with respect to $\mathcal{F}_\star \vee \sigma(\varepsilon)$, it follows that $G^\perp(X)$ is also independent of Y. □

D.5. The Exact Rule–Residual Factorization

We can now state the main representation theorem of this appendix. The theorem is existential and constructive: it builds a specific causal decoder backbone whose attention representation splits into a rule block and a Gaussian residual block.

Theorem D.9 (Rule–residual factorization of a trainable-feature causal decoder)**.** *Assume Theorems D.1 and D.2 and let $U_\star$ be the teacher content subspace from Theorem D.4. Fix an integer $d_\perp$ satisfying*

$$1 \le d_\perp \le d - p_\star,$$

and suppose the number of heads satisfies

$$H \ge r + d_\perp.$$

Choose any orthonormal family

$$u_1, \dots, u_{d_\perp} \in U_\star^\perp.$$

Then there exists a tuple of head parameters

$$\bar{\Theta}^{\text{split}} = (\theta_1^{\text{split}}, \dots, \theta_H^{\text{split}})$$

with the following properties:

(i) *For every $X \in \mathcal{X}_d$, the attention representation equals*

$$A_{\bar{\Theta}^{\text{split}}}(X) = \Big(S_1^\star(X), \dots, S_r^\star(X), \gamma_\psi \langle u_1, x_1 \rangle, \dots, \gamma_\psi \langle u_{d_\perp}, x_1 \rangle, 0, \dots, 0\Big). \tag{48}$$

(ii) *Defining*

$$Z^\perp(X) := \big(\gamma_\psi \langle u_1, x_1 \rangle, \dots, \gamma_\psi \langle u_{d_\perp}, x_1 \rangle\big) \in \mathbb{R}^{d_\perp}, \tag{49}$$

we have

$$Z^\perp(X) \sim \mathcal{N}(0, \gamma_\psi^2 I_{d_\perp}) \quad \textit{and} \quad Z^\perp(X) \perp\!\!\!\perp \mathcal{F}_\star.$$

In particular,

$$Z^\perp(X) \perp\!\!\!\perp S^\star(X), \qquad Z^\perp(X) \perp\!\!\!\perp f^\star(X), \qquad Z^\perp(X) \perp\!\!\!\perp Y.$$

(iii) *There exists a linear readout vector*

$$w^\star = (\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0) \in \mathbb{R}^H$$

such that, with zero bias and all MLP coefficients set to zero, the resulting predictor satisfies

$$f_{\Theta^{\text{rule}}}(X) = f^\star(X) \qquad \textit{for every } X \in \mathcal{X}_d. \tag{50}$$

Equivalently, the Bayes rule is represented exactly using only the first r coordinates of Equation (48).

(iv) *Let*

$$B_{\text{split}} := \max\big\{B_\star,\ 1,\ |\rho_\psi|,\ \|\beta^\star\|_2\big\}, \qquad \beta^\star := (\beta_1^\star, \dots, \beta_r^\star). \tag{51}$$

Then the predictor in (iii) *belongs to $\mathcal{T}_{\text{dec}}(H, W_f, B_{\text{split}})$ for every $W_f \ge 0$.*

Proof. We construct the heads explicitly.

Step 1: Rule heads. For $a \in [r]$, set

$$\theta_a^{\text{split}} := (q_a^\star, k_a^\star, v_a^\star).$$

By Equation (16) and Equation (13), the corresponding head output is exactly

$$A_a(X; \theta_a^{\text{split}}) = S_a^\star(X).$$

Step 2: Residual heads. For $j \in [d_\perp]$, define the $(r+j)$-th head by

$$\theta_{r+j}^{\text{split}} := \big(q_{\text{gate}}^{(1)}, k_{\text{gate}}^{(1)}, v^{(u_j)}\big),$$

where

$$q_{\text{gate}}^{(1)} = (0, e_L), \qquad k_{\text{gate}}^{(1)} = (0, \rho_\psi e_1), \qquad v^{(u_j)} = (u_j, 0).$$

By Theorem D.7,

$$A_{r+j}(X; \theta_{r+j}^{\text{split}}) = \gamma_\psi \langle u_j, x_1 \rangle.$$

Step 3: Remaining heads. For all $h > r + d_\perp$, set

$$\theta_h^{\text{split}} := (0, 0, 0).$$

Since $\psi(0) = 0$, these heads are identically zero.

Combining the three steps yields the exact representation Equation (48), proving (i).

Assertion (ii) follows from Theorem D.8 applied to the orthonormal family $u_1, \ldots, u_{d_\perp}$: the unscaled vector

$$\big(\langle u_1, x_1 \rangle, \ldots, \langle u_{d_\perp}, x_1 \rangle\big)$$

is standard Gaussian and independent of $\mathcal{F}_\star$, hence its scalar multiple $Z^\perp(X)$ is Gaussian with covariance $\gamma_\psi^2 I_{d_\perp}$ and remains independent of $\mathcal{F}_\star$. The displayed independence relations from $S^\star(X)$, $f^\star(X)$, and Y then follow from Theorem D.6 and the proof of Theorem D.8.

For (iii), define

$$w^\star = (\beta_1^\star, \ldots, \beta_r^\star, 0, \ldots, 0) \in \mathbb{R}^H, \qquad c = 0, \qquad \alpha_j = 0 \ \ \forall j \in [W_f].$$

Then, by Equation (18) and Equation (48),

$$\begin{aligned} f_{\Theta^{\text{rule}}}(X) &= \langle w^\star, A_{\bar{\Theta}^{\text{split}}}(X) \rangle \\ &= \sum_{a=1}^{r} \beta_a^\star S_a^\star(X) \\ &= f^\star(X), \end{aligned}$$

which proves Equation (50).

Finally, (iv) follows from the explicit construction. The rule heads satisfy

$$\max_{a \in [r]} \big\{ \|q_a^\star\|_2, \|k_a^\star\|_2, \|v_a^\star\|_2 \big\} \le B_\star$$

by Theorem B.5. The residual heads satisfy

$$\|q_{\text{gate}}^{(1)}\|_2 = 1, \qquad \|k_{\text{gate}}^{(1)}\|_2 = |\rho_\psi|, \qquad \|v^{(u_j)}\|_2 = \|u_j\|_2 = 1,$$

and the unused heads are zero. The readout norm is

$$\|w^\star\|_2 = \|\beta^\star\|_2,$$

while all MLP parameters are zero. Hence the full predictor belongs to $\mathcal{T}_{\text{dec}}(H, W_f, B_{\text{split}})$. □

Remark D.10 (What the theorem actually proves). Theorem D.9 is not a statement about a particular training algorithm discovering the split representation. It is a representation theorem for the trainable-feature decoder class itself. The key structural consequence is that the same causal decoder family contains both an exact Bayes-rule backbone and a statistically orthogonal residual block that is invisible to the teacher sigma-field.

D.6. Consequence for Generic Fact Sets

The next corollary is the precise probabilistic input needed in the sparse fact-localization appendix.

Corollary D.11 (Residual codes of generic fact sets)**.** *Assume the hypotheses of Theorem D.9. Let*

$$F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$$

be a generic fact set in the sense of Theorem B.9, and define

$$S_\mu^\star := S^\star(X^\mu), \qquad Z_\mu^\perp := Z^\perp(X^\mu), \qquad \delta_\mu := \xi^\mu - f^\star(X^\mu).$$

Then:

1. $(Z_\mu^\perp)_{\mu=1}^m$ *are i.i.d. Gaussian vectors with law* $\mathcal{N}(0, \gamma_\psi^2 I_{d_\perp})$.
2. $(Z_\mu^\perp)_{\mu=1}^m$ *are independent of* $(S_\mu^\star, Y^\mu)_{\mu=1}^m$.
3. *If the residual vector* $(\delta_\mu)_{\mu=1}^m$ *is deterministic, or more generally independent of the fact locations* $(X^\mu)_{\mu=1}^m$*, then* $(Z_\mu^\perp)_{\mu=1}^m$ *are independent of* $(S_\mu^\star, \delta_\mu)_{\mu=1}^m$.
4. *Conditional on* $(S_\mu^\star, \delta_\mu)_{\mu=1}^m$*, the vectors* $(Z_\mu^\perp)_{\mu=1}^m$ *remain i.i.d. with law* $\mathcal{N}(0, \gamma_\psi^2 I_{d_\perp})$.

Proof. By Theorem B.9, the locations $X^1, \dots, X^m$ are i.i.d. draws from P_X. Therefore the pairs

$$\big(S^\star(X^\mu), Z^\perp(X^\mu), Y^\mu\big), \qquad \mu \in [m],$$

are i.i.d. copies of $\big(S^\star(X), Z^\perp(X), Y\big)$. By Theorem D.9(ii), each $Z^\perp(X^\mu)$ has law $\mathcal{N}(0, \gamma_\psi^2 I_{d_\perp})$ and is independent of $\big(S^\star(X^\mu), Y^\mu\big)$. This proves (1) and (2).

If the residual vector $(\delta_\mu)_{\mu=1}^m$ is deterministic, or more generally independent of the locations $(X^\mu)_{\mu=1}^m$, then it is independent of $(Z_\mu^\perp, S_\mu^\star)_{\mu=1}^m$, which proves (3). The conditional statement (4) is then immediate from (1)–(3): conditioning on variables independent of the Gaussian residual codes does not alter their law or their independence structure. □

Remark D.12 (Why one token suffices for the residual block). The residual block in Theorem D.9 is extracted from the orthogonal complement of a single token, x_1. This is enough for the later localization argument because, for generic fact sets, the resulting residual codes are already i.i.d. continuous Gaussian vectors in dimension $d_\perp$. That is precisely the regime in which localized fact bumps can be constructed with exponentially small $L^2(P_X)$-mass.

E. Sparse Fact Localization Geometry

This appendix proves the geometric localization theorem that drives the main upper bound. By Theorem D.9, the trainable-feature decoder class contains an exact Bayes-rule backbone together with a residual Gaussian block $Z^\perp(X) \in \mathbb{R}^{d_\perp}$ that is independent of the teacher sigma-field. The goal of this appendix is to show that, for a generic fact set, one can implant an arbitrary bounded residual vector $\delta(F_m) = (\delta_1, \dots, \delta_m)$ by a residual-only one-hidden-layer ReLU correction whose L^2-mass under the rule distribution is exponentially small in $d_\perp$. In contrast to storage-capacity analyses based on associative-memory scaling (Nichani et al., 2024; Xu & Chen, 2025; Dong et al., 2025), the mechanism here is a high-threshold geometric localization in a residual Gaussian block.

Throughout this appendix, we assume all hypotheses of Theorem D.9. In particular:

- $d_\perp \geq 1$ is the residual dimension from Theorem D.9;
- $F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$ is a generic fact set;
- $Z_\mu^\perp := Z^\perp(X^\mu) \in \mathbb{R}^{d_\perp}$ are the residual codes from Theorem D.11;
- $\delta_\mu := \xi^\mu - f^\star(X^\mu)$ are the fact residuals.

We abbreviate

$$\bar{\gamma} := |\gamma_\psi| > 0, \qquad \nu_\perp := \mathcal{N}(0, \bar{\gamma}^2 I_{d_\perp}),$$

so that, by Theorems D.9 and D.11,

$$Z^\perp(X) \sim \nu_\perp \qquad \text{under } P_X,$$

and the random codes $Z_1^\perp, \dots, Z_m^\perp$ are i.i.d. with the same law.

E.1. A Three-ReLU Tent Bump

The basic one-fact localizer is a one-dimensional tent function, realized by three ReLU units. Its only role is to create a compact slab around a chosen high-threshold projection level.

Definition E.1 (Tent bump)**.** For $s \in \mathbb{R}$ and $w > 0$, define

$$\vartheta_{s,w}(t) := \frac{1}{w}\Big(\sigma(t-s+w) - 2\sigma(t-s) + \sigma(t-s-w)\Big), \qquad t \in \mathbb{R}, \tag{52}$$

where $\sigma(u) = u_+ = \max\{u, 0\}$.

