HIERARCHY SHIFT: LEARNING WITH CHANGING SEMANTIC STRUCTURE

Anonymous authors
Paper under double-blind review

ABSTRACT

Hierarchical learning usually assumes that the semantic structure used to judge a prediction is fixed. We study the case in which that structure changes at deployment, possibly while the ordinary data law does not. We call this setting *hierarchy shift*. The resulting excess Bayes risk decomposes exactly into a value recoverable from observed context and a residual hierarchy-information gap. We characterize when this gap is zero and show that it is operational: with a finite action space and bounded risks, an empirical incompatibility statistic from m conditionally iid plausible hierarchies converges to the gap at rate $O(\sqrt{\log|\mathcal{A}|/m})$, without uniqueness, margin, metric, or binary-hierarchy assumptions. Plug-in and kernel-robust versions cover estimated risks and hierarchy posteriors. We then relate residual risk to the Bayes error of identifying the active hierarchy from context. The resulting regime law is simple: hierarchy conditioning can help only when plausible hierarchies favor different decisions and context identifies the relevant semantic state. Controlled experiments recover the exact decomposition, and learned gains track the oracle recoverable value ($r = 0.964$). On locked Visual-WSD, the average MRR gain is small but rises to 0.023 on positive-signal queries; shuffling posterior-to-branch alignment reverses this advantage. A low-conflict CIFAR-100 audit yields only a small gain, as predicted.

1 INTRODUCTION

A hierarchical error has a size, not only a sign. Confusing two neighboring animal classes is different from confusing an animal with a vehicle, and the same distinction matters in medical, scientific, and multimodal taxonomies. Hierarchy-aware losses, hyperbolic representations, selective classifiers, and structured transport all exploit this fact (Nickel & Kiela, 2017; Garg et al., 2022; Goren et al., 2024; Lin et al., 2025). They usually make another assumption: the semantic hierarchy used to score a decision at deployment is the one supplied during training.

That assumption can fail even when the label set and input distribution do not. Consider four disease labels organized by anatomy in one hospital workflow and by etiology in an epidemiological workflow. The same misclassification can be close under one organization and cross-branch under the other. Different institutions can therefore judge the same prediction differently; ontologies can be revised; and a polysemous concept can move between semantic branches as context changes. Multiple-taxonomy learning (Hwang et al., 2012) and taxonomy adaptation (Babbar et al., 2016) acknowledge that one taxonomy need not be universal, but they primarily ask how to share or transfer structure. We ask a different question: *what is the decision risk of not knowing which hierarchy is active, and when can context remove that risk?*

We formalize this uncertainty as *hierarchy shift*. Let P be the law of ordinary observables (X, Y, S), where S is context, and let $\Pi(dH \mid X, S)$ describe the semantic hierarchy used to score or interpret the decision. A pure hierarchy shift changes Π while leaving P fixed. Thus the shift is in the structural decision/evaluation environment, not necessarily in the generative law of (X, Y, S). Let R_0^* be the best structural risk without using S or the realized hierarchy, R_S^* the best risk after observing (X, S) but not H, and R_H^* the risk when the active hierarchy is known. Then

$$R_0^* - R_H^* = (R_0^* - R_S^*) + (R_S^* - R_H^*) =: V_S^* + G_H^*. \tag{1}$$

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

V_S^* is the value recoverable from context and G_H^* is the residual hierarchy-information gap. This decomposition turns a broad robustness question into two precise ones: when is hierarchy information decision-relevant, and how much of that information can context recover?

The answer is governed by conflict and identifiability. If all plausible hierarchies share an optimal action, the residual gap is exactly zero. If they are incompatible, the gap is positive. Our main theorem makes that incompatibility operational: for finite actions and bounded risks, the empirical disagreement cost of m plausible hierarchies estimates $G_Q(x, s)$ with an explicit $m^{-1/2}$ rate and without the uniqueness or margin assumptions used by geometric lower bounds. A plug-in corollary covers estimated risks. Under partial observability, a separate theorem lower-bounds residual structural regret by the Bayes error of inferring the active hierarchy from context. The resulting regime law is simple: conflict determines whether hierarchy information has value; identifiability determines how much of that value context can recover.

The practical response is deliberately minimal. Predictions remain on the leaf/action space, while hierarchy information changes the structural decision rule or the weights assigned to semantic branches. This avoids the root-collapse shortcut that appears when internal nodes can absorb predictive mass. The experiments mirror the theory. A controlled benchmark directly audits the decomposition, finite-sample conflict quantities, and learned regime. Frozen Visual-WSD (Raganato et al., 2023) tests the information mechanism in a real semantic task through matched uniform and context-conditioned branch weighting; a within-query posterior shuffle then breaks semantic alignment while keeping posterior sharpness unchanged. CIFAR-100 supplies the predicted low-conflict boundary. Visual-WSD is therefore a real mechanism test, not a claim that the WordNet ontology itself changes at deployment.

Contributions. (1) We define hierarchy shift as a decision problem distinct from ordinary distribution shift and derive an exact context-value/residual-gap decomposition with a sharp zero-gap boundary. (2) We give a finite-sample incompatibility theorem that turns the central hierarchy-information gap into an estimable quantity, recovers pairwise conflict as a lower bound, and remains stable under estimated risks and hierarchy-kernel misspecification; we then connect residual risk to hierarchy identifiability. (3) We give a leaf/action-space realization and a falsifiable evidence chain: exact controlled audits, a locked real-task comparison, a posterior-shuffle mechanism intervention, and a low-conflict boundary. We do not claim a new Visual-WSD state of the art.

2 Related Work

Learning with supplied hierarchies. Hyperbolic embeddings and networks encode known hierarchies geometrically (Nickel & Kiela, 2017; Ganea et al., 2018; Chami et al., 2019). Hierarchy-aware features, selective prediction, and taxonomy-aware vision–language models use supplied structure in representations, losses, or evaluation (Garg et al., 2022; Goren et al., 2024; Snæbjarnarson et al., 2025; Sani & Anand, 2026; He et al., 2026). Recent variants learn hyperbolic visual hierarchies, discover categories hierarchically, train from labels at different taxonomy levels, or improve robustness across hierarchy trees (Kwon et al., 2024; Liu et al., 2025; Park et al., 2026; Dong et al., 2026; Wang et al., 2026). These methods learn *with* a hierarchy; our question is the decision cost of not knowing which plausible hierarchy is active at deployment.

Multiple taxonomies and structural transfer. Multiple-taxonomy learning shares representations across taxonomies (Hwang et al., 2012), and taxonomy adaptation transfers a supplied structure (Babbar et al., 2016). HierVision further documents that practical hierarchy sources are numerous and heterogeneous (Kasarla et al., 2025). We instead treat the active structure as latent decision information. This distinction matters because hierarchical modeling need not improve prediction simply by being present (Weber et al., 2024): its value depends on whether plausible structures actually change the preferred action.

Structure-aware transport and context-dependent semantics. Hierarchical OT, Tree-Wasserstein, and Gromov–Wasserstein methods use structure as a distance or alignment device (Yurochkin et al., 2019; Lee et al., 2019; Lin et al., 2025; Yang et al., 2024). We ask the prior question of whether missing structural information creates unavoidable decision risk. Context-dependent categorization provides a related semantic setting (Wang et al., 2025). In Visual-WSD, minimal text context selects the relevant sense of an ambiguous word among candidate images (Raganato et al.,

Table 1: Hierarchy shift versus adjacent settings. "Context selects H" means that the decision rule can use context to infer the active semantic state.

Setting	Multiple H	Test-time H shift	Context selects H	Structural risk
Multiple-taxonomy learning	yes	no	no	partial
Taxonomy adaptation	partial	no	no	partial
Hierarchical OT / Tree-Wasserstein	partial	no	no	partial
Hierarchy-aware classifiers	no	no	no	yes
Visual-WSD systems	sense-level	no	yes	no
Hierarchy shift (ours)	yes	yes	yes	yes

2023); prior systems combine CLIP, WordNet, and external knowledge (Zhang et al., 2023). We use Visual-WSD only as a proxy for context-dependent semantic-state selection. Our matched comparison fixes the backbone, candidates, and branch representations and changes only how hierarchy information enters the decision. Table 1 summarizes these distinctions from adjacent hierarchy-aware settings.

