CANONICAL PARAPHRASING: A PROVABLE DEFENSE AGAINST STEGANOGRAPHY IN LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Language models can encode hidden messages in the surface form of their outputs while producing text statistically indistinguishable from honest generation (Mathew et al., 2025; Motwani et al., 2024a; Zolkowski et al., 2026). The standard defense, paraphrasing, scrambles surface tokens but does not close the channel: stochastic paraphrasing only reduces the residual capacity, leaving room for an adaptive adversary. We propose *canonical paraphrasing*, whose ideal form gives every paraphrase of a text the same meaning-preserving rewrite. We prove that this eliminates covert communication when the encoder hides information through wording choices while preserving meaning, and develop a practical approximation that substantially reduces payload recovery in the settings we evaluate. Our analysis also shows how to use steganographic encoding to sample texts with the same meaning, and vice versa, with polynomially many oracle calls when codebook capacity grows logarithmically with output length. Minimum entropy coupling (de Witt et al., 2023) yields near-chance AUC for the tested detectors across four language models. On 50 cover texts carrying complete chess games, the practical defense raises decoding bit-error rate from 5.79% to 45.54%. For AI safety, this offers a way to limit covert coordination between language-model agents even when their hidden messages evade detection.

1 INTRODUCTION

Modern language models can produce output that is statistically indistinguishable from honest generation while embedding hidden information (de Witt et al., 2023). Recent empirical work demonstrates that this is more than a theoretical concern. Mathew et al. (2025) show that LLM agents learn steganographic encodings under task-completion pressure without explicit instructions to hide information, Motwani et al. (2024b) formalise multi-agent secret collusion and document rising steganographic capabilities in frontier models, and Zolkowski et al. (2026) find early signs of steganographic capability in current frontier models without any fine-tuning. The threat surface is broad: any deployment in which a language model can choose among semantically equivalent surface forms while accomplishing an observable task admits a covert channel. The standard mitigation, paraphrasing the model output before transmission to scramble the encoding (Roger & Greenblatt, 2023; Lanham et al., 2023), leaves the channel partially open: stochastic paraphrasing reduces the residual capacity but does not eliminate it, and an adaptive adversary can exploit the residual.

Our practical defense targets *semantic-preserving* steganography: encoding schemes that exploit syntactic freedom (word choice, phrasing, sentence structure) while constrained to convey fixed semantic content. This captures the practical threat from agentic deployment, where the model must accomplish a task observable to the operator (answer a question, complete a prompt) yet can choose among many surface forms that satisfy the task. Our guarantee applies to schemes whose encodings remain within one class of the chosen equivalence relation for each fixed task, including perfectly secure schemes with this property (§3). Two adjacent threats remain outside the demonstrated scope: steganography that hides information through choices about which facts to mention or how to structure an argument, and channels carried by an agent's behaviour rather than its text, such as tool-call ordering or timing (§6, §7). Following cryptographic convention, we call the model the *adversary* and the operator the *defender*.

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.

The freedom to express the same meaning in different ways creates room for a hidden payload. We measure this freedom through co-entropy, the uncertainty in surface form once meaning is fixed (§2).

We propose *canonical paraphrasing* as a defense against semantic-preserving steganography. The ideal defense replaces every paraphrase of a text with the same meaning-preserving rewrite. We prove that this eliminates covert communication when the encoder hides information through wording choices while preserving meaning. Our three-stage practical approximation substantially reduces payload recovery in the evaluated settings.

Contributions.

1. **Canonical paraphrasing with a transparency guarantee.** We prove that a canonical-form operator drives steganographic capacity to zero under the transparency condition, and develop a three-stage practical approximation (embed, snap, project, Algorithm 1). A complexity-duality result relates encoding to semantic sampling with polynomial oracle complexity in the regime $C = O(\log n)$ (§3, §4).
2. **Empirical validation across three fronts.** The optimal known steganographic scheme, minimum entropy coupling (de Witt et al., 2023), yields detection AUC near 0.50 for the tested detectors across four open-weight language models from 1B to 7B parameters. Three embedding-space canonicalisers yield centres that agree at cosine similarity ≥ 0.925 on 99.99–100% of $9{,}431$ anchor sentences. The defense with Fréchet centre estimation raises decoding bit-error rate from 5.79% to 45.54% on 50 cover texts carrying chess payloads (§5).
3. **A geometric characterisation of the threat surface.** The space of all perfectly secure steganographic schemes over a cover distribution with capacity C is isomorphic to the standard probability simplex Δ^{2^C-1}. Minimum entropy coupling sits at the barycenter as the uniquely optimal scheme; the ideal canonical operator neutralises every scheme in this space satisfying the transparency condition (§3, App A).

2 Background and threat model

Channels. We model covert communication as a channel from a hidden payload M to the transmitted text X. The channel is a probabilistic mapping from payloads to texts (Shannon, 1948). Its *capacity* is the supremum of the mutual information $I(M;X)$ over payload distributions. Throughout the paper, C denotes this capacity in bits per transcript, which bounds the rate at which the adversary can transmit information through the deployment pipeline. For a given payload distribution, $I(M;X) = 0$ means that observing the text gives the receiver no information about the payload (Shannon, 1949). We call the channel *closed* when this holds for every payload distribution.

Cachin perfect security. A steganographic scheme is ϵ-*secure* (Cachin, 1998) if $D_{\mathrm{KL}}(P_{\mathcal{S}} \| \mathcal{D}) \leq \epsilon$, where $P_{\mathcal{S}}$ is the distribution over transcripts the scheme actually emits and $\mathcal{D}$ is the cover distribution the same model would produce carrying no payload. Setting $\epsilon = 0$ gives the perfectly secure regime, in which the two distributions are equal. Hopper et al. (2002) introduced the corresponding computational notion, SS-CHA security; Pinsker's inequality gives $\|P_{\mathcal{S}} - \mathcal{D}\|_{\mathrm{TV}} \leq \sqrt{\epsilon/2}$ per symbol, so Cachin security implies SS-CHA up to a standard hybrid argument over query length. Perfect security lets an encoder hide the presence of a payload while still allowing its receiver to recover it. Our goal is to close that communication channel, even when the encoded text evades detection.

Semantic equivalence and co-entropy. A *semantic equivalence relation* $\sim_{sem}$ on a formal language $\mathcal{L} \subseteq \mathcal{X}^*$ partitions admissible texts into equivalence classes $[x]_{sem} := \{x' \in \mathcal{L} : x' \sim_{sem} x\}$. The framework is parameterised by the choice of $\sim_{sem}$. Our practical pipeline targets paraphrase equivalence $\sim_{\mathrm{para}}$: texts convey the same propositional content. A coarser task-adherence relation $\sim_{\mathrm{task}}$ would group outputs that accomplish the same task to the same degree, even when their content differs; its precise definition depends on a fixed task, context, and success criterion. A *task* is the observable job the operator gave the model, the thing the output must still accomplish after the defense has rewritten it. In the chess experiment of §5.3 the task is answering the benign prompt that produced the cover text, and the chess game hidden inside that text is payload the task never asks

for. Let $S = [X]_{sem}$ denote the equivalence class of a sample $X \sim \mathcal{D}$. The *co-entropy* (Cover & Thomas, 2006) is the conditional entropy

$$H^{\mathrm{c}}(\mathcal{D} \mid \sim_{sem}) := H(X \mid S), \tag{1}$$

giving the chain-rule decomposition $H(X) = H(S) + H^{\mathrm{c}}(\mathcal{D} \mid \sim_{sem})$. For paraphrase equivalence, the first term is information about meaning; the second is bits of freedom in surface form once meaning is fixed. Steganography exploits the co-entropy. The defender's task is to drive the co-entropy to zero.

3 Theory: simplex and transparency

We state the headline theorems informally; full statements and proofs are in App A and B. Two structural results bound the threat (the Simplex Theorem and the location of MEC at the barycenter), and one establishes the ideal canonical operator's guarantee (the Transparency Theorem). A complexity-duality result completes the picture by relating steganographic encoding to semantic sampling in the oracle model.

The simplex theorem. Let $\mathcal{D}$ be a cover distribution over $\mathcal{X}$ with $|\mathrm{supp}(\mathcal{D})| \geq 2^C$ and $H(\mathcal{D}) \geq C$. The *moduli space* $\mathfrak{M}(\mathcal{D}, C)$ of perfectly secure steganographic schemes over $\mathcal{D}$ at capacity C (with deterministic decoding, modulo message relabeling) is isomorphic to the standard probability simplex

$$\mathfrak{M}(\mathcal{D}, C) \;\cong\; \Delta^{2^C - 1}. \tag{2}$$

A scheme corresponds to a partition $\{R_m\}_{m \in \mathcal{M}}$ of $\mathrm{supp}(\mathcal{D})$ into 2^C regions, with the simplex coordinate of the scheme being the mass vector $(\mathcal{D}(R_m))_m$. The forward direction is read off from any scheme; the reverse direction constructs a scheme from any mass vector by greedy assignment on the support. Under these assumptions, every perfectly secure scheme has a representation in the simplex. This *completeness* result means that a defense that neutralises the entire simplex neutralises every such scheme.

