MINIMAX LEARNING FROM CORRUPTED CHAIN-OF-THOUGHT SUPERVISION

Anonymous authors
Paper under double-blind review

ABSTRACT

Chain-of-thought (CoT) supervision has been used to train reasoning models, but the intermediate reasoning steps and final answers provided for training can be corrupted. We study learning from corrupted outputs of a deterministic autoregressive teacher model that belongs to a finite candidate class. We establish matching upper and lower bounds on the number of prompt-level training examples needed to achieve a target clean prediction error, in the worst case over classes of a given size. For final-answer supervision, we consider Huber contamination: with probability $\alpha < 1/2$, each answer is replaced by a draw from an arbitrary unknown distribution. We show that empirical risk minimization (ERM) is minimax optimal and characterize how the contamination probability increases the required sample size. For full-CoT supervision, tokens are corrupted independently after the teacher generates the complete sequence, and prediction error is measured by the expected fraction of incorrect tokens. When the token corruption probabilities are known and distinct clean tokens induce distinct observation distributions, maximum likelihood is minimax optimal. Its sample complexity is governed by the distinguishability of the corrupted observations, accumulated over the token positions on which candidate rollouts disagree. For unknown Huber token contamination with probability $\alpha < 1/2$, minimizing token disagreements between candidate-generated sequences and observed sequences remains minimax optimal, without knowing either the corruption probabilities or α. These results quantify the statistical cost of corrupted supervision and show that channel knowledge can sharpen guarantees for particular corruption mechanisms, although it is unnecessary for optimal worst-case performance.

1 INTRODUCTION

Chain-of-thought (CoT) supervision exposes intermediate reasoning steps rather than only final answers and has become an important source of training signal for reasoning models. Existing statistical theories largely study the benefit of observing clean reasoning trajectories (Joshi et al., 2025; Altabaa et al., 2025; Hanneke et al., 2026; Li, 2026). In practice, however, both final answers and intermediate reasoning steps may be inaccurate or corrupted (Havrilla & Iyer, 2024; Zhou et al., 2024). This raises a basic statistical question:

How many corrupted prompt–supervision pairs are necessary and sufficient to recover the clean outputs of a reasoning model?

Corrupted full-CoT supervision poses difficulties beyond ordinary noisy-label learning. The clean trajectory is latent, and each candidate autoregressive model must be evaluated through its own free-running rollout rather than by conditioning on the observed noisy prefix. Moreover, the information in a corrupted trajectory depends on where two candidate rollouts disagree: one disagreement and T disagreements can lead to very different levels of distinguishability after token corruption. The learner may also know the token corruption channel exactly, or know only that it belongs to a broad contamination family. A complete theory should therefore account jointly for the hypothesis-class size, trajectory length, channel geometry, and channel knowledge.

We study these questions for a finite class of deterministic autoregressive models under realizability. An unknown target model, which we call the teacher, generates a clean length-T rollout for each prompt. We consider two forms of supervision. In *Setting A*, only the final answer is observed under

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.

prompt-dependent Huber contamination. In *Setting B*, the complete clean trajectory is first generated and then corrupted independently across token positions through an offline channel. Performance is always evaluated against the corresponding clean output. Our goal is to characterize the minimax prompt-level sample complexity required to achieve an (ϵ, δ)-PAC guarantee.

Contributions. Our main results are as follows.

- For final-answer supervision with contamination level $\alpha < 1/2$, we prove matching upper and lower bounds, up to universal constants, of order $\Theta\left(\frac{\log N + \log(1/\delta)}{\epsilon \gamma_\alpha}\right)$, where $\gamma_\alpha := 1 - 2\sqrt{\alpha(1-\alpha)}$, in the worst case over classes of cardinality at most N. Ordinary empirical risk minimization (ERM) attains this rate without knowing the arbitrary prompt-dependent corruption kernel. Contamination therefore reduces the effective sample size but creates no irreducible clean-error floor in the deterministic realizable setting.
- For full-CoT supervision through a known token channel M, let ρ_M be the largest Bhattacharyya overlap between two distinct rows of M. We show that channel-aware maximum likelihood has the matching minimax rate $\Theta\left(\frac{\log N + \log(1/\delta)}{\epsilon(1-\rho_M^T)}\right)$. The factor $1 - \rho_M^T$ results from the product-channel separation accumulated across disagreeing token positions. Thus, under independent token corruption and for the normalized token-Hamming error, disagreeing positions act as channel redundancy rather than as an intrinsic horizon-complexity penalty.
- We next consider an unknown Huber-type token channel $M_C = (1-\alpha)I + \alpha C$ for full-CoT supervision, where C is an arbitrary row-stochastic matrix. We prove that ordinary tokenwise ERM is minimax-rate optimal uniformly over this family, with separation $1 - (2\sqrt{\alpha(1-\alpha)})^T$. The learning rule uses neither C nor α, although its guarantee depends on the true contamination level. Channel knowledge is therefore unnecessary for worst-case optimality over the Huber family, but enables adaptation to easier channels and learning under more general identifiable corruption mechanisms.
- Fine-tuning experiments on mathematical and medical reasoning data exhibit consistent qualitative behavior: performance degrades with contamination, clean per-position accuracy increases with the state-sequence length under token corruption, in the direction predicted by channel redundancy, and the least-favorable pair-swap channel is particularly challenging without channel knowledge. Controlled finite-class simulations in appendix recover all three predicted sample-complexity scalings up to approximately constant factors.

Together, these results provide a unified account of learning from corrupted final answers and reasoning trajectories. They separate the statistical cost of corruption from the complexity of the underlying autoregressive class and clarify when detailed knowledge of the corruption mechanism is statistically valuable.

2 Related Work

Statistical theories of CoT supervision. Recent work studies when intermediate reasoning supervision improves the learnability of autoregressive models. Joshi et al. (2025) formalize deterministic, time-invariant next-token generators and compare learning from final answers with learning from complete CoT trajectories. Altabaa et al. (2025) derive CoT-information-based bounds in a general framework, but their explicit noisy-channel extension is a Fano-type lower bound for a joint channel acting on the entire answer–trace pair and evaluated by end-to-end 0-1 error. In contrast, under offline tokenwise product corruption and normalized token-Hamming risk, we establish matching worst-case finite-class minimax bounds with explicit separation $\Gamma_{M,T} = 1 - \rho_M^T$ for known channels and prove that channel-agnostic tokenwise ERM is minimax-rate optimal uniformly over the unknown Huber family. Subsequent work shows that clean CoT supervision can have sample complexity independent of the rollout length (Hanneke et al., 2026; Li, 2026); related curriculum-based analyses study compositional learning of long computations (Rajaraman et al., 2026a;b). Our focus is complementary: the observed final answer or trajectory is itself corrupted. For full-CoT supervision, we measure normalized token-Hamming error and identify the exact product-channel separation governing recovery. Thus, the explicit role of the horizon in our results comes from the

accumulation of channel evidence across disagreeing token positions, rather than from an additional complexity penalty for the clean autoregressive class.

Learning with corrupted labels and channels. Our final-answer setting is related to classical classification noise, bounded-noise learning, and Huber contamination (Angluin & Laird, 1988; Massart & Nédélec, 2006; Huber, 1964). Prior work constructs corrected losses under known class-conditional noise, studies learnability without known noise parameters, and characterizes losses that are robust to symmetric corruption (Natarajan et al., 2013; Menon et al., 2015; Van Rooyen et al., 2015). Transition-matrix methods similarly use known or estimated corruption probabilities to correct multiclass training objectives (Patrini et al., 2017; Liu et al., 2023). Our setting differs in two ways. First, the clean final answer is deterministic, while the contaminating distribution may depend arbitrarily on the prompt. Below contamination rate $1/2$, the clean answer retains a uniform majority advantage, so contamination causes a multiplicative sample-complexity penalty rather than an irreducible error floor. Second, full-CoT supervision observes a latent clean trajectory through a tokenwise product channel. The known-channel estimator follows a forward-modeling principle, but the distinguishability of two candidates tensorizes only over token positions on which their clean rollouts disagree. This yields the sharp factor $1 - \rho_M^T$ and a matching lower bound. For the Huber family $M = (1-\alpha)I + \alpha C$, we further show that estimating the unknown row-stochastic matrix C is unnecessary for worst-case optimality: ordinary tokenwise ERM succeeds uniformly because each clean-token disagreement retains a positive majority margin. Channel knowledge nevertheless applies to more general identifiable channels and adapts to corruption mechanisms that are easier than the least favorable Huber channel.

Noisy reasoning traces. Empirical and theoretical studies have also examined imperfect rationales in prompting and fine-tuning. Empirical work Havrilla & Iyer (2024) distinguish static corruption, applied after a trace is generated, from dynamic corruption that changes subsequent reasoning states. Our offline channel model formalizes the static regime: the complete clean trajectory is generated first, and corrupted tokens are never fed back into the target model. Other work studies in-context learning with inaccurate or irrelevant rationales and develops denoising or error-aware demonstrations (Zhou et al., 2024; Cui et al., 2025). In contrast, we study corrupted training supervision and provide prompt-level PAC guarantees together with matching information-theoretic lower bounds.

3 Problem Setting

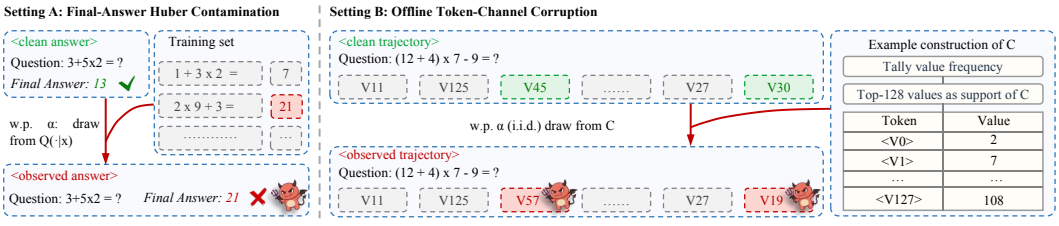

Figure 1: **Two corrupted-supervision settings.** Under final-answer Huber contamination (left), the observation is drawn from $(1-\alpha)\delta_{h^\star(x)} + \alpha Q(\cdot \mid x)$. Under full-CoT corruption (right), the complete clean trajectory is generated first and its tokens are then independently corrupted offline. The figure illustrates the Huber-type channel $M_C = (1-\alpha)I + \alpha C$ and one experimental construction of C. In both settings, evaluation is against the clean output.

We consider two forms of corrupted supervision as shown in Figure 1. Final-answer feedback follows a conditional Huber contamination model. Full-CoT feedback is generated through an offline token corruption channel, which we study under two information regimes: an arbitrary known channel and an unknown Huber-type channel.

Notation. Let Σ be a finite token vocabulary, let $\Sigma^\star := \bigcup_{\ell \geq 0} \Sigma^\ell$ be the set of token strings. Let λ denote the empty string and set $\Sigma^0 := \{\lambda\}$. Let $\mathcal{X} \subseteq \Sigma^\star$ be the prompt space. Prompts are drawn from a distribution $\rho \in \Delta(\mathcal{X})$, where $\Delta(\mathcal{A})$ denotes the set of probability distributions on $\mathcal{A}$. Fix a

generation horizon $T \in \mathbb{N}_+$, write $[T] := \{1, \dots, T\}$, and define $\Sigma^{<T} := \bigcup_{t=0}^{T-1} \Sigma^t$. For a sequence $u_{1:T}$, let $u_{<t} := (u_1, \dots, u_{t-1})$. We write δ_a for the point mass at a.

Deterministic autoregressive models. Let Π be a finite class of deterministic autoregressive models $\pi : \mathcal{X} \times \Sigma^{<T} \to \Sigma$. For $\pi \in \Pi$ and $x \in \mathcal{X}$, define the clean rollout recursively by $y_t^\pi(x) := \pi(x, y_{<t}^\pi(x)), y_{<1}^\pi(x) := \varnothing, t \in [T]$. The induced rollout trajectory and terminal final answer are $R_\pi(x) := y_{1:T}^\pi(x) \in \Sigma^T, \pi_T(x) := y_T^\pi(x) \in \Sigma$. Thus, $y_{1:T-1}^\pi(x)$ is the intermediate reasoning trace. The base class induces $\Pi_{1:T} := \{R_\pi : \pi \in \Pi\} \subseteq (\Sigma^T)^{\mathcal{X}}, \Pi_T := \{\pi_T : \pi \in \Pi\} \subseteq \Sigma^{\mathcal{X}}$, with $|\Pi_T| \le |\Pi_{1:T}| \le |\Pi|$. Deterministic autoregressive models are also considered in Joshi et al. (2025); Rajaraman et al. (2026a).

Assumption 3.1 (Realizability)**.** *There exists an unknown target model $\pi^\star \in \Pi$, which we refer to as the teacher. Write $R^\star := R_{\pi^\star}, h^\star := \pi_T^\star$ for its clean rollout trajectory and final answer.*

Setting A: Final-answer feedback with conditional Huber contamination. For a contamination level $\alpha \in [0, 1/2)$, the learner observes $S_n^{\mathrm{FA}} := \{(x_i, \widetilde{y}_{i,T})\}_{i=1}^n, x_i \overset{\text{i.i.d.}}{\sim} \rho$, where, conditionally on x_i, $\widetilde{y}_{i,T} \sim (1-\alpha)\,\delta_{h^\star(x_i)} + \alpha\, Q(\cdot \mid x_i)$. Here $Q : \mathcal{X} \to \Delta(\Sigma)$ is an arbitrary and unknown corruption kernel and may depend on the prompt. For any final-answer predictor $h : \mathcal{X} \to \Sigma$, define the clean error $L_{\mathrm{FA}}^\star(h) := \Pr_{x\sim\rho}(h(x) \neq h^\star(x))$.

Setting B: Full-CoT feedback through a token channel. For each prompt x_i, the teacher first generates the complete clean trajectory $R^\star(x_i) = \big(R_1^\star(x_i), \dots, R_T^\star(x_i)\big) \in \Sigma^T$. After the clean trajectory has been generated, its tokens are corrupted offline through a row-stochastic matrix $M \in [0,1]^{|\Sigma|\times|\Sigma|}$, where $M(a,z)$ is the probability of observing token z when the clean token is a. Conditional on x_i, the observed trajectory $z_{i,1:T}$ satisfies $\Pr(z_{i,1:T} \mid x_i) = \prod_{t=1}^T M\left(R_t^\star(x_i), z_{i,t}\right)$. Conditional on the prompts and clean trajectories, the corrupted tokens are independent across examples and token positions. Since corruption is applied only after the complete clean rollout has been generated, the observed tokens are not fed back into the teacher's autoregressive generation.