3 Problem Formulation

Let $\mathcal{Y} = \{1, \dots, L\}$ be a common leaf-label set and take the action space to be $\mathcal{A} = \mathcal{Y}$ in classification. A hierarchy is $H = (V_H, E_H, A_H, C_H)$, where A_H denotes the ancestry relation and $C_H \in [0, M]^{L \times L}$ is a symmetric leaf-to-leaf cost with zero diagonal. Internal nodes define ancestry and semantic paths but are not prediction targets. For the decision-theoretic results below, two hierarchies that induce the same C_H are equivalent; the graph structure matters through the supplied leaf-level cost. In applications, C_H can come from normalized tree/path distance in a supplied taxonomy, an expert severity matrix, an ontology-derived semantic distance, or an institution-specific decision cost. We treat this cost as part of the supplied structural environment: the theory optimizes relative to it, but does not certify that the ontology or cost itself is correct. Let S be observable context and $\Pi(dH \mid X, S)$ a hierarchy kernel. The deployment law is

$$Q(dx, dy, ds, dH) = P(dx, dy, ds)\Pi(dH \mid x, s), \tag{2}$$

so this formulation treats H as a scoring/decision state and implies $H \perp Y \mid (X, S)$. Cases in which H also changes the data-generating law belong to the joint-shift extension in Appendix B.3. Throughout, all probability kernels and conditional risks are assumed measurable. Since $\mathcal{A}$ is finite, whenever an optimal action must be selected we use a fixed deterministic tie-breaking rule, which yields a measurable selector. We use the convention

$$\mathrm{TV}(P, Q) := \sup_B |P(B) - Q(B)|, \tag{3}$$

which equals $\frac{1}{2}\|P - Q\|_1$ when P and Q admit densities with respect to a common dominating measure.

Definition 1 (Hierarchy shift)**.** *A hierarchy shift occurs when* $\Pi_{\mathrm{tr}}(\cdot \mid x, s) \neq \Pi_{\mathrm{te}}(\cdot \mid x, s)$ *on a set of positive* $P_{\mathrm{te}}(X, S)$*-measure. It is* pure *if* $P_{\mathrm{tr}} = P_{\mathrm{te}}$.

For action a and hierarchy H, define

$$r_H(a \mid x, s) = \mathbb{E}[C_H(a, Y) \mid X = x, S = s], \qquad \mathcal{A}_H^*(x, s) = \arg\min_a r_H(a \mid x, s). \tag{4}$$

The hierarchy-aware, context-aware, and hierarchy-blind Bayes risks are

$$R_H^* = \mathbb{E}_{X,S,H} \min_a r_H(a \mid X, S), \tag{5}$$

$$R_S^* = \mathbb{E}_{X,S} \min_a \mathbb{E}_{H|X,S} r_H(a \mid X, S), \tag{6}$$

$$R_0^* = \mathbb{E}_X \min_a \mathbb{E}_{S,H|X} r_H(a \mid X, S). \tag{7}$$

Thus $R_0^* \geq R_S^* \geq R_H^*$.

4 Risk from Missing Hierarchy Information

4.1 Exact gap and the zero boundary

For fixed (x, s) define

$$G_Q(x,s) = \min_a \mathbb{E}_{H|x,s} r_H(a \mid x, s) - \mathbb{E}_{H|x,s} \min_a r_H(a \mid x, s). \tag{8}$$

Proposition 2 (Exact decomposition)**.** $R_S^* - R_H^* = \mathbb{E}G_Q(X,S) \geq 0$*, randomized* (X,S)*-measurable actions do not improve* R_S^**, and*

$$R_0^* - R_H^* = V_S^* + G_H^*, \qquad V_S^* := R_0^* - R_S^*, \quad G_H^* := R_S^* - R_H^*. \tag{9}$$

In particular, $G_H^* = \mathbb{E}G_Q(X,S)$*. We write* $T^* := R_0^* - R_H^*$ *for the total structural value.*

Proposition 3 (Zero-gap characterization)**.** *If* $\mathcal{A}$ *is finite, then* $G_Q(x,s) = 0$ *if and only if there exists an action* $a^\star(x,s)$ *that belongs to* $\mathcal{A}_H^*(x,s)$ *for* $\Pi(\cdot \mid x, s)$*-almost every* H*.*

Hierarchy variation is not harmful by itself. It matters only when plausible hierarchies cannot share a decision that is optimal for all of them. This is the decision-theoretic meaning of hierarchy conflict.

4.2 Finite-sample hierarchy incompatibility

Define the hierarchy-specific excess risk

$$\Delta_H(a \mid x, s) = r_H(a \mid x, s) - \min_b r_H(b \mid x, s), \tag{10}$$

so $G_Q(x,s) = \min_a \mathbb{E}_H \Delta_H(a \mid x, s)$. Let $H_1, \ldots, H_m$ be conditionally iid draws from $\Pi(\cdot \mid x, s)$ and set

$$\widehat{G}_m(x,s) = \min_{a \in \mathcal{A}} \frac{1}{m} \sum_{i=1}^m \Delta_{H_i}(a \mid x, s). \tag{11}$$

Theorem 4 (Finite-sample hierarchy incompatibility)**.** *Assume only that* $\mathcal{A}$ *is finite and* $0 \leq r_H(a \mid x, s) \leq M$*. Then*

$$\mathbb{E}\widehat{G}_m(x,s) \leq G_Q(x,s), \tag{12}$$

and for every $0 < \delta < 1$*, with probability at least* $1 - \delta$*,*

$$\left|\widehat{G}_m(x,s) - G_Q(x,s)\right| \leq M\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{2m}}. \tag{13}$$

Moreover $\widehat{G}_m(x,s) \to G_Q(x,s)$ *almost surely as* $m \to \infty$*.*

Given conditionally iid hierarchy samples and access to hierarchy-specific risks, $\widehat{G}_m$ estimates the hierarchy-information gap at the standard $m^{-1/2}$ rate. The result needs no unique Bayes action, margin, metric on actions, convexity, or binary hierarchy state. When the risks are not available exactly, the plug-in bound below quantifies the additional estimation error.

Corollary 5 (Simultaneous control on a deployment set)**.** *Fix any deployment set* $\{(x_j, s_j)\}_{j=1}^n$*. For each point, draw* m *conditionally iid hierarchies from* $\Pi(\cdot \mid x_j, s_j)$ *and form* $\widehat{G}_m(x_j, s_j)$*. Under the assumptions of Theorem 4, with probability at least* $1 - \delta$*,*

$$\max_{1 \leq j \leq n} \left|\widehat{G}_m(x_j, s_j) - G_Q(x_j, s_j)\right| \leq M\sqrt{\frac{\log(2n|\mathcal{A}|/\delta)}{2m}}. \tag{14}$$

This adds only a logarithmic dependence on the number of deployment points. It makes the pointwise theorem directly usable for a finite evaluation batch without introducing smoothness or uniformity assumptions over (x, s).

For two iid hierarchies define their incompatibility

$$\kappa(H, H') = \min_a \{\Delta_H(a) + \Delta_{H'}(a)\}. \tag{15}$$

Corollary 6 (Pairwise conflict lower bound)**.** *Without additional assumptions,*

$$G_Q(x,s) \geq \frac{1}{2}\mathbb{E}_{H,H'|x,s}\kappa(H,H'). \tag{16}$$

For the geometric specialization, let $d_{\mathcal{A}}$ be a metric on $\mathcal{A}$. If, in addition, each hierarchy has a unique optimum and there exists $\mu(x,s) > 0$ such that, for $\Pi(\cdot \mid x,s)$-almost every H and every $a \in \mathcal{A}$, $\Delta_H(a) \geq \mu(x,s)d_{\mathcal{A}}(a,a_H^*)$, then

$$G_Q(x,s) \geq \frac{\mu(x,s)}{2}\mathbb{E}d_{\mathcal{A}}(a_H^*, a_{H'}^*), \tag{17}$$

recovering the geometric decision-dispersion bound as a special case.

The theorem also tolerates estimated risks. If $\widehat{r}_H$ satisfies $\sup_a |\widehat{r}_H(a) - r_H(a)| \leq \varepsilon_r$ on the sampled hierarchies and $\widetilde{G}_m$ is formed by replacing r_H with $\widehat{r}_H$, then the same event gives

$$|\widetilde{G}_m - G_Q| \leq 2\varepsilon_r + M\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{2m}}. \tag{18}$$

For classification costs in $[0, M]$, a predictive distribution within total variation ε_p of the truth gives $\varepsilon_r \leq M\varepsilon_p$. When $\widehat{r}_H$ is learned from data, the theorem conditions on the stated uniform error bound; in practice this bound should be assessed on held-out or cross-fitted risk estimates rather than on the same observations used to fit them.