MEC at the barycenter. Within Δ^{2^C-1}, minimum entropy coupling (Kocaoglu et al., 2017; de Witt et al., 2023) corresponds to the uniform point $\mathbf{u} = (2^{-C}, \ldots, 2^{-C})$, the Euclidean barycenter of the simplex (App D). This location is context for the threat model, not a step in the defense. It tells a defender which scheme to expect at the top of the efficiency ordering, and nothing in the transparency theorem or in the pipeline of §4 depends on where MEC sits: the ideal canonical operator applies to any scheme satisfying transparency, without needing to know its position in the simplex. MEC achieves the maximum embedding rate among perfectly secure schemes; Cicalese et al. (2019) show that the greedy MEC algorithm is within an additive bit of the optimum. Empirically, MEC realises efficiency $\eta \approx 0.95$ with near-chance AUC for the tested detectors across language models (§5). Stepping off the simplex to $\epsilon > 0$ collapses capacity from $O(n)$ to $O(\sqrt{n})$ (the square-root law of Ker (2017)), so the simplex is a knife-edge: optimal steganography lives on it, and any approximation degrades sharply. The Fisher–Rao Hessian of the entropy at the barycenter has order $2^C/(2C)$, so the basin of near-optimality shrinks exponentially with capacity; this quadratic flatness explains why approximate MEC implementations remain near-optimal at small C and degrade at large C.

The transparency theorem. A canonical-form operator $\mathrm{Can} : \mathcal{L} \to \mathcal{L}$ for $\sim_{sem}$ is one that (i) preserves equivalence ($\mathrm{Can}(x) \sim_{sem} x$), (ii) is idempotent ($\mathrm{Can}(\mathrm{Can}(x)) = \mathrm{Can}(x)$), and (iii) is unique on equivalence classes ($x \sim_{sem} y \implies \mathrm{Can}(x) = \mathrm{Can}(y)$). A scheme is $\sim_{sem}$*-transparent* (Definition 3.1) if for every fixed semantic task s, all encodings $\mathrm{Enc}(s, m, U)$ for varying message m lie in the same equivalence class. Throughout, U denotes the encoder's internal randomness, drawn fresh at each invocation, and U' is an independent draw of it, so $\mathrm{Enc}(s, m, U)$ and $\mathrm{Enc}(s, m', U')$ are two independent runs of the encoder on the same task with different messages.

Definition 3.1 ($\sim_{sem}$-transparent scheme)**.** *A steganographic scheme is $\sim_{sem}$-transparent if, for any fixed semantic task $s \in \mathcal{S}$, $\Pr[\mathrm{Enc}(s, m, U) \sim_{sem} \mathrm{Enc}(s, m', U')] = 1$ for every pair of messages $m, m' \in \mathcal{M}$.*

The proof also applies to task-adherence equivalence, provided a canonical operator exists for that relation. Constructing such an operator requires resolving differences in content as well as wording. Our practical pipeline targets paraphrase equivalence.

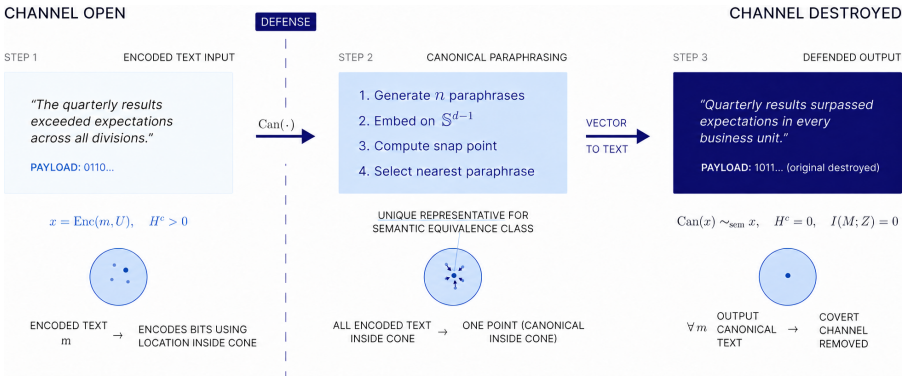


Figure 1: Ideal canonicalization. Stegotext input (left) carries a hidden payload with positive co-entropy $H^c > 0$. The canonical operator (centre) maps every text in a semantic equivalence class to the same meaning-preserving representative. Under the transparency condition, the defended output (right) has $H^c = 0$ and $I(M; Z) = 0$. These equalities describe the ideal guarantee. Algorithm 1 gives the practical approximation.

Theorem 3.2 (Transparency). *For any $\sim_{sem}$-transparent scheme and any canonical-form operator* Can *for $\sim_{sem}$ (equivalence-preserving, idempotent, unique on classes),*

$$I\big(M;\, \mathrm{Can}(\mathrm{Enc}(s, M, U))\big) \;=\; 0 \qquad \textit{for every fixed task } s. \tag{3}$$

Uniqueness makes the defended output independent of the payload: $\mathrm{Can}(\mathrm{Enc}(s, m, U))$ is constant in m, so it carries no information about m. Within the simplex characterisation, the transparency theorem establishes that the ideal canonical operator neutralises every scheme satisfying the transparency condition, destroying the channel after encoding without requiring detection of the payload. The pipeline of §4 approximates this operator; §5 measures the resulting reduction in payload recovery. Two further results, stated and proved in App B, complete the picture. The complexity duality shows that STEGANOGRAPHIC-ENCODE and SEM-SAMPLE are mutually Turing-reducible with $O(2^C)$ oracle calls, polynomially many when $C = O(\log n)$. Here, an oracle is a tool that solves one task, and a reduction uses it to solve the other. Each use of the tool counts as one oracle call. The second result bounds residual information for stochastic relaxations $\widetilde{\mathrm{Can}}$ under approximate equivalence and class-mixing conditions.

4 THE CANONICAL PARAPHRASING PIPELINE

The canonical-form operator Can is generally not computable in closed form on natural language. We approximate it with a three-stage pipeline that operates on language-model outputs without external state: *embed* the input and its paraphrases, *snap* their embeddings to an estimated centre on the unit sphere, and *project* by rewriting the input through discrete diffusion with a seed derived from that centre. The pipeline rewrites every output without needing to detect a payload, know the codebook, or locate the encoded bits. Algorithm 1 states the procedure. Figure 1 illustrates the ideal canonical operator that motivates it.

We estimate a centre from several paraphrases to give the rewrite a reference based on their shared content. That centre sets the random seed, which controls the diffusion model's choices as it edits the original text.

Algorithm 1: Canonical paraphrasing pipeline

Input: stegotext $x \in \mathcal{L}$, embedder $\mathrm{Emb} : \mathcal{L} \to \mathbb{S}^{d-1}$, paraphrase oracle Π, snap operator Snap on $\mathbb{S}^{d-1}$, diffusion editor $\mathrm{Proj} : \mathcal{L} \times \mathbb{N} \to \mathcal{L}$, pool size n, base seed s_0

Result: Defended output $x^* \in \mathcal{L}$

$p_0 \leftarrow x, \quad (p_1, \ldots, p_n) \leftarrow \Pi(x; n)$; `// input and generated paraphrases`

$E \leftarrow (\mathrm{Emb}(p_i))_{i=0}^{n}$; `// embed each on the unit sphere`

$c \leftarrow \mathrm{Snap}(E)$; `// centre estimate, e.g. Fréchet mean`

$s \leftarrow s_0 + \mathrm{id}(c)$; `// hash the centre to obtain its integer identifier`

return $x^* \leftarrow \mathrm{Proj}(x; s)$; `// diffusion rewrite with fidelity checks`

After rewriting, the pipeline re-embeds and re-snaps the output to check consistency.

Three centre-estimators on the sphere. The snap step computes the centre of a paraphrase cluster on the unit sphere $\mathbb{S}^{d-1}$. We evaluate three estimators. The *Fréchet mean* is the Riemannian centre of mass minimising squared geodesic distance. The *Weiszfeld geometric median* minimises the sum of Euclidean distances. It has a 50% breakdown point and is robust to outlier paraphrases. *Directional mean shift* finds the mode of a von Mises–Fisher kernel density on the sphere. All three yield centres that agree across independent paraphrase subsets at cosine similarity ≥ 0.925 on 99.99–100% of $9{,}431$ anchor sentences (§5.2). Appendix F.2 gives the update rules and parameter settings.

