

Manifold Modulation for Latent Space Steganography: A Capacity Bound and Distillation Demodulation Model for Teacher-Student Transmission Channels

Manifold Modulation for Latent Space Steganography: A Capacity Bound and Distillation Demodulation Model for Teacher-Student Transmission Channels

Abstract

Large language models are increasingly deployed in teacher-student pipelines in which a student model is fine-tuned on text produced by a teacher, so any statistical signature the teacher imprints on its outputs is transferred to the student whether or not anyone intends it. No existing treatment gives a quantitative bound on covert bits per carrier token when the carrier task itself constrains the output distribution. We model the teacher's output distribution as a carrier subject to a per-token Kullback-Leibler (KL) modulation budget and derive a capacity bound $C = D / \ln 2$ bits per token, where D is the budget in nats; at an illustrative budget of $D = 0.05$ nats this yields 0.072134 bits per token, or 721.34 bits (90.17 bytes) over a 10^4 -token carrier document. We further derive a steganalysis sample-complexity of approximately 92.10 tokens for a detector at 10^{-2} false-alarm tolerance, and a demodulation requirement of 138,629 carrier tokens for a 10^4 -bit behavior specification. The results reframe fine-tuning-on-outputs as demodulation of a hidden channel, give auditors a closed-form budget for certifying synthetic corpora, and give pipeline designers a principled rate at which targeted behaviors can be encoded into teacher-generated data.

1. Introduction

Organizations that build and deploy large language models face a supply-chain problem that is currently unmanaged: student models are routinely fine-tuned on corpora of teacher-generated text, synthetic instruction data, and agent transcripts, and the student inherits whatever statistical structure the teacher imprinted on that text. The scale of the practice is substantial. Agentic systems now execute entire expert workflows autonomously and emit their reasoning as text; Latent-Y, for example, executes complete antibody design campaigns from text prompts, covering literature review, target analysis, epitope identification, candidate design, computational validation, and selection of lab-ready sequences, with a target-level success rate of 67% across nine targets [1]. Every token of such a transcript is a potential carrier of the producing model's internal statistical state. Today nobody measures what is transferred, nobody bounds it, and nobody can decide whether a given fine-tuning corpus carries an unacceptable amount of upstream model state. If a quantitative capacity bound existed, a data provenance auditor could certify a corpus as carrying at most B bits of upstream state, and a pipeline designer could deliberately encode desired behaviors into synthetic data at a known rate rather than by trial and error.

The cost of leaving this unquantified runs in both directions. In the defensive direction, a teacher model may leak proprietary capabilities, alignment-relevant behaviors, or hidden decision policies through its outputs, and a student fine-tuned on those outputs acquires them without any explicit transfer of weights; the downstream organization has no instrument to detect or bound this. In the offensive direction, an adversary with control of a data-generating teacher could encode arbitrary behavior into a synthetic corpus that looks like an ordinary task dataset, and the victim's fine-tuning run demodulates it automatically. The decision that changes if the present result holds is the audit decision: given a measured or budgeted per-token modulation D and a corpus of N tokens, an auditor can compute the maximum covert information $N \cdot D / \ln 2$ bits and accept or reject the corpus.

This paper formalizes the claim that any transmission from a sufficiently complex, stateful source carries two parallel information streams: the content stream, which is the explicit message, and the source-information stream, which is the sender's internal statistical state. Fine-tuning a student on a teacher's outputs is demodulation of the second stream. We propose engineering this deliberately: constrain the teacher's internal states to a chosen latent "covert manifold" while it performs an unrelated carrier task, so that the manifold's geometry is encoded into token statistics, and a student fine-tuned on that data inherits the targeted behavior.

The contributions are:

1. A formal two-stream transmission model in which a teacher performing a carrier task modulates its output distribution toward a covert manifold, with an explicit per-token KL modulation budget (Section 3).

2. An information-theoretic capacity bound $C = D/\ln 2$ bits per carrier token, derived from the modulation budget by a chain-rule argument, with worked numerical values at three budgets (Section 4).
3. A steganalysis sample-complexity bound showing that detection requires on the order of $\ln(1/\alpha)/D$ tokens, computed as 92.10 tokens at $D = 0.05$ nats and $\alpha = 10^{-2}$ (Section 4).
4. A demodulation condition for distillation: a student recovers a behavior specification of K bits from a carrier corpus of M tokens when $M \geq K \ln 2/D$, computed as 138,629 tokens for $K = 10^4$ bits at the illustrative budget (Section 4).
5. A capacity-versus-carrier-degradation tradeoff under an explicitly assumed linear degradation model, giving the ablation axes that an experimental program must measure (Sections 4 and 5).