The learner observes $S_n^{\mathrm{CoT}} := \{(x_i, z_{i,1:T})\}_{i=1}^n, x_i \overset{\text{i.i.d.}}{\sim} \rho$, but never observes the clean trajectories $R^\star(x_i)$. For a candidate trajectory predictor $g : \mathcal{X} \to \Sigma^T$, write $g_t(x) := [g(x)]_t$. For a fixed channel M, the conditional distribution induced by g is $P_{g,M}(z_{1:T} \mid x) := \prod_{t=1}^T M\big(g_t(x), z_t\big)$.

We study two information regimes:

- **Known-channel regime.** The channel M is fixed and known to the learner. It may be an arbitrary row-stochastic matrix; in particular, it need not be symmetric or diagonally dominant. This regime includes a known Huber-type channel $M = (1-\alpha)I + \alpha C$, when the row-stochastic corruption matrix C is known.

- **Unknown Huber-channel regime.** Let $\mathcal{C}_\Sigma := \big\{C \in [0,1]^{|\Sigma|\times|\Sigma|} : C\mathbf{1} = \mathbf{1}\big\}$ denotes the set of row-stochastic matrices on Σ. For a contamination level $\alpha \in [0, 1/2)$, the true channel is $M_C := (1-\alpha)I + \alpha C, C \in \mathcal{C}_\Sigma$, but the learner is not given C. Thus, $M_C(a,z) = (1-\alpha)\mathbf{1}\{z = a\} + \alpha C(a,z)$. The data are generated by the same offline token-corruption process as above, with $M = M_C$.

Minimax PAC criterion and sample complexity. For $u, v \in \Sigma^T$, define the normalized token-Hamming distance $d_{\mathrm{tok}}(u,v) := \frac{1}{T}\sum_{t=1}^T \mathbf{1}\{u_t \neq v_t\}$. The clean trajectory risk of a predictor $g : \mathcal{X} \to \Sigma^T$ is $L_{\mathrm{tok}}^\star(g) := \mathbb{E}_{X\sim\rho}\left[d_{\mathrm{tok}}\left(g(X), R^\star(X)\right)\right]$. For each $n \in \mathbb{N}_+$, let $\mathfrak{A}_n^{\mathrm{FA}} := \big\{\mathsf{A}_n : (\mathcal{X}\times\Sigma)^n \rightsquigarrow \Sigma^{\mathcal{X}}\big\}$ and $\mathfrak{A}_n^{\mathrm{CoT}} := \big\{\mathsf{A}_n : (\mathcal{X}\times\Sigma^T)^n \rightsquigarrow (\Sigma^T)^{\mathcal{X}}\big\}$ denote the classes of all possibly randomized learning rules, where $\rightsquigarrow$ denotes a Markov kernel. The output spaces are not restricted to Π_T or $\Pi_{1:T}$, so improper learners are included.

For final-answer feedback, define the minimax PAC sample complexity by

$$n_{\mathrm{FA}}^\star(\epsilon,\delta) := \inf\left\{n \in \mathbb{N} : \inf_{\mathsf{A}_n \in \mathfrak{A}_n^{\mathrm{FA}}} \sup_{\substack{\rho\in\Delta(\mathcal{X}),\ h^\star\in\Pi_T \\ Q:\mathcal{X}\to\Delta(\Sigma)}} \Pr\left(L_{\mathrm{FA}}^\star\big(\mathsf{A}_n(S_n^{\mathrm{FA}})\big) > \epsilon\right) \le \delta\right\}. \tag{1}$$

For a fixed known channel M, define

$$n^{\star,\mathrm{CoT}}_{\mathrm{known}}(\epsilon,\delta) := \inf\left\{ n\in\mathbb{N} : \inf_{\mathsf{A}_{n,M}\in\mathfrak{A}_n^{\mathrm{CoT}}} \sup_{\rho\in\Delta(\mathcal{X}),\ R^\star\in\Pi_{1:T}} \Pr\left(L^\star_{\mathrm{tok}}\big(\mathsf{A}_{n,M}(S_n^{\mathrm{CoT}})\big) > \epsilon \right) \le \delta \right\}. \quad (2)$$

Here $\mathsf{A}_{n,M}$ may use the known channel M.

For the unknown Huber-channel regime, define

$$n^{\star,\mathrm{CoT},}_{\mathrm{unknown}}(\epsilon,\delta) := \inf\left\{ n\in\mathbb{N} : \inf_{\mathsf{A}_n\in\mathfrak{A}_n^{\mathrm{CoT}}} \sup_{\substack{\rho\in\Delta(\mathcal{X}),\ R^\star\in\Pi_{1:T}\\ C\in\mathcal{C}_\Sigma}} \Pr\left(L^\star_{\mathrm{tok}}\big(\mathsf{A}_n(S_n^{\mathrm{CoT}})\big) > \epsilon \right) \le \delta \right\}. \quad (3)$$

Because the infimum over learning rules precedes the supremum over C, the same learning rule must satisfy the guarantee uniformly over the entire Huber channel family.

In all three definitions, the probability is taken under the corresponding data-generating process and includes the learner's internal randomness. Throughout, n counts prompt-level examples; hence a full-CoT sample of size n contains nT observed tokens.

4 Warm-Up: Final-Answer Learning under Huber Contamination

We begin with final-answer supervision, which isolates the statistical effect of contamination before turning to full trajectories. Although the corruption kernel $Q(\cdot \mid x)$ is arbitrary, unknown, and prompt dependent, deterministic realizability and $\alpha < 1/2$ ensure that the clean answer retains a strict majority advantage whenever two predictors disagree.

To identify the relevant statistical separation, consider two different clean answers $a, b \in \Sigma$. Under the least favorable Huber contamination, they induce the observed-label distributions $(1-\alpha)\delta_a+\alpha\delta_b$ and $\alpha\delta_a + (1-\alpha)\delta_b$. Their Bhattacharyya overlap is $2\sqrt{\alpha(1-\alpha)}$. We therefore define

$$\gamma_\alpha := 1 - 2\sqrt{\alpha(1-\alpha)} = \left(\sqrt{1-\alpha} - \sqrt{\alpha}\right)^2. \quad (4)$$

The quantity γ_α measures the residual separation supplied by a clean-answer disagreement after contamination.

The minimax quantity $n^\star_{\mathrm{FA}}(\epsilon,\delta)$ defined in Section 3 is associated with a fixed induced class Π_T and contamination level α, whose dependence is suppressed in the notation. To state a cardinality-based lower bound, define

$$n^\star_{\mathrm{FA,wc}}(\epsilon,\delta;N,\alpha) := \sup_{\substack{\mathcal{X}_0\subseteq\Sigma^\star,\ \Pi \text{ finite}\\ |\Pi_T|\le N}} n^\star_{\mathrm{FA}}(\epsilon,\delta), \quad (5)$$

where $n^\star_{\mathrm{FA}}(\epsilon,\delta)$ is evaluated for the induced final-answer class Π_T on $\mathcal{X}_0$ at contamination level α.

We first establish an information-theoretic lower bound and then show that ordinary ERM attains the matching rate.

Theorem 4.1 (Finite-class minimax lower bound)**.** *Suppose $|\Sigma| \ge 2$ and $\alpha \in [0, 1/2)$. There exist universal constants $c, c_0 > 0$ such that, for every $N \ge 4$, $\epsilon \in (0, 1/16)$, and $\delta \in (0, c_0)$,*

$$n^\star_{\mathrm{FA,wc}}(\epsilon,\delta;N,\alpha) \ge c\,\frac{\log N + \log(1/\delta)}{\epsilon\gamma_\alpha}. \quad (6)$$

More precisely, the $\log N$ term is necessary at a constant failure probability for a class satisfying $\frac{N}{2} \le |\Pi_T| \le N$, whereas the $\log(1/\delta)$ term is necessary even for a two-element final-answer class. The result holds for all possibly randomized and improper learners.

Theorem 4.1 shows that no learning rule can uniformly avoid the information contraction caused by contamination. We next show that this is the only statistical price: ordinary ERM achieves the matching dependence on γ_α without knowing the corruption kernel Q.

Given S_n^{FA}, define

$$\widehat{h}_n \in \arg\min_{h\in\Pi_T} \widehat{L}_\alpha(h), \qquad \widehat{L}_\alpha(h) := \frac{1}{n}\sum_{i=1}^n \mathbf{1}\left\{h(x_i) \ne \widetilde{y}_{i,T}\right\}, \quad (7)$$

with ties broken arbitrarily. The following theorem gives its uniform PAC guarantee.

Theorem 4.2 (Finite-class upper bound for final-answer ERM)**.** *Suppose Π_T is finite and $\alpha \in [0, 1/2)$. Under Assumption 3.1, for every $n \in \mathbb{N}_+$ and $\epsilon \in (0, 1)$, the estimator equation 7 satisfies*

$$\sup_{\substack{\rho\in\Delta(\mathcal{X}),\ h^\star\in\Pi_T \\ Q:\mathcal{X}\to\Delta(\Sigma)}} \Pr\left(L^\star_{\mathrm{FA}}\left(\widehat{h}_n\right) > \epsilon\right) \le |\Pi_T|\,(1-\epsilon\gamma_\alpha)^n\,. \tag{8}$$

Consequently, the minimax PAC sample complexity defined in equation 1 satisfies

$$n^\star_{\mathrm{FA}}(\epsilon,\delta) \le \left\lceil \frac{\log(|\Pi_T|/\delta)}{-\log\left(1-\epsilon\gamma_\alpha\right)} \right\rceil \le \left\lceil \frac{\log(|\Pi_T|/\delta)}{\epsilon\gamma_\alpha} \right\rceil \tag{9}$$

for every $\epsilon, \delta \in (0, 1)$.

Remark 4.3 (Matching minimax rate)**.** *Together, Theorems 4.1 and 4.2 yield*

$$n^\star_{\mathrm{FA,wc}}(\epsilon,\delta;N,\alpha) = \Theta\left(\frac{\log N + \log(1/\delta)}{\epsilon\gamma_\alpha}\right) \tag{10}$$

up to universal constants, in the parameter regime of Theorem 4.1. Hence ordinary ERM is minimax-rate optimal in the worst case over finite final-answer classes of cardinality at most N. The factor γ_α has an effective-sample-size interpretation: n contaminated examples provide approximately $n\gamma_\alpha$ clean examples' worth of statistical separation. Since $1-\sqrt{1-u} \in [u/2, u]$ for $u \in [0, 1]$, we have $\frac{1}{2}(1-2\alpha)^2 \le \gamma_\alpha \le (1-2\alpha)^2$. Hence, up to universal constants, equation 10 coincides with the classical finite-class rate under Massart-type noise with margin $1-2\alpha$ (Massart & Nédélec, 2006), measured in disagreement with $h^\star$; conditional Huber contamination with $\alpha < 1/2$ implies a multiclass Massart condition. We therefore view Section 4 as a baseline rather than a new rate. We state it in Bhattacharyya form because $1-\gamma_\alpha$ is exactly the channel overlap ρ_M of the least-favorable Huber channel: this makes final-answer learning the $T = 1$ case of Section 5, where the relevant separation is governed by the Hellinger geometry of M rather than by a majority margin (e.g., a known permutation channel has negative margin but $\rho_M = 0$).

5 Full-CoT Learning under Token Corruption

Unlike final-answer feedback, a corrupted CoT is an observation of a latent clean trajectory through a product channel. Each candidate must therefore be evaluated through its own free-running rollout rather than by conditioning on the observed noisy prefix. Moreover, the information carried by a corrupted trajectory depends on the positions at which the candidate and target rollouts disagree, rather than on the horizon alone.

We study the two information regimes introduced in Section 3. When the channel is known, the learner can score each candidate through its induced corrupted-output distribution. When the channel belongs to an unknown Huber family, this likelihood is unavailable; nevertheless, the identity component preserves a uniform tokenwise majority signal.

5.1 Estimators and upper bounds

Known channel. For two clean tokens $a, b \in \Sigma$, define the Bhattacharyya overlap between the corresponding rows of M by $\mathrm{BC}_M(a,b) := \sum_{z\in\Sigma}\sqrt{M(a,z)M(b,z)}$. The least distinguishable pair of distinct channel inputs determines

$$\rho_M := \max_{a\neq b} \mathrm{BC}_M(a,b), \qquad \Gamma_{M,T} := 1-\rho_M^T. \tag{11}$$

We assume $\rho_M < 1$, equivalently, that no two distinct clean tokens induce identical channel-output distributions. For $g \in \Pi_{1:T}$ and $x \in \mathcal{X}$, let $K_g(x) := \sum_{t=1}^T \mathbf{1}\{g_t(x) \neq R^\star_t(x)\} = T\, d_{\mathrm{tok}}\left(g(x), R^\star(x)\right)$. Product-channel tensorization gives $\mathrm{BC}\left(P_{g,M}(\cdot \mid x), P_{R^\star,M}(\cdot \mid x)\right) \le \rho_M^{K_g(x)}$. Since $s \mapsto \rho_M^s$ is convex on $[0, T]$,

$$\mathrm{BC}\left(P_{g,M}(\cdot \mid x), P_{R^\star,M}(\cdot \mid x)\right) \le 1-\Gamma_{M,T} d_{\mathrm{tok}}\left(g(x), R^\star(x)\right). \tag{12}$$

Thus, $\Gamma_{M,T}$ is the usable separation contributed by one corrupted trajectory per unit of clean normalized token-Hamming error. Given S_n^{CoT}, define the channel-aware maximum-likelihood estimator

$$\widehat{g}_n^M \in \arg\max_{g\in\Pi_{1:T}} \prod_{i=1}^{n} P_{g,M}\left(z_{i,1:T} \mid x_i\right), \tag{13}$$

with ties broken arbitrarily. Each candidate is therefore scored only after its clean rollout is passed through the known channel.

