CAN ACCURATE PREDICTION TELL SPARSE ATTENTION WHAT TO KEEP IN LLMS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Sparse attention reduces long-document cost by passing only a few blocks to later computation. We ask whether accurate next-token prediction from one source identifies what a router should retain under distribution shift. In general, it cannot because labels from one source and unlabeled target documents can leave several deployment worlds compatible with the observations. Repeated facts and visible route positions also prevent a unique mask from being the right target. We characterize the resulting possible and necessary selector sets under a fixed retention budget and declared source, selector, predictor, and loss classes. Labeled source variation recovers the optimal set when suboptimal rules are uniformly separated in worst-source risk, and source mixtures give exact population certificates under convex predictor classes. Label-preserving masking gives a second route with multi-view and adaptive guarantees, whose cost is controlled by informative-example rates and squared performance gaps. Matching lower bounds clarify what extra data are needed before pruning long context and why predictive success alone does not establish causal evidence or computational savings.

1 INTRODUCTION

A long company report says that the company was founded in Paris, serves French customers, and later moved its headquarters to Berlin. A question asks where the headquarters is now. A sparse attention router should keep the statement about Berlin. Some training collections contain only French companies with headquarters in Paris. In such a collection, selecting the customer description gives perfect next-token predictions. The same rule fails on a new collection that includes French companies headquartered in Berlin. This simple case raises the question of whether predictive supervision tells a long-context LLM what it should keep.

The target cannot always be one uniquely correct mask. The move to Berlin sometimes appears in an executive summary and a legal notice. Each copy can support the answer. A different problem appears when the router reads the answer and encodes it in the position of the selected block. The downstream model then predicts correctly without using the retained words. These cases separate four claims that are often combined in discussions of sparse attention. A selection can preserve prediction, identify human-readable evidence, capture a causal relation, or reduce computation. None of these claims follows automatically from another.

We formalize the claim about prediction. A *data source* is a joint distribution of documents, queries, and answers, such as a domain, publisher, time period, or document template. A declared collection of evaluation sources defines which selection rules should be compared. The observation interface specifies which source identities and answers the learner sees. The object of interest is the set of budgeted selection rules that achieve the best worst-source prediction risk with one shared predictor. We call these rules robustly optimal. Identification asks whether unlimited observations determine this set. Partial identification records the rules that remain optimal in some or all deployment worlds compatible with the observations. Finite-sample recovery asks how many documents are needed to estimate the target with high probability. Labels from one source and unlabeled target documents can leave a shortcut and a stable fact indistinguishable. The first obstacle is missing information rather than neural optimization.

The positive results follow an information ladder. Prediction labels from one source reveal only the losses that occur there, even when unlabeled documents are available from the target. Labeled

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.

variation provides document, query, and answer triples from each known source. It reveals a bad rule when the shortcut fails somewhere, as in a collection of French companies headquartered outside Paris. Label-preserving masking offers a direct alternative by removing candidate content and testing whether the answer becomes harder to predict. The two routes depend on how often an observation is informative and on the size of the resulting performance difference. Our analysis also separates independent documents from masked views evaluated on them. Causal interpretation and the cost of searching the full document remain outside the prediction guarantee. Our contributions are summarized below.

1. **Prediction alone leaves a partially identified selector set (Section 3).** We characterize the rules that are optimal in some or all deployment worlds compatible with the observations. A sharp two-world construction gives different optimal sets from identical data. Transcript-capacity results also quantify how visible positions can carry answers without evidential text.
2. **Labeled variation identifies failures through revealing source mixtures (Section 4).** A minimax duality result shows that every suboptimal rule has a source-mixture certificate under convex predictor classes. Uniform recovery yields exact set recovery under a uniform positive separation gap, metric convergence under an error-bound condition for continuous selector classes, and allocation bounds with matching rare-source dependence.
3. **Masking separates document information from view computation (Section 5).** Multi-view importance weighting handles document-dependent and history-dependent allocation with correlated views on each document. Adaptive upper and lower bounds match in their inverse-squared-gap dependence. A variance-sensitive informative-example law and a multi-candidate lower bound quantify the remaining document cost.

2 Problem Setup

We first translate the headquarters example into mathematical objects. One observation is $O = (D, Q, Y)$. The document $D = (X_1, \ldots, X_m)$ contains m addressable blocks, the query or preceding token context is Q, and the next token or task answer is Y. A data source $e \in \mathcal{E}$ has a joint distribution P_e over these observations. Examples include a domain, publisher, time period, or document template. Statistical learning theory often calls such a source an *environment*. A declared collection of evaluation sources defines the population target even when the observation interface withholds some of their answers.

A *selection rule*, or selector, $g \in \mathcal{G}_b$ maps (D, Q) to at most b block indices. The integer b is the retention budget. A router implements the rule. A mask is the index set returned on one document. A predictor $h \in \mathcal{H}$ receives the query, selected contents, and positions in canonical document order. We consider zero-one loss or a clipped and normalized token loss, with $\ell \in [0, 1]$. We assume standard measurability, take $0 < \delta < 1$, and assign value $+\infty$ to an infimum or suboptimal-selector gap over an empty set. The risk of one selector and predictor on source e is

$$R_e(g, h) = \mathbb{E}_{P_e} \ell(h(Q, D_{g(D,Q)}, g(D, Q)), Y). \tag{1}$$

In words, this is the average prediction error after rule g selects the blocks and positions available to predictor h. The visible position pattern is included because it can itself transmit information.

Let $[K] = \{1, \ldots, K\}$ index the declared evaluation sources. We evaluate one shared predictor on all of them and define

$$V(g) = \inf_{h \in \mathcal{H}} \max_{e \in [K]} R_e(g, h), \qquad V^* = \inf_{g \in \mathcal{G}_b} V(g). \tag{2}$$

Thus $V(g)$ is the smallest worst-source error obtainable after fixing selector g, and V^* is the best such error under budget b. The shared predictor represents one downstream model used across all declared sources. A source-specific predictor or a predictor that receives an explicit source identifier defines a different class. Its analysis requires a corresponding change to $\mathcal{H}$. We assume that the infima in formal statements are attained. An approximate optimizer adds its stated tolerance.

Definition 1 (Robustly optimal and approximately optimal selectors)**.** For tolerance $\tau \geq 0$, define

$$\mathcal{G}_\star(\tau) = \{g \in \mathcal{G}_b \mid V(g) \leq V^* + \tau\}. \tag{3}$$

The robustly optimal selector set is $\mathcal{G}_\star = \mathcal{G}_\star(0)$.

This set contains every budgeted selection rule that is as good as the best one on the declared sources. If V^* equals the risk obtained from the full document, the rules in $\mathcal{G}_\star$ are also performance-preserving. Otherwise they are only optimal within budget, and the approximation caused by discarding content remains visible. The excess-risk gap $V(g) - V^*$ is the performance difference between rule g and the best budgeted rule.

Identification and finite-sample recovery answer different questions. Identification asks what can be determined with unlimited observations from the stated interface. Finite-sample recovery asks whether a procedure returns the correct set with high probability from finitely many independent documents. Section 3 withholds target answers, Section 4 supplies labeled document, query, and answer triples with known source identities, and Section 5 supplies controlled masked views in one source. A revealing data source is one on which a suboptimal rule incurs detectable extra risk. The informative-example rate is the probability that one sampled view actually exposes a disputed comparison.

The value in Eq. (2) refits the predictor within $\mathcal{H}$ after a selector is fixed, and every main-text result uses this interface. A frozen audit instead evaluates one predictor h_0 through

$$V_{h_0}(g) = \max_e R_e(g, h_0). \tag{4}$$

These targets need not rank selectors in the same order. Appendix Proposition 13 bounds the discrepancy through the predictor adaptation gap and gives a strict ranking reversal. Appendix Proposition 14 makes the same distinction between one shared predictor and source-specific predictors. This distinction also appears in remove-and-retrain audits of feature importance (Hooker et al., 2019).

For a block-sparse LLM, g abstracts the router, one mask is its output, and h abstracts the downstream attention layers or decoder. The routed transcript contains retained content, positions, and metadata. Token-level routing is the special case of one token per block. The router can inspect the full document, so b limits retained content without limiting search or guaranteeing end-to-end savings. Appendix A separates prediction, performance preservation, evidence, and efficiency. Our results address statistical identification and recovery. Experiments on a concrete LLM test whether an implementation satisfies this interface. They do not alter the identification results.

The resource accounting uses two units. A document is an independent draw of (D, Q, Y). A view is one selector evaluation on that document. A design that tests at most q rules per document uses n independent documents and at most nq views. This distinction matters because several correlated views can share one document without creating additional independent evidence. Router access cost remains separate from both units.

3 Why Prediction Alone Cannot Tell What to Keep

Suppose a sparse attention router predicts every available next token correctly. This success can have three explanations. The retained block can contain a stable fact. A shortcut can be perfect in the observed source. The selected position can encode the answer. We first characterize the information left unresolved by the observation interface. We then separate redundancy from position-based communication.

3.1 Compatible worlds give a partial target

Let $\mathfrak{P}(\mathsf{I})$ be the collection of deployment worlds compatible with an observation interface I and the declared model assumptions. This follows the set-identification perspective (Manski, 2003; Chernozhukov et al., 2007). A world $P \in \mathfrak{P}(\mathsf{I})$ specifies every evaluation-source law and induces a robust value $V_P(g)$. Write $V_P^* = \inf_a V_P(a)$ and $d_P(g) = V_P(g) - V_P^*$. We assume that $\mathfrak{P}(\mathsf{I})$ is nonempty and compact. We also assume continuity of $d_P(g)$ in P for the next result. These conditions guarantee attainment of the regret extrema.

Definition 2 (Possible and necessary optimality)**.** Define the regret envelopes

$$\underline{d}_{\mathsf{I}}(g) = \inf_{P \in \mathfrak{P}(\mathsf{I})} d_P(g), \qquad \overline{d}_{\mathsf{I}}(g) = \sup_{P \in \mathfrak{P}(\mathsf{I})} d_P(g). \tag{5}$$

The possibly optimal set is $\mathcal{G}_\exists(\mathsf{I}) = \bigcup_{P \in \mathfrak{P}(\mathsf{I})} \mathcal{G}_\star(P)$. The necessarily optimal set is $\mathcal{G}_\forall(\mathsf{I}) = \bigcap_{P \in \mathfrak{P}(\mathsf{I})} \mathcal{G}_\star(P)$.

Theorem 3 (Partial identification of sparse selections)**.** *The observation interface identifies*

$$\mathcal{G}_\exists(\mathsf{I}) = \{g : \underline{d}_\mathsf{I}(g) = 0\}, \qquad \mathcal{G}_\forall(\mathsf{I}) = \{g : \overline{d}_\mathsf{I}(g) = 0\}.$$

The robustly optimal selector set is point identified exactly when these two sets agree. If a stronger interface I' leaves fewer compatible worlds, then

$$\mathcal{G}_\exists(\mathsf{I}') \subseteq \mathcal{G}_\exists(\mathsf{I}), \qquad \mathcal{G}_\forall(\mathsf{I}') \supseteq \mathcal{G}_\forall(\mathsf{I}).$$

The theorem turns nonidentification into a set-valued conclusion. A rule outside the possible set is ruled out by every compatible deployment law. A rule inside the necessary set survives every such law. Additional labels shrink uncertainty in the stated directions. The proof uses nonnegative regret, attainment, and set inclusion. It does not require a neural optimization model.

The compatible-world class is part of the model. Appendix Proposition 20 shows that a uniform regret perturbation of size ζ expands the tolerance of the possible and necessary sets by at most ζ. This separates a genuine identification failure from sensitivity near a zero gap.

3.2 One source can hide the wrong reason

Consider a generous interface that supplies unlimited labeled documents from one source and unlimited unlabeled documents from the deployment source. Missing deployment labels can still leave different answer rules compatible with all observations.