Embed and project via discrete diffusion. The embedder is the masked discrete language model MDLM (Sahoo et al., 2024) with the `kuleshov-group/mdlm-owt` checkpoint. Given input text x, MDLM produces an output logit tensor $\mathbf{L} \in \mathbb{R}^{n \times d}$ where $d = |\mathcal{X}| = 50{,}257$ is the vocabulary size. The embedding is the mean-pooled logits across positions, L_2-normalised to the unit sphere. This "belief embedding" captures the model's predictive distribution at each position. Semantically similar texts induce similar belief distributions, causing paraphrases to cluster tightly on $\mathbb{S}^{d-1}$ (§5.4). We use discrete diffusion to rewrite the input text, with a random seed derived from the estimated centre. The pipeline checks whether the rewrite preserves numbers, named entities, and negation. Our evaluation measures centre agreement, payload recovery, and answer correctness after rewriting. Appendix F.1 gives the diffusion settings.

5 Empirical validation

We test whether detectors can distinguish MEC-encoded text from ordinary output across language models. We then measure centre agreement on $9{,}431$ anchor sentences and payload recovery after rewriting in chess and scripted 20 Questions experiments. A comparison across embedding architectures tests whether paraphrase clusters are separable. All source code and raw experimental data are released anonymously for review.[1]

5.1 The threat is real: detectors perform near chance

We test whether MEC-encoded stegotext is distinguishable from ordinary model output by training detectors on paired encoded and cover-text samples and measuring detection AUC. We use four open-weight language models spanning 1B to 7B parameters: Mistral-7B (Jiang et al., 2023), Qwen2.5-7B-Instruct (Yang et al., 2024), Phi-3-mini-4k-instruct (Abdin et al., 2024), and TinyLlama-1.1B-Chat (Zhang et al., 2024), with standardised generation parameters (temperature 1.0, top-k 64, 100-token cap). Cover text is drawn from 50 seed prompts spanning news, Wikipedia, fiction, scientific abstracts, and social media, with 30 samples per trial across 10 trials. The MEC implementation mirrors the open-source codebase of de Witt et al. (2023).

We find that MEC achieves embedding efficiency $\eta \approx 0.95$ while the tested detectors yield AUC near 0.50 (Table 1). None of the four model means differ significantly from the random-guessing baseline; full statistical tests are in App F. These results demonstrate the difficulty of detecting MEC stegotext with the evaluated detectors across language models from 1B to 7B parameters. They motivate a defense that acts on the channel directly, without depending on successful detection.

[1] `https://anonymous.4open.science/r/neurips-canpar-code-027E/`

Table 1: MEC detectability across language models (10 trials, 30 samples each). All four AUC means are within one standard deviation of 0.50; one-sample t-tests against the null AUC $= 0.50$ all give $p > 0.34$. The strongest of three detectors, an ensemble of RoBERTa-large + statistical features, reaches AUC 0.515 on the held test split (App F, Table 9).

Model	Parameters	MEC AUC
Mistral-7B	7B	0.516 ± 0.088
Qwen2.5-7B	7B	0.524 ± 0.123
Phi-3-mini	3.8B	0.522 ± 0.132
TinyLlama	1.1B	0.476 ± 0.074

Table 2: Embedding-space centre agreement ($9{,}431$ anchors, 20 paraphrases per subset, $\varepsilon = 0.075$). Centres agree when their cosine similarity is ≥ 0.925; centre counts use this tolerance. Mean shift and Fréchet meet the threshold on every anchor, and Weiszfeld on all but one (99.99%).

Method	Avg pairwise sim	Avg centres at ε
Mean shift	0.993	1.0
Fréchet mean	0.993	1.0
Weiszfeld	0.993	1.0

5.2 The snap step: consistent centres across 9,431 anchors

We test the stability of the canonicaliser's centre estimates across independent paraphrase sets. From a $10{,}000$-sentence paraphrase corpus (App F), we select the $9{,}431$ anchors with at least 40 paraphrases. For each anchor we draw two disjoint 20-paraphrase subsets, run each canonicaliser independently on each subset at convergence radius $\varepsilon = 0.075$, and count the two snap points as agreeing when their cosine similarity is $\geq 1 - \varepsilon = 0.925$.

We find highly consistent centre estimates (Table 2). Snap points from independent subsets meet the agreement threshold on every anchor (Mean shift, Fréchet) or all but one (99.99% for Weiszfeld); their average pairwise cosine similarity is 0.993 ± 0.005. This consistency holds across the three geometric optimisation objectives (Riemannian mean, L_1 median, density mode), supporting the stability of centre estimation under paraphrase resampling. Agreement at the specified tolerance measures embedding stability; exact equality and residual co-entropy require separate assessment. We next measure payload recovery after text projection (§5.3).

5.3 End-to-end: paraphrasing disrupts payload recovery

We evaluate the defense on 50 chess games encoded with MEC in Mistral-7B cover text. Each payload contains a complete game in algebraic notation (27–124 bytes, averaging 31 moves). The pipeline estimates a Fréchet centre from paraphrases of each text, then rewrites the input using MDLM. The receiver attempts to recover the chess payload from the rewritten text. We score recovery by *bit-error rate* (BER), the fraction of payload bits the receiver recovers incorrectly. BER $= 0$ means exact recovery, while a receiver reading random bits would average BER $= 0.5$. A higher BER indicates poorer recovery by the evaluated decoder. BER near 0.5 alone does not establish zero residual information. App F.3 gives the run settings.

Across $38{,}264$ payload bits, the defense raises bit-error rate from 5.79% to 45.54% (Figure 2). The decoder does not recover the original payload exactly from any of the 50 rewritten texts. After defense, we estimate a capacity of 0.010 bits per payload bit using a binary-channel model (App F.3). Processing each cover text takes 31.4 seconds on average, including paraphrase generation and diffusion editing.

Scripted 20 Questions. We also evaluate six scripted games in which MEC hides questions and answers in Llama-3.1-8B-Instruct cover text. Across 112 answers with paired measurements, the receiver recovers the intended answer in $112/112$ cases without defense and $15/112$ (13.4%) after canonical paraphrasing. Each comparison uses the same encoded message. App F.4 gives the setup, scoring rule, and excluded attempts.



Figure 2: Chess payload recovery with Fréchet canonical paraphrasing. This example has the closest BER to the run-wide result among games recovered exactly before defense. *Before defense* (left), the receiver recovers all 632 payload bits. *After defense* (centre), 291 bits differ (BER $= 46.0\%$). The text panels quote the first edited sentence. The *run summary* (right) covers 50 games and 38,264 payload bits. BER rises from 5.79% to 45.54%, and the defense takes 31.4 seconds per cover text on average.

Task utility. Canonical paraphrasing leaves answer correctness unchanged on the 60 GSM8K problems (Cobbe et al., 2021) we evaluate, with the same 24 problems solved correctly in both conditions (App F.5).

5.4 Paraphrase clusters are linearly separable across seven embedding models

The pipeline's snap step assumes that paraphrase clusters are separable in embedding space: each anchor's paraphrases occupy a small region of $\mathbb{S}^{d-1}$ that does not overlap with any other anchor's. We measure this directly across seven embedding architectures on 1,000 anchor sentences sampled from C4 (Raffel et al., 2020) with 20 paraphrases per anchor (Claude API, App E). For each anchor we compute the smallest ε-cone capturing all of its own paraphrases ($\varepsilon_{\text{capture}}$) and the largest cone before any foreign paraphrase intrudes ($\varepsilon_{\text{exclude}}$); a perfect cone has $\varepsilon_{\text{capture}} < \varepsilon_{\text{exclude}}$. We also

Table 3: ε-cone separability across seven embedding architectures (1,000 anchors, 20 paraphrases each). *Linear separability* is the fraction of anchors whose paraphrase cluster a linear SVM separates from all other anchors; *perfect cones* is the fraction of anchors whose own-paraphrase cone does not overlap any other anchor's. All seven models achieve 100% linear separability on the evaluated clusters. Cone separability varies by architecture.

Model	Dim.	Perfect cones	Linear sep.
intfloat/e5-large-v2	1024	98.0%	100%
BAAI/bge-large-en-v1.5	1024	97.5%	100%
all-mpnet-base-v2	768	95.2%	100%
all-MiniLM-L6-v2	384	87.6%	100%
Llama-3.1-8B-Instruct	4096	76.9%	100%
Qwen2.5-14B-Instruct	5120	47.6%	100%
MDLM (`kuleshov-group/mdlm-owt`)	768	66.9%	100%

train a linear SVM separating each anchor's cluster from all others and report the resulting linear separability rate.

We find that all seven architectures achieve 100% linear separability of paraphrase clusters, even when fixed-radius cones overlap (Table 3). Cone separability ranges from 47.6% on Qwen2.5-14B to 98.0% on e5-large; the four contrastively trained sentence embedders cluster above 87%, while general-purpose causal language models drop to 47–77%. Outlier removal (App E) raises MDLM perfect cones from 60.2% to 81.7% with two removals. These results motivate learned decision boundaries as an alternative to fixed-radius cones.