Lemma E.2 (Elementary properties of the tent bump)**.** *For every $s \in \mathbb{R}$ and $w > 0$, the function $\vartheta_{s,w} : \mathbb{R} \to \mathbb{R}$ satisfies:*

1. $0 \le \vartheta_{s,w}(t) \le 1$ *for all* $t \in \mathbb{R}$*;*
2. $\vartheta_{s,w}(s) = 1$*;*
3. $\vartheta_{s,w}(t) = 0$ *whenever* $|t - s| \ge w$*;*
4. $\vartheta_{s,w}$ *is piecewise linear, supported on* $[s-w, s+w]$*, and admits the explicit form*

$$\vartheta_{s,w}(t) = \begin{cases} 0, & t \le s-w, \\ \dfrac{t-(s-w)}{w}, & s-w \le t \le s, \\ \dfrac{(s+w)-t}{w}, & s \le t \le s+w, \\ 0, & t \ge s+w. \end{cases}$$

Proof. Fix $s \in \mathbb{R}$ and $w > 0$. The formula follows by checking the four regions $t \le s-w$, $s-w \le t \le s$, $s \le t \le s+w$, and $t \ge s+w$ directly from Equation (52). In particular, the function is supported on $[s-w, s+w]$, equals 1 at $t = s$, and is affine on each subinterval. Since it linearly interpolates between 0 and 1 on $[s-w, s]$ and between 1 and 0 on $[s, s+w]$, it takes values in $[0, 1]$. □

E.2. A High-Probability Geometric Separation Event

The key geometric fact is that independent Gaussian residual codes are large in norm and almost orthogonal to one another at the scale relevant for high-threshold projection tests.

Definition E.3 (Geometric separation event)**.** Define the event $\mathcal{E}_{\text{geo}}$ by

$$\begin{aligned} \mathcal{E}_{\text{geo}} := \Big\{ & \tfrac{3}{4}\bar{\gamma}\sqrt{d_\perp} \le \|Z_\mu^\perp\|_2 \le \tfrac{5}{4}\bar{\gamma}\sqrt{d_\perp} \quad \forall \mu \in [m], \\ & \big|\langle \widehat{Z}_\mu^\perp, Z_\nu^\perp \rangle\big| \le \tfrac{1}{8}\bar{\gamma}\sqrt{d_\perp} \quad \forall \mu \ne \nu \Big\}, \end{aligned} \tag{53}$$

where

$$\widehat{Z}_\mu^\perp := \frac{Z_\mu^\perp}{\|Z_\mu^\perp\|_2} \qquad \text{whenever } Z_\mu^\perp \ne 0.$$

Lemma E.4 (High-probability separation of Gaussian residual codes)**.** *There exists $d_0 \in \mathbb{N}$ such that for all $d_\perp \geq d_0$,*

$$\mathbb{P}\big(\mathcal{E}_{\text{geo}}\big) \geq 1 - 2m\, e^{-d_\perp/128} - 2m(m-1)\, e^{-d_\perp/128}. \tag{54}$$

In particular, if $\log m = o(d_\perp)$, then

$$\mathbb{P}\big(\mathcal{E}_{\text{geo}}\big) \to 1.$$

Moreover, on $\mathcal{E}_{\text{geo}}$, for every $\mu \neq \nu$,

$$\big|\langle \widehat{Z}_\mu^\perp, \widehat{Z}_\nu^\perp \rangle\big| \leq \frac{1}{6}. \tag{55}$$

Proof. By Theorem D.11, the vectors $Z_1^\perp, \dots, Z_m^\perp$ are i.i.d. with law $\mathcal{N}(0, \bar{\gamma}^2 I_{d_\perp})$. Write

$$Z_\mu^\perp = \bar{\gamma} G_\mu, \qquad G_\mu \sim \mathcal{N}(0, I_{d_\perp})$$

independently.

Norm concentration. The Euclidean norm is 1-Lipschitz on $\mathbb{R}^{d_\perp}$, so by Gaussian concentration,

$$\mathbb{P}\Big(\big|\|G_\mu\|_2 - \mathbb{E}\|G_\mu\|_2\big| \geq t\Big) \leq 2e^{-t^2/2} \qquad \forall t > 0.$$

Since $\mathbb{E}\|G_\mu\|_2 \in [\sqrt{d_\perp} - 1, \sqrt{d_\perp}]$, there exists d_0 such that for all $d_\perp \geq d_0$,

$$\sqrt{d_\perp} - 1 - \frac{3}{4}\sqrt{d_\perp} \geq \frac{1}{8}\sqrt{d_\perp}.$$

Hence, for $d_\perp \geq d_0$,

$$\mathbb{P}\Big(\|G_\mu\|_2 \notin \big[\tfrac{3}{4}\sqrt{d_\perp}, \tfrac{5}{4}\sqrt{d_\perp}\big]\Big) \leq 2e^{-d_\perp/128}.$$

Rescaling by $\bar{\gamma}$ gives

$$\mathbb{P}\Big(\|Z_\mu^\perp\|_2 \notin \big[\tfrac{3}{4}\bar{\gamma}\sqrt{d_\perp}, \tfrac{5}{4}\bar{\gamma}\sqrt{d_\perp}\big]\Big) \leq 2e^{-d_\perp/128}.$$

A union bound over $\mu \in [m]$ yields the first failure term $2me^{-d_\perp/128}$.

Cross-projection control. Fix $\mu \neq \nu$. Conditional on $\widehat{Z}_\mu^\perp$, the scalar $\langle \widehat{Z}_\mu^\perp, Z_\nu^\perp \rangle$ is Gaussian with mean zero and variance $\bar{\gamma}^2$, because $Z_\nu^\perp$ is independent of $Z_\mu^\perp$ and isotropic. Therefore,

$$\mathbb{P}\Big(\big|\langle \widehat{Z}_\mu^\perp, Z_\nu^\perp \rangle\big| \geq \tfrac{1}{8}\bar{\gamma}\sqrt{d_\perp} \;\Big|\; \widehat{Z}_\mu^\perp\Big) \leq 2e^{-d_\perp/128}.$$

Removing the conditioning leaves the same bound. A union bound over all ordered pairs (μ, ν), $\mu \neq \nu$, gives the second failure term $2m(m-1)e^{-d_\perp/128}$.

Combining the two bounds proves Equation (54). Finally, on $\mathcal{E}_{\text{geo}}$,

$$\big|\langle \widehat{Z}_\mu^\perp, \widehat{Z}_\nu^\perp \rangle\big| = \frac{\big|\langle \widehat{Z}_\mu^\perp, Z_\nu^\perp \rangle\big|}{\|Z_\nu^\perp\|_2} \leq \frac{(\bar{\gamma}/8)\sqrt{d_\perp}}{(3\bar{\gamma}/4)\sqrt{d_\perp}} = \frac{1}{6},$$

which proves Equation (55). □

E.3. One-Fact Tent Localizers

We now define the bump attached to each fact code.

Definition E.5 (Residual tent localizer)**.** Fix the deterministic width

$$w_\perp := \frac{1}{8}\bar{\gamma}\sqrt{d_\perp}. \tag{56}$$

For each $\mu \in [m]$, define the residual tent localizer

$$b_\mu(z) := \vartheta_{\|Z_\mu^\perp\|_2,\, w_\perp}\big(\langle \widehat{Z}_\mu^\perp, z \rangle\big), \qquad z \in \mathbb{R}^{d_\perp}. \tag{57}$$

Lemma E.6 (Exact pointwise localization on the fact codes)**.** *On the event $\mathcal{E}_{\text{geo}}$, the localizers $\{b_\mu\}_{\mu=1}^m$ satisfy:*

1. $b_\mu(Z_\mu^\perp) = 1$ *for every* $\mu \in [m]$*;*
2. $b_\mu(Z_\nu^\perp) = 0$ *for every* $\mu \neq \nu$*;*
3. $0 \le b_\mu(z) \le 1$ *for all* $z \in \mathbb{R}^{d_\perp}$*.*

Consequently, for every coefficient vector $a = (a_1, \dots, a_m) \in \mathbb{R}^m$, the function

$$g_a(z) := \sum_{\mu=1}^m a_\mu b_\mu(z) \tag{58}$$

satisfies

$$g_a(Z_\nu^\perp) = a_\nu \qquad \forall \nu \in [m].$$

Proof. Fix $\mu \in [m]$. Since

$$\langle \widehat{Z}_\mu^\perp, Z_\mu^\perp \rangle = \|Z_\mu^\perp\|_2,$$

we have, by Theorem E.2,

$$b_\mu(Z_\mu^\perp) = \vartheta_{\|Z_\mu^\perp\|_2, w_\perp}(\|Z_\mu^\perp\|_2) = 1.$$

Now fix $\nu \neq \mu$. On $\mathcal{E}_{\text{geo}}$,

$$\left|\langle \widehat{Z}_\mu^\perp, Z_\nu^\perp \rangle\right| \le \frac{1}{8}\bar{\gamma}\sqrt{d_\perp} = w_\perp,$$

whereas

$$\|Z_\mu^\perp\|_2 - w_\perp \ge \frac{3}{4}\bar{\gamma}\sqrt{d_\perp} - \frac{1}{8}\bar{\gamma}\sqrt{d_\perp} = \frac{5}{8}\bar{\gamma}\sqrt{d_\perp}.$$

Therefore

$$\langle \widehat{Z}_\mu^\perp, Z_\nu^\perp \rangle \le w_\perp < \|Z_\mu^\perp\|_2 - w_\perp,$$

so $\langle \widehat{Z}_\mu^\perp, Z_\nu^\perp \rangle$ lies outside the support interval $[\|Z_\mu^\perp\|_2 - w_\perp,\ \|Z_\mu^\perp\|_2 + w_\perp]$ of the tent function. Hence $b_\mu(Z_\nu^\perp) = 0$.

The bound $0 \le b_\mu \le 1$ is immediate from Theorem E.2. The interpolation identity for g_a follows from

$$g_a(Z_\nu^\perp) = \sum_{\mu=1}^m a_\mu b_\mu(Z_\nu^\perp) = a_\nu \cdot 1 + \sum_{\mu \neq \nu} a_\mu \cdot 0 = a_\nu.$$

□

E.4. Population Mass of One-Fact and Two-Fact Overlaps

The next two lemmas quantify the $L^2(\nu_\perp)$-mass of a single localizer and the overlap of two distinct localizers.

Lemma E.7 (Single-localizer Gaussian mass)**.** *Let $Z \sim \nu_\perp = \mathcal{N}(0, \bar{\gamma}^2 I_{d_\perp})$, independent of the fact set. On $\mathcal{E}_{\text{geo}}$, for every $\mu \in [m]$,*

$$\mathbb{E}\big[b_\mu(Z)^2 \mid F_m\big] \le \exp\Big(-\frac{25}{128} d_\perp\Big). \tag{59}$$

Proof. Fix $\mu \in [m]$. By Theorem E.2, $0 \le b_\mu \le 1$ and the support of b_μ is contained in the slab

$$\left\{ z \in \mathbb{R}^{d_\perp} : \left|\langle \widehat{Z}_\mu^\perp, z \rangle - \|Z_\mu^\perp\|_2\right| \le w_\perp \right\}.$$

Hence, on $\mathcal{E}_{\text{geo}}$,

$$\begin{aligned}\mathbb{E}[b_\mu(Z)^2 \mid F_m] &\le \mathbb{P}\Big(\langle \widehat{Z}_\mu^\perp, Z \rangle \ge \|Z_\mu^\perp\|_2 - w_\perp \;\Big|\; F_m\Big) \\ &\le \mathbb{P}\Big(\langle \widehat{Z}_\mu^\perp, Z \rangle \ge \frac{5}{8}\bar{\gamma}\sqrt{d_\perp} \;\Big|\; F_m\Big),\end{aligned}$$

because $\|Z_\mu^\perp\|_2 - w_\perp \geq \frac{5}{8}\bar{\gamma}\sqrt{d_\perp}$ on $\mathcal{E}_{\text{geo}}$.