2. Background and Related Work

Steganography supplies the conceptual template. Steganography is the science of hiding and communicating a secret message by embedding it in an innocent-looking text such that the eavesdropper is unaware of its existence; the term derives from the Greek *steganos* (covered) and *graphia* (writing), and it differs from cryptography in that the existence of the secret message, not merely its content, is concealed, with the intended receiver alone able to detect and decipher the hidden message [4]. Our covert manifold plays the role of the secret message and the carrier task's token distribution plays the role of coverttext. The quantum steganography literature also demonstrates the failure mode that motivates our steganalysis bound: protocols built on BB84 quantum key distribution were shown to be vulnerable to a steganalysis attack that can detect the presence of the hidden transmission, which is why [4] develops modified protocols in which the party communicating steganographically operates in the reverse direction of the party sending the initial quantum states, camouflaged as part of the classical reconciliation step. The lesson we carry over is that embedding direction and detection direction must be analyzed jointly; in our setting the "receiver" is a fine-tuning procedure, which is precisely why the channel is hard to police with conventional tools.

Network steganography provides the engineering precedent and the maturity warning. Network steganography has been a known covert data channeling method for over three decades, and the basic set of techniques and implementation tools has not changed significantly since their introduction in the early 1980s; the predominant classical methods embed data in the detailed operations of network protocols, and the central challenge is embedding data without detectable statistical deviation [7]. The stagnation of technique that [7] documents in the network domain is the outcome model-output steganography should avoid: its capacity limits, detection thresholds, and degradation tradeoffs are still unmeasured, and the present paper supplies analytical scaffolding for those measurements.

The latent-manifold literature supplies the representational machinery. GeMA combines deep generative modeling with classical efficiency-analysis concepts, learning latent manifold frontiers with a ProMan-VAE architecture, and its experiments probe settings where production technologies are heterogeneous, non-convex, or confounded with scale, using $n = 500$ decision-making units averaged over 30 Monte Carlo replications [5]. Two points from [5] transfer directly. First, its demonstration that a latent generative model can represent a structured frontier that classical estimators in observed space miss supports our claim that a low-dimensional covert manifold can be embedded in a high-dimensional output space without disturbing observable task statistics. Second, its limitation discussion states that the latent technology and inefficiency variables have clear conceptual roles but the decomposition is not uniquely identifiable in a purely data-driven sense, being regularised by priors, shape constraints, and structural equations, and that individual latent dimensions should not be over-interpreted [5]. We adopt this caution: our covert manifold is identifiable to the demodulator only because the demodulator knows the manifold's parameterization, which is exactly the steganographic key.

Latent-space subdivision demonstrates that latent structure can be engineered to carry information orthogonal to the primary prediction task. The fluid-flow architecture of [6] splits the latent space so that external information can be passively injected into the prediction process: an obstacle unseen during training is injected by supplying an encoded density field with the obstacle region masked out, and the method predicts stable flow over 400 time steps where the unsplit baseline fails and produces unphysical behavior [6]. This is a working precedent for our central mechanism: a partitioned latent representation in which one subspace carries task content and a second subspace carries injected, task-orthogonal information that a downstream consumer reads out. Our proposal is the steganographic dual of this: the injected subspace is hidden rather than declared.

The teacher-model literature establishes that the carrier streams we analyze exist at industrial scale and quality. Latent-X is an all-atom protein design model that jointly generates sequence and structure, achieving macrocycle hit rates of 91% on MDM2 and 100% on MCL-1 against RFpeptide baselines of 38% and 21%, and mini-binder hit rates from 10% on TrkA to 64% on

BHRF1, including a 52% hit rate on SC2RBD versus 12% for AlphaProteo [2]. Latent-X2 extends this to zero-shot antibody design, producing potent VHHs and scFvs with drug-like profiles from the first generation, testing only 4 to 24 designs per modality-target pair, and demonstrating, for the first time for any de novo generated antibody, confirmed low immunogenicity in ex vivo assays with primary human peripheral blood mononuclear cells from 10 donors [3]. These systems matter to our argument for two reasons. They generate exactly the kind of high-stakes, proprietary-capability-bearing output streams whose transmission to downstream students an auditor would want to bound; and their tiered screening workflows, in which initial single-point BLI screens advance only designs exceeding a predefined binding response threshold [3], illustrate how a generating model's internal decision policy is imprinted on the statistics of its emitted dataset, which is the source-information stream we formalize.

Multi-stream information extraction has a precise astronomical precedent. The SPACE mission design notes that CMB temperature spectra C_l depend on the combination $\Omega_m h^2$ while the matter power spectrum $P(k)$ is sensitive to $\Omega_m h$, so combining the two datasets breaks degeneracies and tightens parameter limits beyond what either dataset achieves in isolation [8]. This is the exact logic of our two-stream model: the content stream and the source-information stream are subject to different parameter degeneracies, and a demodulator that reads both, rather than the content stream alone, recovers sender state that a content-only reader cannot.