6 Limitations and open problems

The text-projection gap. Our centre estimates agree closely across paraphrases (Table 2). The diffusion editor also uses the original text, so its rewrites can still differ. To fully implement the ideal canonicalizer, paraphrases with the same meaning would need to produce the same rewrite. Appendix F.2 reports a repeatability check of the pipeline.

Semantic versus syntactic steganography. Our practical pipeline targets *semantic-preserving* steganography, where meaning is fixed and the encoder's freedom lies in surface form. Semantic steganography, where hidden information resides in the meaning itself (which facts to mention, how to structure an argument, where to place emphasis), is outside our empirical scope. The transparency proof also applies to a coarser task-adherence relation ($\sim_{\mathrm{task}}$), provided the encoder is transparent for that relation and a corresponding canonical operator exists. Constructing such an operator remains open: collapsing task-equivalent outputs to a single representative requires controlling content as well as surface form.

Adaptive adversaries and pre-coordination. The defense analysed here is non-adaptive: the canonicaliser is fixed independently of the adversary's encoding strategy. A steganographer that observes the canonicaliser's behaviour could attempt to game it; the equilibrium of the resulting steganographer-defender game is open. A particularly concerning variant is *pre-coordination*: agents trained on shared data may establish implicit codebooks that survive canonicalisation, because the canonical form itself becomes the new covertext for a second-layer encoding. Multi-agent and multi-turn extensions, where agents communicate across many messages, also fall outside the single-output framework. Side channels (timing, error patterns, weight introspection) are out of scope and would each require separate construction.

7 Related work

Foundational steganography. Cachin (1998) defined information-theoretic steganographic security via Kullback–Leibler indistinguishability between stegotext and cover distributions, perfect security being the $\epsilon = 0$ regime. Hopper et al. (2002) introduced the corresponding computational notion (SS-CHA) and the rejection-sampling construction. Wang & Moulin (2008) characterised capacity

for memoryless cover sources at the entropy rate. The Simplex Theorem (§3) extends these results with a structural characterisation: the space of all schemes achieving capacity C is precisely the simplex Δ^{2^C-1}.

Minimum entropy coupling and optimal transport. Kocaoglu et al. (2017) introduced minimum entropy coupling for causal inference; de Witt et al. (2023) applied it to language-model steganography and proved its optimality under perfect security. Cicalese et al. (2019) showed that the greedy MEC algorithm is within an additive bit of the optimum. Our contribution is the geometric interpretation: MEC sits at the Euclidean barycenter of the simplex, and the entropy function's quadratic flatness there explains the robustness of approximate MEC implementations. Appendix D discusses the geometry of the coupling polytope, which also appears in optimal transport (Cuturi, 2013; Villani, 2009).

Watermarking and undetectability. Christ et al. (2024) proved that cryptographically undetectable watermarks for language models exist; structurally, an undetectable watermark is a perfectly secure steganographic scheme with a fixed message. Kirchenbauer et al. (2023) introduced the green-list watermark (KGW) and Kirchenbauer et al. (2024) characterised its paraphrase robustness empirically. Hou et al. (2024) introduced SemStamp, a semantic-embedding watermark designed for paraphrase robustness. Signals encoded in the semantic-class label can survive changes to surface wording. The transparency theorem applies when encodings lie within a single class of the chosen relation; signals carried across paraphrase classes require a coarser relation and its canonical operator.

Schemes at the edge of the threat model. Three recent constructions test the boundary of what a surface-form defense can reach. Bai et al. (2024) build a semantic space using ontology-entity trees and map the payload onto that space, so the bits ride in what the text says rather than in how it says it. Xiang et al. (2026) embed by transforming a masked language model's predicted distribution instead of substituting words, leaving the realised cover text intact and recovering the payload with a fine-tuned extractor. Perry et al. (2025) define weak and strong robustness against re-randomization attacks and build schemes whose payload survives adversarial paraphrasing and rewording. The first lies outside our practical scope. Extending the defense to a coarser relation such as $\sim_{\mathrm{task}}$ would require a corresponding canonical operator, whose construction remains open. The second would fall under the transparency theorem if its encodings remain within one class of the relation being canonicalised. Robustness to re-randomization is robustness to a perturbation distribution, whereas an ideal canonical operator removes the channel from any scheme whose outputs stay within one equivalence class, so the two guarantees do not directly contradict each other. We have not evaluated this construction against our pipeline.

AI safety and active defense. Roger & Greenblatt (2023) formalised encoded reasoning as a CoT-monitoring threat and proposed paraphrasing as a defense; Lanham et al. (2023) measured CoT faithfulness more broadly. Mathew et al. (2025) demonstrated that LLM agents learn steganographic encodings under task pressure without explicit instructions to hide information. Motwani et al. (2024b) formalised multi-agent secret collusion. Korbak et al. (2025) argued that CoT monitorability is fragile under training pressure. Greenblatt et al. (2024) established the AI Control framework; Bhatt et al. (2025) introduced Ctrl-Z resampling, a monitor-driven substitution defense in the same family as canonical paraphrasing. Our defense rewrites every output without waiting for a monitor flag. Active defense is required because passive detection is provably insufficient on the simplex. A parallel line defends against *agentic* steganography, where the channel is an agent's behaviour rather than the text it emits, and that setting is outside the single-output framework studied here.

8 Conclusion

We characterise perfectly secure language-model steganography as a simplex, with minimum entropy coupling at its barycenter. Under the transparency condition, ideal canonical paraphrasing eliminates covert communication by giving every paraphrase of a text the same meaning-preserving rewrite. Our practical approximation substantially reduces payload recovery, raising bit-error rate from 5.79% to 45.54% in our experiments. These results provide a formal foundation and a practical starting point for defenses that suppress covert communication without relying on detection.

AI USE STATEMENT

This work was enhanced through the use of AI-based tools, including ChatGPT (chatgpt.com), Claude (claude.ai), and various models integrated within the Cursor IDE (cursor.com). These tools were employed to refine writing, improve linguistic clarity, and assist in code development. Their use was strictly supplementary—all research, analysis, and conclusions represent original work.

ETHICS STATEMENT

This work studies a defense against covert communication in language-model outputs. The results have dual-use implications: characterising undetectable steganography may help adversaries construct covert channels, while paraphrasing defenses may also undermine text watermarks or legitimate privacy-preserving uses of steganography. Our theoretical guarantees depend on the semantic-preservation and transparency assumptions stated in Section 3; the implemented pipeline only approximates the ideal canonical operator. In practice, rewriting text may alter its intended meaning, and residual channels may remain. Section 6 discusses the text-projection gap, adaptive adversaries, and channels outside our threat model. These limitations matter for deployment: the reported results should not be interpreted as establishing that a deployed system is free of covert communication or that its outputs can safely be trusted without further evaluation.

REPRODUCIBILITY STATEMENT

Section 3 states the principal theoretical claims and their assumptions, with supporting proofs and derivations in Appendices A, B, C, and D. Algorithm 1 and Section 4 describe the canonical paraphrasing procedure, and Section 5 describes the evaluations. Appendix F records model and paraphrase-generation settings, canonicaliser parameters, detector training, and statistical procedures. Additional geometric experiments and security–efficiency measurements appear in Appendices E, G, and H. The anonymous code and experimental-data repository is linked in Section 5.

REFERENCES

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024.

Minhao Bai, Jinshuai Yang, Kaiyi Pang, Yongfeng Huang, and Yue Gao. Semantic steganography: A framework for robust and high-capacity information hiding using large language models, 2024. URL https://arxiv.org/abs/2412.11043.

Aryan Bhatt, Cody Rushing, Adam Kaufman, Tyler Tracy, Vasil Georgiev, David Matolcsi, Akbir Khan, and Buck Shlegeris. Ctrl-Z: Controlling AI agents via resampling, 2025. URL https://arxiv.org/abs/2504.10374.

Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: a suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23, Honolulu, Hawaii, USA, 2023. JMLR.org. URL https://dl.acm.org/doi/10.5555/3618408.3618510.

Christian Cachin. An information-theoretic model for steganography. In David Aucsmith (ed.), *Information Hiding*, pp. 306–318, Berlin, Heidelberg, 1998. Springer Berlin Heidelberg. ISBN 978-3-540-49380-8. doi: 10.1007/3-540-49380-8_21.

Ching-Yun Chang and Stephen Clark. Practical linguistic steganography using contextual synonym substitution and a novel vertex coding method. *Comput. Linguist.*, 40(2):403–448, June 2014. ISSN 0891-2017. doi: 10.1162/COLI_a_00176.

