A Note on k-NN Gating in RAG

By Gérard BIAU

Sorbonne Université, Institut universitaire de France

gerard.biau@sorbonne-universite.fr

and Claire BOYER

Université Paris-Saclay, Institut Universitaire de France

claire.boyer@universite-paris-saclay.fr

Summary

We develop a statistical proxy framework for retrieval-augmented generation (RAG), designed to formalize how a language model (LM) should balance its own predictions with retrieved evidence. For each query x, the system combines a frozen base model $q_0(\cdot \mid x)$ with a k-nearest neighbor retriever $\hat{r}^{(k)}(\cdot \mid x)$ through a measurable gate $\lambda(x)$. A retrieval-trust weight $w_{\text{fact}}(x)$ quantifies the geometric reliability of the retrieved neighborhood and penalizes retrieval in low-trust regions. We derive the Bayes-optimal per-query gate and analyze its effect on a discordance-based hallucination criterion that captures disagreements between LM predictions and retrieved evidence. We further show that this discordance admits a deterministic asymptotic limit governed solely by the structural agreement (or disagreement) between the Bayes rule and the LM. To account for distribution mismatch between queries and memory, we introduce a hybrid geometric-semantic model combining covariate deformation and label corruption. Overall, this note provides a principled statistical foundation for factuality-oriented RAG systems.

Some key words: Adaptive gating; Hallucination control; Nearest neighbors; Retrieval-augmented generation; Statistical learning.

1. Introduction

Modern language models (LMs) are impressively fluent and versatile, yet they can hallucinate, producing outputs that sound convincing but are factually wrong [Ji et al., 2023, Kalai and Vempala, 2024]. Retrieval-augmented generation (RAG) mitigates this issue by enriching model predictions with information drawn from an external memory, such as curated documents, code repositories, or previously answered queries. These sources are intended to provide relevant and potentially reliable contextual evidence [Lewis et al., 2020]. At inference time, the system retrieves items close to the query in an embedding space and conditions its response on them. While effective in improving accuracy and grounding, this mechanism raises a central design question: for each query, how much should the system rely on the LM vs. the retrieved evidence?

Most current RAG systems address this balance through heuristics. A common approach concatenates the top-k retrieved items into the prompt [Lewis et al., 2020], while others interpolate model and retrieval outputs using a fixed mixture weight, as in cache-based or kNN language models [Grave et al., 2017, Khandelwal et al., 2020, Xu et al., 2023]. Although often effective, these strategies lack adaptive control: when retrieved neighbors are noisy or off-topic, excessive reliance on retrieval may degrade accuracy, while dominant reliance on the LM can lead to underuse of factual evidence and increased hallucination risk. A principled, query-dependent mechanism for balancing the two sources of information is therefore needed.

This note develops a simple and mathematically explicit framework for studying this balancing problem. In our model, the system consists of a frozen base predictor representing the LM, a k-nearest neighbor retriever built from an external memory of labeled examples, and a gate that mixes the two. To make the gating decision interpretable and data-driven, we introduce a *retrieval-trust weight* that quantifies how well the retrieved neighborhood geometrically supports the query. This weight acts as a penalty in training, discouraging retrieval when local evidence is unreliable and yielding an adaptive, geometry-aware gating rule amenable to statistical analysis.

A central theme of the analysis is hallucination control. We introduce a quantitative discordance criterion that captures disagreements between LM predictions and retrieved evidence, and show that the resulting gating rule controls this discordance by activating retrieval only where the local evidence appears reliable. In this way, our results make explicit how the geometry of the retrieved neighborhood and the gating mechanism jointly determine when retrieval improves factual reliability.

To understand retrieval reliability at scale, we analyze the asymptotic regime in which both the memory size n and the number of neighbors k grow with $k/n \to 0$. In this setting we establish consistency for the k-NN retrieval estimator and derive limits for the local discordance signal that governs hallucination variation. These results show that, asymptotically, disagreements between retrieval and the LM reflect genuine structural differences rather than finite-sample noise. We further consider the effect of a mismatch between the query distribution and the retrieval memory.

We emphasize that the model we study is not meant to mirror the architectural details of modern RAG systems, which typically rely on prompt construction, attention over retrieved documents, or other integration mechanisms specific to large LMs. Rather, our goal is to provide an analytically tractable proxy that captures the essential statistical forces governing retrieval-based correction, with the aim of clarifying these mechanisms and stimulating further theoretical work on the foundations of retrieval-augmented generation.

Related work. Our approach connects three lines of research. First, RAG architectures have been extensively explored in NLP and information retrieval [Lewis et al., 2020, Borgeaud et al., 2022, Izacard et al., 2023, Shi et al., 2024]. Second, the retrieval component is rooted in classical nearest neighbor theory [Györfi et al., 2006, Biau and Devroye, 2015], providing a rigorous statistical basis for our analysis. Third, the gating mechanism is related to mixture-of-experts models [Jacobs et al., 1991, Shazeer et al., 2017, Fedus et al., 2022], although here it is explicitly guided by geometric trust rather than latent specialization. Previous approaches to hallucination control include self-consistency checks [Manakul et al., 2023], factuality-oriented evaluation [Min et al., 2023], and LM calibration [Ulmer et al., 2024], but they lack a clear statistical interpretation. The present note provides such an interpretation by linking retrieval geometry, probabilistic gating, and factual reliability within a unified theoretical framework.

2. Model setup

We consider an input-output pair (X, Y) drawn from an unknown distribution on $\mathbb{R}^d \times \mathscr{Y}$, where $\mathscr{Y} = \{1, \dots, C\}$ is a finite label set. An external memory

$$\mathscr{M}_n = \{(U_i, V_i)\}_{i=1}^n, \qquad (U_i, V_i) \sim Q_{UV},$$

stores i.i.d. reference pairs with $U \in \mathbb{R}^d$ and $V \in \mathscr{Y}$. Throughout the main analysis we assume that the memory is drawn from the same distribution as (X, Y)—the *aligned setting*, in which

$P_X = Q_U$ and $P_{Y|X} = Q_{V|U}$. In practice, however, the memory may come from a related but distinct source, such as previously answered queries or labeled documents. For simplicity, we focus on the aligned setting in the main text and return to this more general situation in Appendix B. Retrieval operates in the feature space $\mathbb{R}^d$ equipped with a norm $\|\cdot\|$. For any query $x \in \mathbb{R}^d$, let $U_{(1)}(x), \ldots, U_{(k)}(x)$ denote its k nearest neighbors among $\{U_i\}_{i=1}^n$, with corresponding labels $V_{(1)}(x), \ldots, V_{(k)}(x)$. (Ties are broken deterministically by index order.)