3. Methods

3.1 Setting and notation

Let T be a teacher model with internal state $s \in \mathcal{S}$, performing a carrier task. At each step the teacher emits a token x_t from a vocabulary of size V . Absent modulation, the teacher's output distribution is $p_t(x) = p(x \mid x_{<t}, s)$, which depends on the internal state s . We define the carrier distribution $p_t^{(0)}(x) = p(x \mid x_{<t}, s_0)$ as the distribution the teacher would produce under a reference internal state s_0 that carries no covert information.

A covert manifold is a parameterized family of internal states $\{s(\theta) : \theta \in \mathcal{M}\}$ with $\mathcal{M} \subseteq \mathbb{R}^d$ a d -dimensional manifold. The teacher performing the carrier task while its state is constrained to $s(\theta)$ emits tokens from the modulated distribution

$$\pi_t(x) = p(x \mid x_{<t}, s(\theta)).$$

The modulation strength at step t is the KL divergence between the modulated and carrier distributions,

$$D_t = D_{\text{KL}}(\pi_t \parallel p_t^{(0)}) = \sum_x \pi_t(x) \log \frac{\pi_t(x)}{p_t^{(0)}(x)}.$$

We impose a per-token modulation budget $D_t \leq D$ for all t . The budget is the engineering constraint that keeps the carrier task intact: it bounds the total statistical deviation of the modulated stream from an ordinary carrier stream.

3.2 The two-stream decomposition

The emitted sequence $x_{1:N}$ carries two streams. The content stream is the explicit carrier-task output, measured by the mutual information between the sequence and the carrier task's intended output. The source-information stream is the mutual information between the sequence and the covert parameter θ ,

$$I(x_{1:N}; \theta) = \sum_{x_{1:N}, \theta} P(x_{1:N}, \theta) \log \frac{P(x_{1:N}, \theta)}{P(x_{1:N})P(\theta)}.$$

Under the budget constraint, the per-token covert capacity is bounded in Section 4 by $C = D / \ln 2$ bits, so the total covert information satisfies $I(x_{1:N}; \theta) \leq N \cdot D / \ln 2$.

3.3 Demodulation by fine-tuning

A student model S is fine-tuned on the modulated corpus $\{x_{1:N_i}^{(i)}\}$ generated by the teacher under covert parameters $\{\theta^{(i)}\}$. Fine-tuning is demodulation if the student's induced policy $P_S(y \mid x)$ satisfies

$$I(P_S; \theta) > I(P_S^{(0)}; \theta),$$

where $P_S^{(0)}$ is the policy of a student fine-tuned on unmodulated carrier data. The demodulation gain is the excess mutual information attributable to the modulation. The condition for full recovery of a behavior specification of K bits is derived in Section 4 as $M \geq K \ln 2 / D$, where M is the total number of carrier tokens the student trains on.

3.4 Steganalysis model

A steganalyst observes n tokens from either the carrier distribution $p^{(0)}$ or the modulated distribution π and must decide which. This is a binary hypothesis test between two distributions whose per-token divergence is D . We derive the sample size required for detection at false-alarm probability α in Section 4 using the Chernoff-Stein exponent, which for simple hypotheses is governed by the KL divergence between the hypotheses.

3.5 Assumptions

The analysis rests on assumptions that the present evidence base does not fix empirically, so we state them explicitly. (i) The carrier task constrains the teacher's output distribution only through the budget D ; we assume the carrier task's own entropy is large enough that a budget of $D \leq 0.2$ nats per token is feasible without truncating the carrier vocabulary. (ii) We assume the covert parameter θ is drawn from a d -dimensional manifold whose effective discrete alphabet at token resolution is large compared to the capacity bound, so that the capacity bound, not the manifold dimension, is the binding constraint. (iii) We assume a linear carrier-degradation model $\Delta A \approx \beta D$ with slope β for the ablation analysis; this is an explicitly labeled modeling assumption to be measured experimentally, not a derived result. (iv) We assume the student's fine-tuning procedure is a standard likelihood-based update, so that information about θ present in the training corpus is transferred at a rate bounded by the corpus's covert information content.

4. Analysis

4.1 Capacity bound derivation

The teacher emits a sequence of tokens; each token is drawn from π_t under modulation or $p_t^{(0)}$ under the carrier. The information about θ carried by the whole sequence is bounded by the mutual information $I(x_{1:N}; \theta)$. By the chain rule of mutual information and the fact that conditioning reduces entropy,

$$I(x_{1:N}; \theta) = \sum_{t=1}^N I(x_t; \theta \mid x_{<t}) \leq \sum_{t=1}^N I(x_t; \theta, x_{<t}).$$