Theorem 4 (One-source ambiguity)**.** *There are two worlds with the same labeled source law $P_S(D, Q, Y)$ and the same unlabeled target law $P_T(D, Q)$. Take two binary blocks, a constant query, the selector class $\mathcal{G}_b = \{g_1, g_2\}$, and all binary predictors on one retained bit. Let g_1 retain the first block and let g_2 retain the second. In the source, set $X_1 = X_2 = Y$ for a fair bit. In the target, let X_1 and X_2 be independent fair bits. World A uses $Y = X_1$, and world B uses $Y = X_2$. For the evaluation family containing the source and target, the following statements hold.*

1. *World A has $\mathcal{G}_\star = \{g_1\}$, and world B has $\mathcal{G}_\star = \{g_2\}$. Both worlds have $V^* = 0$.*
2. *Every randomized set estimator has error probability at least $1/2$ in one world.*
3. *Every randomized prediction learner has expected target risk at least $1/4$ in one world. This bound also applies to dense predictors, and the minimax value equals $1/4$.*

For this interface, Theorem 3 gives $\mathcal{G}_\exists = \{g_1, g_2\}$ and $\mathcal{G}_\forall = \varnothing$. The source makes both blocks perfectly predictive. The unlabeled target reveals only that they no longer agree. It does not reveal which block determines the answer. The worlds disagree on half of the target documents, which produces the one-quarter minimax risk. The construction rules out identification rather than one training method.

3.3 Equivalent transcripts and route communication

Nonuniqueness is sometimes harmless. A report can state the Berlin headquarters in an executive summary and a legal notice. Appendix Proposition 15 defines transcript equivalence through a reversible transformation of content and visible positions. Equivalent masks have equal robust value. Appendix Corollary 16 gives an η risk bound when equivalence fails with probability at most η. These results explain why the target is a selector set rather than one physical location.

A different problem appears when a router reads the document before choosing visible positions. The index pattern can transmit a computed answer to the downstream predictor.

Proposition 5 (Selected indices form a side channel)**.** *Suppose Y takes M values and is a deterministic function of (D, Q). If the number of admissible index sets is at least M, there are a selector and predictor with zero loss. The predictor ignores all selected block values and decodes Y from the selected indices.*

The proof assigns one index pattern to each answer. This is a worst-case statement about the full-access interface. It does not assert that trained routers deliberately use such a code. Appendix Theorem 17 gives the exact communication alphabet for visible and hidden positions. For one

retained block, visible positions alone provide $\log_2 m$ bits across a document of length m. The general capacity grows as $b \log_2(m/b)$ plus the bits carried by retained values.

A small retention budget therefore limits information only after the access and transcript interfaces are fixed. Appendix Proposition 18 gives a parity task where value-independent selection below the required budget has risk one half. The results separate a wrong shortcut, harmless redundancy, and route-based answer coding. Low prediction loss cannot distinguish them. The next section studies the labeled source variation that removes compatible worlds.

4 How Labeled Data Sources Expose Bad Selections

A data source can represent a domain, publisher, time period, or document template. Labeled variation means that every sampled document has a query, answer, and known source identity. Such variation is useful only when the labels reveal a failure. In the headquarters example, French companies based outside Paris expose the customer-description shortcut. More samples that preserve the shortcut only estimate its risk more precisely.

The key quantifier ranges over predictors. For every predictor paired with a suboptimal selector, some source needs to expose excess loss. The source can change with the predictor. We first express this mechanism through risk geometry. We then estimate the resulting robust values and identify the complete optimal set.

4.1 Risk geometry and revealing mixtures

For a fixed selector g, each predictor defines a risk vector $\mathbf{r}_g(h) = (R_1(g,h), \ldots, R_K(g,h))$. The robust value is the smallest ℓ_∞ norm in the attainable risk set. A single source need not expose every predictor. A mixture of sources provides the exact population certificate under standard convexity conditions.

Theorem 6 (Revealing-mixture duality)**.** *Fix a selector g. Suppose $\mathcal{H}$ is compact and convex. Assume $h \mapsto R_e(g,h)$ is lower semicontinuous and convex for every source. For λ in the probability simplex Δ_K, define*

$$\phi_g(\lambda) = \inf_{h \in \mathcal{H}} \sum_{e=1}^{K} \lambda_e R_e(g,h).$$

Then

$$V(g) = \max_{\lambda \in \Delta_K} \phi_g(\lambda).$$

For every $g \notin \mathcal{G}_\star$ and every $g^* \in \mathcal{G}_\star$, a maximizing mixture λ_g satisfies

$$\phi_g(\lambda_g) - \phi_{g^*}(\lambda_g) \geq V(g) - V^*.$$

The first equality is a minimax identity, and the second display certifies each bad selector. Different sources can expose different predictors. Their mixture still separates the selector after predictor optimization. Convexity means closure under risk-level predictor mixtures, not convexity of neural-network parameters. The proof applies the Sion minimax theorem (Sion, 1958). Appendix Proposition 23 quantifies the certificate loss for a finite deterministic class through its convexification gap.

4.2 Uniform recovery of worst-source risk

Let $S_e = (O_{e,1}, \ldots, O_{e,n_e})$ contain independent documents from P_e. Samples are independent across sources. Define empirical pair risks $\widehat{R}_e(g,h)$ and

$$\widehat{V}(g) = \inf_{h \in \mathcal{H}} \max_{e} \widehat{R}_e(g,h). \tag{6}$$

The composed loss class is

$$\mathcal{F} = \{O \mapsto \ell(h(Q, D_{g(D,Q)}, g(D,Q)), Y) \mid g \in \mathcal{G}_b, h \in \mathcal{H}\}. \tag{7}$$

It includes selection, visible positions, and prediction. Let $\mathfrak{R}_{n_e}^{(e)}(\mathcal{F})$ be its expected absolute Rademacher complexity under P_e, and define

$$\epsilon_{\mathcal{E}}(\delta) = \max_e \left[2\mathfrak{R}_{n_e}^{(e)}(\mathcal{F}) + \sqrt{\frac{\log(2K/\delta)}{2n_e}} \right]. \tag{8}$$

Theorem 7 (Recovery of worst-source selector values)**.** *With probability at least* $1-\delta$*,*

$$\sup_{g \in \mathcal{G}_b} |\widehat{V}(g) - V(g)| \le \epsilon_{\mathcal{E}}(\delta).$$

An empirical optimizer has excess robust value at most $2\epsilon_{\mathcal{E}}(\delta)$*. Let* $\Delta_{\mathcal{E}} = \inf_{g \notin \mathcal{G}_\star}[V(g) - V^*]$*. If* $\Delta_{\mathcal{E}} > 2\epsilon_{\mathcal{E}}(\delta)$*, every empirical optimizer belongs to* $\mathcal{G}_\star$*.*

Uniform concentration first controls every risk coordinate. The maximum over sources and infimum over predictors are both nonexpansive under a uniform perturbation. Comparing empirical and population optima then gives the two-radius term.

Define the candidate set

$$\widehat{\mathcal{G}} = \{g : \widehat{V}(g) \le \inf_a \widehat{V}(a) + 2\epsilon_{\mathcal{E}}(\delta)\}. \tag{9}$$

Corollary 8 (Set and metric recovery)**.** *On the event in Theorem 7,*

$$\mathcal{G}_\star \subseteq \widehat{\mathcal{G}} \subseteq \mathcal{G}_\star(4\epsilon_{\mathcal{E}}(\delta)).$$

Suppose $\mathcal{G}_b$ *has a metric* d *and* $V(g) - V^* \ge c\, d(g, \mathcal{G}_\star)^\alpha$ *for constants* $c > 0$ *and* $\alpha > 0$*. Here* $d(g, A) = \inf_{a \in A} d(g, a)$*, and* d_{H} *denotes Hausdorff distance. Then*

$$d_{\mathrm{H}}(\widehat{\mathcal{G}}, \mathcal{G}_\star) \le \left(\frac{4\epsilon_{\mathcal{E}}(\delta)}{c} \right)^{1/\alpha}.$$

Exact set recovery holds when $\Delta_{\mathcal{E}} > 4\epsilon_{\mathcal{E}}(\delta)$*.*

The first inclusion retains all exact optima. The second removes every selector beyond the estimated risk resolution. The metric condition converts risk resolution into distance from the optimal set. It is useful for continuous router classes, where an exact positive gap often disappears.

4.3 Where information and samples come from

Appendix Proposition 21 gives source-specific confidence intervals. A selector is rejected only after some source raises its lower robust-risk certificate above the best upper certificate. Appendix Theorem 25 provides a sharper result for predeclared revealing mixtures. Its radius depends on $\sum_e \lambda_e^2/n_e$. Allocating samples in proportion to the mixture weights minimizes this quantity. Corollary 27 approximates the full source simplex with a finite certificate library and quantifies the discretization error. Appendix Proposition 28 gives a matching rare-source obstruction proportional to the inverse revealing-source frequency and the inverse squared gap.

Several bad selectors can require different revealing mixtures. Appendix Proposition 26 gives a convex joint allocation whose root-sum-square form reuses each source sample across several certificates.

For finite classes of sizes G and H, the uniform result gives exact recovery when every source has more than

$$n_e > \frac{8\log(2KGH/\delta)}{\Delta_{\mathcal{E}}^2} \tag{10}$$

independent documents. Correlated queries from one document do not create independent samples.

Corollary 9 (Mixtures of observed data sources)**.** *If a fixed pair* (g, h) *has risk at most* r *in every observed source, then it has risk at most* r *under every mixture* $P_w = \sum_e w_e P_e$ *of those sources.*

The corollary covers deployment changes in source proportions. It does not cover a target outside the observed convex hull. Appendix Theorem 24 gives a conditional discrepancy bound for that case. Unlabeled target documents cannot certify a label-rule discrepancy, as Theorem 4 shows. When labeled source variation is unavailable, masking creates controlled comparisons directly.

5 How Masking Tests What Can Be Removed

Natural data sources do not always break the shortcut under study. Masking supplies a controlled evaluation design that passes candidate views to the predictor and retains the original label. The guarantee treats label preservation as an audit-design condition rather than a theorem conclusion. A view that changes the question or answer defines a different task.

Throughout this section, write $R(g, h)$ for risk in one fixed source. Let $V(g) = \inf_h R(g, h)$ and $V^* = \inf_g V(g)$. These are the $K = 1$ cases of Section 2. The infimum evaluates the best declared predictor for each view. Appendix Proposition 13 relates this target to a frozen predictor. Independent documents are the statistical units. Several masks of one document count as several views but one independent unit.

5.1 Multi-view masking estimates every candidate

Assume finite selector and predictor classes of sizes G and H. For document O_i, a randomized design selects a subset $B_i \subseteq \mathcal{G}_b$ of rules to evaluate. The design can depend on the unmasked document, query, and previous audit history. Conditional on $(D_i, Q_i, \mathcal{H}_{i-1})$, its randomization is independent of the current label and losses. Let

$$\pi_{i,g} = \Pr(g \in B_i \mid D_i, Q_i, \mathcal{H}_{i-1}) \tag{11}$$

be the known inclusion propensity. Assume $\pi_{i,g} \geq \rho_g > 0$ and $|B_i| \leq q$. For the masked loss $L_{i,g,h} \in [0, 1]$, define

$$\widehat{R}_\pi(g, h) = \frac{1}{n} \sum_{i=1}^{n} \frac{\mathbf{1}\{g \in B_i\}}{\pi_{i,g}} L_{i,g,h}. \tag{12}$$

Suppose a deterministic κ_g bounds the average predictable second moment almost surely for every h:

$$\frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[\frac{\mathbf{1}\{g \in B_i\} L_{i,g,h}^2}{\pi_{i,g}^2} \,\middle|\, \mathcal{H}_{i-1} \right] \leq \kappa_g. \tag{13}$$

The crude bound $\kappa_g \leq 1/\rho_g$ always holds.