Corollary 7 (Robustness to hierarchy-kernel misspecification)**.** *At fixed (x,s), let the true hierarchy kernel be Π and let $\widetilde{\Pi}$ satisfy $\mathrm{TV}(\Pi, \widetilde{\Pi}) \leq \varepsilon_\Pi$. Draw $H_1, \ldots, H_m$ from $\widetilde{\Pi}$ and form $\widetilde{G}_m$ from risks with uniform error at most ε_r. Then, with probability at least $1-\delta$,*

$$|\widetilde{G}_m - G_Q| \leq 2\varepsilon_r + M\varepsilon_\Pi + M\sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{2m}}. \tag{19}$$

Thus hierarchy sampling, risk estimation, and kernel misspecification contribute additively. Proofs are in Appendix A.

4.3 Identifiability controls what context can recover

For this subsection, fix x and impose the side-information condition

$$r_H(a \mid x,s) = r_H(a \mid x) \quad \text{for every } a, H \text{ and } P_{S|x,H}\text{-a.e. } s. \tag{20}$$

Thus, once $X = x$ is fixed, S may reveal which hierarchy is active but does not otherwise change the hierarchy-specific risk. We write $r_H(a \mid x)$ for this common value. A sufficient condition is $Y \perp S \mid X = x$. Define $\mathcal{A}_H^*(x) = \arg\min_a r_H(a \mid x)$ and

$$\Delta_S^*(x) = \inf_{g:S\to\mathcal{A}} \mathbb{E}\left[r_H(g(S) \mid x) - \min_a r_H(a \mid x) \mid X = x\right]. \tag{21}$$

Theorem 8 (Binary hierarchy identifiability)**.** *Let $H \in \{H_1, H_2\}$ have arbitrary conditional priors $\pi_i = \mathbb{P}(H = H_i \mid X = x) > 0$, $\pi_1 + \pi_2 = 1$, and let $P_i = P(S \in \cdot \mid X = x, H = H_i)$. Under Eq.* (20)*, assume the optimal sets $\mathcal{A}_{H_1}^*(x)$ and $\mathcal{A}_{H_2}^*(x)$ are disjoint and every non-optimal action under either hierarchy incurs regret at least $\gamma > 0$. Then*

$$\Delta_S^*(x) \geq \gamma e_H^*(x; \pi_1, \pi_2), \tag{22}$$

where e_H^* is the Bayes error for identifying H from S under the actual prior. For equal priors this reduces to

$$e_H^*(x) = \frac{1}{2}\big(1 - \mathrm{TV}(P_1, P_2)\big). \tag{23}$$

Conflict and identifiability are distinct. Disjoint optimal sets make hierarchy information useful; the Bayes error e_H^* measures how much of that value context cannot recover when context acts only through hierarchy information. The finite-state information-risk sandwich, a transport interpretation, and a joint data–hierarchy shift bound are given in Appendix B; none is required by the empirical claims below.

5 HIERARCHY-CONDITIONED STRUCTURAL LEARNING

The theory requires a decision rule that uses the available information about H; it does not prescribe a particular network architecture.

Controlled classification. For the finite hierarchy-state realization used below, let $\rho_\phi(H \mid x, s)$ summarize hierarchy information available from context. The learner predicts only on leaves,

$$p_\theta(a \mid x, s) = \mathrm{softmax}\big(h_\theta(f_\theta(x), \rho_\phi(\cdot \mid x, s))\big)_a, \tag{24}$$

and minimizes

$$\mathcal{L} = \mathcal{L}_{\mathrm{task}} + \lambda_{\mathrm{sev}} \frac{1}{B} \sum_{i=1}^{B} \sum_{H} \rho_\phi(H \mid x_i, s_i) \sum_{a} p_\theta(a \mid x_i, s_i) C_H(a, y_i). \tag{25}$$

Internal nodes define costs but receive no predictive mass, so a shared root cannot absorb probability as a shortcut. In the experiments reported here, ρ_ϕ is not identified by jointly minimizing this structural objective with the predictor: the controlled study uses the exact synthetic hierarchy posterior, while Visual-WSD uses a frozen context encoder. More generally, a learned ρ_ϕ must be estimated or calibrated from hierarchy supervision or a separate objective; the structural loss alone is not intended to identify $\Pi(H \mid x, s)$.

Frozen retrieval. For Visual-WSD, let q be the ambiguous query, j a candidate image, and $\mathcal{S}(w)$ the WordNet senses of the target word (Miller, 1995). With fixed CLIP image/text encoders (Radford et al., 2021),

$$b_j = \cos(E_T(q), E_I(I_j)), \qquad a_{j,s} = \cos(E_T(t_s), E_I(I_j)), \tag{26}$$

where t_s contains the lemma, definition, and available parent/grandparent names. Let E denote the frozen semantic-state text encoder (SimCSE in our experiments). Here a WordNet sense/branch plays the role of a semantic state; this is a proxy realization rather than a literal sample from a deployment-time taxonomy kernel. The encoder gives

$$\rho_{\beta,E}(s \mid q) = \mathrm{softmax}_s\big(\beta \cos(E(q), E(g_s))\big). \tag{27}$$

For queries with at least one candidate sense, we compare the retrieval scores

$$\mathrm{Score}_{\mathrm{avg}}(j \mid q) = b_j + \lambda_{\mathrm{avg}} |\mathcal{S}(w)|^{-1} \sum_{s} a_{j,s}, \tag{28}$$

$$\mathrm{Score}_{\mathrm{cond}}(j \mid q) = b_j + \lambda_{\mathrm{cond}} \sum_{s} \rho_{\beta,E}(s \mid q) a_{j,s}. \tag{29}$$

If $|\mathcal{S}(w)| = 0$, both methods reduce to the base score b_j. The methods share the same base score, branches, candidates, and CLIP checkpoint. Their scientific difference is uniform versus context-dependent branch weighting.

5.1 WHEN IS HIERARCHY CONDITIONING WARRANTED?

The theory yields a simple conditioning principle. Hierarchy conditioning is warranted only when plausible hierarchies materially change the preferred decision and the available context carries information about which state is active. The finite-sample incompatibility theorem addresses the first condition; the identifiability theorem addresses the second. This principle determines when the additional structural machinery in the two realizations above is useful, before any benchmark comparison is considered.

Decision relevance. Given plausible hierarchy states, or samples from a posterior over them, estimate $\widehat{G}_m$ or pairwise κ. If the incompatibility is near zero, the active hierarchy can change without materially changing the preferred action, so a hierarchy-agnostic rule is sufficient. On n deployment points, simultaneous error at most ε with confidence $1 - \delta$ is guaranteed once

$$m \geq \frac{M^2}{2\varepsilon^2} \log \frac{2n|\mathcal{A}|}{\delta}. \tag{30}$$

Corollary 7 also quantifies how estimated risks and a misspecified hierarchy posterior weaken this guarantee.

Table 2: Theory-derived decision principle for hierarchy conditioning.

Condition	Operational quantity	Learning decision
Do plausible hierarchies change the decision?	$\widehat{G}_m$ or pairwise κ	Near zero: use the hierarchy-agnostic rule; little structural value is available.
Can context identify the active state?	held-out hierarchy error or calibrated posterior	Poor identification: collect better context or abstain; the bottleneck is informational.
Does conditioning exploit the information?	conditioned − uniform; aligned − shuffled	Positive aligned gain supports a hierarchy-conditioned rule for that regime.

Identifiability and conditioning. If incompatibility is non-negligible, use held-out hierarchy-prediction error or a calibrated posterior to assess whether context identifies the active state. Poor identifiability makes the bottleneck informational rather than architectural. When both conditions are favorable, a hierarchy-conditioned rule is warranted. The experiments below then ask a separate question: whether the learned or frozen realization uses that information in the predicted regime.

The empirical score $\mathrm{Sig}(q)$ used later has a deliberately narrower role: it is neither an estimator of G_H^* or G_Q nor an estimator of the Bayes hierarchy-identification error e_H^* in Theorem 8. It is a label-free regime indicator that combines observable branch conflict with posterior concentration. Table 2 summarizes the resulting decision rule and the quantity associated with each stage.