Base predictor and retriever distribution. The base language model (LM) is frozen throughout and provides a conditional probability distribution $q_0(\cdot \mid x)$ on $\mathscr{Y}$, interpreted as an estimate of the law of Y given $X = x$. On the retrieval side, the external memory induces its own conditional structure: for any query x, the retriever constructs a local, nonparametric estimate of the distribution of V given $U = x$ by averaging the labels of the k nearest neighbors of x in the memory:

$$\hat{r}_y^{(k)}(x) = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{V_{(j)}(x)=y\}}, \qquad y \in \mathscr{Y}.$$

We write $\hat{r}^{(k)}(y \mid x) = \hat{r}_y^{(k)}(x)$ for $y \in \mathscr{Y}$, and refer to $\hat{r}^{(k)}(\cdot \mid x)$ as the *retriever distribution*, which approximates the conditional distribution $Q_{V|U}(\cdot \mid x)$ in the neighborhood of x. In the aligned setting considered here, where $P_X = Q_U$ and $P_{Y|X} = Q_{V|U}$, $\hat{r}^{(k)}(\cdot \mid x)$ coincides with the target conditional distribution $P_{Y|X}(\cdot \mid x)$.

Retrieval-trust weight. To quantify how well the retrieved neighborhood reflects the query, we define the retrieval-trust weight

$$w_{\text{fact}}(x) = \frac{1}{k} \sum_{j=1}^{k} \exp\big(- \|x - U_{(j)}(x)\|^2 \big). \tag{1}$$

This scalar measures geometric fidelity between x and its retrieved neighbors. When the neighborhood is compact, the distances $\|x - U_{(j)}(x)\|$ are small, the exponential terms are close to one, and $w_{\text{fact}}(x) \approx 1$, indicating high trust in retrieval. When neighbors are far or inconsistent, $w_{\text{fact}}(x)$ decreases toward zero, signaling that the memory provides unreliable support. Thus $w_{\text{fact}}(x)$ provides a continuous, geometry-aware assessment of the reliability of retrieval at the query location x.

Mixture model. At the core of our framework lies a gated mixture that blends the base LM with the retriever. For each query x, predictions are produced by a convex combination of the LM and the retriever distribution:

$$p_\lambda(y \mid x) = (1 - \lambda(x))\, q_0(y \mid x) + \lambda(x)\, \hat{r}_y^{(k)}(x), \qquad y \in \mathscr{Y}, \tag{2}$$

where the (measurable) gate $\lambda : \mathbb{R}^d \longrightarrow [0,1]$ controls the relative reliance on retrieval. Small values of $\lambda(x)$ favor the LM (fluency and generalization), while values near one defer to retrieved evidence (grounding and factuality). Intermediate values achieve an adaptive trade-off.

Population objective. The population-level loss balances predictive accuracy with trust-dependent regularization:

$$\mathscr{L}(\lambda) = \mathbb{E}\Big[\sum_{y \in \mathscr{Y}} P_{Y|X}(y \mid X)\, (-\log p_\lambda(y \mid X)) \Big] + \zeta\, \mathbb{E}\Big[\lambda(X)\, (1 - w_{\text{fact}}(X)) \Big], \tag{3}$$

where $P_{Y|X}$ is the true conditional distribution and $\zeta \geqslant 0$. The first term is the expected cross-entropy, enforcing predictive fit. The second penalizes retrieval in regions with weak geometric support (where $w_{\text{fact}}(x)$ is small). The hyperparameter ζ controls how strongly the gate penalizes retrieval when the geometric support around the query is weak.

When the query x is surrounded by close and coherent neighbors, $w_{\text{fact}}(x) \approx 1$, so the penalty term $\lambda(x)(1 - w_{\text{fact}}(x))$ becomes negligible and the gate's decision is governed primarily by the cross-entropy comparison between the LM and the retriever. In low-density or out-of-distribution regions, $w_{\text{fact}}(x)$ becomes small, amplifying the penalty and discouraging reliance on retrieval. The mechanism therefore self-regulates: retrieval is driven by predictive fit, while geometric support acts as a safeguard, allowing retrieval to compete freely with the LM when support is strong and otherwise pushing $\lambda(x)$ toward zero.

3. Per-query optimization and hard gating

The population loss (3) can be written as an expectation over the query space, $\mathcal{L}(\lambda) = \mathbb{E}[J(\lambda; X)]$, where, for each fixed $x \in \mathbb{R}^d$,

$$J(\lambda; x) = \ell(\lambda; x) + \zeta\, \lambda(x)\, (1 - w_{\text{fact}}(x)) \quad \text{and} \quad \ell(\lambda; x) = \sum_{y \in \mathcal{Y}} P_{Y|X}(y \mid x)\, (-\log p_\lambda(y \mid x)).$$

Because $\mathcal{L}(\lambda)$ is an expectation of the pointwise objective $J(\lambda; x)$, and λ enters $J(\lambda; x)$ only through its value at the query x, minimizing $\mathcal{L}(\lambda)$ reduces to minimizing $J(\lambda; x)$ independently for each x.

In the simplest and most interpretable implementation, the gate takes binary values $\lambda(x) \in \{0, 1\}$, corresponding to a sharp choice between the LM and the retriever. The mixture model (2) then selects one of its two components:

$$p_\lambda(\cdot \mid x) = \begin{cases} q_0(\cdot \mid x), & \lambda(x) = 0, \\ \hat{r}^{(k)}(\cdot \mid x), & \lambda(x) = 1. \end{cases}$$

Define the corresponding local cross-entropies:

$$\ell_0(x) = \sum_{y \in \mathcal{Y}} P_{Y|X}(y \mid x)\, (-\log q_0(y \mid x)) \quad \text{and} \quad \ell_r(x) = \sum_{y \in \mathcal{Y}} P_{Y|X}(y \mid x)\, (-\log \hat{r}_y^{(k)}(x)).$$

Because under hard gating the decision $\lambda(x) \in \{0, 1\}$ selects either the LM or the retriever, the per-query objective

$$J(\lambda; x) = \begin{cases} \ell_0(x), & \lambda(x) = 0, \\ \ell_r(x) + \zeta\, (1 - w_{\text{fact}}(x)), & \lambda(x) = 1, \end{cases}$$

reduces to a simple two-point comparison. The optimal gate at x is therefore the choice that yields the smaller of these two costs.

Proposition 1 (Optimal hard gate). *For each query $x \in \mathbb{R}^d$, the Bayes-optimal hard gate is*

$$\lambda^\star(x) = \begin{cases} 0, & \text{if } \ell_0(x) \leqslant \ell_r(x) + \zeta\, (1 - w_{\text{fact}}(x)), \\ 1, & \text{if } \ell_r(x) + \zeta\, (1 - w_{\text{fact}}(x)) < \ell_0(x). \end{cases}$$

The rule of Proposition 1 states that retrieval is selected exactly when its improvement in predictive cross-entropy over the LM exceeds the geometric penalty $\zeta(1 - w_{\text{fact}}(x))$. The decision therefore depends jointly on model fit and neighborhood quality. When the retriever distribution $\hat{r}^{(k)}(\cdot \mid x)$ achieves a substantially lower cross-entropy than the LM *and* $w_{\text{fact}}(x)$ is large (i.e., the neighborhood of x is dense), the gate switches to retrieval. Conversely, if neighbors are far,

$w_{\mathrm{fact}}(x)$ is small, the penalty dominates, and the gate keeps $\lambda^\star(x) = 0$. The parameter ζ acts as a global regularizer: large values suppress retrieval and favor the LM, while small values permit more aggressive grounding in memory.