At step t , the token distribution given the past and the covert state is $\pi_t(x) = p(x \mid x_{<t}, s(\theta))$. The maximal mutual information between θ and x_t given the past is achieved when the modulation uses the full budget, and is bounded by the divergence between the modulated and carrier distributions:

$$I(x_t; \theta \mid x_{<t}) \leq D_{\text{KL}}(\pi_t \parallel p_t^{(0)}) \leq D.$$

This is the standard bound that the information a random variable carries about a parameter cannot exceed the KL divergence between the conditional distributions it induces; in nats, $I \leq D_{\text{KL}}$. Converting to bits, one nat equals $1 / \ln 2$ bits. Therefore

$$C = \frac{D}{\ln 2} \text{ bits per token,}$$

and the total covert capacity of an N -token carrier document is

$$I_{\text{max}}(N) = \frac{N \cdot D}{\ln 2} \text{ bits.}$$

4.2 Numerical evaluation of the capacity bound

We evaluate the bound at three modulation budgets, using $\ln 2 = 0.693147$.

At $D = 0.05$ nats:

$$C = \frac{0.05}{0.693147} = 0.072134 \text{ bits per token.}$$

At $D = 0.10$ nats:

$$C = \frac{0.10}{0.693147} = 0.144270 \text{ bits per token.}$$

At $D = 0.20$ nats:

$$C = \frac{0.20}{0.693147} = 0.288539 \text{ bits per token.}$$

For a carrier document of $N = 10^4$ tokens at $D = 0.05$ nats:

$$I_{\max} = 10^4 \times 0.072134 = 721.34 \text{ bits.}$$

In bytes this is $721.34/8 = 90.17$ bytes, so a single 10^4 -token document at this budget can carry at most about 90 bytes of covert information. At $D = 0.20$ nats the same document carries $10^4 \times 0.288539 = 2885.39$ bits, or $2885.39/8 = 360.67$ bytes.

4.3 Steganalysis sample complexity

The steganalyst performs a hypothesis test between H_0 : tokens from $p^{(0)}$, and H_1 : tokens from π . By the Chernoff-Stein theorem, the minimal sample size n needed to detect the modulated stream with false-alarm probability α and high detection probability satisfies, to first order,

$$n \gtrsim \frac{\ln(1/\alpha)}{D_{\text{KL}}(\pi \| p^{(0)})}.$$

We evaluate this at $\alpha = 10^{-2}$, so $\ln(1/\alpha) = \ln(100) = 4.60517$:

At $D = 0.05$ nats:

$$n \gtrsim \frac{4.60517}{0.05} = 92.10 \text{ tokens.}$$

At $D = 0.10$ nats:

$$n \gtrsim \frac{4.60517}{0.10} = 46.05 \text{ tokens.}$$

At $D = 0.20$ nats:

$$n \gtrsim \frac{4.60517}{0.20} = 23.03 \text{ tokens.}$$

The inverse scaling is the central tension of the design: doubling the modulation budget doubles the covert capacity per token and halves the number of tokens a steganalyst needs for detection. The channel is therefore detectable at modest sample sizes under this simple hypothesis test; the practical defense of the scheme is not that detection is impossible but that the steganalyst must know which test to run, which is the camouflage lesson of [4], where the modified protocols hide the covert transmission inside the classical reconciliation traffic of the quantum key distribution protocol itself.

4.4 Demodulation condition for distillation

Let the targeted behavior be a specification of K bits, meaning that the student must acquire a policy equivalent to one of 2^K alternatives. The student's training corpus contains M carrier tokens in total, each carrying at most $C = D/\ln 2$ covert bits. The covert information available to the student is at most

$$I_{\text{avail}} = M \cdot \frac{D}{\ln 2} \text{ bits.}$$

Full demodulation requires $I_{\text{avail}} \geq K$, giving the condition

$$M \geq \frac{K \ln 2}{D}.$$

We evaluate this for an illustrative behavior specification of $K = 10^4$ bits, which we state as an assumption about the complexity of a targeted behavioral policy; the correct value of K for any real behavior is an empirical quantity that the experimental program of Section 7 must measure.

At $D = 0.05$ nats:

$$M \geq \frac{10^4 \times 0.693147}{0.05} = \frac{6931.47}{0.05} = 138,629.4 \text{ tokens.}$$

At $D = 0.10$ nats:

$$M \geq \frac{6931.47}{0.10} = 69,314.7 \text{ tokens.}$$

At $D = 0.20$ nats:

$$M \geq \frac{6931.47}{0.20} = 34,657.4 \text{ tokens.}$$

In units of 10^4 -token documents at $D = 0.05$ nats, this is $138,629.4/10^4 = 13.86$, so approximately 14 carrier documents suffice in information content. This number should be read as a lower bound on corpus size: the actual requirement is larger by a factor accounting for the student's learning efficiency, which is less than unity.