6 Experiments

The experiments return to the two questions posed in the introduction: when does hierarchy uncertainty change the preferred decision, and when can context recover that decision value? Controlled hierarchy shift tests both factors directly; Visual-WSD tests the context-to-semantic-state mechanism in a real retrieval setting; and CIFAR-100 tests the predicted low-conflict boundary. In classification we report semantic error severity $\mathrm{SES} = n^{-1}\sum_i C_{H_i}(\hat{y}_i, y_i)$ (lower is better). In retrieval, $\mathrm{MRR} = n^{-1}\sum_i 1/\mathrm{rank}_i$ and HIT@1 is top-1 retrieval accuracy. For Visual-WSD, when $|\mathcal{S}(w)| > 1$ we define

$$\mathrm{Conflict}(q) = \binom{|\mathcal{S}(w)|}{2}^{-1} \sum_{s<t} \mathbf{1}\left\{ \arg\max_j a_{j,s} \neq \arg\max_j a_{j,t} \right\}, \tag{31}$$

$$\mathrm{Ident}(q) = 1 - \frac{\mathsf{H}(\rho(\cdot \mid q))}{\log |\mathcal{S}(w)|}, \qquad \mathrm{Sig}(q) = \mathrm{Conflict}(q)\,\mathrm{Ident}(q), \tag{32}$$

where H is Shannon entropy. For $|\mathcal{S}(w)| \leq 1$, we set $\mathrm{Conflict}(q) = \mathrm{Ident}(q) = \mathrm{Sig}(q) = 0$. Define $\Delta\,\mathrm{RR}(q) = \mathrm{RR}_{\mathrm{cond}}(q) - \mathrm{RR}_{\mathrm{avg}}(q)$. The label-free score $\mathrm{Sig}(q)$ is fixed before retrieval outcomes are evaluated and is neither an estimator of G_H^* or G_Q nor an estimator of e_H^*; it is an operational regime indicator for the two qualitative prerequisites isolated by the theory. The all-locked comparison is confirmatory, Sig stratification is theory-motivated secondary analysis, and posterior shuffling is a post-hoc mechanism control.

6.1 Controlled hierarchy shift

The synthetic problem uses two hierarchies, six latent groups, and four leaf actions. Two groups have different Bayes actions under H_0 and H_1; the others share the same optimum. Conflict mass $\alpha \in \{0, .25, .5, .75, 1\}$ mixes these groups. Context is a symmetric binary channel $S = H \oplus B$, $B \sim \mathrm{Bernoulli}(\eta)$, with $\eta \in \{0, .1, .2, .3, .4, .5\}$. The construction yields

$$T^* = \frac{\alpha}{15}, \qquad V_S^* = \frac{\alpha(1-2\eta)}{15}, \qquad G_H^* = \frac{2\alpha\eta}{15}. \tag{33}$$

Across all 30 settings, exact enumeration agrees with these formulas to at most 1.11×10^{-16} and the decomposition error is zero. A one-hidden-layer MLP is trained on five formal seeds after selecting $\lambda_{\mathrm{sev}} = 4$ on a separate development seed. The learned gain increases with conflict, decreases with context noise, and tracks V_S^* with Pearson $r = 0.964$; Figure 2 and Table 3 show the corresponding trend and aggregated sweep.

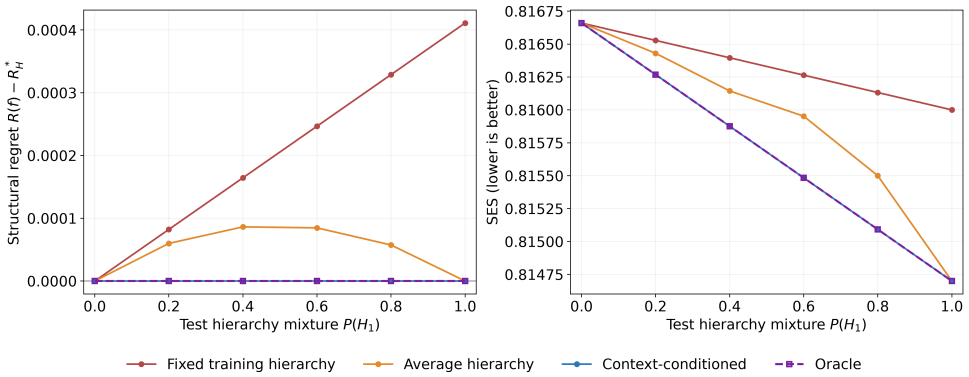

Figure 1: Pure hierarchy shift. We fix $P(X, Y, S)$ and vary only the deployment prior $P(H_1)$. Left: structural regret relative to the active-hierarchy Bayes risk; right: semantic error severity (SES).

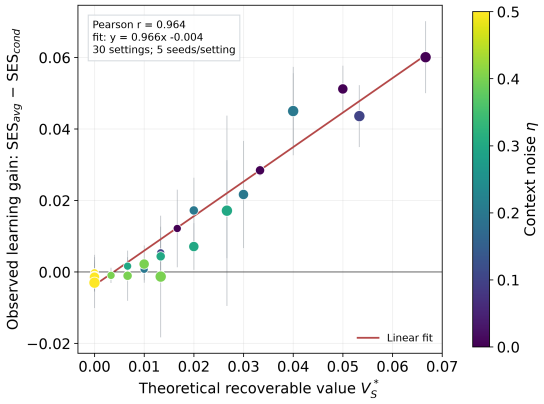

Figure 2: Theoretical recoverable value predicts observed learning gain.

Table 3: Controlled hierarchy-shift gains. Gain rises with conflict α and falls with context noise η.

Sweep	Setting	Gain ΔSES
Conflict	$\alpha = 0.00$	0.0000
	$\alpha = 0.25$	0.0031
	$\alpha = 0.50$	0.0109
	$\alpha = 0.75$	0.0210
	$\alpha = 1.00$	0.0269
Context noise	$\eta = 0.0$	0.0304
	$\eta = 0.1$	0.0223
	$\eta = 0.2$	0.0169
	$\eta = 0.3$	0.0061
	$\eta = 0.4$	-0.0002
	$\eta = 0.5$	-0.0012

To isolate hierarchy shift from ordinary distribution shift, a separate pure-shift evaluation fixes $P(X, Y, S)$ and changes only the deployment hierarchy prior $\Pi(H)$. Figure 1 shows that a fixed training hierarchy accumulates structural regret as the deployment hierarchy moves away from training, while context-conditioned decisions remain close to the oracle. The maximum numerical change in $P(X, Y, S)$ across the shift settings is 3.47×10^{-18}.

6.2 Visual-WSD: context-conditioned branch selection

We use the English Visual-WSD track. The 12,869 silver queries are split deterministically into 3,861 calibration and 9,008 locked confirmation queries. Calibration selects SimCSE (Gao et al., 2021) as the frozen posterior encoder, $\beta = 10$, and $\lambda_{\text{avg}} = \lambda_{\text{cond}} = 2$; the locked split is not used for model selection. Uniform branch averaging in Eq. (28) and conditioned weighting in Eq. (29) share the backbone, candidates, and branch representations and differ only in how hierarchy information enters the decision.

On all 9,008 locked queries, conditioning improves MRR from 0.9231 to 0.9250 (Table 4). The effect is localized: the 8,237 queries with $\text{Sig} = 0$ show essentially no change, whereas the 771 with $\text{Sig} > 0$ gain 0.0229. Figure 3 shows the same regime dependence and the shuffle intervention. The association between $\text{Sig}(q)$ and $\Delta\,\text{RR}(q)$ is positive but modest (Spearman 0.105), consistent with a regime indicator rather than a gap estimator.

Posterior shuffling preserves posterior values and sharpness but breaks context–branch alignment. On $\text{Sig} > 0$, Conditioned $>$ Uniform $>$ Shuffle, with a 0.0439 MRR advantage over Shuffle; on

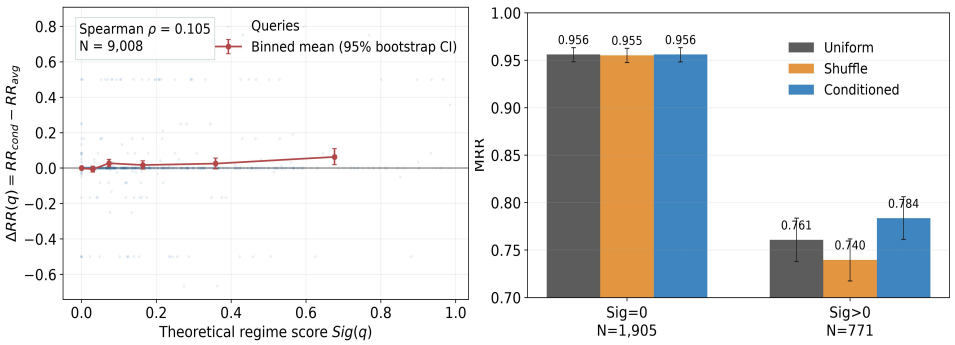

Figure 3: Locked Visual-WSD mechanism analysis: regime association (left) and posterior-shuffle control (right).

Table 4: Locked Visual-WSD results: primary comparison (A) and shuffle control (B).