Soft gating. Although binary switching offers direct interpretability, a continuous gate $\lambda(x) \in [0, 1]$ may also be considered. Since $-\log p_\lambda(y \mid x)$ is convex in λ, the per-query objective $J(\lambda; x)$ is convex, and in fact strictly convex whenever $q_0(\cdot \mid x)$ and $\hat{r}^{(k)}(\cdot \mid x)$ differ on the support of $P_{Y|X}(\cdot \mid x)$. The optimal soft gate satisfies the first-order condition

$$\sum_{y \in \mathcal{Y}} P_{Y|X}(y \mid x) \, \frac{\hat{r}_y^{(k)}(x) - q_0(y \mid x)}{p_{\lambda^\star}(y \mid x)} \; + \; \zeta \, (1 - w_{\mathrm{fact}}(x)) = 0,$$

which admits an efficient one-dimensional numerical solution. In practice, however, the hard decision rule of Proposition 1 captures the essential structure of the gating mechanism and facilitates theoretical analysis of how retrieval, geometry, and factual reliability interact.

4. Hallucination and discordance analysis

Hallucination arises when the LM produces confident predictions that contradict evidence present in retrieved neighbors. To quantify this phenomenon, we combine the k-NN retrieval distribution $\hat{r}_y^{(k)}(x)$ with the retrieval-trust weight $w_{\mathrm{fact}}(x)$ defined in (1). For each $x \in \mathbb{R}^d$, let

$$y_r(x) \in \arg\max\nolimits_{y \in \mathcal{Y}} \hat{r}_y^{(k)}(x)$$

denote the retriever's modal label (ties broken deterministically). We define the local discordance score as

$$\mathscr{H}_{\mathrm{disc}}(q_0; x) = w_{\mathrm{fact}}(x) \, (1 - q_0(y_r(x) \mid x)),$$

and the associated population measure

$$\mathscr{H}_{\mathrm{disc}}(q_0) = \mathbb{E}[\mathscr{H}_{\mathrm{disc}}(q_0; X)].$$

This criterion is large when the retrieval neighborhood is geometrically reliable ($w_{\mathrm{fact}}(x) \approx 1$) but the LM assigns low probability to the label favored by retrieval. In this regime, the LM prediction conflicts with locally supported evidence, which we interpret as a risk of hallucination.

4.1. *Change under optimal gating*

Under the mixture model (2), the hallucination score becomes

$$\mathscr{H}_{\mathrm{disc}}(p_\lambda; x) = w_{\mathrm{fact}}(x) \, (1 - p_\lambda(y_r(x) \mid x)).$$

Thus, the variation relative to the frozen LM is

$$\begin{aligned} \Delta\mathscr{H}(x; \lambda) &= \mathscr{H}_{\mathrm{disc}}(q_0; x) - \mathscr{H}_{\mathrm{disc}}(p_\lambda; x) \\ &= \lambda(x) \, w_{\mathrm{fact}}(x) \, (\hat{r}_{y_r(x)}^{(k)}(x) - q_0(y_r(x) \mid x)), \end{aligned} \tag{4}$$

which is linear in $\lambda(x)$ and satisfies $|\Delta\mathscr{H}(x; \lambda)| \leqslant \lambda(x) \, w_{\mathrm{fact}}(x)$. So, the sign of $\Delta\mathscr{H}(x; \lambda)$ indicates whether gating reduces ($\geqslant 0$) or increases ($\leqslant 0$) local discordance.

Recall that, under hard gating, the optimal decision rule from Proposition 1 is

$$\lambda^\star(x) = \mathbb{1}_{\{\ell_r(x) + \zeta(1 - w_{\mathrm{fact}}(x)) < \ell_0(x)\}}, \tag{5}$$

where $\ell_0(x)$ and $\ell_r(x)$ are the LM and retriever local cross-entropy, respectively. Substituting (5) into (4) yields the realized pointwise change

$$\Delta\mathscr{H}(x;\lambda^\star) = \mathbb{1}_{\{\ell_r(x)+\zeta(1-w_{\text{fact}}(x))<\ell_0(x)\}}\, w_{\text{fact}}(x)\, (\hat{r}^{(k)}_{y_r(x)}(x) - q_0(y_r(x) \mid x)). \tag{6}$$

Interpretation via three regimes. Equation (6) reveals three qualitatively distinct behaviors governing how gating affects hallucination.

(i) Gain region. On

$$\mathscr{A} = \{\ell_r + \zeta(1-w_{\text{fact}}) < \ell_0,\ \hat{r}^{(k)}_{y_r} \geqslant q_0(y_r)\},$$

the retriever achieves a lower cross-entropy and assigns higher (or equal) mass to its own top label than the LM. Retrieval therefore improves predictive fit while reinforcing factual evidence, implying $\Delta\mathscr{H}(x;\lambda^\star) \geqslant 0$. The improvement is largest when $w_{\text{fact}}(x) \approx 1$ and is bounded above by $\mathscr{H}_{\text{disc}}(q_0;x)$.

(ii) Trade-off region. On

$$\mathscr{B} = \{\ell_r + \zeta(1-w_{\text{fact}}) < \ell_0,\ \hat{r}^{(k)}_{y_r} < q_0(y_r)\},$$

the retriever improves the cross-entropy but places less mass on its modal label than the LM. Here gating increases discordance ($\Delta\mathscr{H}(x;\lambda^\star) \leqslant 0$), exhibiting a fundamental tension between likelihood and factual alignment. The penalty $\zeta(1-w_{\text{fact}}(x))$ mitigates these cases: when retrieval is geometrically unreliable, the penalty rises and suppresses harmful switches.

(iii) No-switch region. On

$$\mathscr{C} = \{\ell_r + \zeta(1-w_{\text{fact}}) \geqslant \ell_0\},$$

the LM remains active ($\lambda^\star(x) = 0$) and $\Delta\mathscr{H}(x;\lambda^\star) = 0$. This region corresponds to sparse or out-of-distribution queries where retrieval cannot overcome its geometric penalty.

In summary, $w_{\text{fact}}(x)$ plays a dual role: it boosts the benefit of switching in high-confidence regions through its multiplicative factor and simultaneously reduces the risk of harmful switches by amplifying the penalty where retrieval is unreliable.

4.2. *Asymptotic analysis of discordance*

The decomposition in the previous section showed that the pointwise change

$$\Delta\mathscr{H}(x;\lambda^\star) = \lambda^\star(x)\, w_{\text{fact}}(x)\, (\hat{r}^{(k)}_{y_r(x)}(x) - q_0(y_r(x) \mid x))$$

is entirely governed by the inner quantity

$$\Delta(x) = \hat{r}^{(k)}_{y_r(x)}(x) - q_0(y_r(x) \mid x),$$

whose sign determines whether the optimal gate reduces or increases the local hallucination score. Its interpretation differs sharply across the switching regions $\mathscr{A}$ and $\mathscr{B}$. On $\mathscr{A}$, $\Delta(x) \geqslant 0$ corresponds to a desirable gain: retrieval improves the local cross-entropy and assigns higher mass to its own top label. On $\mathscr{B}$, however, we have $\Delta(x) < 0$ despite a cross-entropy improvement, and it is unclear whether this reflects a genuine semantic disagreement between Bayes and the LM, or merely finite-sample variability of the k-NN estimator.