Theorem 10 (Recovery from adaptive multi-view masking)**.** *Let* $t = \log(2GH/\delta)$ *and*

$$r_g = \sqrt{\frac{2\kappa_g t}{n}} + \frac{2t}{3n\rho_g}, \qquad r_{\max} = \max_g r_g.$$

With probability at least $1 - \delta$*, every pair satisfies* $|\widehat{R}_\pi(g, h) - R(g, h)| \leq r_g$*. An empirical optimizer has excess risk at most* $2r_{\max}$*. The empirical selector set with tolerance* $2r_{\max}$ *contains* $\mathcal{G}_\star$ *and contains only selectors in* $\mathcal{G}_\star(4r_{\max})$*. The audit uses* n *independent documents and at most* nq *evaluated views.*

Inverse propensity weighting gives the correct expectation under document-dependent and history-dependent allocation. A martingale Bernstein bound (Freedman, 1975) controls deviations across independent documents. Dependence among masks evaluated on the same document is allowed. The theorem separates document information from view computation. The fixed one-view design in V11 is recovered by taking $q = 1$ and constant propensities. Candidate-specific intervals and that exact reduction appear in Appendix B.

Positive coverage remains necessary. Uniformly sampling q of G rules gives propensity q/G and maximizes the smallest candidate propensity without gap information. Appendix Proposition 33 proves this minimax statement. The same proposition gives the gap-aware allocation when comparison gaps are known. Corollaries 34 and 35 show that parallel views reduce document cost but leave the order of total view cost unchanged.

Label preservation is also quantitative. Appendix Proposition 32 allows the masked-task label to differ from the original label with probability at most η. This mismatch adds at most 2η to the final excess-risk guarantee. It is a bias term and does not vanish through more samples.

5.2 Adaptive masking stops easy comparisons early

A fixed allocation continues to sample a rule after its inferiority is clear. We instead retain an active set and allocate fresh documents only to active candidates. Evaluating rule g reveals the selected view and label, so the loss of every fixed predictor in $\mathcal{H}$ can be computed on that view.

At round r, let $\varepsilon_r = 2^{-r}$. Every active rule is evaluated until it has

$$N_r = \left\lceil \frac{\log(4GHr^2/\delta)}{2\varepsilon_r^2} \right\rceil \tag{14}$$

samples. Let $\widehat{V}_r(g)$ be its minimum empirical predictor risk. Remove g when

$$\widehat{V}_r(g) - \varepsilon_r > \min_{a \text{ active}} [\widehat{V}_r(a) + \varepsilon_r]. \tag{15}$$

The procedure stops when $\varepsilon_r \le \tau/4$ for a requested resolution $0 < \tau \le 1$.

Theorem 11 (Adaptive masking upper and lower bounds)**.** *With probability at least $1-\delta$, the procedure never removes a selector in $\mathcal{G}_\star$ and returns*

$$\mathcal{G}_\star \subseteq \mathcal{A}_\tau \subseteq \mathcal{G}_\star(\tau).$$

Writing $\Delta_g = V(g) - V^*$, its number of views is at most a universal constant times

$$\sum_{\Delta_g > \tau} \frac{\log[GH \log(4/\Delta_g)/\delta]}{\Delta_g^2} + \frac{|\{g : \Delta_g \le \tau\}| \log[GH \log(4/\tau)/\delta]}{\tau^2}.$$

For $0 < \delta < 1/2$, consider procedures that are δ-correct for every unique-best Bernoulli mean vector in $[1/4, 3/4]^G$. At an instance with best mean $1/2$ and suboptimal means $1/2 + \Delta_g$, where $0 < \Delta_g \le 1/4$, every such procedure satisfies

$$\mathbb{E}T \ge \frac{\mathrm{kl}(1-\delta, \delta)}{4} \sum_{g \ne g^*} \frac{1}{\Delta_g^2}.$$

The upper proof uses confidence intervals that hold for all rounds. A rule stops once the radius is below one quarter of its gap. The lower proof changes one Bernoulli candidate at a time and applies a bandit change-of-measure inequality (Kaufmann et al., 2016). The two statements match in their inverse-squared-gap dependence. Logarithmic factors remain between them.

5.3 A unified law for informative examples

Labeled source variation and masking reveal a disputed comparison only on some observations. Let I mark such an observation, with $\Pr(I = 1) = \omega$. Conditional on $I = 1$, let $W \in [-1, 1]$ be the loss difference between two fixed selector and predictor pairs. Write $\theta = \mathbb{E}[W \mid I = 1]$, $|\theta| \ge \gamma$, and $s^2 = \mathbb{E}[W^2 \mid I = 1]$. The observed comparison is $Z = IW$.

Theorem 12 (Informative-example law)**.** *The empirical mean of Z has the correct sign with probability at least $1-\delta$ if*

$$n\omega\gamma^2 \ge 2\left(s^2 + \frac{2\gamma}{3}\right)\log(1/\delta).$$

The simpler condition $n\omega\gamma^2 \ge 8\log(4/\delta)$ is sufficient without knowledge of s^2. For every $0 < \omega \le 1$ and $0 < \gamma \le 1/4$, there are two bounded-loss worlds with different optimal masks and error at least $3/8$ in one world when

$$n\omega\gamma^2 \le 1/128.$$

Bernstein concentration gives the variance-sensitive upper bound because $\mathbb{E}Z = \omega\theta$ and $\mathbb{E}Z^2 = \omega s^2$. The worst-case corollary retains the earlier, simpler constant. The lower bound uses two rare Bernoulli worlds, KL tensorization, and the Pinsker inequality. Appendix Theorem 37 adds a Fano construction. It proves that simultaneous identification among J candidates also requires a logarithmic factor in J.

Figure 1 displays the simpler upper condition and the two-world lower condition. The left panel plots sufficient document counts against the conditional gap. The right panel uses the normalized information $n\omega\gamma^2$. The middle region reflects only the gap between proved constants.

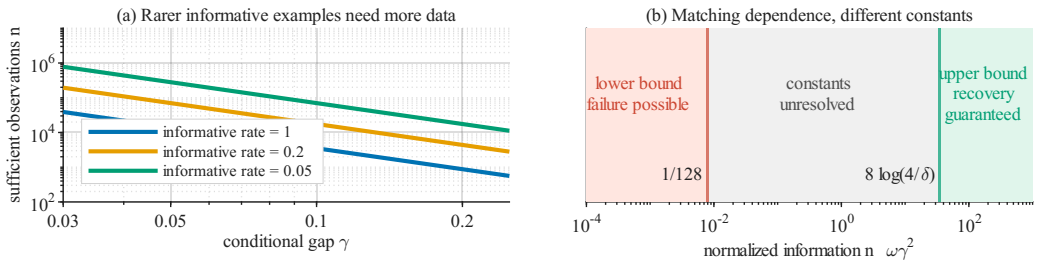


Figure 1: The worst-case analytic bounds in Theorem 12. The left panel uses $\delta = 0.05$. The right panel displays the shared $n\omega\gamma^2$ dependence and the unresolved constants. The curves are theorem-derived illustrations rather than neural network measurements.

Halving the conditional gap multiplies the worst-case sufficient document count by four. Dividing the informative-example rate by five multiplies it by five. Extra samples with $I = 0$ cannot replace the missing information. A valid audit therefore specifies its candidate class, label-preservation condition, document count, view budget, propensities, and confidence level. The certificate concerns prediction risk. It does not establish causal evidence, cheap routing, or neural training convergence.

6 Related Work

Sparse attention and long context. Sparse systems use fixed patterns, content routing, cache selection, and learned gates (Child et al., 2019; Zaheer et al., 2020; Beltagy et al., 2020; Roy et al., 2021; Zhang et al., 2023b; Tang et al., 2024; Jiang et al., 2024; Yuan et al., 2025). SeerAttention is a learned-gating preprint (Gao et al., 2024). Theory studies length generalization and compressibility (Golowich et al., 2025; Vasylenko et al., 2026; Yang & Li, 2026). We study route identification across sources.

Evidence selection and rationalization. Attention explanations remain disputed (Jain & Wallace, 2019; Wiegreffe & Pinter, 2019), and fixed-Transformer weights can be algebraically unidentifiable (Brunner et al., 2020). We instead identify sparse rules by task risk. Rationalization studies short predictive inputs and annotated evidence (Lei et al., 2016; Yu et al., 2019; Guerreiro & Martins, 2021; Vafa et al., 2021; DeYoung et al., 2020). Invariant and causal variants study shifts or explanatory targets (Chang et al., 2020; Zhang et al., 2023a; Liu et al., 2025). Frozen-LLM masking and router co-adaptation provide related interfaces (Allchin, 2026; Aquino-Michaels, 2026), but our set includes route indices and has no causal interpretation.

Stability and identification. Worst-group learning optimizes declared groups (Sagawa et al., 2020). We study a risk-defined selector set using revealing sources and masks, with separate document and view costs. Invariant and domain-adaptation methods study stability and transfer (Peters et al., 2016; Arjovsky et al., 2019; Ben-David et al., 2010). Partial identification studies compatible sets (Manski, 2003; Chernozhukov et al., 2007). Best-arm tools guide masking allocation and lower bounds (Karnin et al., 2013; Garivier & Kaufmann, 2016; Kaufmann et al., 2016; Bhattacharjee et al., 2023).

7 Conclusion

Accurate prediction alone cannot generally tell sparse attention what to keep in an LLM. Under declared source, selector, predictor, and loss classes, compatible worlds yield possible and necessary selector sets. Labeled source variation resolves ambiguity through revealing mixtures, with exact recovery under a uniform positive gap. Multi-view masking separates independent documents from evaluated views, with bounds controlled by informative-example rates and squared gaps. These are conditional prediction-risk certificates. They do not establish causal evidence, frozen-LLM accuracy without the adaptation-gap condition, cheap router access, dependent-document guarantees, or neural training convergence. A next step is to analyze jointly learned routes under dependent document streams.

AI USE STATEMENT

In this work, we used generative AI tools for checking the consistency of mathematical notation and identifying possible typographical or formatting errors in formulas. We have not used generative AI tools for developing the theoretical framework, formulating mathematical claims, deriving results, or generating proofs, and tasks involving datasets, experiments, implementation, qualitative data analysis, and the interpretation of experimental results are not applicable to this work. Additionally, we used generative AI tools for improving the grammar and readability of the manuscript. We reviewed all AI-assisted revisions and manually verified that they preserved the intended mathematical meaning and did not alter the technical claims, derivations, or proofs. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All assumptions and core formal statements appear in Sections 2 through 5. Additional results and complete proofs are in Appendices B and E. Figure 1 visualizes the analytic bounds in Theorem 12 using the parameters stated in its caption. No model training, external dataset, or executable artifact is needed to verify the formal claims. No code or data accompany the submission.

REFERENCES

James E. Allchin. Learning what matters: Supervising global context pruning with causal evidence sets. *CoRR*, abs/2607.21692, 2026.

Keston Aquino-Michaels. Routing absorption in sparse attention: Why random gates are hard to beat. *CoRR*, abs/2603.02227, 2026.

Martin Arjovsky, Leon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *CoRR*, abs/1907.02893, 2019.

Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Mach. Learn.*, 79(1–2):151–175, 2010.

Anirban Bhattacharjee, Sushant Vijayan, and Sandeep Juneja. Best arm identification in rare events. In *UAI*, pp. 163–172. PMLR, 2023.

Gino Brunner, Yang Liu, Damián Pascual, Oliver Richter, Massimiliano Ciaramita, and Roger Wattenhofer. On identifiability in transformers. In *ICLR*, 2020.

Shiyu Chang, Yang Zhang, Mo Yu, and Tommi Jaakkola. Invariant rationalization. In *ICML*, pp. 1448–1458, 2020.

Victor Chernozhukov, Han Hong, and Elie Tamer. Estimation and confidence regions for parameter sets in econometric models. *Econometrica*, 75(5):1243–1284, 2007.

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers. *CoRR*, abs/1904.10509, 2019.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. ERASER: A benchmark to evaluate rationalized NLP models. In *ACL*, pp. 4443–4458. Association for Computational Linguistics, 2020.

David A. Freedman. On tail probabilities for martingales. *Ann. Probab.*, 3(1):100–118, 1975.

Yizhao Gao, Zhichen Zeng, Dayou Du, Shijie Cao, Peiyuan Zhou, Jiaxing Qi, Junjie Lai, Hayden Kwok-Hay So, Ting Cao, Fan Yang, and Mao Yang. SeerAttention: Learning intrinsic sparse attention in your LLMs. *CoRR*, abs/2410.13276, 2024.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *COLT*, pp. 998–1027. PMLR, 2016.