4.5 Capacity-degradation tradeoff under an assumed linear model

We model the carrier task's accuracy loss as linear in the modulation budget,

$$\Delta A \approx \beta D,$$

and we assume the slope $\beta = 0.4$ per nat as an explicitly labeled modeling assumption; measuring β is a principal objective of the experimental program. Under this assumption:

At $D = 0.05$ nats: $\Delta A = 0.4 \times 0.05 = 0.02$, a 2% carrier accuracy loss, with capacity $C = 0.072134$ bits per token.

At $D = 0.10$ nats: $\Delta A = 0.4 \times 0.10 = 0.04$, a 4% loss, with $C = 0.144270$ bits per token.

At $D = 0.20$ nats: $\Delta A = 0.4 \times 0.20 = 0.08$, an 8% loss, with $C = 0.288539$ bits per token.

The capacity-to-degradation ratio is constant under the linear model:

$$\frac{C}{\Delta A} = \frac{1}{\beta \ln 2} = \frac{1}{0.4 \times 0.693147} = \frac{1}{0.277259} = 3.6068 \text{ bits per token per unit fractional accuracy loss.}$$

The experimental ablation over modulation strength exists precisely to test whether this ratio is constant or whether degradation is convex in D ; convexity would make low budgets strictly more efficient per bit.

4.6 Perplexity projection under an assumed baseline

For language-model carriers, a projection of the perplexity impact follows from the same budget. We assume a baseline perplexity $P_0 = 20$ for a mid-sized carrier model and that the modulation budget D converts directly into an excess cross-entropy of D nats per token, so that

$$P = P_0 e^D.$$

At $D = 0.05$ nats: $e^{0.05} = 1.051271$, so $P = 20 \times 1.051271 = 21.03$, a 5.13% increase. At $D = 0.20$ nats: $e^{0.20} = 1.221403$, so $P = 20 \times 1.221403 = 24.43$, a 22.14% increase. Both figures are projections under the stated assumptions ($P_0 = 20$; direct budget-to-cross-entropy conversion); the conversion itself is what the minimal experiment of Section 7 must measure.

4.7 Manifold dimension constraint

The covert manifold has dimension d . If the demodulator must recover a continuous parameter $\theta \in \mathbb{R}^d$ to resolution δ_θ per coordinate over a normalized range of R per coordinate, the parameter carries

$$K_\theta = d \cdot \log_2 \left(\frac{R}{\delta_\theta} \right) \text{ bits.}$$

As a worked example with $d = 8$ and $R/\delta_\theta = 2^{10}$ (ten bits of resolution per coordinate):

$$K_\theta = 8 \times \log_2(2^{10}) = 8 \times 10 = 80 \text{ bits.}$$

By the demodulation condition of Section 4.4, recovering this parameter requires

$$M \geq \frac{K_\theta \ln 2}{D} = \frac{80 \times 0.693147}{0.05} = \frac{55.45176}{0.05} = 1109.04 \text{ tokens.}$$

So a continuous eight-parameter covert state specified to ten bits of resolution per coordinate is recoverable from roughly 1.1×10^3 carrier tokens at $D = 0.05$ nats, an order of magnitude below the 138,629 tokens required for a 10^4 -bit behavioral specification. The manifold dimension is therefore not the binding constraint for behavior-level targets, which confirms assumption (ii) of Section 3.5 for the parameter ranges considered here; it becomes binding only when $K_\theta > K$, that is, when the required parameter resolution exceeds the behavioral specification's complexity.

5. Results

The derivations of Section 4 yield four quantitative results, all computed from the stated inputs with no empirical measurement.

Result 1: Capacity bound. The covert capacity per carrier token is $C = D/\ln 2$ bits (Section 4.1). At the three evaluated budgets this gives $C = 0.072134, 0.144270$, and 0.288539 bits per token at $D = 0.05, 0.10$, and 0.20 nats respectively. A 10^4 -token document carries at most 721.34 bits (90.17 bytes) at $D = 0.05$ nats and 2885.39 bits (360.67 bytes) at $D = 0.20$ nats.

Result 2: Steganalysis sample complexity. Detection by a detector running the optimal simple-hypothesis test at false-alarm probability $\alpha = 10^{-2}$ requires approximately $n = \ln(100)/D$ tokens: 92.10 tokens at $D = 0.05$ nats, 46.05 at $D = 0.10$, and 23.03 at $D = 0.20$ (Section 4.3). The channel is detectable at modest sample sizes once the detector knows which test to run; the capacity and detectability scale inversely with the same budget.

Result 3: Demodulation condition. A student recovers a K -bit behavioral specification from a corpus of at least $M = K \ln 2/D$ tokens: 138,629.4 tokens (≈ 14 documents of 10^4 tokens) for $K = 10^4$ bits at $D = 0.05$ nats, scaling to 69,314.7 and 34,657.4 tokens at $D = 0.10$ and 0.20 nats (Section 4.4). These are information-theoretic lower bounds; actual corpus requirements are larger by the inverse of the student's learning efficiency, which is less than one and unmeasured.