To clarify this, we analyze the asymptotic behavior of $\Delta\mathscr{H}(x;\lambda^\star)$. Recall that we consider the aligned setting, where the query and memory distributions coincide, $P_X = Q_U$ and $P_{Y|X} = Q_{V|U}$. The finite-sample mode stability results established below imply that, when $k \longrightarrow \infty$ and $k/n \longrightarrow 0$, the retriever distribution $\hat{r}^{(k)}(\cdot \mid x)$ converges in probability to the Bayes conditional distribution $P_{Y|X}(\cdot \mid x)$, and the induced modal label $y_r(x)$ converges in probability to the Bayes label $y^\star(x)$. Because $w_{\text{fact}}(x) \to 1$ on the support, the asymptotic behavior of $\Delta\mathscr{H}(x;\lambda^\star)$ is then determined solely by the structural difference

$$P_{Y|X}(y^\star(x) \mid x) \; - \; q_0(y^\star(x) \mid x).$$

As a consequence, any persistent negativity of $\Delta\mathscr{H}(x;\lambda^\star)$ on $\mathscr{B}$ must be structural—stemming from a true mismatch between the Bayes rule and the LM at x—rather than a by-product of k-NN fluctuations.

For notational convenience, let $C := |\mathscr{Y}|$. For $x \in \mathbb{R}^d$ and $k \geq 1$, we denote by

$$R_k(x) \;=\; \max_{1 \leqslant j \leqslant k} \|U_{(j)}(x) - x\|$$

the k-nearest-neighbor radius of the query x among the database points.

Proposition 2 (Finite-sample mode stability). *Fix $x \in \operatorname{supp}(Q_U)$ and assume that the conditional distribution $P_{Y|X}(\cdot \mid \cdot)$ is L-Lipschitz in its second argument. Then, for all $\delta \in (0,1)$,*

$$\begin{aligned} &\mathbb{P}\Big(\max_{y\in\mathscr{Y}} \big|\hat{r}_y^{(k)}(x) - P_{Y|X}(y \mid x)\big| > \delta\Big) \\ &\qquad \leqslant \; 2C \, \exp\Big(-2k\Big(\frac{\delta}{2}\Big)^2\Big) \; + \; \mathbb{P}\Big(R_k(x) > \frac{\delta}{2L}\Big). \end{aligned}$$

In particular, if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, then

$$\max_{y\in\mathscr{Y}} \big|\hat{r}_y^{(k)}(x) - P_{Y|X}(y \mid x)\big| \;\xrightarrow{\mathbb{P}}\; 0.$$

This proposition provides a uniform finite-sample bound on the deviation $\hat{r}^{(k)}(\cdot \mid x) - P_{Y|X}(\cdot \mid x)$ at a fixed query x, valid whenever $P_{Y|X}$ is locally Lipschitz and x lies in the support of the retrieval distribution. In particular, the proposition shows that, as soon as k grows while $k/n \to 0$, the k-NN estimate concentrates around its Bayes target uniformly over labels. This uniform control is precisely what is needed to guarantee that, for large samples, the empirical ordering of the coordinates of $\hat{r}^{(k)}(\cdot \mid x)$ matches the ordering of the Bayes vector $P_{Y|X}(\cdot \mid x)$. The next corollary makes this consequence explicit by showing that the empirical top label $y_r(x)$ converges in probability to the Bayes-optimal label $y^\star(x)$ whenever the latter is unique.

Corollary 1 (Asymptotic mode stability). *Fix $x \in \operatorname{supp}(Q_U)$ and assume that the Bayes label $y^\star(x) \in \arg\max_{y\in\mathscr{Y}} P_{Y|X}(y \mid x)$ is unique, so that*

$$\gamma(x) \;:=\; \max_{y\in\mathscr{Y}} P_{Y|X}(y \mid x) \;-\; \max_{y\neq y^\star(x)} P_{Y|X}(y \mid x) \;>\; 0.$$

Then, under the conditions of Proposition 2, if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$,

$$\mathbb{P}(y_r(x) \neq y^\star(x)) \;\longrightarrow\; 0.$$

Asymptotic behavior of the local discordance $\Delta(x)$. We are in a position to analyze the large-sample behavior of the key quantity $\Delta(x)$, which determines the sign and magnitude of the hallucination variation $\Delta\mathscr{H}(x;\lambda^\star)$ in (6).

Fix $x \in \operatorname{supp}(Q_U)$ and assume that the Bayes label $y^\star(x)$ is unique, so that $\gamma(x) > 0$. By Proposition 2, if $k \to \infty$ and $k/n \to 0$, then

$$\max_{y\in\mathscr{Y}} \big|\hat{r}_y^{(k)}(x) - P_{Y|X}(y \mid x)\big| \;\xrightarrow{\mathbb{P}}\; 0.$$

Therefore

$$\hat{r}^{(k)}_{y_r(x)}(x) - P_{Y|X}(y_r(x) \mid x) \xrightarrow{\mathbb{P}} 0.$$

Recalling $\Delta(x) = \hat{r}^{(k)}_{y_r(x)}(x) - q_0(y_r(x) \mid x)$, this implies

$$\Delta(x) - \big(P_{Y|X}(y_r(x) \mid x) - q_0(y_r(x) \mid x)\big) \xrightarrow{\mathbb{P}} 0.$$

Moreover, Corollary 1 yields $\mathbb{P}\big(y_r(x) = y^\star(x)\big) \longrightarrow 1$. On the event $\{y_r(x) = y^\star(x)\}$, $P_{Y|X}(y_r(x) \mid x) - q_0(y_r(x) \mid x) = P_{Y|X}(y^\star(x) \mid x) - q_0(y^\star(x) \mid x)$. Hence,

$$P_{Y|X}(y_r(x) \mid x) - q_0(y_r(x) \mid x) \xrightarrow{\mathbb{P}} P_{Y|X}(y^\star(x) \mid x) - q_0(y^\star(x) \mid x),$$

and combining with the previous display gives

$$\Delta(x) \xrightarrow{\mathbb{P}} P_{Y|X}(y^\star(x) \mid x) - q_0(y^\star(x) \mid x). \tag{7}$$

In particular, the sign of $\Delta(x)$ coincides with the sign of $P_{Y|X}(y^\star(x) \mid x) - q_0(y^\star(x) \mid x)$ with probability tending to one.

Asymptotic behavior of the local hallucination variation. We are now equipped to describe the asymptotic behavior of the local hallucination variation $\Delta\mathscr{H}(x;\lambda^\star)$ in (6). The result below shows that, for fixed x in the support, the sign and magnitude of $\Delta\mathscr{H}(x;\lambda^\star)$ converge in probability to a deterministic quantity governed solely by the Bayes predictor and the LM. We let

$$\ell_{\mathrm{Bayes}}(x) = \sum_{y \in \mathscr{Y}} P_{Y|X}(y \mid x)\,(-\log P_{Y|X}(y \mid x))$$

and recall that the LM local cross-entropy is $\ell_0(x) = \sum_{y \in \mathscr{Y}} P_{Y|X}(y \mid x)\,(-\log q_0(y \mid x))$.