Noah Golowich, Samy Jelassi, David Brandfonbrener, Sham M. Kakade, and Eran Malach. The role of sparsity for length generalization in LLMs. In *ICML*, pp. 19809–19840, 2025.

Nuno M. Guerreiro and André F. T. Martins. SPECTRA: Sparse structured text rationalization. In *EMNLP*, pp. 6534–6550. Association for Computational Linguistics, 2021.

Sara Hooker, Dumitru Erhan, Pieter-Jan Kindermans, and Been Kim. A benchmark for interpretability methods in deep neural networks. In *NeurIPS*, pp. 9734–9745, 2019.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *NAACL-HLT*, pp. 3543–3556. Association for Computational Linguistics, 2019.

Huiqiang Jiang, Yucheng Li, Chengruidong Zhang, Qianhui Wu, Xufang Luo, Surin Ahn, Zhenhua Han, Amir H. Abdi, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. MInference 1.0: Accelerating pre-filling for long-context LLMs via dynamic sparse attention. In *NeurIPS*, pp. 52481–52515, 2024.

Zohar Karnin, Tomer Koren, and Oren Somekh. Almost optimal exploration in multi-armed bandits. In *ICML*, pp. 1238–1246, 2013.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *JMLR*, 17(1):1–42, 2016.

Tao Lei, Regina Barzilay, and Tommi Jaakkola. Rationalizing neural predictions. In *EMNLP*, pp. 107–117. Association for Computational Linguistics, 2016.

Wei Liu, Zhongyu Niu, Lang Gao, Zhiying Deng, Jun Wang, Haozhao Wang, and Ruixuan Li. Adversarial cooperative rationalization: The risk of spurious correlations in even clean datasets. In *ICML*, pp. 39126–39146, 2025.

Charles F. Manski. *Partial Identification of Probability Distributions*. Springer, 2003.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *J. R. Stat. Soc. Ser. B*, 78(5):947–1012, 2016.

Aurko Roy, Mohammad Saffar, Ashish Vaswani, and David Grangier. Efficient content-based sparse attention with routing transformers. *Trans. Assoc. Comput. Linguistics*, 9:53–68, 2021.

Shiori Sagawa, Pang Wei Koh, Tatsunori B. Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. In *ICLR*, 2020.

Maurice Sion. On general minimax theorems. *Pac. J. Math.*, 8(1):171–176, 1958.

Jiaming Tang, Yilong Zhao, Kan Zhu, Guangxuan Xiao, Baris Kasikci, and Song Han. QUEST: Query-aware sparsity for efficient long-context LLM inference. In *ICML*, pp. 47901–47911, 2024.

Keyon Vafa, Yuntian Deng, David Blei, and Alexander Rush. Rationales for sequential predictions. In *EMNLP*, pp. 10314–10332. Association for Computational Linguistics, 2021.

Pavlo Vasylenko, Hugo Pitorro, André F. T. Martins, and Marcos Treviso. Long-context generalization with sparse attention. In *ICLR*, 2026.

Sarah Wiegreffe and Yuval Pinter. Attention is not not explanation. In *EMNLP-IJCNLP*, pp. 11–20. Association for Computational Linguistics, 2019.

Chiwun Yang and Xiaoyu Li. Beyond sparse weights: When is attention compressible? *CoRR*, abs/2608.21541, 2026.

Mo Yu, Shiyu Chang, Yang Zhang, and Tommi Jaakkola. Rethinking cooperative rationalization: Introspective extraction and complement control. In *EMNLP-IJCNLP*, pp. 4094–4103. Association for Computational Linguistics, 2019.

Jingyang Yuan, Huazuo Gao, Damai Dai, Junyu Luo, Liang Zhao, Zhengyan Zhang, Zhenda Xie, Yuxing Wei, Lean Wang, Zhiping Xiao, Yuqing Wang, Chong Ruan, Ming Zhang, Wenfeng Liang, and Wangding Zeng. Native sparse attention: Hardware-aligned and natively trainable sparse attention. In *ACL*, pp. 23078–23097. Association for Computational Linguistics, 2025.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontañón, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. Big Bird: Transformers for longer sequences. In *NeurIPS*, pp. 17283–17297, 2020.

Wenbo Zhang, Tong Wu, Yunlong Wang, Yong Cai, and Hengrui Cai. Towards trustworthy explanation: On causal rationalization. In *ICML*, pp. 41715–41736, 2023a.

Zhenyu Zhang, Ying Sheng, Tianyi Zhou, Tianlong Chen, Lianmin Zheng, Ruisi Cai, Zhao Song, Yuandong Tian, Christopher Ré, Clark W. Barrett, Zhangyang Wang, and Beidi Chen. H2O: Heavy-hitter oracle for efficient generative inference of large language models. In *NeurIPS*, pp. 34661–34710, 2023b.

A Claim, Interface, and Resource Summary

Table 1: Four claims separated by the model.

Claim	Meaning
Prediction	the observed label is predicted well
Performance preservation	selected information matches full-document task risk
Evidence	selected content explains or causes the answer
Efficiency	selection and reading use the claimed resources

Table 1 separates four conclusions that sparse-routing evaluations can conflate. A selector value also depends on the predictor interface. The main text uses a predictor that is refitted within $\mathcal{H}$ for each selector and shared across sources. The next two results quantify changes to this interface.

Proposition 13 (Bridge to a fixed predictor)**.** *Fix $h_0 \in \mathcal{H}$ and define*

$$V_{h_0}(g) = \max_e R_e(g, h_0), \qquad A_{h_0}(g) = V_{h_0}(g) - V(g), \qquad \beta = \sup_g A_{h_0}(g) - \inf_g A_{h_0}(g).$$

Then $A_{h_0}(g) \geq 0$. Every $\widehat{g}_0 \in \arg\min_g V_{h_0}(g)$ satisfies

$$V(\widehat{g}_0) - V^* \leq \beta.$$

If $\inf_{g \notin \mathcal{G}_\star}[V(g) - V^*] > \beta$, every fixed-predictor minimizer belongs to $\mathcal{G}_\star$. A constant adaptation gap gives $\beta = 0$ and preserves the optimizer set. Strict ranking reversal is possible when the gap condition fails.

Proof. The inequality $A_{h_0}(g) \geq 0$ follows because h_0 is feasible in the infimum that defines $V(g)$. Choose $g^* \in \mathcal{G}_\star$. Optimality of $\widehat{g}_0$ under the fixed predictor gives

$$V(\widehat{g}_0) - V^* \leq A_{h_0}(g^*) - A_{h_0}(\widehat{g}_0) \leq \sup_g A_{h_0}(g) - \inf_g A_{h_0}(g) = \beta. \tag{16}$$

The gap condition rules out every selector outside $\mathcal{G}_\star$. A constant adaptation gap adds the same value to every selector objective.

For strict reversal, one source and two selectors suffice. Let $\mathcal{H} = \{h_0, h_1\}$. Set the risks of (h_0, h_1) to $(0.4, 0)$ for selector g_a and to $(0.1, 0.2)$ for selector g_b. Refitting gives $V(g_a) = 0 < V(g_b) = 0.1$. The fixed predictor gives $V_{h_0}(g_b) = 0.1 < V_{h_0}(g_a) = 0.4$. □

Proposition 14 (Shared and source-specific predictors)**.** *Define*

$$V_{\text{shared}}(g) = \inf_h \max_e R_e(g, h), \qquad V_{\text{specific}}(g) = \max_e \inf_h R_e(g, h).$$

Then $V_{\text{specific}}(g) \leq V_{\text{shared}}(g)$ for every selector. The inequality can be strict, and the two values can rank selectors in opposite orders.

Proof. The minimax inequality gives the first claim. For a strict ranking reversal, use two sources and a convex class generated by predictors h_1, h_2. For selector g_a, let their risk vectors be $(0, 1)$ and $(1, 0)$. Randomization gives $V_{\text{shared}}(g_a) = 1/2$, and source-specific prediction gives $V_{\text{specific}}(g_a) = 0$. For selector g_b, let every predictor have risk vector $(0.4, 0.4)$. Then both values equal 0.4. Shared prediction selects g_b, and source-specific prediction selects g_a. □

Table 2: The masking designs separate independent documents from evaluated views.

Design	Allocation rule	Resource pair
One-view fixed	one randomized rule per document	(n, n)
Multi-view fixed	at most q rules per document	(n, nq)
Adaptive elimination	fresh documents for active rules	(n, T)

The resource pairs in Table 2 list independent documents first and evaluated views second. Multiple views on one document can reduce view-selection overhead. They do not increase the number of independent documents. Router access is a third resource and remains outside these pairs.

B Additional Formal Results

This appendix states supporting results in the order of the three core sections. Complete proofs appear in Appendix E.

B.1 Prediction-only limits

Proposition 15 (Transcript-equivalent evidence)**.** *Let $S \subseteq [m]$ be a fixed mask with $i \in S$ and $j \notin S$, and set $S' = (S \setminus \{i\}) \cup \{j\}$. Suppose $X_j = \tau(X_i)$ almost surely in every evaluation source for a measurable bijection τ. Let T replace the visible pair (i, X_i) by $(j, \tau(X_i))$ and preserve every other selected pair. If $\mathcal{H}$ is closed under precomposition by T and T^{-1}, then $V(g_S) = V(g_{S'})$.*

Corollary 16 (Approximate transcript equivalence)**.** *Retain the masks, transcript map, and closure assumptions of Proposition 15. Replace exact transcript equality by $\max_e \Pr_e[X_j \neq \tau(X_i)] \leq \eta$. Then $|V(g_S) - V(g_{S'})| \leq \eta$.*

Theorem 17 (Capacity of the routed transcript)**.** *Consider an abstract channel in which an encoder selects at most b positions from m and passes symbols from an alphabet of size A in canonical order. If positions are visible, the output alphabet has size*

$$N^{\text{vis}}_{m,b,A} = \sum_{j=0}^{b} \binom{m}{j} A^j.$$

If positions are hidden, its size is

$$N^{\text{hid}}_{b,A} = \sum_{j=0}^{b} A^j.$$

Both cardinalities are attainable as channel alphabets. Zero-error transmission of M messages requires M no larger than the corresponding cardinality. For $1 \leq b \leq m/2$,

$$b \log_2(m/b) + b \log_2 A \leq \log_2 N^{\text{vis}}_{m,b,A} \leq b \log_2(em/b) + b \log_2 A + \log_2(b + 1).$$

With one retained constant symbol and a nonempty mask, visible positions transmit exactly $\log_2 m$ bits.

The theorem describes the downstream transcript channel after a router has computed a message. Attainability as a communication channel does not assert that a given document distribution supplies all messages to the router.

Proposition 18 (Some queries require every supporting block)**.** *Let $X_1, \ldots, X_r$ be independent fair bits, let Y be their parity, and let Q be constant. Suppose the selector chooses a fixed or query-dependent set independent of the block values, and the predictor receives only selected values. Every selector with budget $b < r$ has zero-one risk $1/2$. Budget $b = r$ permits zero risk.*

Proposition 19 (Minimax regret under partial identification). *Assume $\mathcal{G}_b$ is compact and $g \mapsto \overline{d}_{\mathsf{I}}(g)$ is lower semicontinuous. Define*

$$\rho_{\mathsf{I}}^* = \min_{g \in \mathcal{G}_b} \overline{d}_{\mathsf{I}}(g).$$

This is the smallest worst-compatible-world regret achievable by one selector. It equals zero exactly when $\mathcal{G}_\forall(\mathsf{I})$ is nonempty. A stronger interface cannot increase ρ_{I}^* or the width $\overline{d}_{\mathsf{I}}(g) - \underline{d}_{\mathsf{I}}(g)$ for any fixed g.