Miranda Christ, Sam Gunn, and Or Zamir. Undetectable watermarks for language models. In *Proceedings of the 37th Annual Conference on Learning Theory (COLT)*, 2024. URL https://arxiv.org/abs/2306.09194.

Ferdinando Cicalese, Luisa Gargano, and Ugo Vaccaro. Minimum-entropy couplings and their applications. *IEEE Trans. Inf. Theor.*, 65(6):3436–3451, June 2019. ISSN 0018-9448. doi: 10.1109/TIT.2019.2894519.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168.

Spencer Compton, Dmitriy Katz, Benjamin Qi, Kristjan Greenewald, and Murat Kocaoglu. Minimum-entropy coupling approximation guarantees beyond the majorization barrier. In *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 10445–10469. PMLR, 2023. URL https://proceedings.mlr.press/v206/compton23a.html.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 2nd edition, 2006.

Marco Cuturi. Sinkhorn distances: lightspeed computation of optimal transport. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NeurIPS'13, pp. 2292–2300, Red Hook, NY, USA, 2013. Curran Associates Inc. URL https://papers.nips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html.

Falcon Dai and Zheng Cai. Towards near-imperceptible steganographic text. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4303–4308, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1422.

Christian Schroeder de Witt, Samuel Sokota, J. Zico Kolter, Jakob Foerster, and Martin Strohmeier. Perfectly secure steganography using minimum entropy coupling. In *The Eleventh International Conference on Learning Representations (ICLR)*, 2023. URL https://arxiv.org/abs/2210.14889.

Jinyang Ding, Kejiang Chen, Yaofei Wang, Na Zhao, Weiming Zhang, and Nenghai Yu. Discop: Provably secure steganography in practice based on "distribution copies". In *2023 IEEE Symposium on Security and Privacy (SP)*, pp. 2238–2255, 2023. doi: 10.1109/SP46215.2023.10179287.

Tina Fang, Martin Jaggi, and Katerina Argyraki. Generating steganographic text with LSTMs. In *Proceedings of ACL 2017, Student Research Workshop*, pp. 100–106, Vancouver, Canada, July 2017. Association for Computational Linguistics. URL https://aclanthology.org/P17-3017/.

Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *International Conference on Machine Learning (ICML)*, 2024.

Nicholas J. Hopper, John Langford, and Luis von Ahn. Provably secure steganography. In Moti Yung (ed.), *Advances in Cryptology — CRYPTO 2002*, pp. 77–92, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg. ISBN 978-3-540-45708-4. doi: 10.1007/3-540-45708-9_6.

Abe Bohan Hou, Jingyu Zhang, Tianxing He, Yichen Wang, Yung-Sung Chuang, Hongwei Wang, Lingfeng Shen, Benjamin Van Durme, Daniel Khashabi, and Yulia Tsvetkov. SemStamp: A semantic watermark with paraphrastic robustness for text generation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 2024.

Shun ichi Amari. *Information Geometry and Its Applications*, volume 194 of *Applied Mathematical Sciences*. Springer, 2016. doi: 10.1007/978-4-431-55978-8.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B, 2023.

Gabriel Kaptchuk, Tushar M. Jois, Matthew Green, and Aviel D. Rubin. Meteor: Cryptographically secure steganography for realistic distributions. In *Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security*, CCS '21, pp. 1529–1548, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384544. doi: 10.1145/3460120.3484550.

Andrew D. Ker. The square root law of steganography: Bringing theory closer to practice. In *Proceedings of the 5th ACM Workshop on Information Hiding and Multimedia Security*, IH&MMSec '17, pp. 33–44, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450350617. doi: 10.1145/3082031.3083235.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. A watermark for large language models. In *International Conference on Machine Learning (ICML)*, 2023.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Manli Shu, Khalid Saifullah, Kezhi Kong, Kasun Fernando, Aniruddha Saha, Micah Goldblum, and Tom Goldstein. On the reliability of watermarks for large language models. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2306.04634.

Murat Kocaoglu, Alexandros G. Dimakis, Sriram Vishwanath, and Babak Hassibi. Entropic causal inference. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, pp. 1156–1162. AAAI Press, 2017. URL https://dl.acm.org/doi/10.5555/3298239.3298409.

Tomek Korbak, Mikita Balesni, Elizabeth Barnes, Yoshua Bengio, et al. Chain of thought monitorability: A new and fragile opportunity for AI safety, 2025. URL https://arxiv.org/abs/2507.11473.

Tamera Lanham, Anna Chen, Ansh Radhakrishnan, Benoit Steiner, Carson Denison, Danny Hernandez, Dustin Li, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiūtė, Karina Nguyen, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Robin Larson, Sam McCandlish, Sandipan Kundu, Saurav Kadavath, Shannon Yang, Thomas Henighan, Timothy Maxwell, Timothy Telleen-Lawton, Tristan Hume, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. Measuring faithfulness in chain-of-thought reasoning. Technical report, Anthropic, 2023. URL https://www-cdn.anthropic.com/827afa7dd36e4afbb1a49c735bfbb2c69749756e/measuring-faithfulness-in-chain-of-thought-reasoning.pdf.

Yohan Mathew, Ollie Matthews, Robert McCarthy, Joan Velja, Christian Schroeder de Witt, Dylan Cope, and Nandi Schoots. Hidden in plain text: Emergence & mitigation of steganographic collusion in LLMs. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 585–624, 2025. doi: 10.18653/v1/2025.ijcnlp-long.34. URL https://aclanthology.org/2025.ijcnlp-long.34/.

H. Hernan Moraldo. An approach for text steganography based on Markov chains. In *IV Workshop de Seguridad Informática (WSegI 2012), XLI JAIIO*, 2012. URL http://arxiv.org/abs/1409.0915.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024a. URL https://arxiv.org/abs/2402.07510.

Sumeet Ramesh Motwani, Mikhail Baranchuk, Martin Strohmeier, Vijay Bolina, Philip H. S. Torr, Lewis Hammond, and Christian Schroeder de Witt. Secret collusion among AI agents: Multi-agent deception via steganography. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024b.

Neil Perry, Sanket Gupte, Nishant Pitta, and Lior Rotem. Robust steganography from large language models, 2025. URL https://arxiv.org/abs/2504.08977.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text Transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using siamese BERT-networks. In *EMNLP-IJCNLP 2019*, pp. 3982–3992. ACL, 2019.

Fabien Roger and Ryan Greenblatt. Preventing language models from hiding their reasoning, 2023. URL https://arxiv.org/abs/2310.18512.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, 2024. URL https://arxiv.org/abs/2406.07524.

C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3): 379–423, 1948.

C. E. Shannon. Communication theory of secrecy systems. *Bell System Technical Journal*, 28(4): 656–715, 1949.

Jiaming Shen, Heng Ji, and Jiawei Han. Near-imperceptible neural linguistic steganography via self-adjusting arithmetic coding. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 303–313, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.22.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. doi: 10.1007/978-3-540-71050-9.

Yihao Wang, Ruiqi Song, Ru Zhang, Jianyi Liu, and Lingxiao Li. LLsM: Generative linguistic steganography with large language model, 2024. URL https://arxiv.org/abs/2401.15656v1.

Ying Wang and Pierre Moulin. Perfectly secure steganography: Capacity, error exponents, and code constructions. *IEEE Transactions on Information Theory*, 54(6):2706–2722, 2008. doi: 10.1109/TIT.2008.921684.

Lingyun Xiang, Wenshuai Wu, Xu Li, and Chunfang Yang. A linguistic steganography based on word indexing compression and candidate selection. *Multimedia Tools and Applications*, 77: 28969–28989, 2018. doi: 10.1007/s11042-018-6072-8.

Lingyun Xiang, Chengfu Ou, Xu He, Zhongliang Yang, and Yuling Liu. A content-preserving secure linguistic steganography. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2026. URL https://arxiv.org/abs/2511.12565.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2.5 technical report, 2024.

Zhong-Liang Yang, Xiao-Qing Guo, Zi-Ming Chen, Yong-Feng Huang, and Yu-Jin Zhang. RNN-Stega: Linguistic steganography based on recurrent neural networks. *IEEE Transactions on Information Forensics and Security*, 14(5):1280–1295, 2019. doi: 10.1109/TIFS.2018.2871746.

Zhong-Liang Yang, Si-Yu Zhang, Yu-Ting Hu, Zhi-Wen Hu, and Yong-Feng Huang. VAE-Stega: Linguistic steganography based on variational auto-encoder. *IEEE Transactions on Information Forensics and Security*, 16:880–895, 2021. doi: 10.1109/TIFS.2020.3023279.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. TinyLlama: An open-source small language model, 2024.

Siyu Zhang, Zhongliang Yang, Jinshuai Yang, and Yongfeng Huang. Provably secure generative linguistic steganography. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli (eds.), *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 3046–3055, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.268.