A. Uniform vs. conditioned weighting					
Subset	N	Uniform MRR	Cond. MRR	ΔMRR	95% CI
All locked	9008	0.9231	0.9250	+0.0020	$[0.0009, 0.0030]$
$\mathrm{Sig} = 0$	8237	0.9383	0.9383	-0.0000	$[-0.0003, 0.0003]$
$\mathrm{Sig} > 0$	771	0.7607	0.7836	+0.0229	$[0.0108, 0.0349]$
B. Posterior-shuffle control					
Subset	N	Uniform	Shuffle	Conditioned	Cond.−Shuffle
Eligible full	2676	0.8998	0.8932	**0.9064**	+0.0132
$\mathrm{Sig} = 0$	1905	0.9561	0.9553	0.9561	+0.0008
$\mathrm{Sig} > 0$	771	0.7607	0.7397	**0.7836**	**+0.0439**

$\mathrm{Sig} = 0$, the methods are indistinguishable. HIT@1, 50-seed stability, gloss-only, and descriptive-gold checks are in Appendix C.

6.3 Low-conflict boundary on CIFAR-100

CIFAR-100 (Krizhevsky, 2009) is a boundary test. Five-fold out-of-fold calibration gives 2.876% conflict and a 0.000630 SES gain; on a fixed high-entropy subset, conflict rises to 9.590% and the gain to 0.002244 without changing the hierarchy or retraining. Full intervals are in Appendix C.

7 Discussion

Across settings, conflict and identifiability give the same answer: hierarchy information matters when plausible structures favor different actions, and context recovers that value when it identifies the active state. The controlled study tests this directly; Visual-WSD provides qualitative mechanism evidence; CIFAR-100 gives the low-conflict boundary; shuffling shows that alignment, not sharpness alone, drives the informative-regime gain.

The roles are not interchangeable. Visual-WSD has query-specific actions and no ground-truth hierarchy kernel, so $\mathrm{Sig}(q)$ is neither G_Q nor e^*_H and Figure 3 is not a quantitative test of Theorem 4 or Theorem 8. The theory assumes a common leaf/action space and supplied cost C_H, does not validate that cost or cover label-space shift, and Theorem 8 additionally requires Eq. (20); further limitations are in Appendix D.

8 Conclusion

Hierarchy shift asks whether changing semantic structure changes the preferred decision. Its excess Bayes risk splits into context-recoverable value and a residual hierarchy-information gap. Theory and experiments give one criterion: condition on hierarchy when plausible structures favor different actions and context identifies the active state; otherwise the structure has little decision value. This separates harmless semantic reorganization from hierarchy-dependent deployment risk.

Reproducibility Statement

All reproducibility material submitted with this paper is contained in the paper itself. Sections 3–5 state the formal assumptions and estimators; Appendix A gives complete proofs and Appendix B gives the theoretical extensions. Appendix C records the controlled construction and full sweep, Visual-WSD split and frozen hyperparameters, scoring rules, bootstrap and posterior-shuffle protocols, and the CIFAR-100 out-of-fold calibration procedure. Appendix D records the protocol history and the separation between confirmatory and exploratory analyses. No separate code or data archive is submitted with this version.

Ethics Statement

The paper treats a hierarchy and its leaf-level cost as part of the supplied decision environment; it does not certify that the supplied structure is normatively correct. Ontologies, severity matrices, or institution-specific costs may encode domain-specific or institutional biases, and hierarchy conditioning can propagate those biases. We therefore interpret the method as decision conditioning under a supplied structure, not as a procedure for validating or correcting that structure.

AI Use Statement

Generative AI tools were used for language revision, manuscript organization, and interactive assistance in checking mathematical exposition and experimental design/presentation. The authors independently verified the mathematical statements and proofs, implementation choices, data processing, and numerical results reported in the paper. The authors take full responsibility for the final manuscript.

REFERENCES

Rohit Babbar, Ioannis Partalas, Eric Gaussier, Massih-Reza Amini, and Cécile Amblard. Learning taxonomy adaptation in large-scale classification. *Journal of Machine Learning Research*, 17(98): 1–37, 2016.

Ines Chami, Rex Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. In *Advances in Neural Information Processing Systems*, 2019.

Junhao Dong, Yifei Zhang, Hao Zhu, Yew-Soon Ong, and Piotr Koniusz. Hierarchically robust zero-shot vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37642–37652, 2026.

Octavian Ganea, Gary Becigneul, and Thomas Hofmann. Hyperbolic neural networks. In *Advances in Neural Information Processing Systems*, 2018.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 6894–6910, 2021.

Ashima Garg, Depanshu Sani, and Saket Anand. Learning hierarchy aware features fornbsp;reducing mistake severity. In *Computer Vision – ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XXIV*, pp. 252–267, Berlin, Heidelberg, 2022. Springer-Verlag. ISBN 978-3-031-20052-6. doi: 10.1007/978-3-031-20053-3_15.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *Advances in Neural Information Processing Systems*, 2024.

Hulingxiao He, Zhi Tan, and Yuxin Peng. Taxonomy-aware representation alignment for hierarchical visual recognition with large multimodal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31124–31134, 2026.

Sung Ju Hwang, Kristen Grauman, and Fei Sha. Semantic kernel forests from multiple taxonomies. In *Advances in Neural Information Processing Systems*, 2012.

Tejaswi Kasarla, Ruthu Hulikal Rooparaghunath, Stefano D'Arrigo, Gowreesh Mago, Abhishek Jha, Melika Ayoughi, Swasti Shreya Mishra, Ana Manzano Rodríguez, Teng Long, Mina Ghadimi Atigh, Max van Spengler, and Pascal Mettes. Hiervision: Standardized and reproducible hierarchical sources for vision datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pp. 682–695, 2025.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Hyeongjun Kwon, Jinhyun Jang, Jin Kim, Kwonyoung Kim, and Kwanghoon Sohn. Improving visual recognition with hyperbolical visual hierarchy mapping. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 17364–17374, 2024.

John Lee, Max Dabagia, Eva Dyer, and Christopher Rozell. Hierarchical optimal transport for multimodal distribution alignment. In *Advances in Neural Information Processing Systems*, 2019.

Ya-Wei Eileen Lin, Ronald R. Coifman, Gal Mishne, and Ronen Talmon. Tree-wasserstein distance for high dimensional data with a latent feature hierarchy. In *International Conference on Learning Representations*, 2025.

Yuanpei Liu, Zhenqi He, and Kai Han. Hyperbolic category discovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9891–9900, 2025.

George A. Miller. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11): 39–41, 1995.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in Neural Information Processing Systems*, 2017.

Seulki Park, Zilin Wang, and Stella X. Yu. Free-grained hierarchical visual recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32767–32776, 2026.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 8748–8763, 2021.

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. Semeval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, pp. 2227–2234, 2023.

Depanshu Sani and Saket Anand. Hier-cos: Making deep features hierarchy-aware via composition of orthogonal subspaces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11263–11272, 2026.

Vésteinn Snæbjarnarson, Kevin Du, Niklas Stoehr, Serge Belongie, Ryan Cotterell, Nico Lang, and Stella Frank. Taxonomy-aware evaluation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9109–9120, 2025.

Renzhen Wang, De Cai, Kaiwen Xiao, Xixi Jia, Xiao Han, and Deyu Meng. Label hierarchy transition: Delving into class hierarchies to enhance deep classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48(8):10112–10128, 2026.

Zilin Wang, Sangwoo Mo, Stella X. Yu, Sima Behpour, and Liu Ren. Open ad-hoc categorization with contextualized feature learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15108–15117, 2025.

Simon Weber, Barış Zöngür, Nikita Araslanov, and Daniel Cremers. Flattening the parent bias: Hierarchical semantic segmentation in the poincaré ball. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28223–28232, 2024.

Yifei Yang, Wonjun Lee, Dongmian Zou, and Gilad Lerman. Improving hyperbolic representations via gromov–wasserstein regularization. In *European Conference on Computer Vision*, 2024.

Mikhail Yurochkin, Sebastian Claici, Edward Chien, Farzaneh Mirzazadeh, and Justin Solomon. Hierarchical optimal transport for document representation. In *Advances in Neural Information Processing Systems*, 2019.

Shibingfeng Zhang, Shantanu Nath, and Davide Mazzaccara. Gpl at semeval-2023 task 1: Wordnet and clip to disambiguate images. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 1592–1597, 2023.

A Proofs

A.1 Proof of Proposition 2

Condition on $(X, S) = (x, s)$. Before observing H, the best deterministic action has risk $\min_a \mathbb{E}_{H|x,s} r_H(a \mid x, s)$; after observing H, the best risk is $\mathbb{E}_{H|x,s} \min_a r_H(a \mid x, s)$. Their difference is $G_Q(x, s)$. Averaging over (X, S) gives $R_S^* - R_H^* = \mathbb{E}G_Q(X, S) = G_H^*$. A randomized action has risk linear in its mixing distribution, so minimizing over the probability simplex attains an extreme point and cannot improve R_S^*. Finally, $R_0^* - R_H^* = (R_0^* - R_S^*) + (R_S^* - R_H^*) = V_S^* + G_H^*$.