Conditional on F_m, the scalar $\langle \widehat{Z}_\mu^\perp, Z\rangle$ is Gaussian with law $\mathcal{N}(0, \bar{\gamma}^2)$. Therefore, by the standard Gaussian tail bound $\mathbb{P}(N(0,1) \geq t) \leq e^{-t^2/2}$,

$$\mathbb{E}[b_\mu(Z)^2 \mid F_m] \leq \exp\Big(-\frac{1}{2}\Big(\frac{5}{8}\sqrt{d_\perp}\Big)^2\Big) = \exp\Big(-\frac{25}{128}d_\perp\Big).$$

□

Lemma E.8 (Pairwise overlap bound)**.** *Let $Z \sim \nu_\perp = \mathcal{N}(0, \bar{\gamma}^2 I_{d_\perp})$, independent of the fact set. On $\mathcal{E}_{\text{geo}}$, for every $\mu \neq \nu$,*

$$\mathbb{E}\big[b_\mu(Z)b_\nu(Z) \mid F_m\big] \leq \exp\Big(-\frac{75}{224}d_\perp\Big). \tag{60}$$

Proof. Fix $\mu \neq \nu$. Define

$$U := \langle \widehat{Z}_\mu^\perp, Z\rangle, \qquad V := \langle \widehat{Z}_\nu^\perp, Z\rangle.$$

Conditional on F_m, the pair (U, V) is centered Gaussian with

$$\mathrm{Var}(U) = \mathrm{Var}(V) = \bar{\gamma}^2, \qquad \mathrm{Cov}(U, V) = \bar{\gamma}^2\langle \widehat{Z}_\mu^\perp, \widehat{Z}_\nu^\perp\rangle.$$

Hence, by Equation (55),

$$\mathrm{Var}(U+V) = 2\bar{\gamma}^2\Big(1 + \langle \widehat{Z}_\mu^\perp, \widehat{Z}_\nu^\perp\rangle\Big) \leq 2\bar{\gamma}^2\Big(1 + \frac{1}{6}\Big) = \frac{7}{3}\bar{\gamma}^2.$$

By support inclusion and the same lower-threshold bound used in the proof of Theorem E.7, on $\mathcal{E}_{\text{geo}}$,

$$b_\mu(Z)b_\nu(Z) \neq 0 \quad \Longrightarrow \quad U \geq \frac{5}{8}\bar{\gamma}\sqrt{d_\perp} \text{ and } V \geq \frac{5}{8}\bar{\gamma}\sqrt{d_\perp}.$$

Therefore,

$$b_\mu(Z)b_\nu(Z) \neq 0 \quad \Longrightarrow \quad U + V \geq \frac{5}{4}\bar{\gamma}\sqrt{d_\perp}.$$

Using $0 \leq b_\mu, b_\nu \leq 1$, we obtain

$$\mathbb{E}[b_\mu(Z)b_\nu(Z) \mid F_m] \leq \mathbb{P}\Big(U + V \geq \frac{5}{4}\bar{\gamma}\sqrt{d_\perp} \;\Big|\; F_m\Big).$$

Since $U + V$ is centered Gaussian with variance at most $\frac{7}{3}\bar{\gamma}^2$, the Gaussian tail bound yields

$$\begin{aligned}\mathbb{E}[b_\mu(Z)b_\nu(Z) \mid F_m] &\leq \exp\left(-\frac{\big(\frac{5}{4}\bar{\gamma}\sqrt{d_\perp}\big)^2}{2\cdot\frac{7}{3}\bar{\gamma}^2}\right)\\ &= \exp\left(-\frac{25}{16}\cdot\frac{3}{14}d_\perp\right) = \exp\Big(-\frac{75}{224}d_\perp\Big).\end{aligned}$$

□

E.5. The Sparse Fact-Bump Theorem

We now assemble the exact interpolation result.

Theorem E.9 (Sparse fact-bump theorem)**.** *Assume the hypotheses of Theorem D.9. Let $F_m = \{(X^\mu, \xi^\mu)\}_{\mu=1}^m$ be a generic fact set, let $\delta(F_m) = (\delta_1, \ldots, \delta_m)$ be its residual vector, and define*

$$\|\delta(F_m)\|_\infty := \max_{\mu\in[m]} |\delta_\mu|.$$

Then, on the event $\mathcal{E}_{\text{geo}}$, there exists a residual-only one-hidden-layer ReLU correction

$$g_F(z) := \sum_{\mu=1}^m \delta_\mu\, b_\mu(z), \qquad z \in \mathbb{R}^{d_\perp}, \tag{61}$$

with the following properties:

(i) *g_F exactly interpolates the fact residuals on the residual codes:*

$$g_F(Z_\mu^\perp) = \delta_\mu \qquad \forall \mu \in [m]. \tag{62}$$

(ii) *g_F is realized by $3m$ ReLU units:*

$$g_F(z) = \sum_{\mu=1}^{m} \frac{\delta_\mu}{w_\perp} \Big[\sigma\big(\langle \widehat{Z}_\mu^\perp, z\rangle - (\|Z_\mu^\perp\|_2 - w_\perp)\big) - 2\sigma\big(\langle \widehat{Z}_\mu^\perp, z\rangle - \|Z_\mu^\perp\|_2\big) + \sigma\big(\langle \widehat{Z}_\mu^\perp, z\rangle - (\|Z_\mu^\perp\|_2 + w_\perp)\big)\Big]. \tag{63}$$

(iii) *If $Z \sim \nu_\perp = \mathcal{N}(0, \bar{\gamma}^2 I_{d_\perp})$ is independent of F_m, then*

$$\mathbb{E}\big[g_F(Z)^2 \mid F_m\big] \le \|\delta(F_m)\|_\infty^2 \Big[m\, e^{-25d_\perp/128} + m(m-1)\, e^{-75d_\perp/224}\Big]. \tag{64}$$

Proof. Define g_F as in Equation (61). Property (i) follows immediately from Theorem E.6: for each $\nu \in [m]$,

$$g_F(Z_\nu^\perp) = \sum_{\mu=1}^{m} \delta_\mu\, b_\mu(Z_\nu^\perp) = \delta_\nu.$$

Property (ii) follows directly from the definition of b_μ via the tent formula Equation (52). Each b_μ is a linear combination of exactly three ReLU units sharing the same input weight vector $\widehat{Z}_\mu^\perp$, with thresholds $\|Z_\mu^\perp\|_2 - w_\perp$, $\|Z_\mu^\perp\|_2$, and $\|Z_\mu^\perp\|_2 + w_\perp$. Summing over $\mu \in [m]$ yields Equation (63), which uses $3m$ units.

For (iii), expand the square:

$$\mathbb{E}[g_F(Z)^2 \mid F_m] = \sum_{\mu=1}^{m} \delta_\mu^2 \mathbb{E}[b_\mu(Z)^2 \mid F_m] + 2 \sum_{1\le\mu<\nu\le m} \delta_\mu \delta_\nu\, \mathbb{E}[b_\mu(Z) b_\nu(Z) \mid F_m].$$

Using $|\delta_\mu| \le \|\delta(F_m)\|_\infty$ for every μ, together with Theorems E.7 and E.8, we obtain

$$\begin{aligned}\mathbb{E}[g_F(Z)^2 \mid F_m] &\le \|\delta(F_m)\|_\infty^2 \sum_{\mu=1}^{m} \mathbb{E}[b_\mu(Z)^2 \mid F_m] \\ &\quad + 2\|\delta(F_m)\|_\infty^2 \sum_{1\le\mu<\nu\le m} \mathbb{E}[b_\mu(Z) b_\nu(Z) \mid F_m] \\ &\le \|\delta(F_m)\|_\infty^2 \Big[m\, e^{-25d_\perp/128} + m(m-1)\, e^{-75d_\perp/224}\Big].\end{aligned}$$

This proves Equation (64). □

Corollary E.10 (Simplified asymptotic form)**.** *Assume the hypotheses of Theorem E.9. If, in addition,*

$$m \le \exp(d_\perp/256), \tag{65}$$

then for all sufficiently large $d_\perp$, on $\mathcal{E}_{\text{geo}}$,

$$\mathbb{E}\big[g_F(Z)^2 \mid F_m\big] \le 2\, \|\delta(F_m)\|_\infty^2\, m\, e^{-25d_\perp/128}. \tag{66}$$

Consequently, if $me^{-25d_\perp/128} \to 0$, then

$$\mathbb{E}\big[g_F(Z)^2 \mid F_m\big] \to 0 \qquad \text{on } \mathcal{E}_{\text{geo}}.$$

Proof. By Equation (65),

$$m(m-1)e^{-75d_\perp/224} \le m^2 e^{-75d_\perp/224} = me^{-25d_\perp/128} \cdot \Big(me^{-(75/224-25/128)d_\perp}\Big).$$

Now

$$\frac{75}{224} - \frac{25}{128} = \frac{125}{896} > \frac{1}{256},$$

so Equation (65) implies

$$me^{-(75/224-25/128)d_\perp} \le e^{-d_\perp/256} \to 0.$$

Thus, for all sufficiently large $d_\perp$,

$$m(m-1)e^{-75d_\perp/224} \le me^{-25d_\perp/128}.$$

Substituting this into Equation (64) yields Equation (66). The final statement is immediate. □

Remark E.11 (Lift to the full decoder architecture). The function g_F acts only on the residual coordinates $z \in \mathbb{R}^{d_\perp}$. In the full decoder architecture from Sections B and D, these coordinates occupy a designated block of the attention representation

$$A_{\bar{\Theta}^{\text{split}}}(X) = \big(S^\star(X),\, Z^\perp(X),\, 0\big).$$

Hence each residual tent localizer b_μ can be implemented by three hidden ReLU units whose input weight vectors are supported only on the residual block, with zeros on the rule coordinates and on the unused heads. The resulting correction therefore perturbs only the residual subspace and leaves the Bayes rule block untouched. This structural decoupling is precisely what will be combined with Theorem C.5 in the next appendix.

Remark E.12 (Interpretation). The key point is not merely that finitely many facts can be interpolated. In unstructured measurable classes this would already be trivial by Theorem C.9. What is nontrivial here is that the interpolation is achieved by a *structured residual-only ReLU correction* whose population mass under the rule law decays exponentially in the residual dimension. This is the geometric mechanism behind Bayes-optimal coexistence in the trainable-feature decoder class.

F. Proof of the Main Upper Theorem

This appendix assembles the representation and localization ingredients from Appendices D–E into the main positive coexistence theorem. The proof has a simple logical form:

$$\text{rule–residual factorization} + \text{sparse residual fact bumps}$$
$$+\text{coexistence–localizability identity} \Longrightarrow \text{vanishing Bayes-coexistence gap.}$$

The result is deliberately representation-theoretic: it proves that a trainable-feature causal decoder class contains predictors that simultaneously realize the Bayes rule and exactly memorize arbitrary bounded facts, while incurring only exponentially small excess rule risk. This is the positive half of the separation from lazy/kernelized decoders established later in Appendix G.

Throughout this appendix, all notation from Appendices B–E remains in force. In particular, $\bar{\gamma} = |\gamma_\psi| > 0$,

$$w_\perp = \frac{1}{8}\bar{\gamma}\sqrt{d_\perp},$$

and $\mathcal{E}_{\text{geo}}$ is the high-probability geometric event from Theorem E.3. The fact residual vector is denoted by

$$\delta(F_m) = (\delta_1, \dots, \delta_m), \qquad \delta_\mu = \xi^\mu - f^\star(X^\mu).$$

F.1. Budget Needed to Realize the Fact Bumps

The residual tent construction in Appendix E uses thresholds of order $\sqrt{d_\perp}$. We therefore first record an explicit deterministic parameter budget sufficient to implement the construction inside the decoder class from Theorem B.11.

Definition F.1 (Upper-theorem parameter budget)**.** Fix an admissibility level $B_\delta > 0$. Recall B_{split} from Equation (51). Define

$$B_{\text{bump}}(d_\perp, B_\delta) := \max\left\{1,\, \frac{2B_\delta}{w_\perp},\, \frac{11}{8}\bar{\gamma}\sqrt{d_\perp}\right\}, \tag{67}$$

and

$$B_{\text{upper}}(d_\perp, B_\delta) := \max\left\{B_{\text{split}},\, B_{\text{bump}}(d_\perp, B_\delta)\right\}. \tag{68}$$

Remark F.2 (Polynomial budget). If $B_\delta = O(1)$, then

$$B_{\text{upper}}(d_\perp, B_\delta) = O(\sqrt{d_\perp}),$$

which is compatible with the polynomial-budget convention in Theorem B.3.

F.2. Lifting Residual Bumps into the Full Decoder

We now show that the residual correction g_F from Appendix E is not merely an abstract function of $Z^\perp(X)$, but is exactly representable by the MLP block of the trainable-feature decoder, using the split attention representation from Appendix D.