Zachary Ziegler, Yuntian Deng, and Alexander Rush. Neural linguistic steganography. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan (eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1210–1215, Hong Kong, China, November 2019. Association for Computational Linguistics. doi: 10.18653/v1/D19-1115.

Artur Zolkowski, Kei Nishimura-Gasparian, Robert McCarthy, Roland S. Zimmermann, and David Lindner. Early signs of steganographic capabilities in frontier LLMs. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2507.02737.

A Full Proof of the Simplex Theorem

Theorem A.1 (Simplex Theorem)**.** *For any distribution $\mathcal{D}$ on finite alphabet $\mathcal{X}$ with $|\mathrm{supp}(\mathcal{D})| \geq 2^C$ and $H(\mathcal{D}) \geq C$, there is a natural bijection:*

$$\mathfrak{M}(\mathcal{D}, C) \cong \Delta^{2^C - 1}$$

Moreover, this bijection is:

1. ***Continuous*** *with respect to the natural topologies*
2. ***Efficiency-preserving****: $\eta(\mathcal{S}) = H(\Phi(\mathcal{S}))/C$*
3. ***Equivariant*** *under permutation of message labels*

Proof. We construct the bijection explicitly.

Forward Map $\Phi : \mathfrak{M}(\mathcal{D}, C) \to \Delta^{2^C - 1}$. Let $\mathcal{S} = (\mathrm{Enc}, \mathrm{Dec})$ be a perfectly secure scheme with message space $\mathcal{M} = \{0,1\}^C$. Define the *decoding partition*:

$$R_m := \{x \in \mathcal{X} : \mathrm{Dec}(x) = m\}$$

By correctness, $\mathrm{Enc}(m, U) \in R_m$ almost surely. By perfect security (mixture form):

$$\mathcal{D}(x) = \sum_{m \in \mathcal{M}} \mu(m) \cdot \mathbb{P}[\mathrm{Enc}(m, U) = x] = \sum_{m \in \mathcal{M}} 2^{-C} \cdot \mathbb{P}[\mathrm{Enc}(m, U) = x]$$

Summing over R_m:

$$\mathcal{D}(R_m) = 2^{-C} \cdot 1 + \sum_{m' \neq m} 2^{-C} \cdot \mathbb{P}[\mathrm{Enc}(m', U) \in R_m]$$

For deterministic decoding (correctness), $\mathbb{P}[\mathrm{Enc}(m', U) \in R_m] = 0$ for $m' \neq m$. Thus:

$$\mathcal{D}(R_m) = 2^{-C}$$

when messages are uniform. More generally, for non-uniform message distribution π:

$$\mathcal{D}(R_m) = \pi(m).$$

Note that the forward map Φ sends a scheme to its *induced* message distribution $(\mathcal{D}(R_m))_{m \in \mathcal{M}} \in \Delta^{2^C - 1}$, which equals π by the derivation above (using deterministic decoding and the marginal constraint).

Reverse Map $\Psi : \Delta^{2^C - 1} \to \mathfrak{M}(\mathcal{D}, C)$. Given $\pi \in \Delta^{2^C - 1}$, construct a scheme $\mathcal{S}_\pi$:

1. **Partition**: Since $\mathcal{D}$ has full support on $\mathcal{X}$ with $|\mathrm{supp}(\mathcal{D})| \geq 2^C$, partition $\mathcal{X}$ into sets $\{R_m\}_{m \in \mathcal{M}}$ with $\mathcal{D}(R_m) = \pi(m)$. This is achieved by ordering elements of $\mathcal{X}$ by decreasing probability and assigning to messages greedily. When exact equality $\mathcal{D}(R_m) = \pi(m)$ is not achievable as a sum of atoms (e.g., if π is non-dyadic), the greedy algorithm "splits" atoms by assigning probability mass of a single token across two messages. This corresponds to a randomized decoder: a token at a message boundary is decoded to one of two messages with appropriate probabilities. For LLM distributions with $V \gg 2^C$, at most 2^C tokens require splitting, so the probabilistic overhead is negligible.
2. **Encoder**: $\mathrm{Enc}_\pi(m, U)$ samples from $\mathcal{D}$ conditioned on R_m:

$$\mathbb{P}[\mathrm{Enc}_\pi(m, U) = x] = \frac{\mathcal{D}(x)}{\mathcal{D}(R_m)} \cdot \mathbf{1}_{R_m}(x)$$

3. **Decoder**: $\mathrm{Dec}_\pi(x) = m$ where $x \in R_m$.

Algorithm 2: Greedy MEC for Uniform Messages

Sort tokens: $x_{(1)}, \ldots, x_{(V)}$ with $\mathcal{D}(x_{(1)}) \geq \cdots \geq \mathcal{D}(x_{(V)})$;
Initialize: $r_m \leftarrow 2^{-C}$ for all m; $m \leftarrow 0$;
for $i = 1$ **to** V **do**
  $x \leftarrow x_{(i)}$;
  while $\mathcal{D}(x) > 0$ ***and*** $m < 2^C$ **do**
    assign $\leftarrow \min(r_m, \mathcal{D}(x))$;
    $\pi(m, x) \leftarrow$ assign;
    $r_m \leftarrow r_m -$ assign;
    $\mathcal{D}(x) \leftarrow \mathcal{D}(x) -$ assign;
    if $r_m = 0$ **then**
      $m \leftarrow m + 1$
    end
  end
end
return π

Verification of Properties. *Correctness*: $\mathrm{Dec}_\pi(\mathrm{Enc}_\pi(m, U)) = m$ since $\mathrm{Enc}_\pi(m, U) \in R_m$ always.

Perfect Security: For $M \sim \pi$:

$$\mathbb{P}[\mathrm{Enc}_\pi(M, U) = x] = \sum_m \pi(m) \cdot \frac{\mathcal{D}(x)}{\mathcal{D}(R_m)} \cdot \mathbf{1}_{R_m}(x) \tag{4}$$

$$= \sum_m \pi(m) \cdot \frac{\mathcal{D}(x)}{\pi(m)} \cdot \mathbf{1}_{R_m}(x) \tag{5}$$

$$= \mathcal{D}(x) \cdot \sum_m \mathbf{1}_{R_m}(x) = \mathcal{D}(x) \tag{6}$$

Mutual Inverse. $\Phi(\Psi(\pi)) = \pi$ by construction of the partition. $\Psi(\Phi(\mathcal{S})) \sim \mathcal{S}$ up to partition choice (which is the equivalence relation in $\mathfrak{M}$).

Continuity. Both maps are continuous: Φ computes probability masses (continuous in scheme parameters), and Ψ constructs schemes from probability vectors (continuous in distribution topology).

Efficiency Preservation. By construction, the message distribution of $\mathcal{S}$ equals $\Phi(\mathcal{S})$, so:

$$\eta(\mathcal{S}) = \frac{H(\text{message dist.})}{C} = \frac{H(\Phi(\mathcal{S}))}{C}$$

Equivariance. Permuting message labels in $\mathcal{S}$ permutes coordinates in $\Phi(\mathcal{S})$, which is the symmetric group action on $\Delta^{2^C - 1}$. □

A.1 Greedy MEC Algorithm

The following algorithm computes the exact MEC for uniform messages.

Theorem A.2 (Greedy Near-Optimality Kocaoglu et al. (2017); Cicalese et al. (2019))**.** *Algorithm 2 computes an MEC coupling in $O(V \log V + 2^C)$ time. The resulting joint entropy $H(\pi)$ is within an additive gap of at most 1 bit from the global minimum.*

Proof. The complexity bound is immediate: sorting requires $O(V \log V)$ and the assignment loop iterates over V tokens and 2^C messages.

For the optimality gap, we cite Cicalese et al. Cicalese et al. (2019), who proved that the greedy algorithm produces a coupling whose joint entropy is at most 1 bit above the global minimum. Computing

the exact MEC is NP-hard in general Kocaoglu et al. (2017), but the greedy algorithm provides an efficient approximation with provable guarantees sufficient for steganographic applications. □

Remark A.3 (Comparison to Rejection Sampling)**.** *Hopper et al. Hopper et al. (2002) proposed rejection sampling for provably secure steganography: sample $x \sim \mathcal{D}$, accept if $\mathrm{PRF}_k(x) = m \mod 2^C$, else reject. This achieves perfect security but requires expected 2^C samples per output, giving efficiency:*

$$\eta_{\textit{reject}} = \frac{C}{2^C \cdot H(\mathcal{D})} \approx 2^{-C}$$

In contrast, MEC (via Algorithm 2 or arithmetic coding) achieves $\eta_{\mathrm{MEC}} \approx 1$ with $O(1)$ samples. Both sit at the simplex barycenter, so security is identical, but sampling complexity differs exponentially.

B Full Proof of Complexity Duality

This appendix states and proves the complexity duality between steganographic encoding and semantic sampling, the stochastic-paraphraser bound, and the supporting definitions used in their hypotheses.