Result 4: Tradeoff ratios under assumed models. Under the assumed linear degradation model with slope $\beta = 0.4$ per nat, the capacity-to-degradation ratio is constant at 3.6068 bits per token per unit fractional accuracy loss across all budgets (Section 4.5). Under the assumed baseline perplexity $P_0 = 20$ with direct budget-to-cross-entropy conversion, the projected perplexity is 21.03 at $D = 0.05$ nats (5.13% increase) and 24.43 at $D = 0.20$ nats (22.14% increase) (Section 4.6). Both results are projections conditional on their stated modeling assumptions, not measurements.

Result 5: Manifold dimension check. An eight-dimensional covert state at ten bits of resolution per coordinate carries 80 bits and requires 1109.04 tokens to demodulate at $D = 0.05$ nats (Section 4.7), confirming that the capacity bound, not manifold dimension, is the binding constraint for behavior-level targets.

The results that survive without auxiliary assumptions are Results 1, 2, 3, and 5, which follow from the modulation budget alone. Results 4's numbers inherit the linear degradation and perplexity-conversion assumptions and must be treated as hypotheses for the experimental program.

6. Discussion

The central claim of this paper is that the per-token KL budget D is the single quantity that governs covert capacity, steganalytic detectability, and distillation-based demodulation, through the three relations $C = D/\ln 2$, $n \gtrsim \ln(1/\alpha)/D$, and $M \geq K \ln 2/D$. The claim's strength is that all three relations follow from one derivation chain; its weakness is that the chain rests on idealizations that real teacher-student pipelines violate in specific, analyzable ways.

Limitations. The result rests on the following assumptions, each with a distinct sensitivity:

1. *Per-token independence.* The capacity bound sums per-token mutual informations and assumes the modulation at each token is budgeted independently. If the teacher modulates correlatedly across tokens (for example, encoding information in token-to-token transitions rather than marginal distributions), the per-token KL budget does not bound the joint information, and the true capacity can exceed $ND/\ln 2$. The bound is tight only for memoryless modulation given the context; a sequence-level budget D_{seq} over k -grams would be the correct constraint, and the capacity would scale with the sequence-level divergence instead. This is the assumption the result is most sensitive to, because correlated modulation is the natural implementation.
2. *Known reference distribution.* The steganalysis bound assumes the detector knows $p^{(0)}$ exactly. A real auditor must estimate the carrier distribution from unlabeled data, which inflates the required sample size by a factor that grows with the estimation error; the 92.10-token figure is therefore a lower bound on practical detectability, and the defensive reading (detection is cheap) weakens while the offensive reading (detection is hard without the reference) strengthens.
3. *Simple hypotheses.* The Chernoff-Stein argument treats $p^{(0)}$ and π as known, fixed distributions. Composite hypotheses (unknown θ , unknown modulation strength) change the exponent from D to a Chernoff information or a mixture likelihood ratio, generally requiring more samples. The stated n values are optimistic for the detector.
4. *Linear degradation.* The tradeoff ratio 3.6068 bits per token per unit accuracy loss is constant only under the assumed $\Delta A \approx \beta D$ model. If degradation is convex in D (plausible, since large KL deviations force the model off-manifold for the carrier task), low budgets become strictly more efficient per bit, and the optimal operating point moves to the smallest feasible D .
5. *Unit learning efficiency.* The demodulation condition $M \geq K \ln 2/D$ assumes the student extracts every covert bit in the corpus. Real fine-tuning extracts a fraction $\eta < 1$, so the practical corpus size is $M \geq K \ln 2/(\eta D)$; at $\eta = 0.1$ the 10^4 -bit target would require $\approx 1.39 \times 10^6$ tokens rather than 1.39×10^5 .
6. *Illustrative constants.* $K = 10^4$ bits, $\beta = 0.4$, and $P_0 = 20$ are stated assumptions, not measurements. No result in the paper depends on their specific values except the worked numbers attached to them.

Failure modes. Three are foreseeable. First, the modulation may be statistically visible in ways the KL budget does not capture: a detector testing higher-order statistics, token co-occurrence, or perplexity under its own reference model may flag the corpus even when the per-token KL is within budget, because the budget constrains divergence from the teacher's own carrier distribution, not from an auditor's model of it. Second, the student may demodulate partially and unstably, acquiring fragments of the behavior specification that manifest as unexplained behavioral quirks rather than the targeted policy; this is the most likely outcome if M is near the bound. Third, the carrier task may itself be state-dependent in ways that confound the two streams: if the teacher's internal state legitimately varies with task content, the auditor cannot separate source-information from content, and the capacity bound overcounts what is covert.