Unknown Huber channel. Now consider $M_C = (1-\alpha)I + \alpha C, C \in \mathcal{C}_\Sigma, \alpha \in [0, 1/2)$, where C is unavailable to the learner. Recall $\gamma_\alpha = 1 - 2\sqrt{\alpha(1-\alpha)}$ and define

$$\Gamma_{\alpha,T} := 1 - \left(2\sqrt{\alpha(1-\alpha)}\right)^T = 1 - (1-\gamma_\alpha)^T. \tag{14}$$

For every $C \in \mathcal{C}_\Sigma$ and every $a \neq b$,

$$\mathrm{TV}\left(M_C(a,\cdot), M_C(b,\cdot)\right) \geq 1 - 2\alpha, \qquad \mathrm{BC}_{M_C}(a,b) \leq 2\sqrt{\alpha(1-\alpha)}. \tag{15}$$

Consequently,

$$\rho_{M_C} \leq 2\sqrt{\alpha(1-\alpha)}, \qquad \Gamma_{M_C,T} \geq \Gamma_{\alpha,T}. \tag{16}$$

Both inequalities are tight when C deterministically exchanges two tokens. Because C is unknown, the likelihood in equation 13 cannot be evaluated. We instead use ordinary tokenwise ERM:

$$\widehat{g}_n^{\mathrm{unk}} \in \arg\min_{g\in\Pi_{1:T}} \frac{1}{n}\sum_{i=1}^{n} d_{\mathrm{tok}}\left(g(x_i), z_{i,1:T}\right), \tag{17}$$

with ties broken arbitrarily. The estimator uses neither C nor α; the true value of α only indexes its guarantee. To see why this objective is calibrated, define the population risk against the corrupted trajectory by $\widetilde{L}_{\mathrm{tok},C}(g) := \mathbb{E}\left[d_{\mathrm{tok}}\left(g(X), Z_{1:T}\right)\right]$. Then, uniformly over $C \in \mathcal{C}_\Sigma$,

$$\widetilde{L}_{\mathrm{tok},C}(g) - \widetilde{L}_{\mathrm{tok},C}(R^\star) \geq (1-2\alpha) L^\star_{\mathrm{tok}}(g). \tag{18}$$

Indeed, whenever the target token is a and a candidate predicts $b \neq a$, the identity component favors a with probability $1-\alpha$, whereas the corruption component can favor b with probability at most α.

Theorem 5.1 (Full-CoT PAC upper bounds)**.** *Under Assumption 3.1, the following statements hold.*

(i) *For a fixed known channel M satisfying $\rho_M < 1$, the estimator equation 13 satisfies, for every $n \in \mathbb{N}_+$ and $\epsilon \in (0,1)$,*

$$\sup_{\rho\in\Delta(\mathcal{X}),\ R^\star\in\Pi_{1:T}} \Pr_M\left(L^\star_{\mathrm{tok}}\left(\widehat{g}_n^M\right) > \epsilon\right) \leq |\Pi_{1:T}|\left(1-\epsilon\Gamma_{M,T}\right)^n. \tag{19}$$

Consequently, the minimax PAC sample complexity defined in Section 3 satisfies

$$n^{\star,\mathrm{CoT}}_{\mathrm{known}}(\epsilon,\delta) \leq \left\lceil \frac{\log(|\Pi_{1:T}|/\delta)}{-\log\left(1-\epsilon\Gamma_{M,T}\right)} \right\rceil \leq \left\lceil \frac{\log(|\Pi_{1:T}|/\delta)}{\epsilon\Gamma_{M,T}} \right\rceil. \tag{20}$$

(ii) *For the unknown Huber-channel family, the estimator equation 17 satisfies, for every $n \in \mathbb{N}_+$ and $\epsilon \in (0,1)$,*

$$\sup_{\substack{\rho\in\Delta(\mathcal{X}),\ R^\star\in\Pi_{1:T} \\ C\in\mathcal{C}_\Sigma}} \Pr_{M_C}\left(L^\star_{\mathrm{tok}}\left(\widehat{g}_n^{\mathrm{unk}}\right) > \epsilon\right) \leq |\Pi_{1:T}|\left(1-\epsilon\Gamma_{\alpha,T}\right)^n. \tag{21}$$

Consequently,

$$n^{\star,\mathrm{CoT}}_{\mathrm{unknown}}(\epsilon,\delta) \leq \left\lceil \frac{\log(|\Pi_{1:T}|/\delta)}{-\log\left(1-\epsilon\Gamma_{\alpha,T}\right)} \right\rceil \leq \left\lceil \frac{\log(|\Pi_{1:T}|/\delta)}{\epsilon\Gamma_{\alpha,T}} \right\rceil. \tag{22}$$

Complete proofs are given in Appendix C.

5.2 MATCHING MINIMAX RATES

The quantities $n^{\star,\mathrm{CoT}}_{\mathrm{known}}$ and $n^{\star,\mathrm{CoT}}_{\mathrm{unknown}}$ are defined for the fixed induced trajectory class $\Pi_{1:T}$. To state cardinality-based lower bounds, define

$$n^{\star,\mathrm{CoT}}_{\mathrm{known,wc}}(\epsilon,\delta;N,M) := \sup_{\substack{\mathcal{X}_0\subseteq\Sigma^\star,\ \Pi \text{ finite}\\ |\Pi_{1:T}|\le N}} n^{\star,\mathrm{CoT}}_{\mathrm{known}}(\epsilon,\delta), \tag{23}$$

where the sample complexity on the right is evaluated on the induced class $\Pi_{1:T}$ over $\mathcal{X}_0$ with the fixed channel M. Similarly, define

$$n^{\star,\mathrm{CoT}}_{\mathrm{unknown,wc}}(\epsilon,\delta;N,\alpha) := \sup_{\substack{\mathcal{X}_0\subseteq\Sigma^\star,\ \Pi \text{ finite}\\ |\Pi_{1:T}|\le N}} n^{\star,\mathrm{CoT}}_{\mathrm{unknown}}(\epsilon,\delta), \tag{24}$$

where the contamination level α is fixed.

Theorem 5.2 (Matching worst-case minimax PAC rates)**.** *Assume $|\Sigma| \ge 2$. There exist universal constants $c, C, c_0 > 0$ such that, for every $N \ge 4$, $\epsilon \in (0, 1/64)$, and $\delta \in (0, c_0)$, the following hold.*

For every fixed known channel M with $\rho_M < 1$,

$$c\,\frac{\log N + \log(1/\delta)}{\epsilon\Gamma_{M,T}} \le n^{\star,\mathrm{CoT}}_{\mathrm{known,wc}}(\epsilon,\delta;N,M) \le C\,\frac{\log N + \log(1/\delta)}{\epsilon\Gamma_{M,T}}. \tag{25}$$

For every $\alpha \in [0, 1/2)$,

$$c\,\frac{\log N + \log(1/\delta)}{\epsilon\Gamma_{\alpha,T}} \le n^{\star,\mathrm{CoT}}_{\mathrm{unknown,wc}}(\epsilon,\delta;N,\alpha) \le C\,\frac{\log N + \log(1/\delta)}{\epsilon\Gamma_{\alpha,T}}. \tag{26}$$

The lower bounds hold for all possibly randomized and improper learners.

Remark 5.3 (Channel knowledge and trajectory length)**.** *The two regimes differ in the separation available to the learner. For every Huber channel, $\Gamma_{M_C,T} \ge \Gamma_{\alpha,T}$. Thus, channel-aware maximum likelihood adapts to the actual geometry of M_C, whereas channel-agnostic ERM protects against the least favorable member of the Huber family. At the level of the rates, the price of not knowing the channel is captured by $\frac{\Gamma_{M_C,T}}{\Gamma_{\alpha,T}}$. This ratio equals one for a least-favorable pair-swap channel and can be larger when the actual channel preserves more information. Channel knowledge also permits general identifiable channels that do not preserve a tokenwise majority advantage. For example, a known permutation channel has $\rho_M = 0$ and is as informative as the noiseless channel, even though the observed token can systematically differ from the clean token. The dependence on the horizon satisfies $\Gamma_{M,T} = 1 - \rho_M^T \asymp \min\{T(1-\rho_M), 1\}$. Hence additional disagreement positions provide repeated channel evidence until the separation saturates. This does not mean that an arbitrarily longer trajectory is intrinsically easier: positions on which all candidate rollouts agree provide no information. In the linear regime, prompt-level sample complexity decreases as $1/T$, while the total number nT of observed tokens remains of the same order when the induced class complexity is fixed. Finally, $\Gamma_{\alpha,1} = 1 - 2\sqrt{\alpha(1-\alpha)} = \gamma_\alpha$, so final-answer learning is recovered as the one-token special case. For the binary symmetric channel, $\rho_M = 2\sqrt{\alpha(1-\alpha)}$, and the known- and unknown-channel rates coincide.*

6 EXPERIMENTS

In this section, we focus on language-model fine-tuning under corrupted final-answer and full-CoT supervision. Complete preprocessing rules, training grids, and additional results are provided in Appendix D. Controlled finite-class simulations that directly validate the theoretical sample-complexity factors are reported in Appendix D.1.

Datasets. For mathematical reasoning, we use `SFT_510K` [1], filtered for short elementary-level word problems with a single integer answer, yielding $3,862$ training and 300 test examples for Setting A.

[1] `https://huggingface.co/datasets/RLHF4MATH/SFT_510K`

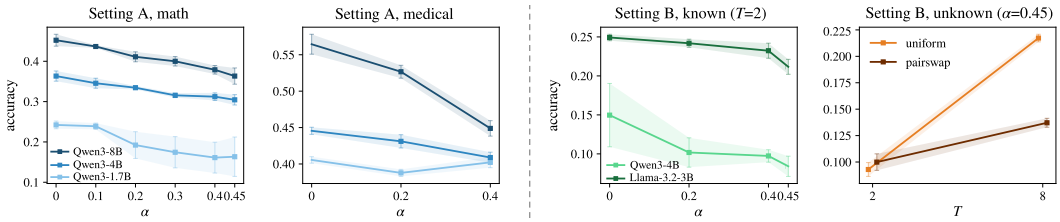

Figure 2: **Left:** Under Setting A with $n = 1000$, clean final-answer accuracy versus contamination level α for Qwen3-1.7B, 4B, and 8B on math data. **Center-left:** The same Setting A protocol applied to medical data, for the same three model scales. **Center-right:** Under Setting B with a known channel ($T = 2$, $n = 1000$), clean state accuracy versus α for Qwen3-4B and Llama-3.2-3B on math data. **Right:** Under an unknown channel ($\alpha = 0.45$, $n = 3000$), channel-agnostic accuracy for Qwen3-4B on math data under uniform and pair-swap Huber corruption for $T \in \{2, 8\}$.

For Setting B, we map each reasoning trace to a discrete state sequence by executing its Python solution and quantizing the intermediate arithmetic assignments into a 128-token vocabulary. For medical reasoning, we use the `en` and `en_mix` configurations of `medical-o1-reasoning-SFT` [2] Chen et al. (2024), restricted to multiple-choice questions with automatically extractable answers, yielding $2,683$ training and 300 test examples. Because these free-text reasoning traces are not readily discretizable, this dataset is used exclusively for Setting A.

Experiment setting. The main results in Figure 2 evaluate Qwen3-1.7B, Qwen3-4B, and Qwen3-8B across model scales, and additional models (e.g. Llama-3.2-3B) for cross-architecture comparison on Setting B. In Setting A, corruption affects only the supervised final answer. In Setting B, the complete clean state trajectory is constructed first and its target tokens are then corrupted independently. Both Setting B objectives are defined in Appendix D.2.1. Unless otherwise stated, all conditions are evaluated over three random seeds using $6 \times$ NVIDIA A100 GPUs.

Metric. For Setting A, we report accuracy against clean final answers. For Setting B, we report clean per-position state accuracy, equivalently one minus the normalized token-Hamming error. Markers in Figure 2 denote means across model seeds, and shaded bands or error bars denote ± 1 standard error. In the unknown-channel panel, every point aggregates two independently generated channel permutations and three model seeds.

Results. Figure 2 exhibits the qualitative trends predicted by our theory. Under Setting A, clean final-answer accuracy decreases with α across model scales and domains, while larger models remain consistently more accurate. Under Setting B with a known channel, the same degradation appears across two model families. With an unknown channel, increasing T improves accuracy under both corruption mechanisms, but substantially less so for pair-swap; at $T = 8$, the gap between uniform and pair-swap becomes pronounced. These trends are in the direction predicted by accumulated channel evidence and the least-favorable role of pair-swap.

7 Conclusion

We developed a minimax PAC theory for learning from corrupted final-answer and full-CoT supervision. For final-answer feedback, we showed that ordinary ERM is minimax-rate optimal under prompt-dependent Huber contamination. For full-CoT feedback, we characterized the optimal sample complexity under a known token channel and showed that, under independent corruption, disagreeing token positions provide channel redundancy for the token-level error. We further proved that, over an unknown Huber channel family, ordinary tokenwise ERM remains worst-case optimal without estimating the corruption mechanism. Channel knowledge is therefore unnecessary for worst-case optimality within this family, but enables adaptation to easier and more general corruption channels. For experiments, controlled simulations recover the predicted sample-complexity trends, while language-model fine-tuning experiments exhibit consistent qualitative behavior. Several directions for future work are deferred in Appendix A.

[2] `https://huggingface.co/datasets/FreedomIntelligence/medical-o1-reasoning-SFT`

AI USE STATEMENT

In this work, we used generative AI tools for language editing, organizing the mathematical notations, checking and debugging the code, and formatting assistance during manuscript preparation. AI-assisted text and code were checked by the authors before inclusion. We take full responsibility for the final content of this work.

ETHICS STATEMENT

This work does not involve human subjects, human-subject experiments, or the collection of personal data. Our empirical study uses publicly available datasets for mathematical and medical reasoning. We use these datasets only for research purposes and follow the corresponding dataset terms and usage requirements. The study focuses on the statistical effects of corrupted supervision and does not introduce new systems that directly interact with or make decisions about individuals.