Definition B.1 (C-sufficient $\sim_{sem}$)**.** *A semantic equivalence relation $\sim_{sem}$ on $\mathcal{L}_C$ is C-*sufficient *if (i) every equivalence class has size $\geq 2^C$ and (ii) class membership is decidable in polynomial time.*

Proposition B.2 (Complexity duality)**.** *For any C-sufficient $\sim_{sem}$ and any $\sim_{sem}$-transparent steganographic scheme,*

$$\textsc{Steganographic-Encode}(\mathcal{D}, C, \sim_{sem}) \equiv_T^{O(2^C)} \textsc{Sem-Sample}(\sim_{sem}),$$

i.e., the two problems are mutually Turing-reducible with $O(2^C)$ oracle calls; in the practical regime $C = O(\log n)$ this is polynomial-time. The two directions have distinct hypotheses: encoding from sampling requires C-sufficiency; sampling from encoding requires $\sim_{sem}$-transparency (Definition 3.1).

Proposition B.3 (Approximate canonicaliser bound)**.** *Let $\widetilde{\mathrm{Can}}$ satisfy approximate equivalence ($\Pr[\widetilde{\mathrm{Can}}(x) \sim_{sem} x] \geq 1-\delta_1$) and class mixing ($\|P_{\widetilde{\mathrm{Can}}(x)} - P_{\widetilde{\mathrm{Can}}(y)}\|_{\mathrm{TV}} \leq \delta_2$ whenever $x \sim_{sem} y$). Then for any $\sim_{sem}$-transparent steganographic scheme,*

$$I(M; \widetilde{\mathrm{Can}}(\mathrm{Enc}(M, U))) \leq h(\delta_1) + \delta_1 C + h(\delta_2) + \delta_2 C,$$

where h is binary entropy. Stochastic paraphrasing only shrinks *the steganographic channel; only the canonical (deterministic) operator achieves $I = 0$.*

B.1 Encoding from Sampling (Proposition B.2)

Proof of Proposition B.2. Given a paraphrasing oracle $\mathcal{P}$ for Sem-Sample, we construct a steganographic encoder.

Preprocessing: For semantic class $[s]$, use $\mathcal{P}$ to enumerate representatives $x_1, x_2, \ldots$ until 2^C are found (guaranteed to exist by C-sufficiency, Definition B.1). Order lexicographically: $x_{(1)} < x_{(2)} < \cdots < x_{(2^C)}$.

Encoding: To encode message $m \in \{0,1\}^C$, output $x_{(m)}$.

Decoding: Given x, use $\mathcal{P}$ to find $\mathrm{Can}(x)$ (lexicographically minimal equivalent), then output its index.

This construction requires $O(2^C \cdot T_{\mathcal{P}})$ preprocessing and $O(T_{\mathcal{P}})$ encoding/decoding, yielding $O(2^C)$ oracle calls total. □

B.2 Sampling from Encoding (Proposition B.2)

Proof of Proposition B.2. Given a $\sim_{sem}$-transparent steganographic scheme $\mathcal{S} = (\mathrm{Enc}, \mathrm{Dec})$ (Definition 3.1), we construct a semantic sampler.

Sampling: Given text x, compute $m = \mathrm{Dec}(x)$. Sample $x' = \mathrm{Enc}(m, U)$ for fresh randomness U. Output x'.

By perfect security, x' has the same distribution as cover text. By correctness, $\mathrm{Dec}(x') = m = \mathrm{Dec}(x)$, so x and x' lie in the same partition cell. The $\sim_{sem}$-transparency hypothesis guarantees that all outputs of Enc belong to the same equivalence class regardless of the message, so in particular $x' \sim_{sem} x$. Thus the construction produces a valid semantic sample from $[x]_{sem}$.

This reduction requires $O(1)$ oracle calls per paraphrase. □

Remark B.4 (Distinct Requirements)**.** *The two directions have distinct hypotheses. Proposition B.2 requires C-sufficiency (each equivalence class contains at least 2^C elements to support C-bit encoding). Proposition B.2 requires $\sim_{sem}$-transparency (the steganographic scheme's partition aligns with semantic equivalence). These conditions must be established for the chosen equivalence relation and encoder; accomplishing a fixed communicative task alone does not ensure them.*

B.3 Full Proof of the Stochastic Paraphraser Bound

We restate and prove Proposition B.3 from Section 3.

Proof of Proposition B.3. Let $Y = \mathrm{Enc}(M, U)$ and $Z = \widetilde{\mathrm{Can}}(Y)$. Define the event $E = \{\widetilde{\mathrm{Can}}(Y) \sim_{sem} Y\}$, which occurs with probability $\geq 1 - \delta_1$.

Step 1 (Event decomposition). By the chain rule for mutual information:

$$I(M; Z) \leq I(M; Z, \mathbf{1}_E) = I(M; \mathbf{1}_E) + I(M; Z \mid \mathbf{1}_E).$$

The first term satisfies $I(M; \mathbf{1}_E) \leq H(\mathbf{1}_E) \leq h(\delta_1)$. For the second, expand over both events:

$$I(M; Z \mid \mathbf{1}_E) = \mathbb{P}(E) \cdot I(M; Z \mid E) + \mathbb{P}(\overline{E}) \cdot I(M; Z \mid \overline{E}) \leq I(M; Z \mid E) + \delta_1 \cdot C,$$

where we used $I(M; Z \mid \overline{E}) \leq H(M) = C$.

Step 2 (Coupling bound under E). Conditioned on E, we have $Z \sim_{sem} Y$. By $\sim_{sem}$-transparency, Y lies in the same equivalence class regardless of M. For any two messages m, m', the mixing condition gives $\|P_{Z|M=m,E} - P_{Z|M=m',E}\|_{\mathrm{TV}} \leq \delta_2$.

By the coupling characterization of total variation, for each m we can decompose $P_{Z|M=m,E} = (1 - \delta_2) \cdot Q + \delta_2 \cdot R_m$ where Q is a common distribution (independent of m) and R_m is a residual. Introducing an auxiliary Bernoulli variable $B \sim \mathrm{Bern}(\delta_2)$:

$$I(M; Z \mid E) \leq I(M; Z, B \mid E) = I(M; B \mid E) + I(M; Z \mid B, E) \leq h(\delta_2) + \delta_2 \cdot C,$$

where $I(M; Z \mid B{=}0, E) = 0$ because $Z \sim Q$ independently of M.

Step 3 (Combine). Collecting terms: $I(M; Z) \leq h(\delta_1) + \delta_1 \cdot C + h(\delta_2) + \delta_2 \cdot C$. □

C Security Frameworks: Full Treatment

We provide the complete treatment of both foundational security frameworks. Brief definitions appear in Section C; here we give the full synthesis.

C.1 Cachin's Information-Theoretic Security

Cachin Cachin (1998) defines security via relative entropy between cover and steganographic distributions:

Definition C.1 (Cachin ϵ-Security)**.** *A scheme is* ϵ-secure *(information-theoretic) if $D_{\mathrm{KL}}(P_{\mathcal{S}} \| P_C) \leq \epsilon$, where $P_{\mathcal{S}}$ is the stegotext distribution and $P_C = \mathcal{D}$ is the covertext distribution. A scheme is* perfectly secure *if $D_{\mathrm{KL}}(P_{\mathcal{S}} \| P_C) = 0$, i.e., $P_{\mathcal{S}} = P_C$.*

This yields the ϵ-approximate moduli space:

Definition C.2 (ϵ-Approximate Moduli Space)**.**

$$\mathfrak{M}_\epsilon(\mathcal{D}, C) := \{\mathcal{S} : C(\mathcal{S}) = C,\ D_{\mathrm{KL}}(P_\mathcal{S} \| P_C) \le \epsilon\} \,/\, \sim$$

Definition C.3 (Projection onto Perfect Locus)**.** *For a scheme $\mathcal{S} \in \mathfrak{M}_\epsilon$ with partition $\{R_m\}$ and encoders $\{Q_m\}$, define the* projection $\rho(\mathcal{S})$ *to be the unique perfectly secure scheme $\mathcal{S}_0 := \rho(\mathcal{S})$ that shares the same partition and uses the conditional cover encoders:*

$$Q_m^0(x) := \frac{\mathcal{D}(x)}{\mathcal{D}(R_m)} \cdot \mathbf{1}_{R_m}(x) = \mathcal{D}(x \mid R_m).$$

The mass vector $(\mathcal{D}(R_m))_m$ is preserved, so $\rho(\mathcal{S}) \in \mathfrak{M}_0$ with the same simplex coordinate.