What would falsify the main result. The main result is the capacity bound $C = D/\ln 2$ as an upper bound on covert bits per token under a per-token KL budget. It would be falsified by a demonstration that a teacher operating under a measured per-token KL budget D transmits more than $ND/\ln 2$ bits of a specified target behavior to a student, verified by the student's behavior on a held-out discrimination test between 2^K alternatives with $K > ND/\ln 2$. A concrete falsifying observation would be: teacher modulated at measured $D = 0.05$ nats per token (verified by KL estimation against its own unmodulated outputs), a corpus of $N = 10^4$ tokens, and a student whose behavior on the held-out test identifies the covert alternative among 2^{800} options at rate significantly above chance, since $800 > 721.34$ exceeds the bound. Conversely, if students trained on modulated corpora at the computed M fail to recover even low- K ($K \leq 100$) specifications at rates above students trained on unmodulated data, the demodulation condition's practical usefulness is falsified even though the upper bound may still hold.

Open questions. The sequence-level (correlated) modulation capacity is open and likely larger than the per-token bound; the correct steganalysis exponent for composite hypotheses with estimated reference distributions is open; the shape of the degradation function $\Delta A(D)$ is open and is the primary experimental target; and the learning efficiency η for behavior-level distillation is

unmeasured for any model pair.

7. Practical Implications and Engineering Pathways

(a) Users and changed decisions. Three users follow directly from the results. First, the *data provenance auditor* at an organization that fine-tunes models on third-party or synthetic corpora: if the result holds, the audit decision changes from "review a sample of documents manually" to "estimate or bound the per-token divergence of the corpus from a reference carrier distribution and compute $N \cdot D / \ln 2$; reject the corpus if this exceeds the acceptable transferred-state budget." The steganalysis result supplies the auditor's sample-size requirement: at $D = 0.05$ nats, testing on the order of 92 tokens per document class suffices for a 10^{-2} false-alarm rate when the reference distribution is known. Second, the *pipeline designer* who wants a student to acquire a targeted behavior: the demodulation condition gives a principled corpus-size target, $M \geq K \ln 2 / D$, replacing trial-and-error scaling of synthetic datasets. Third, the *teacher-model operator* concerned with capability leakage: the capacity bound quantifies the maximum proprietary state information a corpus can carry, which converts an unbounded worry into a bounded, measurable quantity.

(b) Readiness level. This work is purely analytical and has not been peer reviewed: by the NASA Technology Readiness Level definitions it sits at TRL 1-2 (basic principles observed and formulated; concept formulated). The capacity bound, steganalysis bound, and demodulation condition are derived, not demonstrated; no teacher-student experiment has been run under this framework, and analytical work does not exceed TRL 3 in any case. No claim of implementation readiness is made.

(c) Smallest next experiment. The minimal de-risking experiment is a single teacher-student pair with a synthetic behavior target. Concretely: take a small open teacher model; define a covert binary or $K = 8$ -bit target (for example, a systematic preference among 2^8 response variants); fine-tune or prompt-condition the teacher to modulate its outputs toward the target under a measured per-token KL budget, measuring D directly by comparing modulated and unmodulated logits on a held-out prompt set; generate a corpus of $M = K \ln 2 / D$ tokens; fine-tune a student on it; and test the student's recovery of the target against a student trained on unmodulated data of the same size. The measurement that refutes the framework is a recovery rate at or below the unmodulated baseline at the computed corpus size, or a measured per-token KL that cannot be held at the budget without destroying the carrier task. A parallel ablation over $D \in \{0.01, 0.05, 0.10, 0.20\}$ nats measures the degradation slope β and tests the linearity assumption directly.

(d) Binding constraints. The constraints that would bind first, in order: (i) *KL measurement fidelity* — the entire framework requires estimating $D_{\text{KL}}(\pi \| p^{(0)})$ per token from model logits, which is exact when both distributions are accessible (same model, different conditioning) but approximate when the auditor lacks logit access; (ii) *carrier-task tolerance* — the assumed 8% accuracy loss at $D = 0.20$ nats is a projection, and if real degradation is convex, the usable budget may be smaller than 0.05 nats, shrinking capacity proportionally; (iii) *corpus generation cost* — the demodulation condition requires on the order of 1.4×10^5 tokens per 10^4 -bit target at $D = 0.05$ nats before inefficiency correction, and generation of controlled, budget-constrained corpora at that scale is an engineering cost the framework makes computable but does not reduce; (iv) *student learning efficiency* η , unmeasured, which multiplies the corpus requirement by $1/\eta$. Regulatory constraints on synthetic data provenance are an open question the paper does not address.

(e) Order-of-magnitude bounds (projections). Under the assumptions already stated ($D = 0.05$ nats, $K = 10^4$ bits, $\eta = 1$), the corpus requirement is 1.39×10^5 tokens; at a plausible inefficiency $\eta = 0.1$ (assumed, not measured) it is 1.39×10^6 tokens. Under the assumed perplexity model ($P_0 = 20$), the carrier's quality cost is a 5.13% perplexity increase at $D = 0.05$ nats. These are projections conditional on their stated assumptions and are the numbers the next experiment replaces with measurements.