REPRODUCIBILITY STATEMENT

We have made several efforts to ensure that our results are reproducible. For the theory, the problem setting, the data-generating processes for both forms of corrupted supervision, and the minimax PAC criteria are stated in Section 3, and the only structural assumption, realizability, is stated explicitly in Assumption 3.1. Complete proofs of all theoretical claims are given in the appendix: the auxiliary facts on Bhattacharyya tensorization, total variation, and Assouad's lemma are in Appendix B.1; the final-answer lower and upper bounds (Theorems 4.1 and 4.2) are proved in Appendices B.2 and B.3; the channel-separation inequalities, the full-CoT upper bounds (Theorem 5.1), and the matching lower bounds (Theorem 5.2) are proved in Appendices C.1, C.2, and C.3, respectively; and the horizon calculation underlying Remark 5.3 is given in Appendix C.4. For the experiments, the controlled finite-class simulations are fully specified in Appendix D.1, including the hard-instance constructions, the channel families, the exact aggregate sampling scheme, and the definition of the empirical critical sample size in equation 34. For the language-model experiments, Section 6 and Appendix D.2.1 describe the public datasets used, the filtering and discretization steps, the corruption protocols for both settings, the models, and optimization hyperparameters, and the evaluation metrics; and additional ablations are reported in Appendix D.2.2. All reported quantities are averaged over three model seeds, with channel seeds reported where applicable. We have provided implementations of the execution code as a self-contained package to be submitted as supplementary material.

REFERENCES

Awni Altabaa, Omar Montasser, and John D. Lafferty. CoT information: Improved sample complexity under chain-of-thought supervision. In *Advances in Neural Information Processing Systems*, volume 38, pp. 24822–24862. Curran Associates, Inc., 2025. doi: 10.52202/085713-0834. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/235a9abbb6303fac26f205367fe08c0b-Paper-Conference.pdf.

Dana Angluin and Philip Laird. Learning from noisy examples. *Machine learning*, 2(4):343–370, 1988.

Junying Chen, Zhenyang Cai, Ke Ji, Xidong Wang, Wanlong Liu, Rongsheng Wang, Jianye Hou, and Benyou Wang. Huatuogpt-o1, towards medical complex reasoning with llms, 2024. URL https://arxiv.org/abs/2412.18925.

Yingqian Cui, Pengfei He, Xianfeng Tang, Qi He, Chen Luo, Jiliang Tang, and Yue Xing. A theoretical understanding of chain-of-thought: Coherent reasoning and error-aware demonstration. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 3475–3483. PMLR, 2025. URL https://proceedings.mlr.press/v258/cui25b.html.

Steve Hanneke, Idan Mehalel, and Shay Moran. Sample complexity of autoregressive reasoning: Chain-of-thought vs. end-to-end. *arXiv preprint arXiv:2604.12013*, 2026.

Alex Havrilla and Maia Iyer. Understanding the effect of noise in LLM training data with algorithmic chains of thought. *arXiv preprint arXiv:2402.04004*, 2024. URL https://arxiv.org/abs/2402.04004.

Peter J. Huber. Robust estimation of a location parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, 1964. doi: 10.1214/aoms/1177703732.

Nirmit Joshi, Gal Vardi, Adam Block, Surbhi Goel, Zhiyuan Li, Theodor Misiakiewicz, and Nathan Srebro. A theory of learning with autoregressive chain of thought. *arXiv preprint arXiv:2503.07932*, 2025.

Zhiyuan Li. The optimal sample complexity of learning autoregressive chain-of-thought. *arXiv preprint arXiv:2607.07423*, 2026. URL https://arxiv.org/abs/2607.07423.

Yang Liu, Hao Cheng, and Kun Zhang. Identifiability of label noise transition matrix. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 21475–21496. PMLR, 2023. URL https://proceedings.mlr.press/v202/liu23g.html.

Pascal Massart and Élodie Nédélec. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006. doi: 10.1214/009053606000000786.

Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Robert C. Williamson. Learning from corrupted binary labels via class-probability estimation. In *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pp. 125–134. PMLR, 2015. URL https://proceedings.mlr.press/v37/menon15.html.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. *Advances in neural information processing systems*, 26, 2013.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1944–1952, 2017. URL https://openaccess.thecvf.com/content_cvpr_2017/html/Patrini_Making_Deep_Neural_CVPR_2017_paper.html.

Nived Rajaraman, Audrey Huang, Miro Dudik, Robert Schapire, Dylan J Foster, and Akshay Krishnamurthy. Learning to reason with curriculum i: Provable benefits of autocurriculum. *arXiv preprint arXiv:2603.18325*, 2026a.

Nived Rajaraman, Audrey Huang, Miroslav Dudík, Robert Schapire, Dylan Foster, and Akshay Krishnamurthy. Learning to reason with curriculum II: Compositional generalization. *arXiv preprint arXiv:2606.27721*, 2026b. URL https://arxiv.org/abs/2606.27721.

Brendan Van Rooyen, Aditya Krishna Menon, and Robert C. Williamson. Learning with symmetric label noise: The importance of being unhinged. In *Advances in Neural Information Processing Systems*, volume 28, pp. 10–18. Curran Associates, Inc., 2015. URL https://proceedings.neurips.cc/paper_files/paper/2015/file/45c48cce2e2d7fbdea1afc51c7c6ad26-Paper.pdf.

Zhanke Zhou, Rong Tao, Jianing Zhu, Yiwen Luo, Zengmao Wang, and Bo Han. Can language models perform robust reasoning in chain-of-thought prompting with noisy rationales? In *Advances in Neural Information Processing Systems*, volume 37, pp. 123846–123910. Curran Associates, Inc., 2024. doi: 10.52202/079017-3936. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/dfaa29ed28dfa175bcc5e2a54aa199f8-Paper-Conference.pdf.

A Discussion and Future Directions

The results in this paper are obtained under the most restrictive combination of assumptions for which the corrupted-supervision problem is already nontrivial: a finite class, a deterministic teacher, realizability (Assumption 3.1), and offline corruption that is independent across token positions. We regard the resulting rates as a *baseline* for the statistical theory of corrupted CoT supervision in two senses. First, they isolate the price of corruption from every other source of difficulty. The class enters only through its cardinality, the teacher's rollout is a fixed function of the prompt, and, besides the prompt, the only randomness in the data is the corruption itself. Second, the matching lower bounds of Theorems 4.1 and 5.2 show that the separation factors γ_α, $\Gamma_{M,T}$, and $\Gamma_{\alpha,T}$ are unavoidable even in this simplest setting, so that any extension must pay at least these factors. This appendix records what the baseline establishes, and then discusses the four relaxations: stochastic teachers, non-realizable settings, infinite model classes, and corruption that propagates during autoregressive generation. For each, we indicate which parts of the present analysis carry over, and where new phenomena appear.

What the baseline establishes. All three main results have the same product form,

$$n^\star(\epsilon,\delta) \asymp \frac{\text{class complexity} + \log(1/\delta)}{\epsilon \cdot \text{separation}},$$

with class complexity $\log|\Pi_T|$ or $\log|\Pi_{1:T}|$ and separation γ_α, $\Gamma_{M,T}$, or $\Gamma_{\alpha,T}$. Two structural facts underlie this form. The first is *calibration*: for every candidate that disagrees with the teacher, the corrupted objective favors the teacher by a margin proportional to the clean error, uniformly over the corruption mechanism (equation 18 and the Chernoff argument in Appendix B.3). Calibration is why contamination below rate $1/2$ acts as a multiplicative reduction of the effective sample size and creates no clean-error floor. The second is *tensorization over disagreements*: the distinguishability of two candidates through the product channel accumulates only over positions on which their clean rollouts differ. The extensions below differ mainly in whether calibration and tensorization survive, and the discussion is organized around these two properties.

Stochastic teachers. Suppose the teacher is a conditional distribution $\pi^\star(\cdot \mid x, s)$ over next tokens, so that each prompt induces a distribution $P_x^\star$ over clean trajectories, and the observation is drawn from $P_x^\star \circ M^{\otimes T}$. Three aspects of the baseline change. The clean error must be redefined, since there is no longer a single trajectory to disagree with; natural choices are a Hellinger-type distance between the candidate's and the teacher's trajectory laws, or the expected token error of the candidate's own rollout distribution against the teacher's. The likelihood argument of Theorem 5.1(i) is written for general induced laws and extends verbatim as long as one can lower bound the per-prompt separation $1 - \mathrm{BC}\big(P_{g,x} \circ M^{\otimes T},\, P_x^\star \circ M^{\otimes T}\big)$ by the chosen clean error; by the data-processing inequality this separation is at most the clean separation $1 - \mathrm{BC}(P_{g,x}, P_x^\star)$, and the ratio between the two is a Hellinger contraction coefficient of the product channel restricted to the pairs of laws realized by the class. Tensorization over disagreeing positions, however, no longer holds in its present form, because consecutive clean tokens are dependent through the stochastic generation process, so the overlap of the two corrupted laws is an overlap of mixtures over paths rather than a product over positions. Finally, stochasticity of the teacher and stochasticity of the channel can be confounded. A teacher that emits a with probability $1-\alpha$ and b with probability α is indistinguishable from a deterministic teacher emitting a observed through a pair-swap channel. In the unknown-Huber regime, the uniform channel-agnostic guarantee therefore cannot survive without additional structure separating teacher randomness from channel noise, for instance a known channel, or a restriction on the support of the teacher's next-token distribution. The lower-bound constructions carry over with minor changes, since Assouad families can be built from deterministic teachers, which are a special case.

Non-realizable settings. When $\pi^\star \notin \Pi$, the natural target is the best-in-class predictor, for instance $h_\Pi \in \arg\min_{h\in\Pi_T} L^\star_{\mathrm{FA}}(h)$ in Setting A. The main change is that calibration is only partially preserved. Writing the corrupted risk as

$$\widetilde{L}_{\mathrm{FA}}(h) := \Pr\big(h(X) \neq \widetilde{Y}_T\big) = (1-\alpha)L^\star_{\mathrm{FA}}(h) + \alpha\,\mathbb{E}\big[1 - Q(h(X)\mid X)\big],$$

the difference $\widetilde{L}_{\mathrm{FA}}(h) - \widetilde{L}_{\mathrm{FA}}(h_\Pi)$ equals $(1-\alpha)\big(L^\star_{\mathrm{FA}}(h) - L^\star_{\mathrm{FA}}(h_\Pi)\big)$ plus a term bounded in magnitude by $\alpha\Pr(h(X)\neq h_\Pi(X))$. Under realizability this second term is controlled by the clean

error of h, which is what the proof of Theorem 4.2 exploits; in the agnostic case it is not, because a prompt-dependent kernel Q may concentrate its mass exactly on the prompts where h_Π is wrong. A short calculation nevertheless shows that the population minimizer of the corrupted risk cannot be much worse than the best in class: if $\widetilde{L}_{\mathrm{FA}}(h) \le \widetilde{L}_{\mathrm{FA}}(h_\Pi)$, then $(1-\alpha)\big(L^\star_{\mathrm{FA}}(h) - L^\star_{\mathrm{FA}}(h_\Pi)\big) \le \alpha \Pr(h \neq h_\Pi) \le \alpha\big(L^\star_{\mathrm{FA}}(h) + L^\star_{\mathrm{FA}}(h_\Pi)\big)$, and hence $L^\star_{\mathrm{FA}}(h) \le L^\star_{\mathrm{FA}}(h_\Pi)/(1-2\alpha)$. Thus prompt-dependent contamination inflates the achievable clean error by at most the factor $1/(1-2\alpha)$, and this factor is attained by an adversarial Q; consistency in the strong sense, with the learned clean error approaching $L^\star_{\mathrm{FA}}(h_\Pi)$, requires restrictions on Q such as independence from the prompt, under which the ordering of candidates is preserved as in the symmetric-noise literature. The finite-sample analysis then requires agnostic-type deviation bounds in place of the one-sided Chernoff argument, with the familiar transition from $1/\epsilon$ to $1/\epsilon^2$ rates. For full-CoT supervision, the same calculation applies tokenwise, with one additional subtlety: the best-in-class rollout is defined through free-running generation, so the comparator of the rollout-based objective equation 17 can differ from the best-in-class next-token predictor under teacher forcing, and the gap between the two is an exposure-bias effect that is absent under realizability.

Infinite model classes. Replacing $\log|\Pi_T|$ and $\log|\Pi_{1:T}|$ by combinatorial dimensions is the most standard of the four extensions, but it is also the one in which the autoregressive structure becomes essential. In Setting A, the learner only ever interacts with the induced final-answer class Π_T, so if Π_T has Natarajan dimension d_T, one expects a rate of order $d_T\,\mathrm{polylog}/(\epsilon\gamma_\alpha)$, obtained by combining the disagreement-metric analysis of Massart-type noise (Massart & Nédélec, 2006) with our lower-bound constructions. The autoregressive structure enters through the relation between d_T and the dimension d of the base next-token class Π: for time-invariant classes, the end-to-end class can be substantially richer than the base class (Joshi et al., 2025), so contamination multiplies an already inflated complexity. In Setting B the situation is more delicate. Clean CoT supervision achieves sample complexity governed by the base class, and even independent of T (Hanneke et al., 2026; Li, 2026), by treating each $(x, y_{<t}) \mapsto y_t$ as a labeled example of the next-token map. Under corruption the clean contexts $y_{<t}$ are latent, so this device is unavailable, and the rollout-based estimators equation 13 and equation 17 must be analyzed through the growth of the trajectory class $\Pi_{1:T}$ on n prompts, which composition arguments bound only by quantities of order dT. Whether corrupted CoT supervision retains base-class complexity, so that the learner enjoys both the class-complexity benefit of CoT supervision and the channel redundancy of Remark 5.3, or whether corruption necessarily reintroduces a dependence on T in the complexity term, is in our view the central open question left by this work. The finite-class results settle neither direction, since $|\Pi_{1:T}| \le |\Pi|$ trivially, but they fix the separation factor that any answer must contain.