A.2 Proof of Proposition 3

Fix (x, s) and suppress it from notation. If there exists $a^\star \in \mathcal{A}_H^*$ for $\Pi(\cdot \mid x, s)$-almost every H, then

$$\min_a \mathbb{E}_H r_H(a) \le \mathbb{E}_H r_H(a^\star) = \mathbb{E}_H \min_a r_H(a). \tag{34}$$

The reverse inequality holds for every a, hence $G_Q = 0$. Conversely, because $\mathcal{A}$ is finite, some $\bar{a}$ minimizes $\mathbb{E}_H r_H(a)$. If $G_Q = 0$, then

$$0 = \mathbb{E}_H \left[r_H(\bar{a}) - \min_b r_H(b) \right]. \tag{35}$$

The integrand is nonnegative, so it is zero almost surely. Thus $\bar{a} \in \mathcal{A}_H^*$ almost surely.

A.3 Proof of Theorem 4

Fix (x, s) and write $\Delta_H(a)$ for $\Delta_H(a \mid x, s)$. Since $0 \le r_H(a) \le M$, also $0 \le \Delta_H(a) \le M$. Define

$$\mu(a) = \mathbb{E}_H \Delta_H(a), \qquad \widehat{\mu}_m(a) = \frac{1}{m} \sum_{i=1}^{m} \Delta_{H_i}(a). \tag{36}$$

Then $G_Q = \min_a \mu(a)$ and $\widehat{G}_m = \min_a \widehat{\mu}_m(a)$. Because the pointwise minimum is concave,

$$\mathbb{E}\widehat{G}_m = \mathbb{E} \min_a \widehat{\mu}_m(a) \le \min_a \mathbb{E}\widehat{\mu}_m(a) = \min_a \mu(a) = G_Q. \tag{37}$$

For each fixed action, Hoeffding's inequality gives

$$\mathbb{P}(|\widehat{\mu}_m(a) - \mu(a)| > t) \le 2 \exp(-2mt^2/M^2). \tag{38}$$

A union bound over the finite action set implies, with probability at least $1 - \delta$,

$$\sup_a |\widehat{\mu}_m(a) - \mu(a)| \le M \sqrt{\frac{\log(2|\mathcal{A}|/\delta)}{2m}}. \tag{39}$$

Using $|\min_a u_a - \min_a v_a| \le \max_a |u_a - v_a|$ yields the stated concentration inequality. The strong law holds for every action; finiteness of $\mathcal{A}$ makes convergence simultaneous over actions, hence $\widehat{G}_m \to G_Q$ almost surely.

A.4 Proof of Corollary 5

Apply Theorem 4 at deployment point j with failure probability δ/n and take a union bound over $j = 1, \ldots, n$. This gives

$$\max_{1 \le j \le n} |\widehat{G}_m(x_j, s_j) - G_Q(x_j, s_j)| \le M \sqrt{\frac{\log(2n|\mathcal{A}|/\delta)}{2m}}. \tag{40}$$

No independence across deployment points is needed; only the conditional iid hierarchy samples within each point enter Theorem 4.

A.5 Proof of Corollary 6 and the plug-in bound

For $m = 2$,

$$\widehat{G}_2 = \frac{1}{2} \min_a \{\Delta_H(a) + \Delta_{H'}(a)\} = \frac{1}{2} \kappa(H, H'). \tag{41}$$

The downward-bias part of Theorem 4 gives $G_Q \ge \frac{1}{2}\mathbb{E}\kappa(H, H')$. Under structural growth and because $d_{\mathcal{A}}$ is a metric,

$$\kappa(H, H') \ge \mu(x, s) \min_a \{d_{\mathcal{A}}(a, a_H^*) + d_{\mathcal{A}}(a, a_{H'}^*)\} \tag{42}$$

$$\ge \mu(x, s) d_{\mathcal{A}}(a_H^*, a_{H'}^*), \tag{43}$$

where the last step is the triangle inequality.

For the plug-in statement, $\sup_a |\widehat{r}_H(a) - r_H(a)| \le \varepsilon_r$ implies

$$\left| \min_b \widehat{r}_H(b) - \min_b r_H(b) \right| \le \varepsilon_r, \tag{44}$$

so $|\widehat{\Delta}_H(a) - \Delta_H(a)| \le 2\varepsilon_r$ uniformly in a. The empirical minima therefore differ by at most $2\varepsilon_r$, and combining this deterministic perturbation with Theorem 4 proves Eq. (18). If costs lie in $[0, M]$, the TV convention in Section 3 gives $\varepsilon_r \le M\varepsilon_p$ whenever the estimated label distribution is within TV distance ε_p of the truth.

A.6 Proof of Corollary 7

Let $G_{\widetilde{\Pi}} = \min_a \mathbb{E}_{\widetilde{\Pi}} \Delta_H(a)$. Since $0 \le \Delta_H(a) \le M$ and $\mathrm{TV}(\Pi, \widetilde{\Pi}) \le \varepsilon_\Pi$,

$$|\mathbb{E}_\Pi \Delta_H(a) - \mathbb{E}_{\widetilde{\Pi}} \Delta_H(a)| \le M\varepsilon_\Pi \tag{45}$$

for every action, hence $|G_Q - G_{\widetilde{\Pi}}| \le M\varepsilon_\Pi$. Estimated regrets perturb the empirical minimum by at most $2\varepsilon_r$. Applying Theorem 4 under $\widetilde{\Pi}$ and using the triangle inequality yields the claim.

A.7 Proof of Theorem 8

Condition on $X = x$ and use Eq. (20), so the hierarchy-specific regret of an action does not depend directly on S. Because $\mathcal{A}^*_{H_1}(x)$ and $\mathcal{A}^*_{H_2}(x)$ are disjoint, define a decoder $\varphi : \mathcal{A} \to \{1, 2\}$ such that $\varphi(a) = i$ for every $a \in \mathcal{A}^*_{H_i}(x)$, $i \in \{1, 2\}$, and label actions outside both sets arbitrarily. For any S-measurable action rule g, define $\widehat{H}_g(S) = H_{\varphi(g(S))}$. If $\widehat{H}_g(S) \ne H_i$ when $H = H_i$, then $g(S) \notin \mathcal{A}^*_{H_i}(x)$, and the excess risk is at least γ. Therefore

$$\mathbb{E}\Big[r_H(g(S) \mid x) - \min_a r_H(a \mid x) \mid X = x\Big] \ge \gamma\, \mathbb{P}(\widehat{H}_g(S) \ne H \mid X = x). \tag{46}$$

Every estimator of H from S has error at least the Bayes error $e^*_H(x; \pi_1, \pi_2)$. Taking the infimum over g gives $\Delta^*_S(x) \ge \gamma e^*_H(x; \pi_1, \pi_2)$. For equal priors, the binary testing identity gives $e^*_H(x) = \frac{1}{2}(1 - \mathrm{TV}(P_1, P_2))$.

B Theoretical Extensions

B.1 Finite-state information-risk sandwich

Retain the side-information condition Eq. (20). Suppose H is uniform on $K \ge 2$ states with pairwise disjoint optimal-action sets. Assume every non-optimal action incurs at least $\gamma > 0$ excess risk and that choosing an action optimal for another state incurs at most $\Gamma < \infty$. The decoder argument used in Theorem 8 gives the lower bound $\gamma e^*_H(x) \le \Delta^*_S(x)$. For the upper bound, let $\widehat{H}_{\mathrm{Bayes}}(S)$ be a Bayes hierarchy classifier and choose any $a^*_{\widehat{H}_{\mathrm{Bayes}}(S)} \in \mathcal{A}^*_{\widehat{H}_{\mathrm{Bayes}}(S)}(x)$. This rule incurs zero excess risk when the hierarchy is identified correctly and at most Γ otherwise, so

$$\gamma e^*_H(x) \le \Delta^*_S(x) \le \Gamma e^*_H(x). \tag{47}$$

With mutual information measured in bits, Fano's inequality yields

$$\Delta^*_S(x) \ge \gamma \left[1 - \frac{I_2(H; S \mid X = x) + 1}{\log_2 K}\right]_+. \tag{48}$$

For $K = 2$ this Fano term can be vacuous, which is why Theorem 8 is sharper in the binary case.