Theorem 1 (Asymptotic behavior of the local hallucination variation). *Under the aligned setting, let $x \in \operatorname{supp}(Q_U)$. Suppose that the Bayes label $y^\star(x) \in \arg\max_{y\in\mathscr{Y}} P_{Y|X}(y \mid x)$ is unique. Assume moreover that $P_{Y|X}(\cdot \mid \cdot)$ is L-Lipschitz in its second argument and that $\ell_{\mathrm{Bayes}}(x) \neq \ell_0(x)$. Then, if $k \to \infty$ and $k/n \to 0$ as $n \to \infty$,*

$$\Delta\mathscr{H}(x;\lambda^\star) \xrightarrow{\mathbb{P}} \lambda_\infty(x)\,(P_{Y|X}(y^\star(x) \mid x) - q_0(y^\star(x) \mid x)),$$

where $\lambda_\infty(x) := \mathbb{1}_{\{\ell_{\mathrm{Bayes}}(x) < \ell_0(x)\}}$.

Thus, under mode stability, the local hallucination variation $\Delta\mathscr{H}(x;\lambda^\star)$ converges in probability to a deterministic quantity determined solely by the structural relationship between the Bayes rule and the LM at x. In the limit, the randomness of both the k-NN estimator and the factuality weight $w_{\mathrm{fact}}(x)$ disappears, and the gating decision becomes entirely governed by the comparison between the Bayes cross-entropy $\ell_{\mathrm{Bayes}}(x)$ and the LM cross-entropy $\ell_0(x)$.

More precisely, when the switching condition $\ell_{\mathrm{Bayes}}(x) < \ell_0(x)$ holds, the gate eventually activates with probability tending to one, and

$$\Delta\mathscr{H}(x;\lambda^\star) \longrightarrow P_{Y|X}(y^\star(x) \mid x) - q_0(y^\star(x) \mid x).$$

When $\ell_{\mathrm{Bayes}}(x) \geqslant \ell_0(x)$, the gate eventually remains off, so that the variation $\Delta\mathscr{H}(x;\lambda^\star) \to 0$ even if the difference $P_{Y|X}(y^\star(x) \mid x) - q_0(y^\star(x) \mid x)$ is negative.

In either case, any nonvanishing improvement or deterioration of the hallucination score in the large-sample regime reflects a genuine structural relationship between the Bayes predictor and the LM at x, rather than finite-sample instability of the k-NN estimator.

References

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer, Cham, 2015.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, et al. Improving language models by retrieving from trillions of tokens. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240. PMLR, 2022.

William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

Edouard Grave, Moustapha M Cisse, and Armand Joulin. Unbounded cache model for online language modeling with open vocabulary. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna Wallach, Rob Fergus, S. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30, pages 6044–6054. Curran Associates, Inc., 2017.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, New York, 2006.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, et al. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Robert A. Jacobs, Michael I. Jordan, Steven J. Nowlan, and Geoffrey I. Hinton. Adaptive mixtures of local experts. In *Neural Computation*, volume 3, pages 79–87, 1991.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, et al. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55:248,1–38, 2023.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the 56th Annual ACM Symposium on Theory of Computing*, STOC 2024, pages 160–171, New York, 2024. Association for Computing Machinery.

Urvashi Khandelwal, Omer Levy, Dan Jurafsky, Luke Zettlemoyer, and Mike Lewis. Generalization through memorization: Nearest neighbor language models. In *International Conference on Learning Representations*, 2020.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Haibin Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc., 2020.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, page 9004–9017. Association for Computational Linguistics, 2023.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, et al. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100. Association for Computational Linguistics, 2023.

Noam Shazeer, Azalia Mirhoseini, Piotr Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *International Conference on Learning Representations*, 2017.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, et al. REPLUG: Retrieval-augmented black-box language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8371–8384. Association for Computational Linguistics, 2024.

Lakpa Tamang, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. Handling out-of-distribution data: A survey. *arXiv:2507.21160*, 2025.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. Calibrating large language models using their generations only. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459. Association for Computational Linguistics, 2024.

Frank F. Xu, Uri Alon, and Graham Neubig. Why do nearest neighbor language models work? In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 38325–38341. JMLR, 2023.

Kaiyang Zhou, Ziwei Liu, Yu Qiao, Tao Xiang, and Loy Chen Change. Domain generalization: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45:4396–4415, 2023.

A. Proofs

A.1. *Proof of Proposition 2*

Conditionally on the neighbor locations $U_{(1)}(x), \ldots, U_{(k)}(x)$, the variables

$$Z_{j,y} := \mathbb{1}_{\{V_{(j)}(x)=y\}}, \qquad y \in \mathscr{Y},$$

are independent Bernoulli with means $P_{Y|X}(y \mid U_{(j)}(x))$, and $\hat{r}_y^{(k)}(x) = k^{-1} \sum_{j=1}^k Z_{j,y}$. Therefore, by Hoeffding's inequality and a union bound over all classes, for any $\delta \in (0,1)$,

$$\mathbb{P}\Big(\max_{y \in \mathscr{Y}} \Big| \hat{r}_y^{(k)}(x) - \frac{1}{k} \sum_{j=1}^k P_{Y|X}(y \mid U_{(j)}(x)) \Big| > \frac{\delta}{2} \;\Big|\; U_{(1)}(x), \ldots, U_{(k)}(x) \Big) \leqslant 2C \exp\Big(-2k \Big(\tfrac{\delta}{2}\Big)^2 \Big). \tag{8}$$

Dropping the conditioning yields the same bound unconditionally.

Next define the local modulus of continuity

$$\omega_x(r) := \sup_{\|u-x\| \leqslant r} \max_{y \in \mathscr{Y}} |P_{Y|X}(y \mid u) - P_{Y|X}(y \mid x)|.$$

Clearly,

$$\max_{y \in \mathscr{Y}} \Big| \frac{1}{k} \sum_{j=1}^k P_{Y|X}(y \mid U_{(j)}(x)) - P_{Y|X}(y \mid x) \Big| \leqslant \omega_x(R_k(x)).$$

If $P_{Y|X}$ is L-Lipschitz, then $\omega_x(r) \leqslant Lr$, and combining this with (8) yields

$$\begin{aligned} &\mathbb{P}\Big(\max_{y \in \mathscr{Y}} \big| \hat{r}_y^{(k)}(x) - P_{Y|X}(y \mid x) \big| > \delta \Big) \\ &\qquad \leqslant \; 2C \exp\Big(-2k \Big(\tfrac{\delta}{2}\Big)^2 \Big) \; + \; \mathbb{P}\Big(R_k(x) > \tfrac{\delta}{2L} \Big). \end{aligned} \tag{9}$$

Finally, since $x \in \operatorname{supp}(Q_U)$, the local mass $p_{x,\varepsilon} := Q_U(B(x,\varepsilon))$ is strictly positive for every $\varepsilon > 0$. Because

$$\{R_k(x) > \varepsilon\} = \{\operatorname{Bin}(n, p_{x,\varepsilon}) < k\},$$

Chernoff's bound [e.g., Biau and Devroye, 2015, Chapter 20] gives, whenever $k \leqslant \frac{1}{2} n p_{x,\varepsilon}$,

$$\mathbb{P}\big(R_k(x) > \varepsilon \big) \;\leqslant\; \exp\Big(-\tfrac{n p_{x,\varepsilon}}{8} \Big).$$

Because $k/n \to 0$ implies $k \leqslant \frac{1}{2} n p_{x,\varepsilon}$ for large n, the right-hand side tends to zero. Combining this with (9) gives the desired convergence.