Corruption that propagates during generation. Our channel is offline: the complete clean trajectory is generated first, and corrupted tokens are never fed back into the teacher. The alternative dynamic model, studied empirically by Havrilla & Iyer (2024), feeds the corrupted prefix back, $z_t \sim M\big(\pi^\star(x, z_{<t}), \cdot\big)$, so that the observed sequence is a Markov process driven by the teacher on perturbed contexts. This changes the problem in three ways. First, the observed final answer is no longer a noisy copy of the clean answer but the teacher's answer to a perturbed trace, so the final-answer contamination model of Setting A is no longer implied by the trajectory model. Second, the likelihood of a candidate g becomes $\prod_t M\big(g(x, z_{<t}), z_t\big)$, a teacher-forced likelihood on the observed prefixes: in the propagating model, conditioning on the observed prefix is the correct scoring rule, and free-running rollouts are not, which is the opposite of the offline situation analyzed here. Third, the evidence available about the teacher is redistributed across contexts. Contexts visited by the corrupted process are weighted by their visitation probability, so evidence about the teacher's behavior off the clean trajectory accumulates, whereas the clean prefix $R^\star_{<t}(x)$ is visited with probability of order $(1-\alpha)^{t-1}$ under a pair-swap channel, so evidence about on-trajectory behavior at late positions decays exponentially in t. If the evaluation target remains the clean rollout, a candidate that agrees with the teacher on all frequently visited contexts but deviates on the clean trajectory at a late position has nonzero clean error yet is exponentially hard to detect, so prompt-level sample complexity for the clean target can grow exponentially in T, in sharp contrast with the polynomial behavior of the offline model. This is consistent with the empirical observation that dynamic corruption is substantially more harmful than static corruption (Havrilla & Iyer, 2024), and it suggests that under propagating corruption the appropriate target may itself have to change, for instance to the law of the corrupted process or to a robust notion of rollout. Technically, tensoriza-

tion over positions fails because the clean input to the channel at position t depends on the corrupted outputs at earlier positions, and the analysis would require sequential (martingale) versions of the Bhattacharyya arguments in Appendix C, with disagreement sets defined over contexts weighted by visitation probability rather than over positions of a fixed trajectory.

Further directions. Several smaller extensions remain within reach of the present techniques. Offline channels with correlated corruption across positions, such as block or burst corruption, replace the per-position overlap ρ_M by the overlap of the joint channel restricted to disagreement sets, and the $1 - \rho_M^T$ factor by the corresponding joint quantity. Partial observation of a subset of positions, including the final answer alone as the extreme case, interpolates between Settings A and B and can be analyzed by restricting the disagreement count to observed positions. Adaptivity is open in the unknown-Huber regime: our channel-agnostic guarantee is stated in terms of the least-favorable member of the family, and whether an estimator that does not know C can achieve the channel-specific separation $\Gamma_{M_C,T}$ for every C, for instance by joint estimation of the channel and the rollout, is not resolved by our results, although for symmetric channels tokenwise ERM coincides with maximum likelihood and is therefore already adaptive. On the algorithmic side, our language-model experiments use clean-prefix teacher forcing as a surrogate for the finite-class estimators (Appendix D.2.1); rollout-based training objectives that match the estimators analyzed here, and their behavior under corrupted prefixes, are a natural next step.

Taken together, we expect the product structure of complexity and separation to persist across these extensions whenever a positive separation exists and the objective remains calibrated for the intended clean target. The extensions differ in whether these two properties hold, and the baseline established here provides both the yardstick, through its lower bounds, and the proof template, through calibration and tensorization, against which each of them can be measured.

B Proofs for Section 4

We first record three standard facts that will also be used in the full-CoT analysis. For probability mass functions P and Q on a finite set $\mathcal{Z}$, write

$$\mathrm{BC}(P, Q) := \sum_{z \in \mathcal{Z}} \sqrt{P(z)Q(z)}$$

for their Bhattacharyya coefficient.

B.1 Auxiliary facts

Lemma B.1 (Bhattacharyya tensorization)**.** *Let* $P = \bigotimes_{i=1}^n P_i$, $Q = \bigotimes_{i=1}^n Q_i$. *Then* $\mathrm{BC}(P, Q) = \prod_{i=1}^n \mathrm{BC}(P_i, Q_i)$. *In particular,* $\mathrm{BC}(P^{\otimes n}, Q^{\otimes n}) = \mathrm{BC}(P, Q)^n$.

Proof. By direct factorization,

$$\mathrm{BC}(P, Q) = \sum_{z_{1:n}} \sqrt{\prod_{i=1}^n P_i(z_i) \prod_{i=1}^n Q_i(z_i)} = \sum_{z_{1:n}} \prod_{i=1}^n \sqrt{P_i(z_i)Q_i(z_i)} = \prod_{i=1}^n \sum_{z_i} \sqrt{P_i(z_i)Q_i(z_i)}.$$

□

Lemma B.2 (Common-marginal decomposition)**.** *Suppose* $P(x, z) = \rho(x)P_x(z), Q(x, z) = \rho(x)Q_x(z)$ *share the same marginal* ρ *on* $\mathcal{X}$. *Then* $\mathrm{BC}(P, Q) = \mathbb{E}_{X \sim \rho}\left[\mathrm{BC}(P_X, Q_X)\right]$.

Proof. We have

$$\mathrm{BC}(P, Q) = \sum_{x,z} \sqrt{\rho(x)P_x(z)\rho(x)Q_x(z)} = \sum_x \rho(x) \sum_z \sqrt{P_x(z)Q_x(z)}.$$

□

Lemma B.3 (Total variation and Bhattacharyya overlap)**.** *For any probability mass functions P and Q,* $\mathrm{TV}(P,Q) \le \sqrt{1-\mathrm{BC}(P,Q)^2}$.

Proof. By Cauchy–Schwarz,

$$\begin{aligned}
2\,\mathrm{TV}(P,Q) &= \sum_z |P(z)-Q(z)| \\
&= \sum_z \left|\sqrt{P(z)}-\sqrt{Q(z)}\right|\left(\sqrt{P(z)}+\sqrt{Q(z)}\right) \\
&\le \left[\sum_z \left(\sqrt{P(z)}-\sqrt{Q(z)}\right)^2\right]^{1/2} \cdot \left[\sum_z \left(\sqrt{P(z)}+\sqrt{Q(z)}\right)^2\right]^{1/2} \\
&= 2\sqrt{1-\mathrm{BC}(P,Q)^2}.
\end{aligned}$$

Dividing by two proves the claim. □

Lemma B.4 (Assouad's lemma)**.** *Let $\{P_\theta : \theta \in \{0,1\}^d\}$ be a family of distributions, and let $\theta^{\oplus j}$ denote θ with its j-th coordinate flipped. If $\mathrm{TV}\left(P_\theta, P_{\theta^{\oplus j}}\right) \le \tau$ for every $\theta \in \{0,1\}^d$ and $j \in [d]$, then every possibly randomized estimator $\widehat{\theta}$ satisfies $\mathbb{E}_\theta \mathbb{E}_{Z\sim P_\theta}\left[d_{\mathrm{H}}\left(\widehat{\theta}(Z),\theta\right)\right] \ge \frac{d}{2}(1-\tau)$, where θ is uniform on $\{0,1\}^d$.*

Proof. Fix $j \in [d]$. Pair every parameter θ satisfying $\theta_j = 0$ with $\theta^{\oplus j}$. For each such pair, the average error of estimating the j-th coordinate is at least the optimal binary-testing error:

$$\frac{1}{2}P_\theta\left(\widehat{\theta}_j = 1\right) + \frac{1}{2}P_{\theta^{\oplus j}}\left(\widehat{\theta}_j = 0\right) \ge \frac{1}{2}\left[1 - \mathrm{TV}\left(P_\theta, P_{\theta^{\oplus j}}\right)\right] \ge \frac{1}{2}(1-\tau).$$

Averaging over all pairs gives

$$\mathbb{E}_\theta P_\theta\left(\widehat{\theta}_j \ne \theta_j\right) \ge \frac{1}{2}(1-\tau).$$

Summing over $j \in [d]$ proves the result. □

B.2 Proof of Theorem 4.1

We prove the $\log N$ and $\log(1/\delta)$ terms separately. Since the worst-case sample complexity takes a supremum over admissible classes, the two constructions may use different classes.

The $\log N$ term Let $d := \lfloor \log_2 N \rfloor$. Since $N \ge 4$, $2 \le d$, $\frac{N}{2} \le 2^d \le N$. Choose two distinct tokens $a, b \in \Sigma$ and distinct prompts $x_0, x_1, \ldots, x_d \in \Sigma^\star$. Let $\mathcal{X}_0 := \{x_0, x_1, \ldots, x_d\}$, $\beta := 8\epsilon$, and define $\rho(x_0) = 1-\beta$, $\rho(x_j) = \frac{\beta}{d}$, $j \in [d]$. The assumption $\epsilon < 1/16$ guarantees $\beta < 1/2$.

For each $\theta = (\theta_1, \ldots, \theta_d) \in \{0,1\}^d$, define

$$h_\theta(x_0) := a, \qquad h_\theta(x_j) := \begin{cases} a, & \theta_j = 0, \\ b, & \theta_j = 1. \end{cases}$$

Let $\Pi_T := \{h_\theta : \theta \in \{0,1\}^d\}$. Then $|\Pi_T| = 2^d$, $\frac{N}{2} \le |\Pi_T| \le N$. This final-answer class can be induced by a finite class of deterministic autoregressive models. For example, for each θ, define π_θ to output a at the first $T-1$ steps and $h_\theta(x)$ at the terminal step:

$$\pi_\theta(x,s) := \begin{cases} a, & |s| < T-1, \\ h_\theta(x), & |s| = T-1. \end{cases}$$

For target h_θ, define the prompt-dependent corruption kernel by

$$Q_\theta(\cdot \mid x_j) := \begin{cases} \delta_b, & \theta_j = 0, \\ \delta_a, & \theta_j = 1, \end{cases} \qquad j \in [d],$$

and $Q_\theta(\cdot \mid x_0) := \delta_a$. Thus, on an informative prompt x_j, the observed label equals the clean label with probability $1-\alpha$ and the opposite token with probability α.

Let P_θ denote the one-sample distribution of $(X, \widetilde{Y}_T)$, and let P_θ^n denote its n-fold product. Consider any possibly randomized and improper learner A_n, and write $\widehat{h} := \mathsf{A}_n(S_n^{\mathrm{FA}})$. From $\widehat{h}$, define $\widehat{\theta}_j := \mathbf{1}\{\widehat{h}(x_j) = b\}, j \in [d]$. If $\widehat{h}(x_j) \notin \{a, b\}$, this convention can only underestimate the prediction error. Hence

$$L^\star_{\mathrm{FA},\theta}(\widehat{h}) \geq \frac{\beta}{d} d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right). \tag{27}$$

Fix θ and j, and let $\theta^{\oplus j}$ be obtained by flipping the j-th coordinate. The distributions P_θ and $P_{\theta^{\oplus j}}$ differ only when $X = x_j$, which occurs with probability β/d. Conditional on $X = x_j$, the two observed-label distributions are, up to relabeling,

$$(1-\alpha)\delta_a + \alpha\delta_b \qquad \text{and} \qquad \alpha\delta_a + (1-\alpha)\delta_b.$$

Their Bhattacharyya coefficient is $2\sqrt{\alpha(1-\alpha)} = 1 - \gamma_\alpha$. By Lemma B.2, $\mathrm{BC}\left(P_\theta, P_{\theta^{\oplus j}}\right) = 1 - \frac{\beta\gamma_\alpha}{d}$. Lemmas B.1 and B.3 therefore give $\mathrm{TV}\left(P_\theta^n, P_{\theta^{\oplus j}}^n\right) \leq \tau_n$, where $\tau_n := \sqrt{1 - \left(1 - \frac{\beta\gamma_\alpha}{d}\right)^{2n}}$. By Lemma B.4,

$$\mathbb{E}_\theta \mathbb{E}_{P_\theta^n}\left[d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right)\right] \geq \frac{d}{2}(1 - \tau_n). \tag{28}$$

Now suppose that A_n satisfies the PAC guarantee with failure probability at most $\delta \leq c_0$ on this class. Since $\beta = 8\epsilon$, equation 27 implies $d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right) > \frac{d}{8} \implies L^\star_{\mathrm{FA},\theta}(\widehat{h}) > \epsilon$. Consequently, $\mathbb{E}_\theta P_\theta^n\left(d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right) > \frac{d}{8}\right) \leq \delta$. Since the Hamming distance is at most d,

$$\mathbb{E}_\theta \mathbb{E}_{P_\theta^n}\left[d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right)\right] \leq \frac{d}{8} + \delta d. \tag{29}$$

Choose $c_0 \leq 1/16$. Combining equation 28 and equation 29 gives $\frac{d}{2}(1 - \tau_n) \leq \frac{3d}{16}$, and hence $\tau_n \geq \frac{5}{8}$. Therefore, $\left(1 - \frac{\beta\gamma_\alpha}{d}\right)^{2n} \leq \frac{39}{64}$. It follows that

$$n \geq \frac{\log(64/39)}{2\left[-\log\left(1 - \frac{\beta\gamma_\alpha}{d}\right)\right]}.$$

Because $\frac{\beta\gamma_\alpha}{d} \leq \frac{1}{4}$ and $-\log(1-u) \leq 2u, u \in [0, 1/2]$, we obtain $n \geq c_1 \frac{d}{\beta\gamma_\alpha} \geq c_2 \frac{\log N}{\epsilon\gamma_\alpha}$ for universal constants $c_1, c_2 > 0$.

The $\log(1/\delta)$ term Choose two prompts x_0, x_1 and let $\beta := 2\epsilon, \rho(x_1) = \beta, \rho(x_0) = 1 - \beta$. Define two final-answer predictors $h_0(x_0) = h_1(x_0) = a, h_0(x_1) = a, h_1(x_1) = b$. As above, these predictors can be induced by deterministic autoregressive models.

Define the corruption kernels by $Q_0(\cdot \mid x_1) = \delta_b, Q_1(\cdot \mid x_1) = \delta_a$, and $Q_0(\cdot \mid x_0) = Q_1(\cdot \mid x_0) = \delta_a$. Let P_0 and P_1 denote the corresponding one-sample distributions. Their Bhattacharyya coefficient is $\mathrm{BC}(P_0, P_1) = 1 - \beta\gamma_\alpha$. Hence $\mathrm{BC}(P_0^n, P_1^n) = (1 - \beta\gamma_\alpha)^n$.