For $\tau \geq 0$, define approximate partial-identification sets

$$\mathcal{G}_\exists^{\mathfrak{P}}(\tau) = \{g : \inf_{P \in \mathfrak{P}} d_P(g) \leq \tau\}, \qquad \mathcal{G}_\forall^{\mathfrak{P}}(\tau) = \{g : \sup_{P \in \mathfrak{P}} d_P(g) \leq \tau\}. \tag{17}$$

Proposition 20 (Stability of partial identification). *Let $\mathfrak{P}$ and $\widetilde{\mathfrak{P}}$ be two nonempty compatible-world classes. Suppose that every $P \in \mathfrak{P}$ has a $\widetilde{P} \in \widetilde{\mathfrak{P}}$ satisfying*

$$\sup_g |d_P(g) - d_{\widetilde{P}}(g)| \leq \zeta,$$

and the same condition holds with the two classes reversed. Then their lower and upper regret envelopes differ uniformly by at most ζ. Consequently, for every $\tau \geq 0$,

$$\mathcal{G}_\exists^{\mathfrak{P}}(\tau) \subseteq \mathcal{G}_\exists^{\widetilde{\mathfrak{P}}}(\tau + \zeta), \qquad \mathcal{G}_\forall^{\mathfrak{P}}(\tau) \subseteq \mathcal{G}_\forall^{\widetilde{\mathfrak{P}}}(\tau + \zeta),$$

with symmetric inclusions after exchanging the classes. Exact partial sets are therefore stable as approximate sets under a uniform perturbation of compatible regrets.

B.2 Labeled data sources

For finite classes of sizes G and H, define

$$r_e = \sqrt{\frac{\log(2KGH/\delta)}{2n_e}}, \quad L(g) = \inf_h \max_e [\widehat{R}_e(g,h) - r_e], \quad U(g) = \inf_h \max_e [\widehat{R}_e(g,h) + r_e]. \tag{18}$$

Proposition 21 (Revealing-source confidence intervals). *With probability at least $1-\delta$, every selector satisfies $L(g) \leq V(g) \leq U(g)$. The set*

$$\mathcal{C} = \{g : L(g) \leq \inf_a U(a)\}$$

contains $\mathcal{G}_\star$. It excludes every selector whose lower endpoint exceeds the best upper endpoint.

Proposition 22 (Active sources in a finite predictor game). *Fix a selector and a finite predictor class. Allow randomized predictors $p \in \Delta_H$. There is a saddle point (p^*, λ^*) for the source-predictor risk game. Every source with $\lambda_e^* > 0$ satisfies*

$$R_e(g, p^*) = V_{\mathrm{rand}}(g),$$

and every predictor with $p_h^* > 0$ satisfies

$$\sum_e \lambda_e^* R_e(g,h) = V_{\mathrm{rand}}(g),$$

where $V_{\mathrm{rand}}(g) = \min_{p \in \Delta_H} \max_e R_e(g,p)$. Here $R_e(g,p) = \sum_h p_h R_e(g,h)$. Thus a positive dual weight identifies a source that is active at the robust optimum.

Proposition 23 (Dual certificates with a nonconvex predictor class). *Let $\mathcal{H}$ be finite. Define the deterministic value $V_{\mathrm{det}}(g) = \min_h \max_e R_e(g,h)$ and the randomized value*

$$V_{\mathrm{rand}}(g) = \min_{p \in \Delta_H} \max_e \sum_h p_h R_e(g,h).$$

Let $c_g = V_{\mathrm{det}}(g) - V_{\mathrm{rand}}(g) \geq 0$. Then

$$V_{\mathrm{rand}}(g) = \max_{\lambda \in \Delta_K} \min_{h \in \mathcal{H}} \sum_e \lambda_e R_e(g,h).$$

If g^* minimizes the deterministic value and λ_g maximizes the last display, then

$$\phi_g(\lambda_g) - \phi_{g^*}(\lambda_g) \geq V_{\mathrm{det}}(g) - V_{\mathrm{det}}(g^*) - c_g.$$

Thus a positive deterministic certificate remains whenever the deterministic selector gap exceeds c_g.

For the composed class in Eq. (7), define

$$d_{\mathcal{F}}(P_T, P_w) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{P_T} f - \mathbb{E}_{P_w} f|. \tag{19}$$

For an action $a = (g, h)$, write $R_P(a) = \mathbb{E}_P f_a$. Let a_T^* minimize target risk.

Theorem 24 (Conditional target guarantee)**.** *For every learned pair $\widehat{a}$ and every observed-source mixture P_w,*

$$R_T(\widehat{a}) - R_T(a_T^*) \le R_w(\widehat{a}) - R_w(a_T^*) + 2d_{\mathcal{F}}(P_T, P_w).$$

The tightest bound of this form is obtained by taking the infimum over $w \in \Delta_K$.

Let $\Lambda \subset \Delta_K$ be a finite, predeclared set of source mixtures. A mixture chosen on an independent pilot sample also qualifies. Define

$$\widehat{\phi}_g(\lambda) = \inf_h \sum_e \lambda_e \widehat{R}_e(g, h), \qquad r(\lambda) = \sqrt{\frac{1}{2} \log\left(\frac{2GH|\Lambda|}{\delta}\right) \sum_e \frac{\lambda_e^2}{n_e}}. \tag{20}$$

Theorem 25 (Finite-sample revealing-mixture certificates)**.** *With probability at least $1 - \delta$,*

$$|\widehat{\phi}_g(\lambda) - \phi_g(\lambda)| \le r(\lambda)$$

for every g and $\lambda \in \Lambda$. Let $U_\star$ be a simultaneous valid upper bound on V^*. A selector is suboptimal whenever

$$\max_{\lambda \in \Lambda} [\widehat{\phi}_g(\lambda) - r(\lambda)] > U_\star.$$

For a fixed mixture and total source sample budget $N = \sum_e n_e$, the continuous allocation $n_e = N\lambda_e$ minimizes $\sum_e \lambda_e^2 / n_e$ over positive-weight sources. The minimum is $1/N$.

Proposition 26 (Allocation for several revealing mixtures)**.** *Fix mixtures $\lambda^{(1)}, \dots, \lambda^{(L)}$ and desired variance proxies $v_1, \dots, v_L > 0$. Let E_+ contain the sources that receive positive weight in at least one mixture. The continuous minimum-document allocation solves*

$$\min_{n_e > 0} \sum_{e \in E_+} n_e \quad \textit{subject to} \quad \sum_{e \in E_+} \frac{(\lambda_e^{(j)})^2}{n_e} \le v_j \quad \textit{for } j \in [L].$$

This is a convex program. There are multipliers $\alpha_j \ge 0$ such that every optimum satisfies

$$n_e = \sqrt{\sum_{j=1}^{L} \alpha_j (\lambda_e^{(j)})^2},$$

together with primal feasibility and complementary slackness. These conditions are also sufficient. For one mixture, they reduce to $n_e = \lambda_e / v_1$ and total budget $1/v_1$. Rounding every positive n_e upward preserves feasibility and adds fewer than $|E_+|$ documents.

Corollary 27 (Finite grids approximate all revealing mixtures)**.** *Let $K \ge 2$ and $0 < \eta \le 1$. Set $M = \lceil 2(K-1)/\eta \rceil$ and define*

$$\Lambda_M = \left\{ (a_1/M, \dots, a_K/M) : a_e \in \mathbb{Z}_+, \ \sum_e a_e = M \right\}.$$

Then $|\Lambda_M| = \binom{M+K-1}{K-1}$, and every $\lambda \in \Delta_K$ is within η in ℓ_1 distance of a grid point. Under Theorem 6,

$$\max_{\lambda \in \Lambda_M} \phi_g(\lambda) \ge V(g) - \eta.$$

Suppose $U_\star \le V^* + u$ and the event in Theorem 25 holds. A suboptimal selector g is certified whenever

$$V(g) - V^* > \eta + u + 2 \max_{\lambda \in \Lambda_M} r(\lambda).$$

Proposition 28 (Rare revealing sources impose a sampling cost)**.** *Suppose source identities are sampled independently and source $e_\star$ appears with probability $p > 0$. There are two bounded-loss worlds that agree on every other source and have different optimal candidates. Conditional on $e_\star$, a Bernoulli comparison favors a different candidate in each world by $0 < \gamma \le 1/4$. Every learner has error at least $3/8$ in one world when*

$$np\gamma^2 \le 1/128.$$

B.3 MASKING DESIGNS

Corollary 29 (Fixed one-view masking)**.** *Draw one selector independently of each document from a fixed design ν, and let $\rho = \min_g \nu(g) > 0$. Then Theorem 10 holds with $q = 1$, $\rho_g = \nu(g)$, and $\kappa_g \le 1/\nu(g)$. In particular, every pair risk is estimated within*

$$\sqrt{\frac{2\log(2GH/\delta)}{n\rho}} + \frac{2\log(2GH/\delta)}{3n\rho}.$$

Under Theorem 10, let $\widehat{V}_\pi(g) = \inf_h \widehat{R}_\pi(g,h)$, $L_\pi(g) = \widehat{V}_\pi(g) - r_g$, and $U_\pi(g) = \widehat{V}_\pi(g) + r_g$.

Proposition 30 (Candidate-specific masking profile)**.** *With probability at least $1-\delta$, the set*

$$\mathcal{C}_\pi = \{g : L_\pi(g) \le \min_a U_\pi(a)\}$$

contains $\mathcal{G}_\star$. Let $r_* = \min_{g^* \in \mathcal{G}_\star} r_{g^*}$. Every $g \in \mathcal{C}_\pi$ satisfies $V(g) - V^* \le 2(r_g + r_*)$. Thus $\mathcal{C}_\pi = \mathcal{G}_\star$ if every suboptimal selector has gap larger than $2(r_g + r_*)$.

Proposition 31 (Coverage is necessary)**.** *Suppose the interface reveals only the query, label, selected mask, and selected view. If a candidate g_0 has zero inclusion probability, two bounded-loss worlds can have identical feedback and identical risks for every tested mask. The rule g_0 is uniquely optimal in one world and strictly suboptimal in the other.*

Proposition 32 (Sensitivity to label changes under masking)**.** *Let $Y^{(g)}$ be the label for the task induced by view g. Suppose $\Pr(Y^{(g)} \ne Y) \le \eta_g$ and changing the label changes the bounded loss by at most one. Let $V(g)$ use the original label and let $\widetilde{V}(g)$ use $Y^{(g)}$. Then*

$$|V(g) - \widetilde{V}(g)| \le \eta_g.$$

If $\eta_g \le \eta$ for every selector and an estimated rule satisfies $V(\widehat{g}) \le V^* + \epsilon$, then

$$\widetilde{V}(\widehat{g}) \le \widetilde{V}^* + \epsilon + 2\eta.$$

Thus approximate label preservation contributes an additive bias distinct from sampling error.

Proposition 33 (Optimal static allocation)**.** *Suppose every document receives at most $q \le G$ candidate evaluations. Let $\bar{\pi}_g$ be the average inclusion probability. Then*

$$\min_g \bar{\pi}_g \le q/G.$$

Uniformly drawing a subset of size q attains equality. It is minimax optimal for unknown candidate gaps under criteria controlled by the smallest propensity.

For J informative comparisons with known gaps $\gamma_j > 0$, consider the design problem

$$\max_{0 \le \pi_j \le 1,\ \sum_j \pi_j \le q} \min_j \pi_j \gamma_j^2.$$

Its value and one optimal allocation are

$$t^* = \min\left\{ \min_j \gamma_j^2, \frac{q}{\sum_j \gamma_j^{-2}} \right\}, \qquad \pi_j^* = \frac{t^*}{\gamma_j^2}.$$

Unused budget can be assigned arbitrarily when the first term determines t^*.

Corollary 34 (Parallel views trade documents for evaluations)**.** *Under a uniform subset design with $q \le G$ views per document, every candidate has propensity q/G. Let $t = \log(2GH/\delta)$. The radius in Theorem 10 obeys*

$$r_{\max} \le \sqrt{\frac{2Gt}{nq}} + \frac{2Gt}{3nq}.$$

For $0 < \epsilon \le 1$, the condition $nq \ge 8Gt/\epsilon^2$ gives $r_{\max} < \epsilon$. Hence increasing q reduces the required number of independent documents by the same factor. The sufficient total number of views nq keeps the same order.