Lemma F.3 (Decoder realization of sparse residual bumps)**.** *Assume the hypotheses of Theorem D.9. Let F_m be a B_δ-admissible generic fact set and suppose*

$$H \geq r + d_\perp, \qquad W_f \geq 3m, \qquad B \geq B_{\mathrm{upper}}(d_\perp, B_\delta).$$

Then, on $\mathcal{E}_{\mathrm{geo}}$, there exists a parameter vector Θ_F such that

$$f_{\Theta_F} \in \mathcal{T}_{\mathrm{dec}}(H, W_f, B)$$

and, for every $X \in \mathcal{X}_d$,

$$f_{\Theta_F}(X) = f^\star(X) + g_F(Z^\perp(X)), \tag{69}$$

where g_F is the residual correction from Equation (61). In particular,

$$f_{\Theta_F}(X^\mu) = \xi^\mu \qquad \forall \mu \in [m]. \tag{70}$$

Proof. Fix a fact set F_m and work on the event $\mathcal{E}_{\mathrm{geo}}$.

Step 1: Attention backbone. Use the split attention heads from Theorem D.9. Thus

$$A_{\bar{\Theta}^{\mathrm{split}}}(X) = \Big(S_1^\star(X), \dots, S_r^\star(X), Z_1^\perp(X), \dots, Z_{d_\perp}^\perp(X), 0, \dots, 0\Big),$$

where

$$Z^\perp(X) = \big(Z_1^\perp(X), \dots, Z_{d_\perp}^\perp(X)\big) \in \mathbb{R}^{d_\perp}.$$

Set the linear readout to

$$w^\star = (\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0) \in \mathbb{R}^H$$

and set the scalar bias $c = 0$. Then, by Theorem D.9,

$$\langle w^\star, A_{\bar{\Theta}^{\mathrm{split}}}(X)\rangle = f^\star(X) \qquad \forall X \in \mathcal{X}_d.$$

Step 2: MLP implementation of one tent bump. For each fact $\mu \in [m]$, define the vector in the full attention-coordinate space

$$\widetilde{z}_\mu := \big(0_r,\ \widehat{Z}_\mu^\perp,\ 0_{H-r-d_\perp}\big) \in \mathbb{R}^H.$$

Since $\|\widehat{Z}_\mu^\perp\|_2 = 1$, we have

$$\|\widetilde{z}_\mu\|_2 = 1.$$

Moreover,

$$\langle \widetilde{z}_\mu, A_{\bar{\Theta}^{\mathrm{split}}}(X)\rangle = \langle \widehat{Z}_\mu^\perp, Z^\perp(X)\rangle.$$

For each $\mu \in [m]$, allocate three hidden ReLU units with common input direction $\widetilde{z}_\mu$, output weights

$$\alpha_{\mu,1} := \frac{\delta_\mu}{w_\perp}, \qquad \alpha_{\mu,2} := -\frac{2\delta_\mu}{w_\perp}, \qquad \alpha_{\mu,3} := \frac{\delta_\mu}{w_\perp},$$

and thresholds

$$\tau_{\mu,1} := \|Z_\mu^\perp\|_2 - w_\perp, \qquad \tau_{\mu,2} := \|Z_\mu^\perp\|_2, \qquad \tau_{\mu,3} := \|Z_\mu^\perp\|_2 + w_\perp. \tag{71}$$

All unused hidden units, if any, are assigned output coefficient zero.

The contribution of the three units attached to fact μ is

$$\begin{aligned}\frac{\delta_\mu}{w_\perp}\Big[&\sigma\big(\langle \widehat{Z}_\mu^\perp, Z^\perp(X)\rangle - (\|Z_\mu^\perp\|_2 - w_\perp)\big)\\ &- 2\sigma\big(\langle \widehat{Z}_\mu^\perp, Z^\perp(X)\rangle - \|Z_\mu^\perp\|_2\big)\\ &+ \sigma\big(\langle \widehat{Z}_\mu^\perp, Z^\perp(X)\rangle - (\|Z_\mu^\perp\|_2 + w_\perp)\big)\Big]\\ &= \delta_\mu\, \vartheta_{\|Z_\mu^\perp\|_2, w_\perp}\big(\langle \widehat{Z}_\mu^\perp, Z^\perp(X)\rangle\big) = \delta_\mu b_\mu(Z^\perp(X)).\end{aligned}$$

Summing over $\mu \in [m]$, the MLP block realizes

$$g_F(Z^\perp(X)) = \sum_{\mu=1}^{m} \delta_\mu b_\mu(Z^\perp(X)).$$

Together with the linear readout, this proves Equation (69).

Step 3: Exact interpolation of facts. By Theorem E.9,

$$g_F(Z_\mu^\perp) = \delta_\mu \qquad \forall \mu \in [m].$$

Therefore,

$$f_{\Theta_F}(X^\mu) = f^\star(X^\mu) + g_F(Z^\perp(X^\mu)) = f^\star(X^\mu) + \delta_\mu = \xi^\mu,$$

which proves Equation (70).

Step 4: Parameter budget. The split attention heads and the Bayes linear readout are bounded by B_{split} by Theorem D.9. Each MLP input vector $\widetilde{z}_\mu$ has norm 1. Since F_m is B_δ-admissible,

$$|\alpha_{\mu,1}|,\ |\alpha_{\mu,3}| \le \frac{B_\delta}{w_\perp}, \qquad |\alpha_{\mu,2}| \le \frac{2B_\delta}{w_\perp}.$$

On $\mathcal{E}_{\text{geo}}$,

$$\|Z_\mu^\perp\|_2 \le \frac{5}{4}\bar{\gamma}\sqrt{d_\perp},$$

and hence

$$|\tau_{\mu,k}| \le \frac{5}{4}\bar{\gamma}\sqrt{d_\perp} + \frac{1}{8}\bar{\gamma}\sqrt{d_\perp} = \frac{11}{8}\bar{\gamma}\sqrt{d_\perp} \qquad \text{for } k = 1, 2, 3.$$

Thus every attention parameter, linear readout parameter, MLP input weight, MLP output coefficient, and MLP threshold is bounded by $B_{\text{upper}}(d_\perp, B_\delta)$, and therefore by B. Hence $f_{\Theta_F} \in \mathcal{T}_{\text{dec}}(H, W_f, B)$. □

F.3. Main Upper Theorem

We can now state the positive coexistence result in its final form.

Theorem F.4 (Main upper theorem: Bayes-optimal coexistence for trainable-feature decoders)**.** *Assume Theorems B.4, B.5, D.1 and D.2. Fix $d_\perp \le d - p_\star$ and suppose*

$$H \ge r + d_\perp, \qquad W_f \ge 3m.$$

Let F_m be a B_δ-admissible generic fact set. If

$$B \ge B_{\text{upper}}(d_\perp, B_\delta),$$

then, with probability at least

$$1 - 2m\, e^{-d_\perp/128} - 2m(m-1)\, e^{-d_\perp/128}, \tag{72}$$

over the generic fact locations, the trainable-feature decoder class satisfies

$$\Lambda_{F_m}\big(\mathcal{T}_{\text{dec}}(H, W_f, B); P_{\text{rule}}\big) \le \|\delta(F_m)\|_\infty^2 \Big[m\, e^{-25d_\perp/128} + m(m-1)\, e^{-75d_\perp/224}\Big]. \tag{73}$$

Consequently,

$$\Delta_{F_m}\big(\mathcal{T}_{\mathrm{dec}}(H, W_f, B)\big) \le \|\delta(F_m)\|_\infty^2 \Big[m\, e^{-25d_\perp/128} + m(m-1)\, e^{-75d_\perp/224}\Big]. \tag{74}$$

Equivalently, on the same event there exists $f_{\Theta_F} \in \mathcal{T}_{\mathrm{dec}}(H, W_f, B)$ such that

$$f_{\Theta_F}(X^\mu) = \xi^\mu \qquad \forall \mu \in [m],$$

and

$$R_{\mathrm{rule}}(f_{\Theta_F}) - R_{\mathrm{Bayes}} \le \|\delta(F_m)\|_\infty^2 \Big[m\, e^{-25d_\perp/128} + m(m-1)\, e^{-75d_\perp/224}\Big]. \tag{75}$$

Proof. By Theorem E.4, the event $\mathcal{E}_{\mathrm{geo}}$ holds with probability at least the quantity in Equation (72). We prove the claimed bounds on this event.

By Theorem F.3, there exists $f_{\Theta_F} \in \mathcal{T}_{\mathrm{dec}}(H, W_f, B)$ such that

$$f_{\Theta_F}(X) = f^\star(X) + g_F(Z^\perp(X)) \qquad \forall X \in \mathcal{X}_d,$$

and

$$f_{\Theta_F}(X^\mu) = \xi^\mu \qquad \forall \mu \in [m].$$

Thus the function

$$g(X) := g_F(Z^\perp(X))$$

is feasible for the fact-localizability problem

$$\Lambda_{F_m}\big(\mathcal{T}_{\mathrm{dec}}(H, W_f, B); P_{\mathrm{rule}}\big).$$

Under the rule marginal P_X, the residual vector $Z^\perp(X)$ has law $\nu_\perp = \mathcal{N}(0, \bar{\gamma}^2 I_{d_\perp})$ by Theorem D.9. Therefore, conditional on the fact set,

$$\|g\|_{L^2(P_X)}^2 = \mathbb{E}_{X\sim P_X}\big[g_F(Z^\perp(X))^2 \mid F_m\big] = \mathbb{E}_{Z\sim\nu_\perp}\big[g_F(Z)^2 \mid F_m\big].$$

Applying Theorem E.9 gives

$$\|g\|_{L^2(P_X)}^2 \le \|\delta(F_m)\|_\infty^2 \Big[m\, e^{-25d_\perp/128} + m(m-1)\, e^{-75d_\perp/224}\Big].$$

Since g is feasible for the localizability variational problem, this proves Equation (73).

The coexistence bound Equation (74) follows immediately from the coexistence–localizability identity Theorem C.5. Finally, because $f_{\Theta_F} = f^\star + g$, the Bayes Pythagorean identity Theorem C.2 yields

$$R_{\mathrm{rule}}(f_{\Theta_F}) - R_{\mathrm{Bayes}} = \|g\|_{L^2(P_X)}^2,$$

which proves Equation (75). □

Corollary F.5 (Simplified high-dimensional coexistence bound)**.** *Under the assumptions of Theorem F.4, suppose in addition that*

$$m \le \exp(d_\perp/256).$$

Then, with the same probability as in Equation (72),

$$\Delta_{F_m}\big(\mathcal{T}_{\mathrm{dec}}(H, W_f, B)\big) \le 2\, \|\delta(F_m)\|_\infty^2\, m\, e^{-25d_\perp/128}. \tag{76}$$

The same bound holds for the localizability functional

$$\Lambda_{F_m}\big(\mathcal{T}_{\mathrm{dec}}(H, W_f, B); P_{\mathrm{rule}}\big).$$

Proof. This is immediate from Theorems E.10 and F.4. □

Corollary F.6 (Vanishing Bayes-coexistence gap)**.** *Consider a sequence of problem instances indexed by d, satisfying the assumptions of Theorem F.4. Suppose*

$$\log m = o(d_\perp), \qquad \|\delta(F_m)\|_\infty = O(1), \qquad m\, e^{-25d_\perp/128} \to 0,$$

and choose

$$H \geq r + d_\perp, \qquad W_f \geq 3m, \qquad B \geq B_{\mathrm{upper}}(d_\perp, B_\delta).$$

Then

$$\Delta_{F_m}\big(\mathcal{T}_{\mathrm{dec}}(H, W_f, B)\big) \longrightarrow 0$$

with high probability over the generic fact locations. Equivalently, the trainable-feature decoder class admits asymptotically Bayes-optimal coexistence: it can exactly memorize all facts in F_m while paying vanishing excess rule risk.

Proof. The probability of $\mathcal{E}_{\mathrm{geo}}$ tends to one by Theorem E.4 and the assumption $\log m = o(d_\perp)$. On this event, Theorem F.4 gives

$$\Delta_{F_m}\big(\mathcal{T}_{\mathrm{dec}}(H, W_f, B)\big) \leq \|\delta(F_m)\|_\infty^2 \Big[m\, e^{-25d_\perp/128} + m(m-1)\, e^{-75d_\perp/224}\Big].$$

The first term tends to zero by assumption. For the second term, observe that

$$m(m-1)e^{-75d_\perp/224} \leq m^2 e^{-75d_\perp/224}.$$

Since $\log m = o(d_\perp)$, for every fixed $a > 0$,

$$m^2 e^{-ad_\perp} \to 0.$$

Taking $a = 75/224$ proves that the second term also vanishes. The boundedness of $\|\delta(F_m)\|_\infty$ then gives the claim. □

F.4. Uniformity over Arbitrary Bounded Fact Values

The upper theorem is uniform over bounded fact labels. This point is important: the proof does not assume random fact values, random labels, or independence between the residual labels and the rule labels. Randomness is used only to place the fact locations in high-dimensional general position.