B.2 Transport as a severity-consistent realization

For leaf measures $\mu, \nu \in \Delta(\mathcal{Y})$ define $W_H(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \langle C_H, \pi \rangle$. If a fixed metric d_0 is dominated by every C_H and W_0 is its Wasserstein metric, then for hierarchy-conditioned targets q^H_y and any hierarchy-agnostic μ,

$$\mathbb{E}_H W_H(\mu, q^H_y) \ge \frac{1}{2} \mathbb{E}_{H,H'} W_0(q^H_y, q^{H'}_y). \tag{49}$$

Indeed, $W_H \ge W_0$, and the triangle inequality gives $W_0(q^H_y, q^{H'}_y) \le W_0(q^H_y, \mu) + W_0(\mu, q^{H'}_y)$; average over iid H, H'. For a Dirac target, $W_H(\delta_a, \delta_y) = C_H(a, y)$ and $W_H(p, \delta_y) = \sum_a p(a) C_H(a, y)$, so transport to a Dirac target is exactly expected severity rather than a distinct algorithmic contribution.

B.3 Joint data and hierarchy shift

Let $z = (x, y, s)$ and suppose every $\ell_f(z, H) \in [0, 1]$ is uniformly L_Z-Lipschitz in z and L_H-Lipschitz in a hierarchy metric, with constants independent of $f \in \mathcal{F}$ and under the usual measurability conditions. Define

$$D_{Z,H} = \inf_{\xi \in \Gamma(P_{\mathrm{tr}}, P_{\mathrm{te}})} \left\{ L_Z \mathbb{E}_\xi d_Z(z, z') + L_H \mathbb{E}_\xi W_1^{d_H}(\Pi_{\mathrm{tr}}(\cdot \mid z), \Pi_{\mathrm{te}}(\cdot \mid z')) \right\}. \tag{50}$$

A standard Rademacher bound under the training deployment law plus a coupling argument yields, with probability at least $1 - \delta$, simultaneously for all $f \in \mathcal{F}$,

$$R_{\mathrm{te}}(f) \le \widehat{R}_{\mathrm{tr}}(f) + 2\widehat{\mathfrak{R}}_n(\ell \circ \mathcal{F}) + 3\sqrt{\frac{\log(2/\delta)}{2n}} + D_{Z,H}. \tag{51}$$

Thus movement of the ordinary data law and movement of the hierarchy kernel contribute distinct deployment terms.

C Supplementary Experimental Results

C.1 Controlled construction and complete sweep

The controlled costs are

$$C_{H_0} = \frac{1}{3}\begin{bmatrix} 0 & 1 & 3 & 3 \\ 1 & 0 & 3 & 3 \\ 3 & 3 & 0 & 1 \\ 3 & 3 & 1 & 0 \end{bmatrix}, \qquad C_{H_1} = \frac{1}{3}\begin{bmatrix} 0 & 3 & 1 & 3 \\ 3 & 0 & 3 & 1 \\ 1 & 3 & 0 & 3 \\ 3 & 1 & 3 & 0 \end{bmatrix}. \tag{52}$$

The conflict groups are $p(Y \mid g_0) = (.20, .40, .40, 0)$ and $p(Y \mid g_1) = (.40, 0, .20, .40)$; their Bayes actions differ across H_0, H_1 and the minimum normalized margin is $1/15$. The benchmark varies $\alpha \in \{0, .25, .5, .75, 1\}$ and $\eta \in \{0, .1, .2, .3, .4, .5\}$. The learned results use five formal seeds with identical train/validation/test splits. Positive $\mathrm{SES}_{\mathrm{Average}} - \mathrm{SES}_{\mathrm{Conditioned}}$ denotes an advantage for conditioning. Table 5 reports all 30 (α, η) settings, separating exact oracle quantities from learned finite-sample SES.

Pure-shift audit. The separate pure-shift evaluation used in Figure 1 fixes the observable distribution and changes only the hierarchy mechanism. Its absolute SES scale is not used in the 30-setting averages of Table 5.

Table 6 isolates which components are needed in the separately recorded pure-shift evaluation.

Table 7 reports the corresponding no-shift, partial-shift, and strong-shift deployment conditions while keeping $P(X, Y, S)$ fixed.

C.2 Robustness and finite-state audit

To assess imperfect context information, independent Gaussian noise is added to the input representation at fixed $\sigma \in \{0, .05, .10, .20\}$. Each entry in Table 8 averages all 30 (α, η) settings and five formal seeds.

As Table 8 shows, the conditioned policy has the lowest SES at every evaluated noise level, while its advantage over averaging decreases from 0.012383 to 0.005697.

Table 9 gives the finite-state information–risk audit at a common context-noise level. In Table 9, the two-state row is a deliberate boundary case: the recorded construction has no decision conflict there, so hierarchy uncertainty alone produces no regret. The four- and eight-state cases exhibit positive residual regret under imperfect context information.

C.3 Visual-WSD protocol and secondary checks

The final corrected calibration selects SimCSE RoBERTa-large, $\beta = 10$, and $\lambda_{\mathrm{avg}} = \lambda_{\mathrm{cond}} = 2$. All comparisons use the same CLIP backbone, candidate images, branch representations, and

base retrieval scores. Queries without multiple candidate senses use the base CLIP score and have $\mathrm{Sig}(q) = 0$. Of the 9,008 locked queries, conditioning changes reciprocal rank for only 161: 100 improve and 61 worsen. Table 10 shows that the same localization appears in HIT@1 on the shuffle-eligible subset.

Posterior-shuffle stability. The intervention preserves each query's posterior values while permuting their semantic-branch assignment. Across 50 fixed non-identity permutation seeds, Conditioned has higher MRR than Shuffle in all 50 seeds on both the full mechanism subset and the $\mathrm{Sig}(q) > 0$ subset. On the informative subset, the Conditioned–Shuffle MRR difference is 0.043892 with 95% bootstrap CI $[0.029803, 0.057870]$. Table 11 summarizes the seed-level stability.

Query-level effects. Most locked queries remain unchanged, and Table 12 shows that the gain is concentrated on a small subset.

Signal decomposition. Table 13 shows that the combined signal has the largest of the three reported rank correlations, although all are modest, supporting its use as a regime indicator rather than a direct estimator of the hierarchy-information gap.

Gloss-only ablation. A post-hoc gloss-only ablation keeps the same posterior but replaces the structured branch prompt with a gloss-only prompt. Locked MRR is 0.925219 for gloss-only and 0.925047 for structured conditioning; the difference is -0.000171 with CI $[-0.001898, 0.001618]$. We therefore find no evidence that explicitly inserting parent/grandparent names contributes beyond posterior branch conditioning.

Gold and organizer baseline. Gold results are descriptive because gold labels had been observed before the final prompt correction. Uniform and Conditioned obtain gold MRR 0.670817 and 0.729435 and HIT@1 0.483801 and 0.570194. The organizer CLIP baseline is stronger (MRR 0.740203, HIT@1 0.606911), so the paper makes no state-of-the-art claim.

C.4 CIFAR-100 BOUNDARY PROTOCOL

The CIFAR-100 experiment uses deterministic stratified five-fold out-of-fold calibration over all 50,000 training examples. For each fold, one temperature is fit on the other 40,000 examples and

Table 5: Complete controlled hierarchy-shift sweep. Oracle columns are population quantities; learned SES values are means over five formal seeds.

α	η	Oracle V_S^*	Oracle G_H^*	Average SES	Conditioned SES	Gain
0.00	0.0	0.000000	0.000000	0.151411	0.151411	0.000000
0.00	0.1	0.000000	0.000000	0.150011	0.150011	0.000000
0.00	0.2	0.000000	0.000000	0.152278	0.152278	0.000000
0.00	0.3	0.000000	0.000000	0.153056	0.153056	0.000000
0.00	0.4	0.000000	0.000000	0.147922	0.147922	0.000000
0.00	0.5	0.000000	0.000000	0.146711	0.146711	0.000000
0.25	0.0	0.016667	0.000000	0.245689	0.233522	0.012167
0.25	0.1	0.013333	0.003333	0.244833	0.239478	0.005356
0.25	0.2	0.010000	0.006667	0.247367	0.246600	0.000767
0.25	0.3	0.006667	0.010000	0.246367	0.244733	0.001633
0.25	0.4	0.003333	0.013333	0.247378	0.248322	-0.000944
0.25	0.5	0.000000	0.016667	0.245300	0.245533	-0.000233
0.50	0.0	0.033333	0.000000	0.342889	0.314456	0.028433
0.50	0.1	0.026667	0.006667	0.337300	0.319756	0.017544
0.50	0.2	0.020000	0.013333	0.343456	0.326244	0.017211
0.50	0.3	0.013333	0.020000	0.340456	0.336078	0.004378
0.50	0.4	0.006667	0.026667	0.341400	0.342456	-0.001056
0.50	0.5	0.000000	0.033333	0.338389	0.339467	-0.001078
0.75	0.0	0.050000	0.000000	0.439544	0.388344	0.051200
0.75	0.1	0.040000	0.010000	0.441533	0.396511	0.045022
0.75	0.2	0.030000	0.020000	0.433178	0.411500	0.021678
0.75	0.3	0.020000	0.030000	0.436878	0.429767	0.007111
0.75	0.4	0.010000	0.040000	0.438167	0.435956	0.002211
0.75	0.5	0.000000	0.050000	0.433233	0.434689	-0.001456
1.00	0.0	0.066667	0.000000	0.529100	0.469022	0.060078
1.00	0.1	0.053333	0.013333	0.528444	0.484844	0.043600
1.00	0.2	0.040000	0.026667	0.536767	0.491756	0.045011
1.00	0.3	0.026667	0.040000	0.533622	0.516478	0.017144
1.00	0.4	0.013333	0.053333	0.531133	0.532422	-0.001289
1.00	0.5	0.000000	0.066667	0.530878	0.533867	-0.002989

Table 6: Ablation in the separately recorded pure-shift evaluation.