Corollary 35 (Document lower bound under a parallel-view budget)**.** *Assume $0 < \delta < 1/2$. In the Bernoulli subclass of Theorem 11, let candidate losses within each document be independent. Suppose each independent document supplies at most q evaluated views selected before their losses are observed. Every δ-correct procedure satisfies*

$$\mathbb{E}N_{\mathrm{doc}} \geq \frac{\mathrm{kl}(1-\delta,\delta)}{4q} \sum_{g \neq g^*} \frac{1}{\Delta_g^2}.$$

Corollary 36 (Finite informative-example profile)**.** *Consider J comparisons on n independent documents. Comparison j is informative with probability $\omega_j > 0$ and has conditional gap $\gamma_j > 0$. All empirical signs are correct with probability at least $1-\delta$ when the upper condition in Theorem 12 holds with failure probability δ/J for every j. Dependence among comparisons within one document is allowed.*

Theorem 37 (Many-comparison lower bound)**.** *Fix $J \geq 4$, $0 < \omega \leq 1$, and $0 < \gamma \leq 1/4$. There are J bounded-loss worlds. World v has candidate v as its unique optimum, and all other candidates have conditional gap γ. The feedback reveals the informative-event indicator and the complete loss vector on that event. Every estimator still satisfies*

$$\sup_{v \in [J]} \Pr_v(\widehat{v} \neq v) \geq 1 - \frac{4n\omega\gamma^2 + \log 2}{\log J}.$$

Consequently, uniform error at most δ requires

$$n\omega\gamma^2 \geq \frac{(1-\delta)\log J - \log 2}{4}.$$

C Result and Proof Map

Table 3, Table 4, and Table 5 record the roles of the formal results for prediction-only data, labeled sources, and masking. They also separate population identification, finite-sample recovery, and interface sensitivity.

Table 3: Prediction-only results.

Result	Question	Main condition	Proof mechanism
Theorem 3	What remains determined by incomplete labels	compact compatible-world class	regret envelopes and set inclusion
Theorem 4	Can one labeled source determine target evidence	missing target labels	indistinguishable worlds
Proposition 15	How should duplicate evidence be represented	reversible transcript map	predictor transfer
Proposition 5	Can positions carry an answer	full router access	injective index code
Theorem 17	How many messages fit in a route	finite positions and alphabet	exact counting
Proposition 20	Are partial sets stable under model perturbations	uniformly close compatible regrets	envelope comparison

D Worked Examples and Boundary Cases

This section solves small instances of the abstract results. The examples use population risks and do not serve as empirical evidence.

D.1 Partial identification in the two-world construction

Restrict the selector class to g_1 and g_2 from Theorem 4. Their target risks are

	g_1	g_2
world A	0	$1/2$
world B	$1/2$	0

(21)

Table 4: Labeled-source results.

Result	Question	Main condition	Proof mechanism
Theorem 6	What exposes a suboptimal selector	convex compact predictor class	minimax duality
Theorem 7	How accurately are robust values estimated	uniform loss-class control	Rademacher concentration
Corollary 8	How does risk error become set error	positive gap or metric error bound	level-set sandwich
Theorem 25	How should samples estimate a source mixture	fixed finite mixture library	weighted Hoeffding bound
Proposition 26	How can certificates share source samples	several fixed source mixtures	convex allocation and KKT conditions
Proposition 28	What does a rare revealing source cost	two source worlds	KL and total variation

Table 5: Masking results.

Result	Question	Main condition	Proof mechanism
Theorem 10	Can several correlated views share a document	known positive propensities	martingale Bernstein bound
Theorem 11	How many views does adaptive elimination need	fresh documents and finite classes	confidence sequence and change of measure
Proposition 33	How should a static view budget be allocated	bounded views per document	max-min allocation
Theorem 12	What controls one disputed comparison	informative rate and conditional gap	Bernstein and two-world testing
Theorem 37	What does simultaneous identification cost	J alternative worlds	Fano inequality

so both selectors have lower regret envelope zero and upper regret envelope one half. The possibly optimal set contains both rules, and the necessarily optimal set is empty. A deterministic minimax-regret choice incurs regret one half in one compatible world.

Randomization changes the decision problem but not the identified deterministic sets. Let a randomized selector choose g_1 with probability p. Its target risks in worlds A and B are $(1-p)/2$ and $p/2$. The worst-compatible-world regret is

$$\frac{1}{2}\max\{p, 1-p\}. \tag{22}$$

It is minimized at $p = 1/2$ with value $1/4$. This is the same value as the dense-prediction minimax risk in Theorem 4. Randomization hedges unresolved worlds. It does not identify which world generated the target labels.

D.2 A REVEALING MIXTURE WITH NO UNIVERSAL REVEALING SOURCE

Consider one candidate selector with two predictors and two sources. Its risk matrix is

	e_1	e_2
h_1	0.6	0.1
h_2	0.1	0.6

. (23)

Allow randomized predictors and let p be the probability of h_1. The two source risks are $0.1 + 0.5p$ and $0.6 - 0.5p$. Their maximum is minimized at $p = 1/2$, so the robust value is 0.35.

No individual source rejects both predictors relative to a comparator with constant risk 0.2. Source e_1 rejects h_1 but accepts h_2. Source e_2 has the reverse behavior. Let λ be the weight on e_1. The best weighted risk is

$$\phi_g(\lambda) = \min\{0.1 + 0.5\lambda,\ 0.6 - 0.5\lambda\}. \tag{24}$$

Its maximum is 0.35 at $\lambda = 1/2$. The equal mixture therefore certifies a gap of 0.15 against the comparator. This example displays the predictor-dependent revealing-source quantifier in Theorem 6.

D.3 Changing the predictor interface reverses the selection

Take one source and two selectors. The risks under predictors (h_0, h_1) are

$$\begin{array}{c|cc} & h_0 & h_1 \\ \hline g_a & 0.4 & 0 \\ g_b & 0.1 & 0.2 \end{array}. \tag{25}$$

Refitting selects g_a because its best risk is zero. A frozen audit using h_0 selects g_b because $0.1 < 0.4$. The adaptation gaps are 0.4 and zero, so their oscillation is 0.4. Proposition 13 allows a reversal of this size. A constant adaptation gap would shift both selector values equally and preserve their ranking.

D.4 Gap-aware masking allocation

Take three informative comparisons with conditional gaps $\gamma = (0.05, 0.10, 0.20)$ and one expected view per document. The inverse squared gaps are $(400, 100, 25)$, whose sum is 525. Proposition 33 gives

$$t^* = 1/525, \qquad \pi^* = (400, 100, 25)/525 \approx (0.762, 0.190, 0.048). \tag{26}$$

Every product $\pi_j^* \gamma_j^2$ equals $1/525$. Uniform allocation gives a smallest product $(1/3)(0.05)^2 = 1/1200$. Gap-aware allocation spends most views on the smallest gap and improves the worst information rate by a factor of $1200/525$. This is an oracle comparison because the gaps are unknown before the audit.

D.5 Approximate label preservation creates bias

Suppose masking changes the task label on at most two percent of documents. Proposition 32 gives a risk perturbation of at most 0.02 for every selector. An observed-label rule with excess risk 0.01 then has view-specific-label excess risk at most 0.05. Increasing the number of documents reduces the sampling term 0.01 only when it arose from estimation. It does not reduce the 0.04 worst-case contribution from label mismatch.

E Complete Proofs

E.1 Proofs for Section 3

E.1.1 Proof of Theorem 3

For every compatible world P, regret satisfies $d_P(g) \geq 0$. If g is optimal in one compatible world, then $\underline{d}_{\mathsf{I}}(g) = 0$. Conversely, compactness of $\mathfrak{P}(\mathsf{I})$ and continuity of $d_P(g)$ give a world that attains the infimum. A zero infimum therefore gives a world with $d_P(g) = 0$. This proves the characterization of $\mathcal{G}_\exists(\mathsf{I})$.

A nonnegative function has supremum zero exactly when it is zero everywhere. Hence $\overline{d}_{\mathsf{I}}(g) = 0$ exactly when g is optimal in every compatible world. This proves the characterization of $\mathcal{G}_\forall(\mathsf{I})$.

The intersection of a collection of sets equals its union exactly when every set in the collection is the same. Thus the two identified sets agree exactly when all compatible worlds have the same robustly optimal selector set. A stronger interface satisfies $\mathfrak{P}(\mathsf{I}') \subseteq \mathfrak{P}(\mathsf{I})$. A union over the smaller collection can only shrink, and an intersection over the smaller collection can only expand. □

E.1.2 Proof of Theorem 4

The labeled source law and unlabeled target law are identical across the two worlds. Each world has a perfect fixed one-block rule. Under world A, g_1 has zero risk in both sources. The second target block is independent of $Y = X_1$, so every predictor based on g_2 has target error one half. Thus $\mathcal{G}_\star = \{g_1\}$. The symmetric argument gives $\mathcal{G}_\star = \{g_2\}$ in world B.

Every set estimator has the same output distribution in both worlds. The probabilities of returning the two distinct singleton sets sum to at most one. Its average error under a uniform prior on the worlds is at least one half. One world therefore has error at least one half.

For a randomized prediction rule, let $a(x_1, x_2)$ be the probability of predicting one. The worlds agree on target inputs with $x_1 = x_2$. On inputs $(0, 1)$ and $(1, 0)$, their conditional error probabilities sum to one. These inputs have total probability one half. The sum of target risks is at least one half, so the larger risk is at least one quarter. Choosing the predictor based on X_1 or X_2 with equal probability attains one quarter in both worlds. □

E.1.3 Proof of Proposition 15

Take any predictor h for mask S. For a transcript produced by S', define $h' = h \circ T^{-1}$. Closure keeps h' in $\mathcal{H}$. The transformed input to h' equals the complete transcript received by h under S. Their losses agree almost surely in every source. Taking the maximum over sources and the infimum over predictors gives $V(g_{S'}) \leq V(g_S)$. Precomposition by T gives the reverse inequality. □

E.1.4 Proof of Corollary 16

Transfer predictors through the transcript maps from Proposition 15. The losses agree on the event $X_j = \tau(X_i)$. Their difference is at most one on its complement. Every source risk changes by at most η. Maxima over sources and infima over predictors preserve this bound in both directions. □

E.1.5 Proof of Proposition 5

Let $\mathcal{M}_b$ be the admissible index family. Choose an injective code c from the label set into $\mathcal{M}_b$. The selector computes $Y(D, Q)$ and returns $c(Y)$. The predictor ignores selected values and returns the inverse code of the visible pattern. Its output equals Y on every observation. □

E.1.6 Proof of Theorem 17

For a visible transcript of length j, there are $\binom{m}{j}$ index sets and A^j value strings. Summing over admissible lengths gives $N^{\mathrm{vis}}_{m,b,A}$. Hidden positions remove the index choice and leave A^j strings at length j, which gives $N^{\mathrm{hid}}_{b,A}$. An encoder with an abstract message input can assign one message to each output symbol, so both channel cardinalities are attainable. Distinct messages require distinct outputs under zero error.

For the asymptotic bounds, the term with $j = b$ gives

$$N^{\mathrm{vis}}_{m,b,A} \geq \binom{m}{b} A^b \geq (m/b)^b A^b. \tag{27}$$

For $b \leq m/2$, both $\binom{m}{j}$ and A^j are nondecreasing over $0 \leq j \leq b$. Therefore

$$N^{\mathrm{vis}}_{m,b,A} \leq (b+1)\binom{m}{b} A^b \leq (b+1)(em/b)^b A^b. \tag{28}$$

Taking base-two logarithms proves the bounds. With one retained constant symbol and a nonempty mask, the output alphabet is the set of m positions. □

E.1.7 Proof of Proposition 18

If $b < r$, at least one parity bit is unobserved. Conditional on all selected bits, the missing bit remains fair and independent. Flipping it changes the label and leaves the predictor input unchanged. The conditional label is fair, so every predictor has error one half. Retaining all r bits permits exact parity computation. □

E.1.8 Proof of Proposition 19

For a fixed selector, $\overline{d}_{\mathsf{I}}(g)$ is its largest regret over compatible worlds. Minimizing this quantity gives the minimax-regret identity by definition. Compactness and lower semicontinuity give a

minimizer. The minimum is zero exactly when some selector has zero regret in every compatible world. Theorem 3 identifies this condition with a nonempty necessary set.