Any learner with clean error at most ϵ must output the correct token at x_1, because an incorrect prediction there incurs error $\rho(x_1) = 2\epsilon > \epsilon$. Thus, any learner satisfying the PAC guarantee induces a binary test between P_0^n and P_1^n whose average error is at most δ.

On the other hand, every binary test with equal priors has average error at least $\frac{1}{2}\left[1 - \mathrm{TV}(P_0^n, P_1^n)\right]$. Using Lemma B.3 and $1 - \sqrt{1-u} \geq \frac{u}{2}, u \in [0, 1]$, we obtain $\delta \geq \frac{1}{4}\left(1 - \beta\gamma_\alpha\right)^{2n}$. Therefore,

$$n \geq \frac{\log(1/(4\delta))}{2\left[-\log(1 - \beta\gamma_\alpha)\right]}.$$

For $\epsilon < 1/16$ and $\delta \leq 1/16$, $-\log(1 - \beta\gamma_\alpha) \leq 2\beta\gamma_\alpha$ and $\log\frac{1}{4\delta} \geq \frac{1}{2}\log\frac{1}{\delta}$. Hence $n \geq c_3 \frac{\log(1/\delta)}{\epsilon\gamma_\alpha}$ for a universal constant $c_3 > 0$.

Combining the two constructions and using $\max\{r, s\} \geq \frac{r+s}{2}$ proves

$$n^{\star}_{\mathrm{FA,wc}}(\epsilon, \delta; N, \alpha) \geq c\frac{\log N + \log(1/\delta)}{\epsilon\gamma_{\alpha}}.$$

B.3 Proof of Theorem 4.2

Fix $h \in \Pi_T$ and let $r_h := L^{\star}_{\mathrm{FA}}(h)$. For the i-th sample, define the empirical-loss difference

$$D_i(h) := \mathbf{1}\left\{h(X_i) \neq \widetilde{Y}_{i,T}\right\} - \mathbf{1}\left\{h^{\star}(X_i) \neq \widetilde{Y}_{i,T}\right\}.$$

If $h(X_i) = h^{\star}(X_i)$, then $D_i(h) = 0$. Now condition on a prompt x satisfying $h(x) \neq h^{\star}(x)$. Let

$$p_+(x) := \Pr\left(\widetilde{Y}_T = h^{\star}(x) \mid X = x\right), \quad \text{and} \quad p_-(x) := \Pr\left(\widetilde{Y}_T = h(x) \mid X = x\right).$$

Then $p_+(x) \geq 1 - \alpha, p_-(x) \leq \alpha$, and

$$D_i(h) = \begin{cases} 1, & \widetilde{Y}_{i,T} = h^{\star}(x), \\ -1, & \widetilde{Y}_{i,T} = h(x), \\ 0, & \text{otherwise.} \end{cases}$$

For every $\lambda > 0$,

$$\begin{aligned}\mathbb{E}\left[e^{-\lambda D_i(h)} \mid X_i = x\right] &= p_+(x)e^{-\lambda} + p_-(x)e^{\lambda} + 1 - p_+(x) - p_-(x) \\ &\leq (1-\alpha)e^{-\lambda} + \alpha e^{\lambda}.\end{aligned}$$

Averaging over X_i yields $\mathbb{E}\left[e^{-\lambda D_i(h)}\right] \leq 1 - r_h + r_h\left[(1-\alpha)e^{-\lambda} + \alpha e^{\lambda}\right]$.

For $\alpha \in (0, 1/2)$, the expression in brackets is minimized at $\lambda_{\alpha} = \frac{1}{2}\log\frac{1-\alpha}{\alpha}$, and its minimum equals $2\sqrt{\alpha(1-\alpha)} = 1 - \gamma_{\alpha}$. For $\alpha = 0$, the same value is obtained in the limit $\lambda \to \infty$. Hence $\mathbb{E}\left[e^{-\lambda_{\alpha} D_i(h)}\right] \leq 1 - r_h\gamma_{\alpha}$.

By Markov's inequality and independence across examples,

$$\Pr\left(\widehat{L}_{\alpha}(h) \leq \widehat{L}_{\alpha}(h^{\star})\right) = \Pr\left(\sum_{i=1}^{n} D_i(h) \leq 0\right) = \Pr\left(e^{-\lambda_{\alpha}\sum_{i=1}^{n} D_i(h)} \geq 1\right) \leq (1 - r_h\gamma_{\alpha})^n.$$

Define $\mathcal{B}_{\epsilon} := \{h \in \Pi_T : L^{\star}_{\mathrm{FA}}(h) > \epsilon\}$. Since $h^{\star} \in \Pi_T$, if the ERM output has clean error larger than ϵ, then some $h \in \mathcal{B}_{\epsilon}$ has empirical risk no larger than that of $h^{\star}$. Therefore,

$$\Pr\left(L^{\star}_{\mathrm{FA}}(\widehat{h}_n) > \epsilon\right) \leq \sum_{h \in \mathcal{B}_{\epsilon}} \Pr\left(\widehat{L}_{\alpha}(h) \leq \widehat{L}_{\alpha}(h^{\star})\right) \leq |\Pi_T|\,(1 - \epsilon\gamma_{\alpha})^n.$$

The derivation is uniform over ρ, $h^{\star} \in \Pi_T$, and $Q : \mathcal{X} \to \Delta(\Sigma)$, proving equation 8.

The right-hand side is at most δ whenever

$$n \geq \frac{\log(|\Pi_T|/\delta)}{-\log(1 - \epsilon\gamma_{\alpha})}.$$

Since ERM is one of the learning rules over which the minimax definition takes an infimum,

$$n^{\star}_{\mathrm{FA}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\Pi_T|/\delta)}{-\log(1 - \epsilon\gamma_{\alpha})} \right\rceil.$$

Finally, $-\log(1-u) \geq u, u \in [0, 1)$, gives

$$n^{\star}_{\mathrm{FA}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\Pi_T|/\delta)}{\epsilon\gamma_{\alpha}} \right\rceil.$$

C Proofs for Section 5

C.1 Channel-separation inequalities

Lemma C.1. *For every $q \in [0,1)$ and every $k \in \{0,\dots,T\}$, $q^k \le 1 - \frac{k}{T}\left(1-q^T\right)$, with the convention $0^0 = 1$.*

Proof. For $q \in (0,1)$, the function $s \mapsto q^s$ is convex on $[0,T]$ and therefore lies below the chord connecting $(0,1)$ and (T,q^T). Thus, $q^k \le \left(1-\frac{k}{T}\right) + \frac{k}{T}q^T$. The case $q=0$ follows directly. □

C.1.1 Known-channel separation

Fix $g \in \Pi_{1:T}$ and $x \in \mathcal{X}$. By Lemma B.1,

$$\mathrm{BC}\left(P_{g,M}(\cdot \mid x), P_{R^\star,M}(\cdot \mid x)\right) = \prod_{t=1}^{T} \sum_{z\in\Sigma} \sqrt{M(g_t(x),z)M(R_t^\star(x),z)}.$$

A matching position contributes a factor of one. A disagreeing position contributes at most ρ_M. Therefore, $\mathrm{BC}\left(P_{g,M}(\cdot \mid x), P_{R^\star,M}(\cdot \mid x)\right) \le \rho_M^{K_g(x)}$, where $K_g(x) = T\, d_{\mathrm{tok}}\left(g(x), R^\star(x)\right)$. Applying Lemma C.1 gives

$$\rho_M^{K_g(x)} \le 1 - \frac{K_g(x)}{T}\left(1-\rho_M^T\right) = 1 - \Gamma_{M,T} d_{\mathrm{tok}}\left(g(x), R^\star(x)\right).$$

This proves equation 12. Averaging over $X \sim \rho$ yields

$$\mathbb{E}_{X\sim\rho}\,\mathrm{BC}\left(P_{g,M}(\cdot \mid X), P_{R^\star,M}(\cdot \mid X)\right) \le 1 - \Gamma_{M,T} L^\star_{\mathrm{tok}}(g). \tag{30}$$

C.1.2 Uniform separation of Huber channels

Fix $C \in \mathcal{C}_\Sigma$ and distinct tokens $a, b \in \Sigma$. Write $C_a := C(a,\cdot), C_b := C(b,\cdot)$. Then

$$M_C(a,\cdot) = (1-\alpha)\delta_a + \alpha C_a, \quad \text{and} \quad M_C(b,\cdot) = (1-\alpha)\delta_b + \alpha C_b.$$

By the reverse triangle inequality,

$$\begin{aligned} &2\,\mathrm{TV}\left(M_C(a,\cdot), M_C(b,\cdot)\right) \\ &\quad = \|(1-\alpha)(\delta_a - \delta_b) + \alpha(C_a - C_b)\|_1 \\ &\quad \ge (1-\alpha)\|\delta_a - \delta_b\|_1 - \alpha\|C_a - C_b\|_1 \\ &\quad \ge 2(1-\alpha) - 2\alpha. \end{aligned}$$

Thus, $\mathrm{TV}\left(M_C(a,\cdot), M_C(b,\cdot)\right) \ge 1-2\alpha$. Lemma B.3 gives

$$\mathrm{BC}_{M_C}(a,b) \le \sqrt{1-(1-2\alpha)^2} = 2\sqrt{\alpha(1-\alpha)}.$$

This proves equation 15. Taking the maximum over $a \ne b$ gives $\rho_{M_C} \le 2\sqrt{\alpha(1-\alpha)}$, and hence

$$\Gamma_{M_C,T} = 1 - \rho_{M_C}^T \ge 1 - \left(2\sqrt{\alpha(1-\alpha)}\right)^T = \Gamma_{\alpha,T}.$$

This proves equation 16.

The bound is tight. Choose two tokens a, b and let C deterministically exchange them: $C(a,\cdot) = \delta_b, C(b,\cdot) = \delta_a$. Then

$$M_C(a,\cdot) = (1-\alpha)\delta_a + \alpha\delta_b, \quad \text{and} \quad M_C(b,\cdot) = \alpha\delta_a + (1-\alpha)\delta_b,$$

whose Bhattacharyya overlap is exactly $2\sqrt{\alpha(1-\alpha)}$.

C.1.3 CALIBRATION OF CHANNEL-AGNOSTIC ERM

Fix $g \in \Pi_{1:T}$, $x \in \mathcal{X}$, and $t \in [T]$. If $g_t(x) = R^\star_t(x)$, then the candidate and target incur identical tokenwise losses. Otherwise, write $a := R^\star_t(x), b := g_t(x), a \neq b$. For $Z_t \sim M_C(a, \cdot)$,

$$\begin{aligned}\mathbb{E}\left[\mathbf{1}\{b \neq Z_t\} - \mathbf{1}\{a \neq Z_t\} \mid X = x\right] &= M_C(a, a) - M_C(a, b) \\ &= 1 - \alpha + \alpha C(a, a) - \alpha C(a, b) \\ &\geq 1 - 2\alpha.\end{aligned}$$

Summing over positions, dividing by T, and averaging over $X \sim \rho$ yields

$$\widetilde{L}_{\mathrm{tok},C}(g) - \widetilde{L}_{\mathrm{tok},C}(R^\star) \geq (1 - 2\alpha) L^\star_{\mathrm{tok}}(g),$$

which proves equation 18.

C.2 PROOF OF THEOREM 5.1

C.2.1 KNOWN CHANNEL

Fix $g \in \Pi_{1:T}$ and define

$$\mathcal{E}_g := \left\{ \prod_{i=1}^{n} P_{g,M}(Z_{i,1:T} \mid X_i) \geq \prod_{i=1}^{n} P_{R^\star,M}(Z_{i,1:T} \mid X_i) \right\}.$$

Condition on $X_{1:n} = x_{1:n}$. Write $\mathcal{L}_g(z_{1:n} \mid x_{1:n}) := \prod_{i=1}^{n} P_{g,M}(z_{i,1:T} \mid x_i)$ and define $\mathcal{L}_\star$ analogously. On the event $\mathcal{L}_g(z_{1:n} \mid x_{1:n}) \geq \mathcal{L}_\star(z_{1:n} \mid x_{1:n})$, we have $\mathcal{L}_\star(z_{1:n} \mid x_{1:n}) \leq \sqrt{\mathcal{L}_\star(z_{1:n} \mid x_{1:n}) \mathcal{L}_g(z_{1:n} \mid x_{1:n})}$. Therefore,

$$\begin{aligned}\Pr_M\left(\mathcal{E}_g \mid X_{1:n} = x_{1:n}\right) &\leq \sum_{z_{1:n}} \sqrt{\mathcal{L}_\star(z_{1:n} \mid x_{1:n}) \mathcal{L}_g(z_{1:n} \mid x_{1:n})} \\ &= \prod_{i=1}^{n} \mathrm{BC}\left(P_{g,M}(\cdot \mid x_i), P_{R^\star,M}(\cdot \mid x_i)\right).\end{aligned}$$

Averaging over the i.i.d. prompts and applying equation 30,

$$\Pr_M(\mathcal{E}_g) \leq \left[\mathbb{E}_{X \sim \rho} \mathrm{BC}\left(P_{g,M}(\cdot \mid X), P_{R^\star,M}(\cdot \mid X)\right)\right]^n \leq \left(1 - \Gamma_{M,T} L^\star_{\mathrm{tok}}(g)\right)^n.$$

Define $\mathcal{B}_\epsilon := \{g \in \Pi_{1:T} : L^\star_{\mathrm{tok}}(g) > \epsilon\}$. Since $R^\star \in \Pi_{1:T}$, if the maximum-likelihood estimator has clean risk larger than ϵ, then $\mathcal{E}_g$ occurs for some $g \in \mathcal{B}_\epsilon$. Hence

$$\Pr_M\left(L^\star_{\mathrm{tok}}(\widehat{g}^M_n) > \epsilon\right) \leq \sum_{g \in \mathcal{B}_\epsilon} \Pr_M(\mathcal{E}_g) \leq |\Pi_{1:T}| \left(1 - \epsilon \Gamma_{M,T}\right)^n.$$

The derivation is uniform over ρ and $R^\star \in \Pi_{1:T}$, proving equation 19.