Method	Context S	Hierarchy posterior	Hierarchy cost	SES
CE classifier	×	×	×	0.817520
Fixed hierarchy	×	H_0	✓	0.816610
Average hierarchy	×	Uniform H	✓	0.816200
Context-only weighting	✓	✓	×	0.816660
Proposed	✓	✓	✓	**0.815910**

Table 7: Deployment hierarchy shift under a fixed observable data distribution.

Setting	Train H	Test H	$P(X, Y, S)$ changed?	Fixed-H	Average-H	Conditioned	Oracle
No shift	H_0	H_0	No	0.816660	0.816660	0.816660	0.816660
Partial shift	H_0	$0.5H_0 + 0.5H_1$	No	0.816610	0.816200	0.815910	0.815910
Strong shift	H_0	H_1	No	0.816000	0.814700	0.814700	0.814700

Table 8: Robustness under Gaussian input noise. Lower SES is better.

σ	Fixed	Average	Conditioned
0.00	0.341333	0.341156	**0.328773**
0.05	0.341645	0.341630	**0.331427**
0.10	0.341896	0.341745	**0.333601**
0.20	0.343431	0.342689	**0.336992**

Table 9: Finite-state information–risk audit at context-noise level $\epsilon = 0.25$. Mutual information is in bits.

Hierarchy states	Context noise	$I(H; S)$	Bayes error e_H^*	Observed regret Δ_S^*
2	0.25	0.188722	0.250000	0.000000
4	0.25	0.792481	0.250000	0.035714
8	0.25	1.486883	0.250000	0.050000

Table 10: HIT@1 on the posterior-shuffle-eligible locked-silver subset (percent).

Subset	N	Uniform	Shuffle	Conditioned
All queries	2676	84.38	83.39	**85.31**
$\mathrm{Sig}(q) = 0$	1905	93.12	92.98	93.02
$\mathrm{Sig}(q) > 0$	771	62.78	59.70	**66.28**

Table 11: Posterior-shuffle stability on the shuffle-eligible subset.

Method	Mean MRR	Std.	95% CI
Uniform	0.899835	0	$[0.899835, 0.899835]$
Shuffle	0.893208	0.001400	$[0.892810, 0.893606]$
Conditioned	0.906407	0	$[0.906407, 0.906407]$

Table 12: Query-level reciprocal-rank changes on locked silver.

Category	Number of queries	Mean Δ RR
Improved	100	+0.335337
Unchanged	8847	0
Degraded	61	-0.261410

applied to the held-out 10,000, so no example is calibrated using a parameter fit on itself. Fold conflict rates are 2.75%, 2.85%, 2.93%, 2.87%, and 2.98%, giving pooled conflict 2.876%. After

Table 13: Association between signal components and query-level reciprocal-rank improvement.

Signal	Spearman correlation with $\Delta\,\mathrm{RR}$
Conflict	0.102577
Ident	0.057466
Sig	**0.105234**

Table 14: CIFAR-100 boundary evaluation. The entropy criterion is frozen before OOF evaluation.

Subset	Samples	Conflict rate	Realized SES gap [95% CI]
Full OOF	50,000	2.876%	0.000630 $[0.000220, 0.001030]$
High entropy	12,701	9.590%	0.002244 $[0.000787, 0.003661]$

calibration, the hierarchy-aware, average, and fixed-H_0 expected risks are 0.278927, 0.279212, and 0.279609.

The high-entropy subset is determined only from the frozen unit-temperature predictive distribution, using $\mathcal{H}(x) \geq 2.943954$ nats. This selects 12,701 examples. Conflict rises to 9.590%, and the realized aware–agnostic SES improvement is 0.002244 with 95% CI $[0.000787, 0.003661]$. Table 14 reports the full-population and uncertainty-enriched boundary results side by side.

D Extended Scope, Limitations, and Protocol History

D.1 Extended scope and limitations

Our theoretical guarantees are relative to the supplied semantic cost and hierarchy representation. The framework applies when a common leaf/action space and a meaningful hierarchy-specific cost are available, while the choice of ontology or cost remains application dependent. The controlled hierarchy-shift experiments isolate this setting; broader deployments may additionally involve other forms of distributional shift. Visual-WSD evaluates context-dependent semantic-state selection and serves as a boundary study rather than a complete ontology revision benchmark. Direct evaluation under real ontology or taxonomy revision remains future work.

The analysis focuses on finite action spaces and bounded hierarchy-specific risks, with identifiability characterized under the side-information condition stated in Theorem 8. Within this setting, the bounds separately account for hierarchy sampling, risk estimation, and kernel misspecification. In Visual-WSD, $\mathrm{Sig}(q)$ serves as a theory-motivated regime indicator based on two observable signals: conflict and posterior concentration; it is used qualitatively rather than as a plug-in estimate of G_H^*, $G_Q(x, s)$, or the Bayes hierarchy-identification error e_H^*.

D.2 Protocol history

The locked-silver uniform-versus-conditioned comparison is the primary confirmatory real-task result. A punctuation bug in the structured branch prompt was corrected before the final result package; encoder, temperature, and score weights were then reselected using the original calibration partition only, after which locked silver was evaluated under the frozen corrected configuration. The posterior-shuffle experiment was added afterward as a post-hoc mechanism control. Gold labels had been observed before the final prompt correction, so gold is descriptive only. The calibration signal threshold and tercile cut points equal zero; therefore the main analysis uses continuous $\mathrm{Sig}(q)$ and the non-degenerate $\mathrm{Sig} = 0$ versus $\mathrm{Sig} > 0$ split rather than claiming an independent top-30% subgroup. The gloss-only ablation is exploratory.

D.3 Recorded reproducibility details

Implementation Details. All experiments were conducted on a Linux server with two NVIDIA GeForce RTX 4090 GPUs (24 GB each), two Intel Xeon Gold 6326 CPUs, and 384 GB RAM. The software environment used Python 3.10, PyTorch 2.8.0, CUDA 12.9, and Transformers 4.56.2.

Visual-WSD experiments used `openai/clip-vit-large-patch14-336` for image–text scoring and `princeton-nlp/sup-simcse-roberta-large` for posterior estimation. CIFAR-100 experiments used frozen CLIP ViT-B/32 logits. WordNet 3.0 provided the lexical hierarchies. Random seeds and model revisions were fixed and recorded for reproducibility.

The paper records the controlled construction and full parameter grid, the calibration/locked Visual-WSD split sizes, frozen encoders and hyperparameters, scoring rules, 10,000-sample paired bootstrap procedure, posterior-shuffle intervention and seed count, and the five-fold CIFAR-100 out-of-fold calibration protocol. These details are the reproducibility material included in this submission; no separate code or data archive accompanies the paper.

D.4 Interpretive boundary

The three empirical settings play different roles and should not be read as interchangeable benchmarks. The controlled study is the direct test of pure hierarchy shift because it holds the observable data law fixed while changing the hierarchy mechanism. Visual-WSD tests the narrower context-to-semantic-state information mechanism, and posterior shuffling isolates alignment from posterior sharpness. CIFAR-100 tests the predicted low-conflict boundary. None of these experiments validates the supplied semantic cost or constitutes direct evidence for deployment under real ontology replacement; that setting may also alter labels, covariates, or annotation practice and remains outside the present empirical scope.

Accordingly, small gains under low conflict and losses after posterior shuffling are not competing observations: they test different requirements of the same conflict–identifiability regime.