For a stronger interface, the compatible-world collection shrinks. A supremum over this smaller collection cannot increase, and an infimum cannot decrease. Both the minimax value and each difference $\overline{d} - \underline{d}$ therefore cannot increase. □

E.1.9 Proof of Proposition 20

Fix a selector g. For every $P \in \mathfrak{P}$, choose a corresponding $\widetilde{P}$ from the assumption. Then

$$d_P(g) \geq d_{\widetilde{P}}(g) - \zeta \geq \inf_{Q \in \widetilde{\mathfrak{P}}} d_Q(g) - \zeta. \tag{29}$$

Taking the infimum over P gives one direction of the lower-envelope bound. Exchanging the two classes gives the reverse direction. The same argument with suprema gives the upper-envelope bound.

If the lower envelope for g under $\mathfrak{P}$ is at most τ, the lower envelope under $\widetilde{\mathfrak{P}}$ is at most $\tau + \zeta$. This proves the first inclusion. The upper-envelope bound proves the second. Exchanging the classes gives the symmetric inclusions. □

E.2 Proofs for Section 4

E.2.1 Proof of Theorem 6

For fixed g, define $f(h, \lambda) = \sum_e \lambda_e R_e(g, h)$. The simplex is compact and convex. The predictor class is compact and convex by assumption. The function is linear and continuous in λ, and it is lower semicontinuous and convex in h. The Sion minimax theorem gives

$$\inf_{h \in \mathcal{H}} \sup_{\lambda \in \Delta_K} f(h, \lambda) = \sup_{\lambda \in \Delta_K} \inf_{h \in \mathcal{H}} f(h, \lambda). \tag{30}$$

The inner supremum on the left equals $\max_e R_e(g, h)$. The right side equals $\max_\lambda \phi_g(\lambda)$. Compactness and semicontinuity give attainment.

Let λ_g maximize ϕ_g. For an optimal selector g^*,

$$\phi_{g^*}(\lambda_g) = \inf_h \sum_e (\lambda_g)_e R_e(g^*, h) \leq \inf_h \max_e R_e(g^*, h) = V^*. \tag{31}$$

The dual identity gives $\phi_g(\lambda_g) = V(g)$. Subtracting proves the certificate inequality. □

E.2.2 Uniform convergence lemma

Lemma 38 (Expected absolute Rademacher bound)**.** *For a class $\mathcal{F} \subseteq [0,1]^{\mathcal{O}}$ and n independent observations, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}} |\widehat{P}_n f - Pf| \leq 2\mathfrak{R}_n(\mathcal{F}) + \sqrt{\frac{\log(2/\delta)}{2n}}.$$

Proof. Symmetrization bounds the expected supremum by twice the expected absolute Rademacher complexity. Replacing one observation changes the supremum by at most $1/n$. The bounded difference inequality controls upper and lower deviations. A union bound proves the display. □

E.2.3 Proof of Theorem 7

Apply Lemma 38 in every source with failure probability δ/K. A union bound gives

$$\sup_{e,g,h} |\widehat{R}_e(g, h) - R_e(g, h)| \leq \epsilon_{\mathcal{E}}(\delta). \tag{32}$$

The maximum and infimum operators are one-Lipschitz under a uniform perturbation. Hence the same radius bounds $|\widehat{V}(g) - V(g)|$ for every g. If $\widehat{g}$ is an empirical optimizer, then

$$V(\widehat{g}) \leq \widehat{V}(\widehat{g}) + \epsilon_{\mathcal{E}} \leq \inf_g \widehat{V}(g) + \epsilon_{\mathcal{E}} \leq V^* + 2\epsilon_{\mathcal{E}}. \tag{33}$$

A suboptimal-selector gap larger than the final term rules out a suboptimal empirical optimizer. □

E.2.4 Proof of Corollary 8

Work on the uniform event from Theorem 7. For $g \in \mathcal{G}_\star$,

$$\widehat{V}(g) \le V^* + \epsilon_{\mathcal{E}} \le \inf_a \widehat{V}(a) + 2\epsilon_{\mathcal{E}}. \tag{34}$$

Thus $\mathcal{G}_\star \subseteq \widehat{\mathcal{G}}$. If $g \in \widehat{\mathcal{G}}$, then

$$V(g) \le \widehat{V}(g) + \epsilon_{\mathcal{E}} \le \inf_a \widehat{V}(a) + 3\epsilon_{\mathcal{E}} \le V^* + 4\epsilon_{\mathcal{E}}. \tag{35}$$

This proves the risk-set sandwich and exact recovery under the stated gap.

Under the error-bound condition, every $g \in \widehat{\mathcal{G}}$ satisfies

$$c\, d(g, \mathcal{G}_\star)^\alpha \le V(g) - V^* \le 4\epsilon_{\mathcal{E}}. \tag{36}$$

The candidate set contains $\mathcal{G}_\star$, so the reverse directed distance is zero. Taking the supremum over $g \in \widehat{\mathcal{G}}$ proves the Hausdorff bound. □

E.2.5 Proof of Proposition 21

The Hoeffding inequality and a union bound over KGH triples give

$$|\widehat{R}_e(g,h) - R_e(g,h)| \le r_e \tag{37}$$

for every triple with probability at least $1-\delta$. On this event,

$$\max_e [\widehat{R}_e(g,h) - r_e] \le \max_e R_e(g,h) \le \max_e [\widehat{R}_e(g,h) + r_e]. \tag{38}$$

Taking infima over h gives $L(g) \le V(g) \le U(g)$. Every optimal selector has lower endpoint at most V^*, and every upper endpoint is at least V^*. Hence every optimal selector belongs to $\mathcal{C}$. The exclusion rule follows from the definition of $\mathcal{C}$. □

E.2.6 Proof of Proposition 22

The randomized predictor simplex and source simplex are compact and convex, and the payoff is bilinear. The finite minimax theorem gives a saddle point (p^*, λ^*) with value $V_{\mathrm{rand}}(g)$. Every source risk under p^* is at most $V_{\mathrm{rand}}(g)$ because their maximum equals the game value. Their λ^*-weighted average also equals $V_{\mathrm{rand}}(g)$. A source with positive weight cannot have risk below the value, since no source risk exceeds it. This proves the first claim.

Every pure predictor has λ^*-weighted risk at least $V_{\mathrm{rand}}(g)$ because the minimum over randomized predictors equals the value. Their p^*-weighted average equals $V_{\mathrm{rand}}(g)$. A predictor with positive weight therefore has weighted risk equal to the value. □

E.2.7 Proof of Proposition 23

The finite predictor and source simplexes form a finite zero-sum game. Its minimax identity gives

$$V_{\mathrm{rand}}(g) = \max_{\lambda \in \Delta_K} \min_{p \in \Delta_H} \sum_{e,h} \lambda_e p_h R_e(g,h). \tag{39}$$

For a fixed λ, a linear objective over p attains its minimum at a pure predictor. This proves the displayed dual representation. Deterministic predictors are feasible randomized predictors, so $c_g \ge 0$.

At a maximizing mixture λ_g, $\phi_g(\lambda_g) = V_{\mathrm{rand}}(g) = V_{\mathrm{det}}(g) - c_g$. For the deterministic optimum g^*, weak duality gives $\phi_{g^*}(\lambda_g) \le V_{\mathrm{rand}}(g^*) \le V_{\mathrm{det}}(g^*)$. Subtracting proves the certificate bound. □

E.2.8 Proof of Corollary 9

Linearity of expectation gives

$$R_w(g,h) = \sum_e w_e R_e(g,h) \le \max_e R_e(g,h) \le r. \tag{40}$$

□

E.2.9 Proof of Theorem 24

The discrepancy definition gives

$$R_T(\widehat{a}) \le R_w(\widehat{a}) + d_{\mathcal{F}}(P_T, P_w) \tag{41}$$

and

$$R_T(a_T^*) \ge R_w(a_T^*) - d_{\mathcal{F}}(P_T, P_w). \tag{42}$$

Subtracting proves the displayed bound. It holds for every w, so taking an infimum preserves validity. □

E.2.10 Proof of Theorem 25

Fix (g, h, λ). The weighted empirical risk is a sum of independent variables with ranges λ_e/n_e. The Hoeffding inequality gives

$$\Pr\left(\left|\sum_e \lambda_e[\widehat{R}_e(g,h) - R_e(g,h)]\right| > t\right) \le 2\exp\left(-\frac{2t^2}{\sum_e \lambda_e^2/n_e}\right). \tag{43}$$

A union bound over $GH|\Lambda|$ triples gives the radius $r(\lambda)$. Taking an infimum over h changes a uniformly accurate objective by no more than the same radius.

The lower endpoint $\widehat{\phi}_g(\lambda) - r(\lambda)$ is no larger than $\phi_g(\lambda)$, and $\phi_g(\lambda) \le V(g)$ by weak duality. If this lower endpoint exceeds $U_\star \ge V^*$, selector g is suboptimal.

For the allocation claim, the Cauchy–Schwarz inequality gives

$$1 = \left(\sum_e \lambda_e\right)^2 \le \left(\sum_e \frac{\lambda_e^2}{n_e}\right)\left(\sum_e n_e\right). \tag{44}$$

Equality holds exactly when n_e is proportional to λ_e on its positive support. □

E.2.11 Proof of Proposition 26

Every constraint is convex on the positive orthant because $n \mapsto 1/n$ is convex. Strict feasibility follows by taking all sample sizes large enough, so the Karush–Kuhn–Tucker conditions are necessary and sufficient. The Lagrangian is

$$\mathcal{L}(n, \alpha) = \sum_e n_e + \sum_{j=1}^{L} \alpha_j \left[\sum_e \frac{(\lambda_e^{(j)})^2}{n_e} - v_j\right]. \tag{45}$$

Stationarity in n_e gives

$$1 - \frac{\sum_j \alpha_j (\lambda_e^{(j)})^2}{n_e^2} = 0, \tag{46}$$

which is the displayed allocation. Primal feasibility, dual feasibility, and complementary slackness complete the characterization.

For $L = 1$, stationarity makes $n_e = c\lambda_e$ on the positive support. The active constraint gives $1/c = v_1$, so $n_e = \lambda_e/v_1$ and the total is $1/v_1$. Replacing each n_e by its ceiling decreases every inverse-sample constraint. The increase in the total is smaller than the number of active sources. □

E.2.12 Proof of Proposition 28

Let I indicate that the sampled source equals $e_\star$, so $\Pr(I = 1) = p$. Use two candidates and identical constant feedback outside this event. Conditional on $I = 1$, give the first candidate Bernoulli mean $1/2 - \gamma$ in the first world and $1/2 + \gamma$ in the second world. Give the second candidate the known mean $1/2$ in both worlds. The optimal candidate changes across the worlds. The one-observation divergence is at most $16p\gamma^2$. Tensorization and the Pinsker inequality give

$$\mathrm{TV}(P_-^n, P_+^n) \le \sqrt{8np\gamma^2} \le 1/4. \tag{47}$$

Every decision induces a binary test. Its average error is at least $(1 - \mathrm{TV})/2 \ge 3/8$. □

E.2.13 PROOF OF COROLLARY 27

For any $\lambda \in \Delta_K$, set $a_e = \lfloor M\lambda_e \rfloor$ for $e < K$ and $a_K = M - \sum_{e<K} a_e$. This is a grid point. The first $K-1$ coordinate errors are nonnegative and sum to less than $(K-1)/M$. The last coordinate error has the same absolute value as that sum. Therefore

$$\|\lambda - a/M\|_1 < \frac{2(K-1)}{M} \le \eta. \tag{48}$$

The stars-and-bars count gives the stated grid size.

For any predictor, bounded risks give

$$\left|\sum_e (\lambda_e - \lambda'_e) R_e(g,h)\right| \le \|\lambda - \lambda'\|_1. \tag{49}$$

Taking infima over predictors shows that ϕ_g is one-Lipschitz. A grid point within η of a dual maximizer has value at least $V(g) - \eta$.