The right-hand side is at most δ whenever

$$n \geq \frac{\log(|\Pi_{1:T}|/\delta)}{-\log(1 - \epsilon \Gamma_{M,T})}.$$

Since the channel-aware MLE is a valid member of the class of learning rules over which the minimax definition takes an infimum,

$$n^{\star,\mathrm{CoT}}_{\mathrm{known}}(\epsilon, \delta) \leq \left\lceil \frac{\log(|\Pi_{1:T}|/\delta)}{-\log(1 - \epsilon \Gamma_{M,T})} \right\rceil.$$

The simplified bound follows from $-\log(1 - u) \geq u$.

C.2.2 UNKNOWN HUBER CHANNEL

For a fixed $g \in \Pi_{1:T}$, define $D_{i,t}(g) := \mathbf{1}\{g_t(X_i) \neq Z_{i,t}\} - \mathbf{1}\{R^\star_t(X_i) \neq Z_{i,t}\}$, and $D_i(g) := \sum_{t=1}^T D_{i,t}(g)$. Then

$$\frac{1}{n}\sum_{i=1}^n d_{\mathrm{tok}}\left(g(X_i), Z_{i,1:T}\right) \leq \frac{1}{n}\sum_{i=1}^n d_{\mathrm{tok}}\left(R^\star(X_i), Z_{i,1:T}\right)$$

if and only if $\sum_{i=1}^n D_i(g) \leq 0$.

Fix a prompt x. At a matching position, $D_{i,t}(g) = 0$. At a disagreeing position, write

$$a := R^\star_t(x), \qquad b := g_t(x).$$

Let

$$p_a := M_C(a,a), \qquad p_b := M_C(a,b).$$

Then

$$p_a \geq 1-\alpha, \qquad p_b \leq \alpha,$$

and

$$D_{i,t}(g) = \begin{cases} 1, & Z_{i,t} = a, \\ -1, & Z_{i,t} = b, \\ 0, & Z_{i,t} \notin \{a,b\}. \end{cases}$$

Thus, for every $\lambda > 0$,

$$\mathbb{E}\left[e^{-\lambda D_{i,t}(g)} \mid X_i = x\right] = p_a e^{-\lambda} + p_b e^{\lambda} + 1 - p_a - p_b \leq (1-\alpha)e^{-\lambda} + \alpha e^{\lambda}.$$

Conditional independence across positions gives

$$\mathbb{E}\left[e^{-\lambda D_i(g)} \mid X_i = x\right] \leq \left[(1-\alpha)e^{-\lambda} + \alpha e^{\lambda}\right]^{K_g(x)}.$$

By Markov's inequality and independence across examples,

$$\Pr_{M_C}\left(\sum_{i=1}^n D_i(g) \leq 0\right) \leq \left[\mathbb{E}_{X\sim\rho}\left[\left((1-\alpha)e^{-\lambda} + \alpha e^{\lambda}\right)^{K_g(X)}\right]\right]^n.$$

For $\alpha \in (0, 1/2)$, the expression $(1-\alpha)e^{-\lambda} + \alpha e^{\lambda}$ is minimized at $\lambda_\alpha = \frac{1}{2}\log\frac{1-\alpha}{\alpha}$, with minimum $q_\alpha := 2\sqrt{\alpha(1-\alpha)}$. For $\alpha = 0$, the same value follows by taking $\lambda \to \infty$. Therefore,

$$\Pr_{M_C}\left(\sum_{i=1}^n D_i(g) \leq 0\right) \leq \left[\mathbb{E}_{X\sim\rho} q_\alpha^{K_g(X)}\right]^n.$$

By Lemma C.1, $q_\alpha^{K_g(X)} \leq 1 - \Gamma_{\alpha,T}\frac{K_g(X)}{T}$. Hence $\mathbb{E}_{X\sim\rho} q_\alpha^{K_g(X)} \leq 1 - \Gamma_{\alpha,T} L^\star_{\mathrm{tok}}(g)$, and so

$$\Pr_{M_C}\left(g \text{ has observed empirical loss no larger than } R^\star\right) \leq \left(1 - \Gamma_{\alpha,T} L^\star_{\mathrm{tok}}(g)\right)^n.$$

A union bound over $\mathcal{B}_\epsilon = \{g \in \Pi_{1:T} : L^\star_{\mathrm{tok}}(g) > \epsilon\}$ gives, uniformly over $C \in \mathcal{C}_\Sigma$, $\Pr_{M_C}\left(L^\star_{\mathrm{tok}}(\widehat{g}^{\mathrm{unk}}_n) > \epsilon\right) \leq |\Pi_{1:T}|\left(1 - \epsilon\Gamma_{\alpha,T}\right)^n$. This proves equation 21. The exact and simplified minimax PAC bounds follow exactly as in the known-channel case.

C.3 PROOF OF THEOREM 5.2

The upper bounds follow from Theorem 5.1. It remains to prove the two lower bounds.

C.3.1 KNOWN-CHANNEL LOWER BOUND

Since Σ is finite, there exist distinct tokens $a, b \in \Sigma$ such that $\rho_M = \mathrm{BC}_M(a,b)$. Define

$$w_0 := (a, \ldots, a), \qquad w_1 := (b, \ldots, b).$$

Then $d_{\mathrm{tok}}(w_0, w_1) = 1$. Let U_M and V_M be their corresponding corrupted-trajectory distributions. By Bhattacharyya tensorization,

$$\mathrm{BC}(U_M, V_M) = \rho_M^T. \tag{31}$$

We again prove the $\log N$ and $\log(1/\delta)$ terms separately.

The $\log N$ term. Let $d := \lfloor \log_2 N \rfloor, \beta := 16\epsilon$. Choose prompts $x_0, x_1, \ldots, x_d$ and define

$$\rho(x_0) = 1 - \beta, \qquad \rho(x_j) = \frac{\beta}{d}, \quad j \in [d].$$

For every $\theta \in \{0,1\}^d$, define a deterministic prefix-independent autoregressive model by $\pi_\theta(x_0, s) := a$ and

$$\pi_\theta(x_j, s) := \begin{cases} a, & \theta_j = 0, \\ b, & \theta_j = 1, \end{cases} \qquad j \in [d].$$

Its clean trajectory satisfies

$$R_{\pi_\theta}(x_0) = w_0, \qquad R_{\pi_\theta}(x_j) = w_{\theta_j}.$$

Let $\Pi := \{\pi_\theta : \theta \in \{0,1\}^d\}$. Then $|\Pi_{1:T}| = 2^d$, $\frac{N}{2} \le |\Pi_{1:T}| \le N$. Let P_θ denote the one-sample distribution of $(X, Z_{1:T})$ under target π_θ. For an arbitrary possibly randomized and improper learner, write $\widehat{g} := \mathsf{A}_n(S_n^{\mathrm{CoT}})$. Decode

$$\widehat{\theta}_j \in \operatorname*{arg\,min}_{r \in \{0,1\}} d_{\mathrm{tok}}\left(\widehat{g}(x_j), w_r\right).$$

If $\widehat{\theta}_j \neq \theta_j$, then the triangle inequality and $d_{\mathrm{tok}}(w_0, w_1) = 1$ imply $d_{\mathrm{tok}}\left(\widehat{g}(x_j), w_{\theta_j}\right) \ge \frac{1}{2}$. Consequently,

$$L^\star_{\mathrm{tok},\theta}(\widehat{g}) \ge \frac{\beta}{2d} d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right). \tag{32}$$

For adjacent θ and $\theta^{\oplus j}$, the one-sample distributions differ only at x_j. By equation 31,

$$\mathrm{BC}\left(P_\theta, P_{\theta^{\oplus j}}\right) = 1 - \frac{\beta}{d} + \frac{\beta}{d}\rho_M^T = 1 - \frac{\beta \Gamma_{M,T}}{d}.$$

Hence $\mathrm{TV}\left(P_\theta^n, P_{\theta^{\oplus j}}^n\right) \le \tau_n$, where $\tau_n := \sqrt{1 - \left(1 - \frac{\beta\Gamma_{M,T}}{d}\right)^{2n}}$. Assouad's lemma yields

$$\mathbb{E}_\theta \mathbb{E}_{P_\theta^n}\left[d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right)\right] \ge \frac{d}{2}(1 - \tau_n). \tag{33}$$

Suppose the learner satisfies the PAC guarantee with failure probability at most $\delta \le c_0$. Since $\beta = 16\epsilon$, equation 32 implies $d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right) > \frac{d}{8} \implies L^\star_{\mathrm{tok},\theta}(\widehat{g}) > \epsilon$. Therefore, $\mathbb{E}_\theta P_\theta^n\left(d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right) > \frac{d}{8}\right) \le \delta$. Since the Hamming distance is at most d, $\mathbb{E}_\theta \mathbb{E}_{P_\theta^n}\left[d_{\mathrm{H}}\left(\widehat{\theta}, \theta\right)\right] \le \frac{d}{8} + \delta d$. Taking $c_0 \le 1/16$ and comparing with equation 33, we obtain $\tau_n \ge \frac{5}{8}$. The same logarithmic calculation as in Appendix B.2 gives

$$n \ge c \frac{\log N}{\epsilon \Gamma_{M,T}}.$$

The $\log(1/\delta)$ term. Choose two prompts x_0, x_1 and set

$$\beta := 4\epsilon, \qquad \rho(x_1) = \beta, \qquad \rho(x_0) = 1 - \beta.$$

Consider two targets that both generate w_0 at x_0, while at x_1 they generate w_0 and w_1, respectively.

For an arbitrary output $\widehat{g}$, define a test by selecting the closer of w_0 and w_1 to $\widehat{g}(x_1)$. Under target r, if $L^\star_{\mathrm{tok},r}(\widehat{g}) \le \epsilon$, then $d_{\mathrm{tok}}\left(\widehat{g}(x_1), w_r\right) \le \frac{\epsilon}{\beta} = \frac{1}{4}$. Since $d_{\mathrm{tok}}(w_0, w_1) = 1$, the induced test is correct. Thus, a learner with failure probability at most δ induces a binary test with average error at most δ.

The one-sample Bhattacharyya coefficient of the two data distributions is $1 - \beta + \beta\rho_M^T = 1 - \beta\Gamma_{M,T}$. As in the final-answer binary-testing argument, $\delta \ge \frac{1}{4}\left(1 - \beta\Gamma_{M,T}\right)^{2n}$. For sufficiently small universal c_0,

$$n \ge c \frac{\log(1/\delta)}{\epsilon \Gamma_{M,T}}.$$

Combining the two constructions gives

$$n^{\star,\mathrm{CoT}}_{\mathrm{known,wc}}(\epsilon, \delta; N, M) \ge c \frac{\log N + \log(1/\delta)}{\epsilon \Gamma_{M,T}}.$$

C.3.2 UNKNOWN-CHANNEL LOWER BOUND

Choose distinct tokens $a, b \in \Sigma$ and let $C \in \mathcal{C}_\Sigma$ exchange them while fixing all other tokens:

$$C(a,\cdot) = \delta_b, \qquad C(b,\cdot) = \delta_a, \qquad C(c,\cdot) = \delta_c \quad \text{for } c \notin \{a,b\}.$$

Then

$$M_C(a,\cdot) = (1-\alpha)\delta_a + \alpha\delta_b, \quad \text{and} \quad M_C(b,\cdot) = \alpha\delta_a + (1-\alpha)\delta_b.$$

The overlap of these two rows is $2\sqrt{\alpha(1-\alpha)}$, while all other distinct row pairs have no larger overlap. Hence $\rho_{M_C} = 2\sqrt{\alpha(1-\alpha)}$ and $\Gamma_{M_C,T} = \Gamma_{\alpha,T}$. The known-channel lower bound applies even when this channel is revealed to the learner. It therefore also applies to the smaller class of channel-agnostic learning rules, yielding

$$n^{\star,\mathrm{CoT}}_{\mathrm{unknown,wc}}(\epsilon,\delta;N,\alpha) \geq c\frac{\log N + \log(1/\delta)}{\epsilon\Gamma_{\alpha,T}}.$$

Together with the upper bound from Theorem 5.1, this proves Theorem 5.2.

C.4 ADDITIONAL CALCULATION FOR REMARK 5.3

Let $s_M := 1-\rho_M$. Then $\Gamma_{M,T} = 1-(1-s_M)^T$. Bernoulli's inequality gives $\Gamma_{M,T} \leq \min\{Ts_M, 1\}$. Conversely, $(1-s_M)^T \leq e^{-Ts_M}$, so $\Gamma_{M,T} \geq 1-e^{-Ts_M}$. For $x \in [0,1]$, $1-e^{-x} \geq (1-e^{-1})x$, whereas for $x \geq 1$, $1-e^{-x} \geq 1-e^{-1}$. Therefore, $(1-e^{-1})\min\{Ts_M,1\} \leq \Gamma_{M,T} \leq \min\{Ts_M,1\}$, which proves

$$\Gamma_{M,T} \asymp \min\{T(1-\rho_M), 1\}.$$

Remark C.2. *Two caveats qualify this redundancy. First, it relies on corruption being independent across positions: the tensorization in equation equation 11 is what turns K disagreeing positions into K independent pieces of channel evidence, and our analysis does not cover correlated or propagating corruption, such as an early error that consistently alters all subsequent states. Second, it is a statement about normalized token-Hamming error. Measured in final-answer error, the only worst-case guarantee available from Theorem 5.1 is $L^\star_{\mathrm{FA}}(g) \leq T\, L^\star_{\mathrm{tok}}(g)$, so final-answer error ϵ requires token error ϵ/T and hence $\lceil \log(|\Pi_{1:T}|/\delta)\, T/(\epsilon\Gamma_{M,T})\rceil \asymp \log(|\Pi_{1:T}|/\delta)/(\epsilon\min\{1-\rho_M,\, 1/T\})$ samples. In particular, when candidate rollouts differ only in the final token, a corrupted trajectory carries exactly the information of a corrupted final answer, and full-CoT supervision offers no gain over Setting A for final-answer accuracy. The redundancy is therefore a property of classes whose rollouts diverge on many positions, as in the hard instances of Theorem 5.2, rather than of trajectory length per se.*

D EXPERIMENTAL DETAILS

This appendix contains the two parts of our empirical evaluation. Appendix D.1 presents controlled finite-class simulations that directly test the theoretical scaling laws. Appendix D.2 serves as a comprehensive supplement to the real large language model (LLM) experiments. Specifically, Appendix D.2.1 details the main configurations, including datasets, data corruption protocols, and training settings, followed by ablations across model scale and architecture (Appendix D.2.2).