Proposition F.7 (Uniformity over admissible residuals)**.** *Fix the fact locations $X^1, \ldots, X^m$ and suppose $\mathcal{E}_{\mathrm{geo}}$ holds. Then the conclusions of Theorems F.3 and F.4 hold simultaneously for all residual vectors $\delta \in \mathbb{R}^m$ satisfying*

$$\|\delta\|_\infty \leq B_\delta.$$

Proof. On $\mathcal{E}_{\mathrm{geo}}$, the directions $\widehat{Z}_\mu^\perp$, thresholds $\tau_{\mu,k}$, and localizers b_μ are fixed functions of the fact locations. For any residual vector δ, the correction

$$g_F(z) = \sum_{\mu=1}^{m} \delta_\mu b_\mu(z)$$

interpolates the corresponding residuals by Theorem E.6. The only quantities in the parameter budget and in the L^2-bound that depend on δ are the output coefficients $\delta_\mu/w_\perp$, $-2\delta_\mu/w_\perp$, and the scalar $\|\delta\|_\infty$. If $\|\delta\|_\infty \leq B_\delta$, then these coefficients are uniformly bounded by $2B_\delta/w_\perp$, exactly as in the proof of Theorem F.3. The risk and localizability bounds then follow uniformly from the same calculation as in Theorem F.4. □

Remark F.8 (What the upper theorem establishes). The result above should not be read as an optimization theorem. It establishes that the trainable-feature decoder class contains a structured predictor of the form

$$f_{\Theta_F}(X) = f^\star(X) + g_F(Z^\perp(X)),$$

where $f^\star$ is the Bayes rule and g_F is a sparse residual correction. The correction exactly memorizes arbitrary bounded facts, but its mass under the rule distribution is exponentially small in the residual dimension. By Theorem C.5, this is precisely the condition needed for asymptotically Bayes-optimal coexistence.

Remark F.9 (Why feature learning matters). The proof relies on the existence of a learned split representation

$$A_{\bar{\Theta}^{\text{split}}}(X) = (S^\star(X), Z^\perp(X), 0),$$

where $S^\star(X)$ carries the rule and $Z^\perp(X)$ is statistically orthogonal residual structure. The next appendix shows that this localization mechanism is not available in the lazy/kernelized version of the same decoder unless its effective dimension is large enough. Thus the positive theorem is not merely a storage-capacity statement; it identifies a representation-level mechanism by which trainable features separate reusable rule computation from sparse fact memorization.

G. Lazy/Kernel Converse

This appendix proves the negative half of the feature-learning separation. The main upper theorem in Appendix F used trainable attention features to create a rule–residual split

$$A_{\bar{\Theta}^{\text{split}}}(X) = (S^\star(X), Z^\perp(X), 0),$$

and then implanted facts in the residual Gaussian block with exponentially small $L^2(P_X)$-mass. We now show that this phenomenon is not available to a bounded lazy/tangent model unless its kernel has sufficiently large effective dimension.

The converse is stated in a deliberately strong oracle form. Instead of penalizing the lazy model for failing to learn the Bayes rule, we grant it the Bayes rule $f^\star$ for free and ask whether a bounded RKHS correction can memorize random fact residuals while remaining invisible under P_X. The answer is no whenever the kernel effective dimension is small relative to the number of independent facts. This is the precise sense in which feature learning, rather than mere kernel interpolation around a fixed representation, is responsible for the coexistence mechanism of Appendix F. This perspective is aligned with the standard neural tangent and lazy-training view, where the model behaves as a kernel method around initialization (Jacot et al., 2018; Chizat et al., 2019).

Throughout this appendix, fix a base point Θ_0 and write

$$K := K_{\Theta_0}$$

for the tangent kernel from Equation (19). All probabilities below are conditional on Θ_0, unless explicitly stated otherwise.

G.1. Spectral Kernel Preliminaries

We begin with a standard spectral description of the RKHS associated with the tangent kernel.

Assumption G.1 (Trace-class tangent kernel)**.** The kernel $K : \mathcal{X}_d \times \mathcal{X}_d \to \mathbb{R}$ is symmetric, positive semidefinite, and satisfies

$$\mathbb{E}_{X\sim P_X} K(X, X) < \infty.$$

Let $T_K : L^2(P_X) \to L^2(P_X)$ be the associated integral operator

$$(T_K f)(x) := \int K(x, x') f(x')\, dP_X(x').$$

We assume that T_K admits a Mercer decomposition

$$K(x, x') = \sum_{j\geq 1} \lambda_j \phi_j(x)\phi_j(x') \qquad \text{in } L^2(P_X \otimes P_X), \tag{77}$$

where

$$\lambda_1 \geq \lambda_2 \geq \cdots > 0, \qquad \sum_{j\geq 1} \lambda_j < \infty,$$

and $(\phi_j)_{j\geq 1}$ is an orthonormal family in $L^2(P_X)$.

Definition G.2 (Effective dimension)**.** For $\zeta > 0$, define the kernel effective dimension

$$\mathcal{N}_K(\zeta) := \operatorname{tr}\big(T_K(T_K + \zeta I)^{-1}\big) = \sum_{j\geq 1} \frac{\lambda_j}{\lambda_j + \zeta}. \tag{78}$$

Lemma G.3 (RKHS spectral coordinates)**.** *Under Theorem G.1, every $h \in \mathcal{H}_K$ admits a coordinate expansion*

$$h = \sum_{j\geq 1} a_j \phi_j \qquad \text{in } L^2(P_X),$$

with

$$\|h\|^2_{L^2(P_X)} = \sum_{j\geq 1} a_j^2, \qquad \|h\|^2_{\mathcal{H}_K} = \sum_{j\geq 1} \frac{a_j^2}{\lambda_j}. \tag{79}$$

Conversely, every sequence $(a_j)_{j\geq 1}$ satisfying $\sum_j a_j^2/\lambda_j < \infty$ defines an element of $\mathcal{H}_K$.

Proof. This is the standard spectral characterization of the RKHS of a trace-class positive kernel. For completeness, let $T_K^{1/2}$ be the positive square root of T_K. The RKHS is isometrically identified with $\mathrm{Range}(T_K^{1/2})$ equipped with the norm

$$\|T_K^{1/2} u\|_{\mathcal{H}_K} = \|P_{\overline{\mathrm{Range}(T_K)}} u\|_{L^2(P_X)}.$$

Since

$$T_K \phi_j = \lambda_j \phi_j,$$

a function $h = \sum_j a_j \phi_j$ lies in $\mathrm{Range}(T_K^{1/2})$ precisely when

$$\sum_{j\geq 1} \frac{a_j^2}{\lambda_j} < \infty,$$

and then its RKHS norm is the second quantity in Equation (79). The $L^2(P_X)$ identity follows from orthonormality of $(\phi_j)_j$. □

G.2. Oracle-Centered Kernel Correction Class

The cleanest converse is obtained by granting the kernel class the Bayes rule and asking only whether the kernel can supply a localized fact correction.

Definition G.4 (Oracle-centered kernel correction class)**.** For $R \geq 0$, define

$$\mathcal{K}_R^\star := \{f^\star + h : h \in \mathcal{H}_K,\ \|h\|_{\mathcal{H}_K} \leq R\}. \tag{80}$$

Remark G.5 (Why this is a stronger baseline than the lazy class). The class $\mathcal{K}_R^\star$ gives the kernel model the Bayes rule $f^\star$ for free. Therefore, a lower bound for $\mathcal{K}_R^\star$ is not caused by approximation error of the rule. It is a lower bound on the kernel's ability to implant facts invisibly under the rule distribution. The transfer back to the actual affine lazy class $\mathcal{T}_{\mathrm{lazy}}(R; \Theta_0)$ is given in Theorem G.12.

G.3. Random Signed Facts

The upper theorem in Appendix F is uniform over arbitrary bounded residual vectors. To prove that a lazy/kernel class cannot have the same uniform coexistence property, it suffices to exhibit a distribution over bounded facts that it cannot localize. We use independent Rademacher signs.

Definition G.6 (Signed generic fact set)**.** Fix an amplitude $\tau > 0$. A τ-signed generic fact set is generated as follows:

$$X^1, \ldots, X^m \overset{\text{i.i.d.}}{\sim} P_X, \qquad \omega_1, \ldots, \omega_m \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1, +1\},$$

with the signs independent of the locations, and

$$\xi^\mu := f^\star(X^\mu) + \tau\omega_\mu, \qquad \mu \in [m]. \tag{81}$$

Thus the fact residual vector is

$$\delta_\mu = \xi^\mu - f^\star(X^\mu) = \tau\omega_\mu.$$

G.4. A Local Rademacher Anti-Interpolation Bound

The next lemma is the core technical statement. It says that an RKHS ball whose functions have small $L^2(P_X)$-mass cannot interpolate independent signs unless the kernel effective dimension is large.

Lemma G.7 (Local Rademacher complexity of an RKHS ball)**.** *Assume Theorem G.1. Let*

$$X^1, \dots, X^m \overset{\text{i.i.d.}}{\sim} P_X, \qquad \omega_1, \dots, \omega_m \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}\{-1, +1\},$$

independently. For $R, \rho, \zeta > 0$, define the localized RKHS class

$$\mathcal{H}_K(R, \rho) := \left\{ h \in \mathcal{H}_K : \|h\|_{\mathcal{H}_K} \le R,\ \|h\|_{L^2(P_X)} \le \rho \right\}.$$

Then

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}_K(R,\rho)} \sum_{\mu=1}^{m} \omega_\mu h(X^\mu) \right] \le \sqrt{m \mathcal{N}_K(\zeta) \left(\rho^2 + \zeta R^2\right)}. \tag{82}$$

Proof. Let $h \in \mathcal{H}_K(R, \rho)$, and write

$$h = \sum_{j \ge 1} a_j \phi_j$$

as in Theorem G.3. Define

$$S_j := \sum_{\mu=1}^{m} \omega_\mu \phi_j(X^\mu).$$

For any $\zeta > 0$,

$$\begin{aligned}
\sum_{\mu=1}^{m} \omega_\mu h(X^\mu) &= \sum_{j \ge 1} a_j S_j \\
&\le \left(\sum_{j \ge 1} a_j^2 \left(1 + \frac{\zeta}{\lambda_j}\right) \right)^{1/2} \left(\sum_{j \ge 1} \frac{S_j^2}{1 + \zeta/\lambda_j} \right)^{1/2} \\
&= \left(\sum_{j \ge 1} a_j^2 + \zeta \sum_{j \ge 1} \frac{a_j^2}{\lambda_j} \right)^{1/2} \left(\sum_{j \ge 1} \frac{\lambda_j}{\lambda_j + \zeta} S_j^2 \right)^{1/2}.
\end{aligned} \tag{83}$$

Since $h \in \mathcal{H}_K(R, \rho)$, the first factor in Equation (83) is at most

$$(\rho^2 + \zeta R^2)^{1/2}.$$

Therefore,

$$\sup_{h \in \mathcal{H}_K(R,\rho)} \sum_{\mu=1}^{m} \omega_\mu h(X^\mu) \le (\rho^2 + \zeta R^2)^{1/2} \left(\sum_{j \ge 1} \frac{\lambda_j}{\lambda_j + \zeta} S_j^2 \right)^{1/2}.$$

Taking expectations and using Jensen's inequality,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}_K(R,\rho)} \sum_{\mu=1}^{m} \omega_\mu h(X^\mu) \right] \le (\rho^2 + \zeta R^2)^{1/2} \left(\mathbb{E} \sum_{j \ge 1} \frac{\lambda_j}{\lambda_j + \zeta} S_j^2 \right)^{1/2}.$$

For each j,

$$\begin{aligned}
\mathbb{E}[S_j^2] &= \mathbb{E}\left[\sum_{\mu,\nu=1}^{m} \omega_\mu \omega_\nu \phi_j(X^\mu) \phi_j(X^\nu) \right] \\
&= \sum_{\mu=1}^{m} \mathbb{E}[\phi_j(X^\mu)^2] = m,
\end{aligned}$$

because the signs are independent, centered, and $\|\phi_j\|_{L^2(P_X)} = 1$. Hence

$$\mathbb{E}\sum_{j\geq 1}\frac{\lambda_j}{\lambda_j+\zeta}S_j^2 = m\sum_{j\geq 1}\frac{\lambda_j}{\lambda_j+\zeta} = m\mathcal{N}_K(\zeta).$$

Substitution gives Equation (82). □

Lemma G.8 (Anti-interpolation lower bound)**.** *Assume Theorem G.1. Let F_m be a τ-signed generic fact set in the sense of Theorem G.6. Fix $R, \zeta > 0$ and $\eta \in (0,1)$. With probability at least $1-\eta$ over the fact locations and signs, every $h \in \mathcal{H}_K$ satisfying*

$$\|h\|_{\mathcal{H}_K} \leq R, \qquad h(X^\mu) = \tau\omega_\mu \quad \forall \mu \in [m],$$

also satisfies

$$\|h\|^2_{L^2(P_X)} \geq \left(\eta^2\,\frac{\tau^2 m}{\mathcal{N}_K(\zeta)} - \zeta R^2\right)_+, \tag{84}$$

where $a_+ := \max\{a, 0\}$.