On the uniform estimation event, its empirical lower endpoint is at least $V(g) - \eta - 2r(\lambda)$. The gap condition makes this quantity larger than $V^* + u$, which is at least $U_\star$. The certificate in Theorem 25 then rejects g. □

E.3 PROOFS FOR SECTION 5

E.3.1 PROOF OF THEOREM 10

Fix (g,h) and write

$$Y_i = \frac{\mathbf{1}\{g \in B_i\}}{\pi_{i,g}} L_{i,g,h}. \tag{50}$$

Conditional on the complete document, query, label, and past history, randomization of B_i gives $\mathbb{E}[Y_i \mid O_i, \mathcal{H}_{i-1}] = L_{i,g,h}$. The fresh document is independent of the past, so $\mathbb{E}[Y_i \mid \mathcal{H}_{i-1}] = R(g,h)$. Thus $Y_i - R(g,h)$ is a martingale difference.

The increment has absolute value at most $1/\rho_g$. Its predictable variance sum is bounded by $n\kappa_g$. A two-sided martingale Bernstein inequality gives

$$\Pr\left(|\widehat{R}_\pi(g,h) - R(g,h)| > \sqrt{\frac{2\kappa_g t}{n}} + \frac{2t}{3n\rho_g}\right) \le 2e^{-t}. \tag{51}$$

Set $t = \log(2GH/\delta)$ and take a union bound. Infima over h preserve the candidate-specific deviations. Replacing every radius by $r_{\max}$ gives the optimizer and set-recovery conclusions through the same two-radius and four-radius comparisons used in Section 4. The deterministic constraint $|B_i| \le q$ gives at most nq evaluated views. □

E.3.2 PROOF OF COROLLARY 29

For a fixed one-view design, $B_i = \{A_i\}$ and $\pi_{i,g} = \nu(g)$. Since $L_{i,g,h} \le 1$,

$$\mathbb{E}\left[\frac{\mathbf{1}\{A_i = g\} L_{i,g,h}^2}{\nu(g)^2}\right] \le \frac{1}{\nu(g)}. \tag{52}$$

Substitution into Theorem 10 gives the result. □

E.3.3 PROOF OF PROPOSITION 30

Work on the simultaneous event from Theorem 10. Infima over predictors give $L_\pi(g) \le V(g) \le U_\pi(g)$ for all g. An optimal selector g^* satisfies

$$L_\pi(g^*) \le V^* \le \min_a U_\pi(a), \tag{53}$$

so it belongs to $\mathcal{C}_\pi$.

Choose $g^\dagger \in \mathcal{G}_\star$ with radius r_*. For $g \in \mathcal{C}_\pi$,

$$V(g) \le L_\pi(g) + 2r_g \le \min_a U_\pi(a) + 2r_g \le U_\pi(g^\dagger) + 2r_g \le V^* + 2r_* + 2r_g. \tag{54}$$

The gap condition excludes all suboptimal selectors. □

E.3.4 Proof of Proposition 31

Use zero-one loss, a constant query, and a fair label Y. Let one tested view g_1 equal $Y \oplus B$, where B is independent and $\Pr(B = 1) = 1/4$. Its best risk is $1/4$. Let every other tested view be an independent fair bit, with best risk $1/2$. These tested views have the same joint law with Y in both worlds.

Place a bit Z only in the untested view g_0. In world A, set $Z = Y$, so g_0 has risk zero and is uniquely optimal. In world B, let Z be an independent fair bit, so g_0 has risk $1/2$ and is strictly worse than g_1. Zero inclusion probability hides Z in every observation. The complete feedback laws agree across worlds. □

E.3.5 Proof of Proposition 32

Fix (g, h). The two losses agree on the event $Y^{(g)} = Y$ and differ by at most one elsewhere. Hence

$$|R(g, h) - \widetilde{R}(g, h)| \le \eta_g. \tag{55}$$

Taking infima over h preserves this bound.

Let $\widetilde{g}^*$ minimize the view-specific-label risk. Uniform bias at most η gives

$$\widetilde{V}(\widehat{g}) \le V(\widehat{g}) + \eta \le V^* + \epsilon + \eta \le V(\widetilde{g}^*) + \epsilon + \eta \le \widetilde{V}^* + \epsilon + 2\eta. \tag{56}$$

□

E.3.6 Proof of Proposition 33

The per-document budget gives

$$\sum_g \bar{\pi}_g = \frac{1}{n} \sum_i \mathbb{E}|B_i| \le q. \tag{57}$$

Hence the smallest propensity is at most the average q/G. A uniformly random subset of size q includes every candidate with probability q/G.

For the gap-aware problem, a value t is feasible exactly when $\pi_j \ge t/\gamma_j^2$ for every j, with $t \le \min_j \gamma_j^2$ and $t \sum_j \gamma_j^{-2} \le q$. The largest feasible value is the stated t^*. Setting $\pi_j = t^*/\gamma_j^2$ attains it. □

E.3.7 Proof of Corollary 34

Uniform subset sampling gives $\rho_g = q/G$. Bounded loss gives $\kappa_g \le G/q$. Substitution into Theorem 10 proves the radius. If $nq \ge 8Gt/\epsilon^2$, its square-root term is at most $\epsilon/2$. Its linear term is at most $\epsilon^2/12$, which is at most $\epsilon/12$. Their sum is below ϵ. The resource conclusions follow from $n = (nq)/q$. □

E.3.8 Proof of Corollary 35

Independent candidate losses make every evaluated view one Bernoulli observation from its candidate. A parallel procedure can be simulated by a sequential procedure that delays all updates until the end of each document. The change-of-measure lower bound in Theorem 11 therefore applies to its total number T of evaluated views. If N_{doc} is the number of independent documents, the budget gives $T \le qN_{\mathrm{doc}}$ on every sample path. Division by q and expectation of the pathwise inequality prove the claim. □

E.3.9 Proof of Theorem 11, upper statement

For every round, selector, and predictor, the Hoeffding inequality gives

$$\Pr(|\widehat{R}_r(g, h) - R(g, h)| > \varepsilon_r) \le 2e^{-2N_r\varepsilon_r^2} \le \frac{\delta}{2GHr^2}. \tag{58}$$

A union bound over all rounds and pairs has total probability below δ. On the complementary event, $|\widehat{V}_r(g) - V(g)| \le \varepsilon_r$ for every active selector.

An optimal selector has lower endpoint at most V^*, and the smallest upper endpoint is at least V^*. It is never removed. A suboptimal selector has lower endpoint at least $V(g) - 2\varepsilon_r$. An optimal selector has upper endpoint at most $V^* + 2\varepsilon_r$. Thus selector g is removed once $4\varepsilon_r < \Delta_g$. At stopping, every active selector has gap at most $4\varepsilon_r \le \tau$.

A selector with $\Delta_g > \tau$ is sampled only through the first round with $\varepsilon_r < \Delta_g/4$. Substitution into Eq. (14) gives its inverse-squared-gap term up to a universal constant. Selectors with gap at most τ remain through the stopping round and give the second term. □

E.3.10 Proof of Theorem 11, lower statement

Consider the Bernoulli subclass with best mean loss $\mu_{g^*} = 1/2$ and suboptimal means $\mu_g = 1/2 + \Delta_g$. Fix a suboptimal g. Form an alternative instance by changing only its mean to $1/2 - \epsilon$, where $\epsilon > 0$. Candidate g is uniquely optimal in the alternative.

Let E be the event that the procedure returns g^*. A δ-correct procedure satisfies $\Pr(E) \ge 1 - \delta$ in the original instance and $\Pr'(E) \le \delta$ in the alternative. The bandit change-of-measure inequality lower-bounds the left-hand side of the following display by $\mathrm{kl}(\Pr(E), \Pr'(E))$. Since $0 < \delta < 1/2$, monotonicity of the binary relative entropy over these probability ranges lower-bounds this divergence by $\mathrm{kl}(1 - \delta, \delta)$. Hence

$$\mathbb{E}[N_g]\ \mathrm{kl}(1/2 + \Delta_g, 1/2 - \epsilon) \ge \mathrm{kl}(1 - \delta, \delta). \tag{59}$$

Let ϵ decrease to zero. The Bernoulli divergence satisfies

$$\mathrm{kl}(1/2 + \Delta_g, 1/2) \le 4\Delta_g^2. \tag{60}$$

Therefore

$$\mathbb{E}[N_g] \ge \frac{\mathrm{kl}(1 - \delta, \delta)}{4\Delta_g^2}. \tag{61}$$

Summing over suboptimal candidates proves the lower bound on $T = \sum_g N_g$. □

E.3.11 Proof of Theorem 12, variance-sensitive upper statement

Set $Z = IW$. Its mean is $\mu = \omega\theta$, and its variance is at most $\mathbb{E}Z^2 = \omega s^2$. Also $|Z - \mu| \le 2$. Suppose $\theta > 0$. The one-sided Bernstein inequality gives

$$\Pr(\overline{Z}_n \le 0) \le \exp\left[-\frac{n\mu^2}{2(\omega s^2 + 2\mu/3)}\right]. \tag{62}$$

The exponent is increasing in $\mu > 0$. Since $\mu \ge \omega\gamma$, it is at least

$$\frac{n\omega\gamma^2}{2(s^2 + 2\gamma/3)}. \tag{63}$$

The case $\theta < 0$ follows by applying the same argument to $-Z$. The displayed condition in the theorem makes the error probability at most δ. □

E.3.12 Proof of Theorem 12, worst-case upper statement

Let N be the number of informative observations. A multiplicative Chernoff bound gives

$$\Pr(N < n\omega/2) \le e^{-n\omega/8}. \tag{64}$$

The stated worst-case condition and $\gamma \le 1$ make this probability at most $\delta/2$. Conditional on $N = m \ge n\omega/2$, the Hoeffding inequality for an interval of length two gives

$$\Pr(|\widehat{\theta} - \theta| \ge \gamma \mid N = m) \le 2e^{-m\gamma^2/2} \le 2e^{-n\omega\gamma^2/4} \le \delta/2. \tag{65}$$

The sign is correct on the complementary event. □

E.3.13 PROOF OF THEOREM 12, LOWER STATEMENT

Let the informative event occur with probability ω. Outside it, both worlds return identical constant feedback. On the event, a disputed candidate has Bernoulli loss with mean $1/2 - \gamma$ in one world and $1/2 + \gamma$ in the other. A reference candidate has known conditional loss $1/2$. The optimal candidate changes across worlds.

The one-observation divergence satisfies

$$D_{\mathrm{KL}}(P_- \| P_+) = \omega \,\mathrm{kl}(1/2 - \gamma \| 1/2 + \gamma) \le 16\omega\gamma^2. \tag{66}$$

Tensorization and the Pinsker inequality yield

$$\mathrm{TV}(P_-^n, P_+^n) \le \sqrt{8n\omega\gamma^2} \le 1/4. \tag{67}$$

Every candidate choice induces a binary test. Its average error is at least $(1 - \mathrm{TV})/2 \ge 3/8$. □

E.3.14 PROOF OF COROLLARY 36

Apply the variance-sensitive upper statement with failure probability δ/J to every comparison. A union bound gives simultaneous sign recovery. The proof uses independence across documents. It does not use independence among comparisons observed on the same document. □

E.3.15 PROOF OF THEOREM 37

Let I be Bernoulli with mean ω and reveal I to the learner. Outside $I = 1$, every world returns the same zero loss vector. Conditional on $I = 1$, world v returns independent Bernoulli losses. Candidate v has mean $1/2 - \gamma$, and every other candidate has mean $1/2$. Thus v is uniquely optimal with conditional gap γ.

Let P_0 be the reference world in which every conditional mean is $1/2$. The divergence for one document satisfies

$$D_{\mathrm{KL}}(P_v \| P_0) = \omega \,\mathrm{kl}(1/2 - \gamma \| 1/2) \le 4\omega\gamma^2. \tag{68}$$

After n documents, the divergence is at most $4n\omega\gamma^2$. Put a uniform prior on $v \in [J]$. The mutual information between v and the observations is at most the average divergence to P_0, hence at most $4n\omega\gamma^2$. The Fano inequality gives

$$\Pr(\widehat{v} \ne v) \ge 1 - \frac{4n\omega\gamma^2 + \log 2}{\log J} \tag{69}$$

under the uniform prior. The largest world-specific error is at least this average. Rearranging proves the necessary condition for uniform error at most δ. □