A.2. *Proof of Corollary 1*

Fix $x \in \operatorname{supp}(Q_U)$ and assume that the Bayes label $y^\star(x)$ is unique, so that $\gamma(x) > 0$. Set $\varepsilon := \frac{\gamma(x)}{3}$. By Proposition 2, applied with this choice of ε, we obtain

$$\mathbb{P}\Big(\max_{y \in \mathscr{Y}} \big| \hat{r}_y^{(k)}(x) - P_{Y|X}(y \mid x) \big| > \varepsilon \Big) \;\longrightarrow\; 0 \qquad \text{as } n \to \infty,$$

whenever $k \to \infty$ and $k/n \to 0$. Now, define the event

$$A_n \;:=\; \Big\{ \max_{y \in \mathscr{Y}} \big| \hat{r}_y^{(k)}(x) - P_{Y|X}(y \mid x) \big| \leqslant \varepsilon \Big\},$$

so that $\mathbb{P}(A_n^c) \to 0$ as $n \to \infty$. On A_n, for any $y_1, y_2 \in \mathscr{Y}$,

$$\hat{r}_{y_1}^{(k)}(x) - \hat{r}_{y_2}^{(k)}(x) \;\geqslant\; P_{Y|X}(y_1 \mid x) - P_{Y|X}(y_2 \mid x) \;-\; 2\varepsilon.$$

In particular, taking $y_1 = y^\star(x)$ and any $y_2 \neq y^\star(x)$ gives

$$P_{Y|X}(y^\star(x) \mid x) - P_{Y|X}(y_2 \mid x) \;\geqslant\; \gamma(x) \;=\; 3\varepsilon,$$

hence

$$\hat{r}^{(k)}_{y^\star(x)}(x) - \hat{r}^{(k)}_{y_2}(x) \;\geqslant\; 3\varepsilon - 2\varepsilon \;=\; \varepsilon \;>\; 0.$$

Thus, on A_n,

$$\hat{r}^{(k)}_{y^\star(x)}(x) > \hat{r}^{(k)}_{y}(x) \qquad \forall\, y \neq y^\star(x),$$

and therefore the empirical and Bayes top labels coincide:

$$y_r(x) = \arg\max_{y\in\mathscr{Y}} \hat{r}^{(k)}_y(x) = \arg\max_{y\in\mathscr{Y}} P_{Y|X}(y \mid x) = y^\star(x).$$

Consequently,

$$\mathbb{P}(y_r(x) \neq y^\star(x)) \;\leqslant\; \mathbb{P}(A_n^c) \;\longrightarrow\; 0 \qquad \text{as } n \to \infty,$$

which establishes the claim.

A.3. *Proof of Theorem 1*

Since $x \in \operatorname{supp}(Q_U)$ and $k/n \to 0$, Proposition 3 yields

$$w_{\text{fact}}(x) \xrightarrow{\mathbb{P}} 1. \tag{10}$$

Next, recall that

$$\lambda^\star(x) = \mathbb{1}_{\{\ell_r(x) + \zeta(1 - w_{\text{fact}}(x)) < \ell_0(x)\}},$$

where

$$\ell_r(x) = \sum_{y\in\mathscr{Y}} P_{Y|X}(y \mid x)\,(-\log \hat{r}^{(k)}_y(x)).$$

By Proposition 2 and the convention $0 \log 0 = 0$, continuity of log on labels with $P_{Y|X}(y \mid x) > 0$ implies

$$\ell_r(x) \;\xrightarrow{\mathbb{P}}\; \ell_{\text{Bayes}}(x).$$

Using (10), we obtain

$$\ell_r(x) + \zeta(1 - w_{\text{fact}}(x)) \;\xrightarrow{\mathbb{P}}\; \ell_{\text{Bayes}}(x).$$

Since $\ell_{\text{Bayes}}(x) \neq \ell_0(x)$ by assumption, the indicator

$$\mathbb{1}_{\{\ell_r(x) + \zeta(1 - w_{\text{fact}}(x)) < \ell_0(x)\}}$$

converges in probability to

$$\lambda_\infty(x) = \mathbb{1}_{\{\ell_{\text{Bayes}}(x) < \ell_0(x)\}}.$$

Combining this limit with (7) proves the theorem.

B. Query-memory distribution mismatch

Throughout the main text, we worked in the aligned setting, in which the distribution of query inputs coincides with that of the memory inputs ($P_X = Q_U$), and the conditional label mechanisms also agree ($P_{Y|X} = Q_{V|U}$). In realistic retrieval-augmented systems, however, this alignment rarely holds. The retrieval memory $\mathscr{M}_n = \{(U_i, V_i)\}_{i=1}^n$ is typically built from a large and possibly heterogeneous corpus, whereas incoming queries (X, Y) may follow a distinct or temporally evolving distribution. This discrepancy gives rise to two fundamental sources of mismatch: a geometric shift, where the distribution of query embeddings differs from that of stored items, and a semantic drift, where the mapping between inputs and labels in memory no longer reflects that of the current environment.

Such deviations are well known in the broader machine-learning literature. The former corresponds to covariate shift, while the latter is related to semantic shift, both extensively studied in the context of domain adaptation and domain generalization; see, for example, Zhou et al. [2023], Tamang et al. [2025]. In RAG, these mismatches manifest concretely as the retrieval of outdated, irrelevant, or off-distribution items—conditions strongly associated with factual hallucination.

To analyze these effects within a unified framework, we introduce a *hybrid geometric-semantic mismatch model*, in which the distribution of queries may differ from that of the memory both through geometric deformation of the input space and through semantic misalignment between memory labels and the true query labels. This setting allows us to interpret the retrieval-trust weight $w_{\text{fact}}(x)$ as an implicit correction mechanism that naturally downweights unreliable retrievals.

B.1. *A hybrid geometric-semantic mismatch model*

We characterize query-memory mismatch by distinguishing the distribution of query inputs P_X from that of memory inputs Q_U, and the query labeling mechanism $P_{Y|X}$ from the memory labeling mechanism $Q_{V|U}$. This perspective allows us to model both geometric distortion in the embedding space and semantic mismatch in the conditional labels.

Geometric shift. Queries are assumed to be generated from memory inputs through a perturbation model:

$$X = T(U) = U + \xi(U),$$

where $U \sim Q_U$ and $\xi : \mathbb{R}^d \to \mathbb{R}^d$ is a deformation function. The distribution $P_X = T_\# Q_U$ thus represents the law of queries obtained by displacing the memory inputs. When computing the k nearest neighbors of a query x, the retrieved points $U_{(1)}(x), \ldots, U_{(k)}(x)$ are the elements of $\{U_i\}_{i=1}^n$ that lie closest to x in the embedding space. If $\|\xi(u)\|$ is small, the neighborhoods of x and u largely overlap; as $\|\xi(u)\|$ increases, retrieved neighbors become less representative of x, capturing the geometric component of the mismatch.