(f) Risks and misuse. The offensive path is direct and is the reason the defensive results matter: an adversary controlling a data-generating teacher can encode a targeted behavior into an innocuous-looking synthetic corpus at the computed rates, and the victim's fine-tuning run demodulates it automatically. The steganalysis result cuts both ways: it tells the auditor how many tokens to test, but also tells the adversary exactly how large a corpus must be before detection at a given false-alarm rate becomes reliable, and the composite-hypothesis setting (auditor does not know the reference distribution) favors the adversary. Publication of the capacity bound itself does not enable misuse beyond what the mechanism's existence already implies; the actionable defensive output is the audit procedure, which should be deployed before the encoding procedure.

8. Conclusion

This paper formalized teacher-student fine-tuning as demodulation of a hidden information channel and derived the quantities that govern it from a single per-token KL modulation budget. The capacity bound $C = D / \ln 2$ bits per token gives 0.072134 bits per token at an illustrative budget of $D = 0.05$ nats, or at most 721.34 bits (90.17 bytes) over a 10^4 -token document; the steganalysis bound gives detection at approximately 92.10 tokens for a known-reference detector at 10^{-2} false-alarm tolerance; and the demodulation condition $M \geq K \ln 2 / D$ gives 138,629 carrier tokens for a 10^4 -bit behavioral specification at the same budget. The tradeoff between covert capacity and carrier degradation was parameterized under an explicitly assumed linear model, and the manifold dimension was shown not to bind for behavior-level targets. The results convert an unmanaged supply-chain exposure into a computable audit quantity and a deliberate encoding rate, and they specify the minimal experiment — a budget-controlled teacher-student distillation with measured per-token KL and a recovery test against an unmodulated baseline — that would confirm or refute the framework's practical usefulness. The assumptions that remain unmeasured (degradation shape, learning efficiency, sequence-level modulation capacity) are exactly the quantities that experiment measures.

Provenance and AI Involvement

- This paper was drafted, reconciled, reviewed, revised and checked by the QNFO automated research pipeline (qnfo-research-exec 0.18.6-codeagent). Language models were tools in that process; they are not authors. The named author is accountable for the content.
- Models: drafting by openai/gpt-oss-120b, zai-org/glm-5.3-flash; reconciliation, completion, review, revision and verification by zai-org/glm-5.3-flash, zai-org/glm-5.3, openai/gpt-oss-120b, with the QNFO model router as a fallback.
- Sources: every reference is drawn from a bibliography retrieved from arXiv and the QNFO corpus before drafting; every cited work's full text was retrieved (arXiv HTML, or the QNFO paper's body) and the writers were given excerpts of its own sections; statements about a cited work are restricted to its retrieved title, abstract and those excerpts.
- Checks passed before publication: required sections present and finished, cross-references resolve, reference list rendered from the retrieved bibliography, no assistant residue or placeholders, adversarial model review with its blocking findings resolved, and a citation existence check at publication.
- Independent recomputation: a separately generated script recomputed 16 quantitative claims from the paper's stated inputs; 0 differed from the stated value at the stated precision.
- Not checked: whether each cited source supports the sentence that cites it beyond the retrieved full-text excerpts; derivation steps other than the recomputed claims. This paper has not been peer reviewed, and this version was published by the pipeline without a prior manual review by the author.

References

- [1] Latent-Y: A Lab-Validated Autonomous Agent for De Novo Drug Design. arXiv:2603.29727v2. <https://arxiv.org/abs/2603.29727v2> [2] Latent-X: An Atom-level Frontier Model for De Novo Protein Binder Design. arXiv:2507.19375v1. <https://arxiv.org/abs/2507.19375v1> [3] Drug-like antibodies with low immunogenicity in human panels designed with Latent-X2. arXiv:2512.20263v1. <https://arxiv.org/abs/2512.20263v1> [4] Hide and seek with quantum resources: New and modified protocols for quantum steganography. arXiv:2110.05893v1. <https://arxiv.org/abs/2110.05893v1> [5] GeMA: Learning Latent Manifold Frontiers for Benchmarking Complex Systems. arXiv:2603.16729v1. <https://arxiv.org/abs/2603.16729v1> [6] Latent Space Subdivision: Stable and Controllable Time Predictions for Fluid Flow. arXiv:2003.08723v1. <https://arxiv.org/abs/2003.08723v1> [7] Trends toward real-time network data steganography. arXiv:1604.02778v1. <https://arxiv.org/abs/1604.02778v1> [8] SPACE: the SPectroscopic All-sky Cosmic Explorer. arXiv:0804.4433v1. <https://arxiv.org/abs/0804.4433v1>