Proof. Fix $\rho > 0$. If there exists $h \in \mathcal{H}_K(R,\rho)$ satisfying

$$h(X^\mu) = \tau\omega_\mu \qquad \forall \mu \in [m],$$

then

$$\sum_{\mu=1}^m \omega_\mu h(X^\mu) = \tau\sum_{\mu=1}^m \omega_\mu^2 = \tau m.$$

Consequently,

$$\sup_{g\in\mathcal{H}_K(R,\rho)}\sum_{\mu=1}^m \omega_\mu g(X^\mu) \geq \tau m.$$

By Markov's inequality and Theorem G.7,

$$\mathbb{P}\left(\exists h \in \mathcal{H}_K(R,\rho) : h(X^\mu) = \tau\omega_\mu\ \forall\mu\right) \leq \frac{\sqrt{m\mathcal{N}_K(\zeta)\,(\rho^2+\zeta R^2)}}{\tau m}. \tag{85}$$

If

$$\rho^2 < \eta^2\frac{\tau^2 m}{\mathcal{N}_K(\zeta)} - \zeta R^2,$$

then the right-hand side of Equation (85) is strictly less than η. Thus, with probability at least $1-\eta$, no interpolant with RKHS norm at most R and $L^2(P_X)$-norm at most such a ρ exists. Taking the supremum over all admissible ρ below the displayed threshold proves Equation (84). If the threshold is negative, the claim is vacuous and follows from nonnegativity of $\|h\|^2_{L^2(P_X)}$. □

G.5. Kernel Converse for Fact Localizability

We now translate the anti-interpolation statement into a lower bound on the Bayes-coexistence gap.

Theorem G.9 (Oracle kernel converse)**.** *Assume Theorem G.1. Let F_m be a τ-signed generic fact set. Fix $R, \zeta > 0$ and $\eta \in (0,1)$. Then, with probability at least $1-\eta$ over the fact locations and signs,*

$$\Lambda_{F_m}(\mathcal{K}_R^\star; P_{\text{rule}}) \geq \left(\eta^2\,\frac{\tau^2 m}{\mathcal{N}_K(\zeta)} - \zeta R^2\right)_+. \tag{86}$$

Consequently,

$$\Delta_{F_m}(\mathcal{K}_R^\star) \geq \left(\eta^2\,\frac{\tau^2 m}{\mathcal{N}_K(\zeta)} - \zeta R^2\right)_+. \tag{87}$$

Proof. For the oracle-centered class

$$\mathcal{K}_R^\star = \{f^\star + h : h \in \mathcal{H}_K,\ \|h\|_{\mathcal{H}_K} \le R\},$$

a correction g is feasible for $\Lambda_{F_m}(\mathcal{K}_R^\star; P_{\text{rule}})$ if and only if $g \in \mathcal{H}_K$,

$$\|g\|_{\mathcal{H}_K} \le R, \qquad g(X^\mu) = \delta_\mu = \tau\omega_\mu \quad \forall \mu \in [m].$$

Therefore the localizability functional is precisely

$$\Lambda_{F_m}\big(\mathcal{K}_R^\star; P_{\text{rule}}\big) = \inf_{\substack{g \in \mathcal{H}_K \\ \|g\|_{\mathcal{H}_K} \le R \\ g(X^\mu) = \tau\omega_\mu\ \forall \mu}} \|g\|^2_{L^2(P_X)}.$$

Applying Theorem G.8 to every feasible g gives Equation (86). The coexistence-gap lower bound Equation (87) follows immediately from the coexistence–localizability identity Theorem C.5. □

Corollary G.10 (Effective-dimension bottleneck)**.** *Let $\eta_d \in (0,1)$ be any sequence with $\eta_d \to 0$. Suppose that, for some $\zeta_d > 0$ and constant $C_N < \infty$,*

$$\mathcal{N}_K(\zeta_d) \le C_N\, \eta_d^2\, m, \tag{88}$$

and

$$\zeta_d R^2 \le \frac{\tau^2}{2C_N}. \tag{89}$$

Then, for a τ-signed generic fact set F_m,

$$\Lambda_{F_m}\big(\mathcal{K}_R^\star; P_{\text{rule}}\big) \ge \frac{\tau^2}{2C_N} \tag{90}$$

and

$$\Delta_{F_m}\big(\mathcal{K}_R^\star\big) \ge \frac{\tau^2}{2C_N} \tag{91}$$

with probability at least $1 - \eta_d$.

Proof. By Theorem G.9, with probability at least $1 - \eta_d$,

$$\Lambda_{F_m}\big(\mathcal{K}_R^\star; P_{\text{rule}}\big) \ge \eta_d^2 \frac{\tau^2 m}{\mathcal{N}_K(\zeta_d)} - \zeta_d R^2.$$

Using Equation (88),

$$\eta_d^2 \frac{\tau^2 m}{\mathcal{N}_K(\zeta_d)} \ge \frac{\tau^2}{C_N}.$$

Using Equation (89), the right-hand side is at least

$$\frac{\tau^2}{C_N} - \frac{\tau^2}{2C_N} = \frac{\tau^2}{2C_N}.$$

The same lower bound for the Bayes-coexistence gap follows from Theorem C.5. □

G.6. Transfer to the Affine Lazy Decoder Class

We finally translate the oracle-centered kernel bound back to the affine lazy class from Theorem B.13. The argument is a simple containment: if the initialization offset $f_{\Theta_0} - f^\star$ belongs to the RKHS, then the affine lazy class is contained in a larger oracle-centered RKHS ball.

Lemma G.11 (Affine lazy class is contained in an oracle-centered RKHS ball)**.** *Assume Theorem G.1. Suppose that*

$$f_{\Theta_0} - f^\star \in \mathcal{H}_K \qquad \textit{and set} \qquad R_0 := \|f_{\Theta_0} - f^\star\|_{\mathcal{H}_K} < \infty.$$

Then, for every $R \ge 0$,

$$\mathcal{T}_{\text{lazy}}(R; \Theta_0) \subseteq \mathcal{K}_{R+R_0}^\star. \tag{92}$$

Proof. Take

$$f \in \mathcal{T}_{\text{lazy}}(R; \Theta_0).$$

By Theorem B.13, there exists a parameter displacement u with $\|u\|_2 \le R$ such that

$$f(X) = f_{\Theta_0}(X) + \langle \nabla_\Theta f_{\Theta_0}(X), u \rangle .$$

The tangent feature map

$$\Phi_{\Theta_0}(X) := \nabla_\Theta f_{\Theta_0}(X)$$

generates the kernel

$$K(X, X') = \langle \Phi_{\Theta_0}(X), \Phi_{\Theta_0}(X') \rangle.$$

Therefore the function

$$h_u(X) := \langle \Phi_{\Theta_0}(X), u \rangle$$

belongs to $\mathcal{H}_K$ and satisfies

$$\|h_u\|_{\mathcal{H}_K} \le \|u\|_2 \le R.$$

By assumption, $f_{\Theta_0} - f^\star \in \mathcal{H}_K$. Hence

$$f = f^\star + \big(f_{\Theta_0} - f^\star + h_u\big),$$

where

$$\|f_{\Theta_0} - f^\star + h_u\|_{\mathcal{H}_K} \le R_0 + R$$

by the triangle inequality. Thus

$$f \in \mathcal{K}^\star_{R+R_0},$$

which proves the containment. □

Corollary G.12 (Lazy/NTK converse for the decoder tangent class)**.** *Assume Theorem G.1 and the RKHS offset condition of Theorem G.11. Let F_m be a τ-signed generic fact set. Fix $R, \zeta > 0$ and $\eta \in (0, 1)$. Then, with probability at least $1 - \eta$,*

$$\Lambda_{F_m}\big(\mathcal{T}_{\text{lazy}}(R; \Theta_0); P_{\text{rule}}\big) \ge \left(\eta^2 \frac{\tau^2 m}{\mathcal{N}_K(\zeta)} - \zeta(R + R_0)^2\right)_+ . \tag{93}$$

Consequently,

$$\Delta_{F_m}\big(\mathcal{T}_{\text{lazy}}(R; \Theta_0)\big) \ge \left(\eta^2 \frac{\tau^2 m}{\mathcal{N}_K(\zeta)} - \zeta(R + R_0)^2\right)_+ . \tag{94}$$

Proof. By Theorem G.11,

$$\mathcal{T}_{\text{lazy}}(R; \Theta_0) \subseteq \mathcal{K}^\star_{R+R_0}.$$

By monotonicity under class enlargement Theorem C.7,

$$\Lambda_{F_m}\big(\mathcal{T}_{\text{lazy}}(R; \Theta_0); P_{\text{rule}}\big) \ge \Lambda_{F_m}\big(\mathcal{K}^\star_{R+R_0}; P_{\text{rule}}\big).$$

Applying Theorem G.9 with radius $R + R_0$ gives Equation (93). The coexistence-gap bound follows from Theorem C.5. □

Corollary G.13 (Nonvanishing lazy coexistence gap)**.** *Let $\eta_d \to 0$. Suppose the hypotheses of Theorem G.12 hold and that, for some $\zeta_d > 0$ and constant $C_N < \infty$,*

$$\mathcal{N}_K(\zeta_d) \le C_N \eta_d^2 m, \qquad \zeta_d (R + R_0)^2 \le \frac{\tau^2}{2C_N}.$$

Then, for a τ-signed generic fact set,

$$\Delta_{F_m}\big(\mathcal{T}_{\text{lazy}}(R; \Theta_0)\big) \ge \frac{\tau^2}{2C_N}$$

with probability at least $1 - \eta_d$.

Proof. This is Theorem G.10 applied to the containing oracle-centered class $\mathcal{K}^\star_{R+R_0}$, followed by Theorem G.11 and Theorem C.7. □

G.7. Interpretation of the Converse

Remark G.14 (What the converse proves). The upper theorem in Appendix F gives, for trainable-feature decoders, an exponentially small coexistence bound of order

$$\|\delta(F_m)\|_\infty^2 \left[m e^{-25 d_\perp/128} + m(m-1) e^{-75 d_\perp/224} \right].$$

By contrast, Theorem G.13 shows that a bounded lazy/tangent model with insufficient effective dimension has a nonvanishing Bayes-coexistence gap for random signed facts, even when it is granted $f^\star$ as an oracle baseline. Thus the separation is not due to failure to learn the rule; it is due to the inability of a fixed kernel representation to create fact corrections whose population mass vanishes.

Remark G.15 (Why this is not a universal impossibility theorem). The lower bound is an effective-dimension bottleneck, not a claim that no kernel can ever memorize. If $\mathcal{N}_K(\zeta)$ is of order m at a scale ζ compatible with the RKHS radius, then the lower bound may become vacuous. This is the correct behavior: a kernel with sufficiently many high-resolution degrees of freedom can in principle fit many facts. The separation established here is that the trainable-feature decoder construction achieves vanishing coexistence through a learned residual geometry, whereas a bounded lazy/tangent model requires a correspondingly large effective dimension.

H. Structural Localization and Deletability

This appendix proves that the coexistence construction from Appendix F is not only an existence result: the rule and fact components are structurally separated. The Bayes rule is implemented by a low-dimensional attention/readout block, while the memorized fact residuals are implemented by a sparse collection of MLP units whose input weights are supported only on the residual coordinates. This yields an exact selective deletion operation: one can remove any chosen subset of memorized fact residuals by zeroing the corresponding MLP output coefficients, without changing the learned rule block.