D.1 PART I: CONTROLLED FINITE-CLASS SIMULATIONS

Evaluation protocol. *The controlled simulations instantiate the sufficient statistics of the hard families used in the lower-bound proofs. They therefore isolate the statistical effects of contamination without introducing neural optimization error or model misspecification.*

For each experimental condition, we evaluate a grid $\mathcal{N}$ of prompt-level sample sizes. Let $\widehat{p}_n$ denote the empirical probability that the learned predictor has clean risk greater than ϵ. We define the empirical critical sample size by

$$\widehat{n} := \min\{n \in \mathcal{N} : \widehat{p}_n \leq \delta\}. \tag{34}$$

The values of N, ϵ, δ, the number of simulation replications, and the threshold-interpolation rule are held fixed within each comparison and are included in the released configuration.

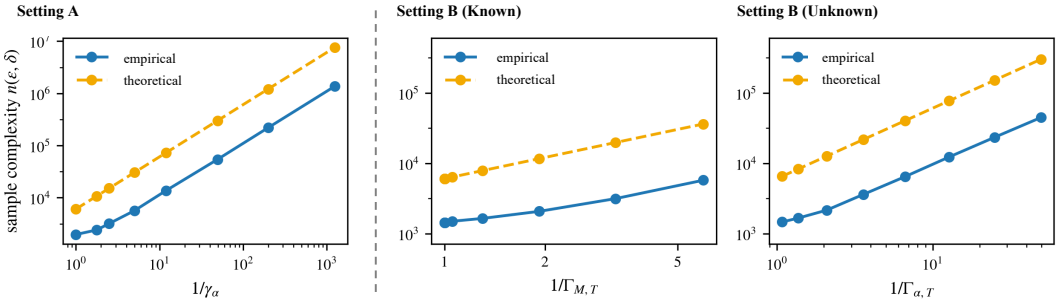

Figure 3: **Part I: Controlled finite-class simulations.** Blue solid curves show empirical critical prompt-level sample sizes, and orange dashed curves show the corresponding theoretical predictions. **Left:** Setting A, plotted against $1/\gamma_\alpha$. **Middle:** Setting B with a known channel, plotted against $1/\Gamma_{M,T}$. **Right:** Setting B with an unknown Huber channel, plotted against $1/\Gamma_{\alpha,T}$. The empirical curves reproduce the predicted dependence on the relevant separation quantities up to approximately constant multiplicative gaps.

Setting A: final-answer contamination. *We instantiate the final-answer hard construction with* $\beta = 8\epsilon$ *and scan* $\alpha \in \{0, 0.05, 0.1, 0.2, 0.3, 0.4, 0.45, 0.48\}$. *For every* α, *the sample-size grid contains* 64 *geometrically spaced values between* 1 *and* $\left\lceil 1.25\, n^{\mathrm{FA}}_{\mathrm{upper}}(\alpha) \right\rceil$. *We compare the empirical critical sample size with* $\frac{1}{\gamma_\alpha}, \gamma_\alpha = 1 - 2\sqrt{\alpha(1-\alpha)}$. *The estimated threshold increases from* 2,108 *at* $\alpha = 0$ *to* 1,588,922 *at* $\alpha = 0.48$ *and is nearly linear in* $1/\gamma_\alpha$. *The residual discrepancy is approximately a constant multiplicative factor.*

Setting B with a known symmetric channel. *The full-CoT hard construction uses* $\beta = 16\epsilon$. *For a vocabulary of size* K, *we consider*

$$M_\alpha(a, z) = (1-\alpha)\mathbf{1}\{a = z\} + \frac{\alpha}{K-1}\mathbf{1}\{a \neq z\}. \tag{35}$$

For distinct $a, b \in \Sigma$, *the Bhattacharyya overlap between the corresponding rows is*

$$\rho_M = 2\sqrt{\frac{\alpha(1-\alpha)}{K-1}} + \frac{K-2}{K-1}\alpha. \tag{36}$$

The primary horizon scan uses $K = 4, \quad \alpha = 0.4, \quad T \in \{1, 2, 4, 8, 16, 32, 64, 128\}$, *for which* $\rho_M \approx 0.832352$. *We compare* $\widehat{n}$ *against* $\frac{1}{\Gamma_{M,T}}, \Gamma_{M,T} = 1 - \rho_M^T$. *The empirical threshold decreases from* 5,864 *at* $T = 1$ *to* 1,588 *at* $T = 64$ *and then saturates, reproducing the predicted linear-gain and saturation regimes.*

As a consistency check, we also use $K = 2, \quad \alpha = 0.4, \quad T \in \{1, 2, 4, 8, 16, 32\}$. *For the binary symmetric channel,* $\rho_M = 2\sqrt{\alpha(1-\alpha)} \approx 0.979796$, *and hence* $\Gamma_{M,1} = 1 - \rho_M = \gamma_\alpha$. *This identity is verified numerically to six decimal places. Since the Setting A and Setting B hard constructions use different values of* β, *their raw critical sample sizes need not coincide.*

Exact aggregate sampling. *For the symmetric channel, we sample aggregate counts using the exact conditional decomposition*

$$N_{\neq a} \sim \mathrm{Binomial}(mT, \alpha),$$

followed by

$$N_b \mid N_{\neq a} \sim \mathrm{Binomial}\left(N_{\neq a}, \frac{1}{K-1}\right).$$

This is an exact decomposition of the corresponding multinomial distribution. For coordinates whose clean token is b, *we apply the symmetry-preserving exchange* $a \leftrightarrow b$.

Setting B with unknown Huber channels. *For a general channel row, let*

$$p_a = M_C(a, a), \qquad p_b = M_C(a, b).$$

We use the exact sequential multinomial decomposition

$$N_a \sim \text{Binomial}(mT, p_a),$$

followed conditionally by

$$N_b \mid N_a \sim \text{Binomial}\left(mT - N_a, \frac{p_b}{1 - p_a}\right).$$

We consider four channel families: 1. Uniform channel: the symmetric channel in equation 35; 2. Swap channel: a selected token pair is exchanged deterministically by C, while the remaining rows are generated from a fixed seed; 3. Global pair-swap channel: all vocabulary symbols are partitioned into disjoint pairs and exchanged; 4. Asymmetric channel: a non-symmetric random channel generated with concentration parameter 8.0.

To test uniform row separation and population calibration, we use $K = 8, \alpha \in \{0, 0.05, 0.10, \ldots, 0.45, 0.48\}$, *and generate* 200 *asymmetric corruption matrices for every value of* α, *together with the corresponding swap channel.*

To test the least-favorable sample-complexity factor, we use the swap channel and scan $\alpha \in \{0.2, 0.4, 0.45\}, T \in \{1, 2, 8, 16\}$. *At the common stress-test point* $\alpha = 0.4, T = 8$, *we compare the uniform, asymmetric, and swap channels.*

Figure 3 shows that the empirical critical sample sizes follow the predicted factors $1/\gamma_\alpha$, $1/\Gamma_{M,T}$, *and* $1/\Gamma_{\alpha,T}$, *respectively.*

D.2 Part II: Language-Model Fine-Tuning Experiments

D.2.1 Main Configuration

Datasets and parameters. *Mathematical reasoning examples are filtered from* `SFT_510K` *to short elementary-arithmetic word problems that contain no LaTeX, contain at most 280 characters, and admit a pure-integer answer: of* $11,280$ *scanned examples,* $4,162$ *pass, split into* $3,862$ *training and* 300 *test examples. Medical reasoning examples are filtered from* `medical-o1-reasoning-SFT` *to multiple-choice questions with an automatically extractable single-letter answer: of* $44,591$ *scanned records,* $2,983$ *qualify, split into* $2,683$ *training and* 300 *test examples. For Setting B, the 128 most frequent extracted values define* $\mathcal{V}_{\text{state}} = \{\texttt{<VAL_0>}, \ldots, \texttt{<VAL_127>}\}$; *these tokens are added to the model vocabulary, with input-embedding and output-head parameters initialized from the mean representations of their corresponding subword realizations. All models are adapted using LoRA with rank* 16, *scaling parameter* 32, *and dropout* 0.05. *We train for four epochs using AdamW, a cosine learning-rate schedule, five warmup steps, and bfloat16 arithmetic.*

Training objectives in Setting B. *For an input* x, *let* $s^{\text{cl}}_{<t}$ *denote the clean teacher-forcing prefix and* z_t *the observed, possibly corrupted, state token at position* t. *The channel-agnostic model minimizes the standard tokenwise cross-entropy*

$$\mathcal{L}_{\text{agn}}(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log p_\theta(z_t \mid x, s^{\text{cl}}_{<t}),$$

whereas, when the channel M *is known, we minimize*

$$\mathcal{L}_M(\theta) = -\frac{1}{T}\sum_{t=1}^{T} \log\left[\sum_{a\in\Sigma} p_\theta(a \mid x, s^{\text{cl}}_{<t}) M(a, z_t)\right],$$

where $M(a, z) = \Pr(Z = z \mid S^{\text{cl}} = a)$. *Both objectives use the same clean-prefix teacher forcing and differ only in their treatment of the corrupted target: the known-channel objective passes the predicted clean-token distribution through* M *before evaluating the observed token, corresponding to standard forward loss correction (Patrini et al., 2017).*

Both Setting B objectives are trained with teacher forcing on the clean state prefix and differ only in their treatment of the corrupted target (Appendix D.2.1); evaluation is free-running, decoding the state sequence greedily from the prompt alone. Clean-prefix conditioning gives the learner oracle access to on-trajectory contexts that the observation model of Section 3 does not provide, so these runs are a favorable surrogate for the finite-class estimators.

Interpreting the clean baselines. *Absolute accuracies in Settings A and B are not directly comparable, even when $\alpha = 0$, because the two protocols use different prediction interfaces. Setting B requires the model to emit only a sequence of quantized intermediate states, without visible scratch space or free-form reasoning, and each state is predicted over a 128-token vocabulary; Setting A instead permits tool-integrated text-and-code reasoning before producing the final answer. We therefore do not interpret the lower clean accuracy in Setting B as evidence of corruption-induced degradation or an optimization failure, nor as a controlled comparison between CoT and final-answer supervision.*

Corruption protocols. *Setting A: for every seed, we sample n examples from the clean training pool. Each answer independently enters the contamination component with probability α; if selected, it is replaced by the answer of another randomly sampled training problem, otherwise the clean answer is retained, and the prompt is never modified. Setting B: each clean reasoning trace is first converted into a length-T state sequence, and the state targets are then corrupted independently across positions. For the known symmetric channel, a state token is retained with probability $1 - \alpha$ and otherwise replaced uniformly by one of the other tokens. For the unknown-channel experiment, the corruption component is instead a global fixed-point-free pair-swap permutation, with the pairing fixed by a channel seed that is not given to the channel-agnostic learner. Evaluation always uses clean labels.*

D.2.2 Ablations

Model-scale ablation *As shown in Figure 2 (Left and Center-left),the accuracy declines with α for Qwen3-1.7B, Qwen3-4B, and Qwen3-8B on both math and medical data, with the larger model uniformly more accurate at every α. While larger models degrade more slowly in relative terms on open-ended math tasks, this robustness is domain-dependent and does not cleanly replicate on multiple-choice medical data. Despite these relative variations, the absolute downward trend persists across all evaluated sample sizes. The theory's qualitative prediction – accuracy falls as contamination rises, with the effective sample size scaling as $n\gamma_\alpha$ – therefore holds consistently across three model scales within the same architecture.*

Cross-architecture ablation. *To verify that our observations are not Qwen-specific artifacts, we repeat the Setting A math protocol on three additional architectures i.e. Phi-3.5-mini, Llama-3.2-3B, and Mistral-7B—at $\alpha \in \{0, 0.2, 0.45\}$ and $n \in \{1000, 3660\}$, using Qwen3-4B as a reference baseline. Figure 4 illustrates the test accuracy against contamination level α for all four models. Despite substantial differences in their absolute performance and variance (notably the high variance for Phi-3.5-mini at $n = 3660$), every architecture exhibits a monotonic decline in accuracy from $\alpha = 0$ to $\alpha = 0.45$ across both sample sizes. This confirms that the qualitative degradation-under-contamination prediction holds consistently across different model architectures as well as scales.*

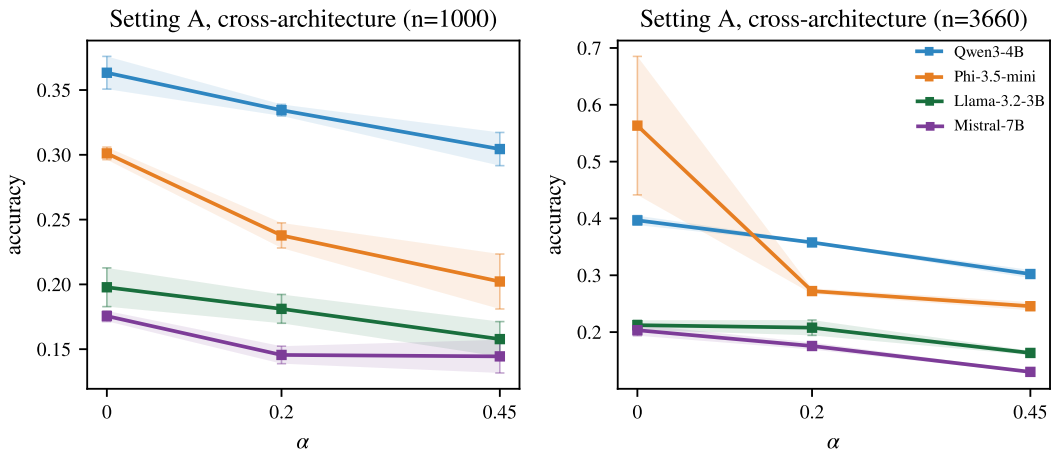

Figure 4: Cross-architecture robustness check in Setting A (math). The plots present clean test accuracy against contamination level α for Qwen3-4B, Phi-3.5-mini, Llama-3.2-3B, and Mistral-7B at sample sizes $n = 1000$ (left) and $n = 3660$ (right). Square markers represent the mean accuracy across three random seeds, while shaded bands and error bars denote ± 1 standard error.