Label drift. Even when geometric distortion is negligible, the conditional relationship between features and labels in memory may differ from that of current queries. We model this semantic deviation as a local corruption process:

$$Q_{V|U}(y \mid u) = (1 - \rho(u))\, P_{Y|X}(y \mid u) + \rho(u)\, s(y \mid u), \qquad 0 \leqslant \rho(u) \leq 1,$$

where $\rho(u)$ is a local corruption rate and $s(\cdot \mid u)$ an arbitrary spurious distribution. Thus the retrieval distribution constructed from $\mathscr{M}_n$ approximates a corrupted version of $P_{Y|X}$: the memory is reliable when $\rho(u)$ is small and increasingly misleading as $\rho(u)$ grows.

Retrieval trust and interpretation. The two mechanisms combine into the coupled model

$$U \sim Q_U, \quad X = T(U), \quad V \mid U \sim Q_{V|U}(\cdot \mid U), \quad Y \mid X \sim P_{Y|X}(\cdot \mid X). \tag{11}$$

Both types of shift influence the reliability of retrieval, but in different ways. The retrieval-trust weight $w_{\text{fact}}(x)$, defined in (1), measures the geometric compatibility between the query x and its retrieved neighbors. Large geometric deformations $\|\xi(u)\|$ inflate the distances $\|x - U_{(j)}(x)\|$ and therefore directly reduce $w_{\text{fact}}(x)$. By contrast, strong semantic corruption $\rho(u)$ does not affect $w_{\text{fact}}(x)$ itself, but makes the retrieved labels unreliable as proxies for the true query labels, even when geometric proximity is high. Thus $w_{\text{fact}}(x)$ should be interpreted as a geometry-aware indicator of retrieval quality, while the retrieval distribution captures the semantic reliability of the retrieved labels under joint geometric and semantic shift.

In the next subsection, we formalize these observations by characterizing the asymptotic behavior of both the retrieval-trust weight $w_{\text{fact}}(x)$ and the k-NN retriever $\hat{r}^{(k)}(\cdot \mid x)$ under mild regularity assumptions on T and ρ. Proposition 3 shows that $w_{\text{fact}}(x)$ converges to a deterministic function of the distance between the query and the memory support, capturing geometric compatibility, while Proposition 4 establishes that the retriever converges to the local memory label distribution at the nearest points of the support. Together, these results give a precise statistical interpretation of the penalty term in (3): retrieval is discouraged precisely in regions where geometric proximity is weak or where the limiting retrieval distribution reflects boundary or corrupted semantics.

B.2. *Geometry of the trust weight under distribution shift*

Let $\{(U_i, V_i)\}_{i=1}^n$ be the memory pairs with $U_i \in \mathbb{R}^d$, drawn i.i.d. from Q_{UV}. Denote by Q_U the marginal law of the U_i's, and by $S = \mathrm{supp}(Q_U)$ its (closed) support. For any query $x \in \mathbb{R}^d$, we let

$$d(x, S) = \inf_{u \in S} \|x - u\|.$$

(Since S is closed, this infimum is attained.)

Proposition 3 (Asymptotic behavior of the trust weight). *Fix $x \in \mathbb{R}^d$. If $k/n \to 0$ as $n \to \infty$, then*

$$w_{\mathrm{fact}}(x) = \frac{1}{k} \sum_{j=1}^{k} \exp\big(- \|x - U_{(j)}(x)\|^2 \big) \xrightarrow{\text{a.s.}} \exp\big(- d(x, S)^2 \big).$$

In particular, if $x \in S$, then $w_{\mathrm{fact}}(x) \xrightarrow{\text{a.s.}} 1$.

Under the geometric component of the hybrid model, $X = T(U) = U + \xi(U)$ with $U \sim Q_U$, a query $x = T(u)$ satisfies $d(x, S) \leqslant \|x - u\| = \|\xi(u)\|$. Therefore, Proposition 3 yields the almost-sure limit

$$w_{\mathrm{fact}}(x) \xrightarrow[n\to\infty]{\text{a.s.}} \exp\big(- d(x, S)^2 \big) \geqslant \exp\big(- \|\xi(u)\|^2 \big),$$

with equality whenever $\|x - u\| = d(x, S)$.

Proposition 4 (Asymptotics of the k-NN retrieval distribution). *Fix $x \in \mathbb{R}^d$. Assume that $k \to \infty$ and $k/n \to 0$ as $n \to \infty$, and define*

$$\hat{r}_y^{(k)}(x) = \frac{1}{k} \sum_{j=1}^{k} \mathbb{1}_{\{V_{(j)}(x) = y\}}, \qquad y \in \mathcal{Y}.$$

Let $S = \mathrm{supp}(Q_U)$ and let

$$N_S(x) = \{u \in S : \|x - u\| = d(x, S)\}$$

be the (possibly non-singleton) set of nearest points in S to x.

(i) **In-support point.** *If $x \in S$ and $Q_{V|U}(y \mid \cdot)$ is continuous at x, then*

$$\hat{r}_y^{(k)}(x) \xrightarrow{\text{a.s.}} Q_{V|U}(y \mid x).$$

(ii) **Unique nearest point off support.** *If $x \notin S$, the nearest set is a singleton $N_S(x) = \{u_x\}$, and $Q_{V|U}(y \mid \cdot)$ is continuous at u_x, then*

$$\hat{r}_y^{(k)}(x) \xrightarrow{\text{a.s.}} Q_{V|U}(y \mid u_x).$$

(iii) **Multiple nearest points.** *If $Q_{V|U}(y \mid \cdot)$ is continuous on $N_S(x)$, then almost surely,*

$$\min_{u \in N_S(x)} Q_{V|U}(y \mid u) \leqslant \liminf_{n\to\infty} \hat{r}_y^{(k)}(x) \leqslant \limsup_{n\to\infty} \hat{r}_y^{(k)}(x) \leqslant \max_{u \in N_S(x)} Q_{V|U}(y \mid u).$$

In particular, if $Q_{V|U}(y \mid u)$ is constant on $N_S(x)$, then $\hat{r}_y^{(k)}(x)$ converges to that common value.

Interpretation under the hybrid shift model. Proposition 4 shows that the empirical retriever $\hat{r}^{(k)}(\cdot \mid x)$ consistently estimates the memory's local label mechanism in the region of the database that is geometrically closest to the query. When $x \in S$, the estimate converges to $Q_{V|U}(\cdot \mid x)$; when $x \notin S$ but admits a unique projection $u_x \in N_S(x)$, it converges to $Q_{V|U}(\cdot \mid u_x)$. Proposition 3 complements this semantic characterization with a geometric one: $w_{\text{fact}}(x) \approx 1$ indicates that x lies in or close to S, whereas small values of $w_{\text{fact}}(x)$ flag out-of-support queries for which retrieval reflects boundary behavior rather than genuine local structure. Thus, under geometric and semantic mismatch, the pair $(\hat{r}^{(k)}, w_{\text{fact}})$ jointly encodes the reliability of retrieved evidence, a property that directly supports and stabilizes the gating rule.