The result should be distinguished from empirical or mechanistic claims that "attention retrieves" and "MLPs memorize." Such phenomena have been suggested and studied in several recent transformer-theory and component analysis works (Xu & Chen, 2025; Nichani et al., 2024; Dong et al., 2025). Here, the claim is a theorem about the explicit coexistence construction: the rule and fact contributions are separated at the level of the represented function, the active coordinates, and the coefficient-space Gram operator. The deletion corollaries are related in spirit to the goal of machine unlearning (Cao & Yang, 2015; Bourtoule et al., 2021), but are stronger and narrower: they are exact for the constructed predictor and the squared-loss population model considered in this paper.

Throughout this appendix, we assume the hypotheses of Theorem F.4 and work on the geometric event $\mathcal{E}_{\mathrm{geo}}$. Let Θ_F denote the parameter vector constructed in Theorem F.3. Thus

$$f_{\Theta_F}(X) = f^\star(X) + g_F(Z^\perp(X)), \qquad g_F(z) = \sum_{\mu=1}^{m} \delta_\mu b_\mu(z), \tag{95}$$

where b_μ is the residual tent localizer from Theorem E.5.

H.1. Rule and Fact Dictionaries

We first isolate the two dictionaries that appear in the construction.

Definition H.1 (Rule and fact feature dictionaries)**.** Define the rule-feature vector

$$\Phi_{\mathrm{R}}(X) := \big(S_1^\star(X), \dots, S_r^\star(X)\big)^\top \in \mathbb{R}^r, \tag{96}$$

and, conditional on the fact locations, define the fact-feature vector

$$\Phi_{\mathrm{F}}(X) := \big(b_1(Z^\perp(X)), \dots, b_m(Z^\perp(X))\big)^\top \in \mathbb{R}^m. \tag{97}$$

The constructed predictor can then be written as

$$f_{\Theta_F}(X) = \langle \beta^\star, \Phi_{\mathrm{R}}(X) \rangle + \langle \delta, \Phi_{\mathrm{F}}(X) \rangle, \tag{98}$$

where

$$\beta^\star = (\beta_1^\star, \dots, \beta_r^\star)^\top, \qquad \delta = (\delta_1, \dots, \delta_m)^\top.$$

Lemma H.2 (Finite rule-feature energy)**.** *Under Theorems B.4, B.5 and D.1,*

$$C_{\mathrm{R}}^2 := \mathbb{E}_{X\sim P_X}\left[\|\Phi_{\mathrm{R}}(X)\|_2^2\right] = \sum_{a=1}^{r} \|S_a^\star\|_{L^2(P_X)}^2 < \infty.$$

Proof. Fix $a \in [r]$. By Equation (13),

$$S_a^\star(X) = \sum_{t=1}^{L-1} \psi\big(\langle q_a^\star, z_L(X)\rangle \langle k_a^\star, z_t(X)\rangle\big)\langle v_a^\star, z_t(X)\rangle.$$

The teacher vectors have bounded Euclidean norm by Theorem B.5. Under Theorem D.1, each inner product

$$\langle q_a^\star, z_L(X)\rangle, \quad \langle k_a^\star, z_t(X)\rangle, \quad \langle v_a^\star, z_t(X)\rangle$$

has finite moments of all orders. Since ψ has at most linear growth by Theorem B.4,

$$|\psi(uv)| \le B_\psi(1 + |uv|).$$

Therefore each summand in $S_a^\star(X)$ is bounded in $L^2(P_X)$ by a finite linear combination of Gaussian moments, for example by Cauchy–Schwarz applied to

$$(1 + |UV|)|W|,$$

where U, V, W are centered Gaussian variables with finite variances depending only on the teacher norm bound and the fixed sequence length. Since L and r are fixed, summing over $t \in [L-1]$ and $a \in [r]$ gives

$$\sum_{a=1}^{r} \mathbb{E}\big[(S_a^\star(X))^2\big] < \infty.$$

□

H.2. Exact Parameter-Block Localization

The next proposition formalizes the structural support of the constructed coexistence predictor.

Proposition H.3 (Sparse-plus-low-rank structural decomposition)**.** *Under the hypotheses of Theorem F.3, the constructed predictor f_{Θ_F} admits the following decomposition.*

(i) *The rule component*

$$f_{\mathrm{rule}}(X) := f^\star(X)$$

is implemented using only the first r attention coordinates and the linear readout

$$w^\star = (\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0).$$

(ii) *The fact component*

$$f_{\mathrm{fact}}(X) := g_F(Z^\perp(X))$$

is implemented using at most $3m$ hidden ReLU units in the final MLP. Each such unit has an input weight vector supported only on the residual attention coordinates

$$\{r+1, \dots, r+d_\perp\}.$$

In particular, every fact-unit input weight is zero on the first r rule coordinates.

(iii) *The two components are additively separated:*

$$f_{\Theta_F}(X) = f_{\mathrm{rule}}(X) + f_{\mathrm{fact}}(X) \qquad \forall X \in \mathcal{X}_d.$$

Proof. Part (i) is exactly the readout construction in Theorem D.9: the split attention representation is

$$A_{\bar{\Theta}^{\text{split}}}(X) = \big(S_1^\star(X), \dots, S_r^\star(X), Z_1^\perp(X), \dots, Z_{d_\perp}^\perp(X), 0, \dots, 0\big),$$

and the linear readout

$$w^\star = (\beta_1^\star, \dots, \beta_r^\star, 0, \dots, 0)$$

therefore yields

$$\langle w^\star, A_{\bar{\Theta}^{\text{split}}}(X)\rangle = \sum_{a=1}^{r} \beta_a^\star S_a^\star(X) = f^\star(X).$$

For part (ii), Theorem F.3 constructs each tent localizer b_μ using three ReLU units with full attention-space input direction

$$\widetilde{z}_\mu = (0_r, \widehat{Z}_\mu^\perp, 0_{H-r-d_\perp}) \in \mathbb{R}^H.$$

Thus the input weights of these ReLU units vanish on the rule coordinates $1, \dots, r$ and on the unused coordinates $r + d_\perp + 1, \dots, H$. Summing over $\mu \in [m]$ uses at most $3m$ units and implements

$$g_F(Z^\perp(X)) = \sum_{\mu=1}^{m} \delta_\mu b_\mu(Z^\perp(X)).$$

Part (iii) is the identity Equation (95). □

H.3. Selective Deletion Operators

We now define deletion at the parameter level. Since the construction attaches exactly three ReLU units to each fact, selective deletion is implemented by zeroing the corresponding output coefficients.

Definition H.4 (Selective deletion mask)**.** Let $U \subseteq [m]$ be a set of facts to delete. The selective deletion operator

$$\mathsf{Del}_U$$

acts on the constructed parameter vector Θ_F as follows:

- all attention-head parameters are left unchanged;
- the linear readout $w^\star$ and bias $c = 0$ are left unchanged;
- for every $\mu \in U$, the three MLP output coefficients

  $$\alpha_{\mu,1}, \alpha_{\mu,2}, \alpha_{\mu,3}$$

  attached to the localizer b_μ are set to zero;
- all remaining MLP parameters are left unchanged.

We write

$$\Theta_F^{(-U)} := \mathsf{Del}_U(\Theta_F).$$

When $U = [m]$, we write simply

$$\Theta_F^{(-\text{all})} := \Theta_F^{(-[m])}.$$

Lemma H.5 (Idempotence and locality of deletion)**.** *For every $U, V \subseteq [m]$,*

$$\mathsf{Del}_U(\mathsf{Del}_V(\Theta_F)) = \mathsf{Del}_{U\cup V}(\Theta_F).$$

In particular,

$$\mathsf{Del}_U(\mathsf{Del}_U(\Theta_F)) = \mathsf{Del}_U(\Theta_F).$$

Moreover, Del_U *modifies no attention-head parameter and no rule-readout parameter.*

Proof. The operator Del_U only sets to zero the MLP output coefficients associated with facts in U. Applying Del_V first zeros the coefficients indexed by V, and applying Del_U afterwards zeros the coefficients indexed by U. The resulting zeroed set is exactly $U \cup V$. The idempotence statement is the special case $V = U$. The final claim follows immediately from Theorem H.4. □

H.4. Exact Selective Deletability

The following theorem is the main result of this appendix. It says that the constructed memorized facts are exactly and selectively removable.

Theorem H.6 (Exact selective deletability)**.** *Assume the hypotheses of Theorem F.4 and work on $\mathcal{E}_{\text{geo}}$. Let $U \subseteq [m]$ be any subset of facts to delete, and let $K := [m] \setminus U$ be the set of facts to keep. Then the deleted predictor satisfies, for every $X \in \mathcal{X}_d$,*

$$f_{\Theta_F^{(-U)}}(X) = f^\star(X) + \sum_{\mu \in K} \delta_\mu b_\mu(Z^\perp(X)). \tag{99}$$

Consequently:

(i) *every kept fact remains exactly memorized:*

$$f_{\Theta_F^{(-U)}}(X^\nu) = \xi^\nu \qquad \forall \nu \in K; \tag{100}$$

(ii) *every deleted fact has its residual correction exactly removed:*

$$f_{\Theta_F^{(-U)}}(X^\nu) = f^\star(X^\nu) \qquad \forall \nu \in U; \tag{101}$$

(iii) *the rule-risk excess after deletion is bounded by*

$$R_{\text{rule}}(f_{\Theta_F^{(-U)}}) - R_{\text{Bayes}} \le \|\delta_K\|_\infty^2 \left[|K|\, e^{-25 d_\perp/128} + |K|(|K|-1)\, e^{-75 d_\perp/224} \right], \tag{102}$$

where $\|\delta_K\|_\infty := \max_{\mu \in K} |\delta_\mu|$, with the convention that the right-hand side is zero if $K = \varnothing$.

Proof. By construction of Θ_F, the three ReLU units attached to fact μ contribute exactly

$$\delta_\mu b_\mu(Z^\perp(X))$$

to the output. Applying Del_U sets these three output coefficients to zero for every $\mu \in U$ and leaves all other parameters unchanged. Therefore the resulting predictor is

$$f_{\Theta_F^{(-U)}}(X) = f^\star(X) + \sum_{\mu \in [m] \setminus U} \delta_\mu b_\mu(Z^\perp(X)),$$

which proves Equation (99).

Now fix $\nu \in K$. By Theorem E.6,

$$b_\mu(Z_\nu^\perp) = \mathbf{1}_{\{\mu = \nu\}} \qquad \forall \mu \in [m]$$

on $\mathcal{E}_{\text{geo}}$. Since $\nu \in K$,

$$f_{\Theta_F^{(-U)}}(X^\nu) = f^\star(X^\nu) + \sum_{\mu \in K} \delta_\mu b_\mu(Z_\nu^\perp) = f^\star(X^\nu) + \delta_\nu = \xi^\nu.$$

This proves Equation (100).

If instead $\nu \in U$, then $\nu \notin K$, and again using Theorem E.6,

$$\sum_{\mu \in K} \delta_\mu b_\mu(Z_\nu^\perp) = 0.$$

Hence

$$f_{\Theta_F^{(-U)}}(X^\nu) = f^\star(X^\nu),$$

which proves Equation (101).

It remains to prove the risk bound. Define

$$g_K(z) := \sum_{\mu \in K} \delta_\mu b_\mu(z).$$

Then

$$f_{\Theta_F^{(-U)}}(X) = f^\star(X) + g_K(Z^\perp(X)).$$

By the Bayes Pythagorean identity Theorem C.2,

$$R_{\text{rule}}(f_{\Theta_F^{(-U)}}) - R_{\text{Bayes}} = \mathbb{E}_{X\sim P_X}\left[g_K(Z^\perp(X))^2\right].$$

The same proof as Theorem E.9, with the index set $[m]$ replaced by K, yields

$$\mathbb{E}_{X\sim P_X}\left[g_K(Z^\perp(X))^2\right] \le \|\delta_K\|_\infty^2 \left[|K|\, e^{-25d_\perp/128} + |K|(|K|-1)\, e^{-75d_\perp/224}\right].$$

This proves Equation (102). □

Corollary H.7 (Full deletion recovers the Bayes rule exactly)**.** *Under the hypotheses of Theorem H.6,*

$$f_{\Theta_F^{(-\text{all})}}(X) = f^\star(X) \qquad \forall X \in \mathcal{X}_d. \tag{103}$$

Consequently,

$$R_{\text{rule}}(f_{\Theta_F^{(-\text{all})}}) = R_{\text{Bayes}}. \tag{104}$$

Proof. Apply Theorem H.6 with $U = [m]$, so that $K = \varnothing$. Then the sum in Equation (99) is empty, hence

$$f_{\Theta_F^{(-\text{all})}}(X) = f^\star(X)$$

for every X. The risk identity follows immediately from Theorem B.15 and Equation (22). □

H.5. Residual Memorization Energy

To formalize what is erased by deletion, we measure memorization relative to the Bayes rule. This isolates the memorized residual from the rule-predicted value at the same location.

Definition H.8 (Residual fact-memorization energy)**.** For a predictor $f \in \mathcal{M}_2(P_X)$ and a fact set F_m, define

$$\mathfrak{M}_{F_m}(f) := \frac{1}{m}\sum_{\mu=1}^{m}\big(f(X^\mu) - f^\star(X^\mu)\big)^2. \tag{105}$$

Proposition H.9 (Deletion erases residual memorization)**.** *Under the hypotheses of Theorem H.6,*

$$\mathfrak{M}_{F_m}(f_{\Theta_F}) = \frac{1}{m}\sum_{\mu=1}^{m}\delta_\mu^2, \tag{106}$$

whereas

$$\mathfrak{M}_{F_m}(f_{\Theta_F^{(-\text{all})}}) = 0. \tag{107}$$

More generally, for $U \subseteq [m]$ and $K = [m] \setminus U$,

$$\mathfrak{M}_{F_m}(f_{\Theta_F^{(-U)}}) = \frac{1}{m}\sum_{\mu\in K}\delta_\mu^2. \tag{108}$$

Proof. For the original predictor, exact interpolation gives

$$f_{\Theta_F}(X^\mu) - f^\star(X^\mu) = \xi^\mu - f^\star(X^\mu) = \delta_\mu,$$

so Equation (106) follows.

For the fully deleted predictor, Theorem H.7 gives

$$f_{\Theta_F^{(-\text{all})}}(X^\mu) - f^\star(X^\mu) = 0$$

for every μ, proving Equation (107).

For a selective deletion set U, Theorem H.6 gives

$$f_{\Theta_F^{(-U)}}(X^\mu) - f^\star(X^\mu) = \begin{cases} \delta_\mu, & \mu \in K, \\ 0, & \mu \in U. \end{cases}$$

Substituting into Equation (105) proves Equation (108). □

H.6. Coefficient-Space Block Diagonalization

We next prove a quantitative block-separation result. This is not a claim about the full nonlinear parameter Hessian of gradient descent. Rather, it is a precise statement about the population quadratic form induced by the rule-feature and fact-feature dictionaries in Theorem H.1. This is the appropriate object for the representation theorem: it measures the functional coupling between the low-rank rule block and the sparse fact block.

Definition H.10 (Rule–fact Gram blocks)**.** Conditional on the fact locations, define the Gram blocks

$$G_{\mathrm{RR}} := \mathbb{E}_{X\sim P_X}\left[\Phi_{\mathrm{R}}(X)\Phi_{\mathrm{R}}(X)^\top\right] \in \mathbb{R}^{r\times r}, \tag{109}$$

$$G_{\mathrm{RF}} := \mathbb{E}_{X\sim P_X}\left[\Phi_{\mathrm{R}}(X)\Phi_{\mathrm{F}}(X)^\top\right] \in \mathbb{R}^{r\times m}, \tag{110}$$

$$G_{\mathrm{FF}} := \mathbb{E}_{X\sim P_X}\left[\Phi_{\mathrm{F}}(X)\Phi_{\mathrm{F}}(X)^\top\right] \in \mathbb{R}^{m\times m}. \tag{111}$$

Equivalently, for coefficient vectors $u \in \mathbb{R}^r$ and $a \in \mathbb{R}^m$,

$$\mathcal{Q}(u,a) := \mathbb{E}_{X\sim P_X}\left[(\langle u, \Phi_{\mathrm{R}}(X)\rangle + \langle a, \Phi_{\mathrm{F}}(X)\rangle)^2\right] \tag{112}$$

has block matrix

$$G := \begin{pmatrix} G_{\mathrm{RR}} & G_{\mathrm{RF}} \\ G_{\mathrm{RF}}^\top & G_{\mathrm{FF}} \end{pmatrix}.$$

The Hessian of $\mathcal{Q}$ with respect to (u,a) is $2G$.

Lemma H.11 (Small rule–fact cross Gram)**.** *On $\mathcal{E}_{\mathrm{geo}}$,*

$$\|G_{\mathrm{RF}}\|_{\mathrm{op}} \le C_{\mathrm{R}}\sqrt{m}\, e^{-25d_\perp/256}, \tag{113}$$

where

$$C_{\mathrm{R}}^2 = \mathbb{E}\big[\|\Phi_{\mathrm{R}}(X)\|_2^2\big] < \infty.$$

Proof. For $a \in [r]$ and $\mu \in [m]$, the (a,μ)-entry of G_{RF} is

$$(G_{\mathrm{RF}})_{a\mu} = \mathbb{E}\big[S_a^\star(X)\, b_\mu(Z^\perp(X))\big].$$

By Cauchy–Schwarz,

$$|(G_{\mathrm{RF}})_{a\mu}| \le \|S_a^\star\|_{L^2(P_X)}\, \|b_\mu(Z^\perp(X))\|_{L^2(P_X)}.$$

Under P_X, $Z^\perp(X) \sim \nu_\perp$, and by Theorem E.7,

$$\|b_\mu(Z^\perp(X))\|_{L^2(P_X)} \le e^{-25d_\perp/256}$$

on $\mathcal{E}_{\mathrm{geo}}$. Therefore

$$|(G_{\mathrm{RF}})_{a\mu}| \le \|S_a^\star\|_{L^2(P_X)} e^{-25d_\perp/256}.$$

Taking the Frobenius norm gives

$$\begin{aligned} \|G_{\mathrm{RF}}\|_{\mathrm{op}} &\le \|G_{\mathrm{RF}}\|_{\mathrm{F}} \\ &\le \left(\sum_{a=1}^{r}\sum_{\mu=1}^{m} \|S_a^\star\|_{L^2(P_X)}^2 e^{-25d_\perp/128}\right)^{1/2} \\ &= C_{\mathrm{R}}\sqrt{m}\, e^{-25d_\perp/256}. \end{aligned}$$

□

Lemma H.12 (Small fact Gram)**.** *On $\mathcal{E}_{\mathrm{geo}}$,*

$$\|G_{\mathrm{FF}}\|_{\mathrm{op}} \le e^{-25d_\perp/128} + (m-1)e^{-75d_\perp/224}. \tag{114}$$

Proof. The diagonal entries of G_{FF} satisfy

$$(G_{\mathrm{FF}})_{\mu\mu} = \mathbb{E}\big[b_\mu(Z^\perp(X))^2\big] \le e^{-25d_\perp/128}$$

by Theorem E.7. For $\mu \neq \nu$,

$$|(G_{\mathrm{FF}})_{\mu\nu}| = \mathbb{E}\big[b_\mu(Z^\perp(X))b_\nu(Z^\perp(X))\big] \le e^{-75d_\perp/224}$$

by Theorem E.8. Since G_{FF} is a symmetric matrix, Gershgorin's circle theorem gives

$$\|G_{\mathrm{FF}}\|_{\mathrm{op}} \le \max_{\mu\in[m]} \left(|(G_{\mathrm{FF}})_{\mu\mu}| + \sum_{\nu\neq\mu} |(G_{\mathrm{FF}})_{\mu\nu}| \right) \le e^{-25d_\perp/128} + (m-1)e^{-75d_\perp/224}.$$

□

Theorem H.13 (Coefficient-space block localization)**.** *On $\mathcal{E}_{\mathrm{geo}}$, the coefficient-space Hessian of $\mathcal{Q}$ satisfies*

$$\nabla^2_{(u,a)}\mathcal{Q} = 2\begin{pmatrix} G_{\mathrm{RR}} & G_{\mathrm{RF}} \\ G_{\mathrm{RF}}^\top & G_{\mathrm{FF}} \end{pmatrix},$$

with

$$\|G_{\mathrm{RF}}\|_{\mathrm{op}} \le C_{\mathrm{R}}\sqrt{m}\, e^{-25d_\perp/256} \tag{115}$$

and

$$\|G_{\mathrm{FF}}\|_{\mathrm{op}} \le e^{-25d_\perp/128} + (m-1)e^{-75d_\perp/224}. \tag{116}$$

Consequently, if

$$m\, e^{-25d_\perp/128} \to 0,$$

then

$$\|G_{\mathrm{RF}}\|_{\mathrm{op}} \to 0 \qquad \textit{and} \qquad \|G_{\mathrm{FF}}\|_{\mathrm{op}} \to 0.$$

Proof. The expression for the Hessian follows by differentiating the quadratic form

$$\mathcal{Q}(u,a) = u^\top G_{\mathrm{RR}} u + 2u^\top G_{\mathrm{RF}} a + a^\top G_{\mathrm{FF}} a.$$

The two operator-norm bounds are exactly Theorems H.11 and H.12.

For the asymptotic statement, the assumption $me^{-25d_\perp/128} \to 0$ implies

$$\sqrt{m}e^{-25d_\perp/256} = \big(me^{-25d_\perp/128}\big)^{1/2} \to 0,$$

so $\|G_{\mathrm{RF}}\|_{\mathrm{op}} \to 0$. It also implies

$$e^{-25d_\perp/128} \to 0.$$

Moreover,

$$(m-1)e^{-75d_\perp/224} = me^{-25d_\perp/128} \cdot e^{-(75/224-25/128)d_\perp} \cdot \frac{m-1}{m} \to 0,$$

because $75/224 - 25/128 = 125/896 > 0$. Hence $\|G_{\mathrm{FF}}\|_{\mathrm{op}} \to 0$. □

Remark H.14 (Why this is the right Hessian object). Theorem H.13 concerns the Hessian of the population squared norm over the fixed rule/fact dictionaries $(\Phi_{\mathrm{R}}, \Phi_{\mathrm{F}})$. It does not assert that the full nonlinear parameter Hessian of the transformer is block diagonal around every parameterization. The theorem instead proves the structural fact needed for this paper: in the constructed coexistence representation, rule variations and fact variations have vanishing population coupling in the high-dimensional regime.

H.7. Summary of the Trustworthiness Consequence

Combining the preceding results yields the exact structural statement used in the main text.

Corollary H.15 (Structural localizability implies exact residual deletion)**.** *Under the hypotheses of Theorem F.4, with the same high probability as in Equation (72), there exists a predictor $f_{\Theta_F} \in \mathcal{T}_{\mathrm{dec}}(H, W_f, B)$ satisfying:*

1. *it exactly memorizes all facts:*
$$f_{\Theta_F}(X^\mu) = \xi^\mu \qquad \forall \mu \in [m];$$
2. *it has vanishing excess rule risk under the scaling of Theorem F.6;*
3. *it decomposes as*
$$f_{\Theta_F} = f^\star + g_F \circ Z^\perp,$$
where $f^\star$ is implemented by the rule block and $g_F \circ Z^\perp$ by at most $3m$ residual-only MLP units;
4. *deleting those residual-only MLP output coefficients gives*
$$f_{\Theta_F^{(-\mathrm{all})}} = f^\star$$
exactly, and therefore restores Bayes-optimal rule risk while erasing all residual fact memorization energy.

Proof. The existence and exact memorization claims follow from Theorem F.4. The decomposition and sparsity claims follow from Theorem H.3. The exact deletion statement follows from Theorem H.7, and the residual-memorization statement follows from Theorem H.9. The high-probability event is the same event $\mathcal{E}_{\mathrm{geo}}$ used in Theorem F.4. □

Remark H.16 (Interpretation for memorization and trustworthiness). The deletion result is exact because the theorem is about the constructed coexistence representation, not because arbitrary trained transformers are guaranteed to possess such a clean decomposition. Its significance is conceptual: it exhibits a regime in which Bayes-optimal rule generalization and exact fact memorization coexist without entanglement. In this regime, memorized residuals are not merely detectable; they are structurally localized and removable without damaging the rule predictor.