Propositions 3 and 4 yield almost-sure pointwise limits for $w_{\text{fact}}(x)$ and—when the limit exists—for $\hat{r}^{(k)}(\cdot \mid x)$. In particular, in cases (i) and (ii) of Proposition 4, the k-NN retriever admits a deterministic limit $r_\infty(\cdot \mid x)$. Whenever $r_\infty(\cdot \mid x)$ exists, continuity of the mixture implies that, for any measurable $\lambda : \mathbb{R}^d \to [0,1]$,

$$\begin{aligned} p_\lambda(y \mid x) &= (1 - \lambda(x))\, q_0(y \mid x) + \lambda(x)\, \hat{r}_y^{(k)}(x) \\ &\xrightarrow[n\to\infty]{\text{a.s.}}\; p_{\lambda,\infty}(y \mid x) := (1 - \lambda(x))\, q_0(y \mid x) + \lambda(x)\, r_\infty(y \mid x). \end{aligned}$$

Thus, under the hybrid shift model (11), we obtain

$$\begin{aligned} & r_\infty(y \mid x) - P_{Y|X}(y \mid x) \\ & = (1 - \rho(u_x))\big(P_{Y|X}(y \mid u_x) - P_{Y|X}(y \mid x)\big) + \rho(u_x)\big(s(y \mid u_x) - P_{Y|X}(y \mid x)\big), \end{aligned}$$

and, writing $\|\cdot\|_1 = \sum_{y\in\mathscr{Y}} |\cdot|$,

$$\|r_\infty(\cdot \mid x) - P_{Y|X}(\cdot \mid x)\|_1 \;\leqslant\; (1 - \rho(u_x))\, \delta_{\text{geom}}(x) + \rho(u_x)\, \delta_{\text{sem}}(x),$$

where

$$\delta_{\text{geom}}(x) = \|P_{Y|X}(\cdot \mid u_x) - P_{Y|X}(\cdot \mid x)\|_1, \quad \delta_{\text{sem}}(x) = \|s(\cdot \mid u_x) - P_{Y|X}(\cdot \mid x)\|_1.$$

If $P_{Y|X}(\cdot \mid x)$ is locally Lipschitz as a map from $\mathbb{R}^d$ to $(\mathbb{R}^C, \|\cdot\|_1)$, then

$$\delta_{\text{geom}}(x) = \|P_{Y|X}(\cdot \mid u_x) - P_{Y|X}(\cdot \mid x)\|_1 \;\leqslant\; L\, \|x - u_x\| \;\leqslant\; L\, d(x,S).$$

Hence the total retrieval bias is jointly driven by the geometric displacement $d(x,S)$ and the local corruption level $\rho(u_x)$. When x lies near the memory support and corruption is small, the limiting retriever $r_\infty(\cdot \mid x)$ is close to the true conditional distribution, and $w_{\text{fact}}(x) \approx 1$ makes the penalty $\lambda(x)(1 - w_{\text{fact}}(x))$ negligible, so retrieval is not discouraged. As x moves away from the support or corruption increases, $\|r_\infty(\cdot \mid x) - P_{Y|X}(\cdot \mid x)\|_1$ grows while $w_{\text{fact}}(x) \approx e^{-d(x,S)^2}$ shrinks, naturally steering the gate toward the base model.

B.3. *Proof of Proposition 3*

Let $D_n^{(j)}(x) = \|x - U_{(j)}(x)\|$ be the j-th nearest-neighbor distance among $\{U_i\}_{i=1}^n$. By Lemma 2.2 of Biau and Devroye [2015], if $k/n \to 0$ then $D_n^{(k)}(x) \xrightarrow{\text{a.s.}} d(x,S)$ as $n \to \infty$. Since $D_n^{(1)}(x) \leqslant \cdots \leqslant D_n^{(k)}(x)$ and $D_n^{(1)}(x) \geqslant d(x,S)$, we obtain

$$0 \leqslant \max_{1\leqslant j\leqslant k} \left|D_n^{(j)}(x) - d(x,S)\right| \leqslant D_n^{(k)}(x) - d(x,S) \xrightarrow{\text{a.s.}} 0,$$

so $D_n^{(j)}(x) \xrightarrow{\text{a.s.}} d(x,S)$ uniformly over $1 \leqslant j \leqslant k$. With $\varphi(t) = \exp(-t^2)$ continuous, uniform convergence gives

$$\max_{1\leqslant j\leqslant k} \left|\varphi(D_n^{(j)}(x)) - \varphi(d(x,S))\right| \xrightarrow{\text{a.s.}} 0.$$

This shows the desired claim.

B.4. *Proof of Proposition 4*

Let $D_n^{(j)}(x) = \|x - U_{(j)}(x)\|$. As in the proof of Proposition 3, when $k/n \to 0$ as $n \to \infty$,

$$0 \;\leqslant\; \max_{1\leqslant j\leqslant k} |D_n^{(j)}(x) - d(x,S)| \;\leqslant\; D_n^{(k)}(x) - d(x,S) \xrightarrow{\text{a.s.}} 0.$$

Thus the neighbor locations satisfy almost surely: if $x \in S$, then $U_{(j)}(x) \longrightarrow x$; if $x \notin S$ and $N_S(x) = \{u_x\}$, then $U_{(j)}(x) \longrightarrow u_x$; in general, every cluster point of the sequence $\{U_{(j)}(x)\}_{j=1}^k$ lies in $N_S(x)$.

Conditionally on the neighbor locations, the variables $Z_{j,y} := \mathbb{1}_{\{V_{(j)}(x)=y\}}$ are independent Bernoulli with means $Q_{V|U}(y \mid U_{(j)}(x))$, and $\hat{r}_y^{(k)}(x) = k^{-1} \sum_{j=1}^k Z_{j,y}$. Thus, Hoeffding's inequality yields

$$\mathbb{P}\left(\left|\hat{r}_y^{(k)}(x) - \frac{1}{k}\sum_{j=1}^k Q_{V|U}(y \mid U_{(j)}(x))\right| > \varepsilon \;\middle|\; U_{(1)}(x), \ldots, U_{(k)}(x)\right) \leqslant 2e^{-2k\varepsilon^2},$$

hence, as $k \longrightarrow \infty$,

$$\hat{r}_y^{(k)}(x) - \frac{1}{k}\sum_{j=1}^k Q_{V|U}(y \mid U_{(j)}(x)) \xrightarrow{\text{a.s.}} 0.$$

For (i), continuity at $x \in S$ implies

$$\max_{1 \leqslant j \leqslant k} |Q_{V|U}(y \mid U_{(j)}(x)) - Q_{V|U}(y \mid x)| \xrightarrow{\text{a.s.}} 0,$$

so the Cesàro mean converges to $Q_{V|U}(y \mid x)$. The same argument gives (ii).

For (iii), continuity on $N_S(x)$ implies the Cesàro averages of $\{Q_{V|U}(y \mid U_{(j)}(x))\}_{j=1}^k$ must lie within the convex hull of the function values on $N_S(x)$, giving the stated bounds.