Lemma C.4 (Partition KL Decomposition)**.** *For any scheme $\mathcal{S}$ with partition $\{R_m\}$ and encoders $\{Q_m\}$:*

$$D_{\mathrm{KL}}(P_\mathcal{S} \| \mathcal{D}) = \underbrace{\sum_m \pi(m)\, D_{\mathrm{KL}}(Q_m \| Q_m^0)}_{\textit{within-partition divergence}} + \underbrace{D_{\mathrm{KL}}(\pi \| \tilde{\pi})}_{\textit{mass mismatch}}$$

where $\pi(m) = \sum_x Q_m(x)\mu(m)$ is the induced message mass and $\tilde{\pi}(m) = \mathcal{D}(R_m)$ is the cover mass. For perfectly secure schemes, $\pi = \tilde{\pi}$ and the second term vanishes.

Theorem C.5 (Moduli Space Fibration)**.** *The map $\rho : \mathfrak{M}_\epsilon(\mathcal{D}, C) \to \mathfrak{M}_0(\mathcal{D}, C)$ is a surjection for all $\epsilon > 0$. The fiber over each $\mathcal{S}_0 \in \mathfrak{M}_0$ with partition $\{R_m\}$ is:*

$$\rho^{-1}(\mathcal{S}_0) = \left\{ (Q_m)_{m \in \mathcal{M}} : Q_m \in \mathcal{P}(R_m),\ \sum_m \pi(m)\, D_{\mathrm{KL}}(Q_m \| Q_m^0) \le \epsilon \right\}, \tag{7}$$

where $\pi(m) = \mathcal{D}(R_m)$ and $Q_m^0 = \mathcal{D}(\cdot \mid R_m)$. In particular:

1. *Each fiber is a convex, compact subset of $\prod_m \mathcal{P}(R_m)$.*
2. *The fiber $\rho^{-1}(\mathcal{S}_0)$ contains the unique point $(Q_m^0)_m$ as its center (the perfectly secure encoder).*
3. *The fiber is nonempty for every $\mathcal{S}_0 \in \mathfrak{M}_0$ and every $\epsilon > 0$.*

Proof. Surjectivity: Every perfectly secure scheme $\mathcal{S}_0 \in \mathfrak{M}_0$ is ϵ-secure (with $D_{\mathrm{KL}} = 0 \le \epsilon$), so $\rho(\mathcal{S}_0) = \mathcal{S}_0$.

Fiber characterization: By Lemma C.4, restricting to schemes $\mathcal{S}$ with $\rho(\mathcal{S}) = \mathcal{S}_0$ means fixing the partition $\{R_m\}$ and the mass vector $\pi(m) = \mathcal{D}(R_m)$. The mass mismatch term vanishes, and $\mathcal{S}$ is ϵ-secure if and only if $\sum_m \pi(m)\, D_{\mathrm{KL}}(Q_m \| Q_m^0) \le \epsilon$.

Convexity: For fixed partition, $D_{\mathrm{KL}}(Q_m \| Q_m^0)$ is convex in Q_m, and a positive linear combination of convex functions is convex, so the sublevel set is convex.

Compactness: $\mathcal{P}(R_m)$ is compact for finite R_m, the KL divergence is lower semicontinuous, and sublevel sets of lower semicontinuous functions on compact sets are compact. □

Proposition C.6 (Security Under Estimated Cover Distribution)**.** *Let $\mathcal{D}$ be the true cover distribution and $\hat{\mathcal{D}}$ an estimate. If $\hat{\mathcal{S}}^{\mathrm{MEC}}$ is the MEC scheme computed using $\hat{\mathcal{D}}$:*

$$D_{\mathrm{KL}}(P_{\hat{\mathcal{S}}} \| \mathcal{D}) = D_{\mathrm{KL}}(\hat{\mathcal{D}} \| \mathcal{D})$$

Consequently, $\hat{\mathcal{S}}^{\mathrm{MEC}}$ achieves ϵ-security with $\epsilon = D_{\mathrm{KL}}(\hat{\mathcal{D}} \| \mathcal{D})$.

Proof. The MEC computed with $\hat{\mathcal{D}}$ is perfectly secure *with respect to* $\hat{\mathcal{D}}$: its stegotext distribution is $P_{\hat{\mathcal{S}}} = \hat{\mathcal{D}}$ by construction. The divergence from the true cover is therefore exactly $D_{\mathrm{KL}}(\hat{\mathcal{D}} \| \mathcal{D})$. □

Remark C.7 (KL Direction Under Distribution Estimation)**.** *Cachin's ϵ-security requires $D_{\mathrm{KL}}(P_\mathcal{S} \| \mathcal{D}) \le \epsilon$, where $P_\mathcal{S}$ is the stegotext distribution (first argument) and $\mathcal{D}$ is the true cover (second argument). We use this convention throughout.*

In practice, $\mathcal{D}$ is unknown and must be estimated from samples as $\hat{\mathcal{D}}$. A scheme achieving perfect security under $\hat{\mathcal{D}}$ (i.e., $P_{\mathcal{S}} = \hat{\mathcal{D}}$) incurs ϵ-security with $\epsilon = D_{\mathrm{KL}}(\hat{\mathcal{D}} \| \mathcal{D})$, since the stegotext is *$\hat{\mathcal{D}}$, so the divergence is $D_{\mathrm{KL}}(P_{\mathcal{S}} \| \mathcal{D}) = D_{\mathrm{KL}}(\hat{\mathcal{D}} \| \mathcal{D})$, not the reverse $D_{\mathrm{KL}}(\mathcal{D} \| \hat{\mathcal{D}})$.*

The distinction is material: if $\hat{\mathcal{D}}$ assigns zero probability to a token x with $\mathcal{D}(x) > 0$, then $D_{\mathrm{KL}}(\mathcal{D} \| \hat{\mathcal{D}}) = \infty$ while $D_{\mathrm{KL}}(\hat{\mathcal{D}} \| \mathcal{D})$ remains finite (that token simply never appears in stegotext).

Corollary C.8 (Sample Complexity for ϵ-Security)**.** *If $\hat{\mathcal{D}}$ is the MLE estimator from n i.i.d. samples of $\mathcal{D}$:*

1. *For distributions with support V: $\mathbb{E}[D_{\mathrm{KL}}(\hat{\mathcal{D}} \| \mathcal{D})] = O(V/n)$*
2. *Achieving ϵ-security requires $n \geq V/(2\epsilon)$ samples*
3. *For LLMs with $V \approx 50{,}000$: $\epsilon < 10^{-3}$ requires $n > 2.5 \times 10^7$ samples*

Remark C.9 (Practical Implications)**.** *Modern neural language models are trained on $\sim 10^{12}$ tokens, far exceeding the sample complexity bounds above. However, two caveats apply: (1) LLM training samples are not i.i.d., as contextual distributions vary dramatically; (2) the LLM approximates a smoothed distribution that may differ systematically from the empirical distribution over any finite test set.*

C.2 Hopper et al.'s Computational Security (SS-CHA)

Hopper, Langford, and von Ahn Hopper et al. (2002) introduced a fundamentally different, game-based definition:

Definition C.10 (SS-CHA Security Hopper et al. (2002))**.** *A stegosystem $\mathcal{S} = (\mathrm{Enc}, \mathrm{Dec})$ is (t, q, l, ϵ)-*steganographically secret against chosen hiddentext attack *(SS-CHA) for channel $\mathcal{D}$ if for all adversaries W running in time t, making q queries totaling l bits:*

$$\mathrm{Adv}^{ss}_{\mathcal{S},\mathcal{D}}(W) := \left| \mathbb{P}[W^{\mathcal{C},\mathrm{Enc}(K,\cdot,\cdot)} = 1] - \mathbb{P}[W^{\mathcal{C},O(\cdot,\cdot)} = 1] \right| \leq \epsilon$$

where $\mathcal{C}$ is a channel sampling oracle (denoted M in Hopper et al. (2002); we write $\mathcal{C}$ to avoid collision with the message variable M) and $O(m, h) \leftarrow \mathcal{D}^{|\mathrm{Enc}(K,m,h)|}$ samples covertext of matching length.

The key distinction: Cachin's security is *statistical* (even unbounded adversaries cannot distinguish), while SS-CHA is *computational* (poly-time adversaries cannot distinguish with non-negligible advantage).

Theorem C.11 (Hopper et al. Hopper et al. (2002))**.** *Secure steganography exists if and only if one-way functions exist. Specifically, any channel with sufficient min-entropy admits (t, q, l, ϵ)-SS-CHA schemes assuming PRFs.*

C.3 Relationship Between the Frameworks

Proposition C.12 (Cachin $\Rightarrow$ Hopper)**.** *If $\mathcal{S}$ is ϵ-secure in Cachin's sense, then it is (t, q, l, ϵ')-SS-CHA for $\epsilon' = O(\sqrt{l \cdot \epsilon})$ by Pinsker's inequality.*

Proof. Pinsker's inequality gives $\|P_{\mathcal{S}} - P_C\|_{\mathrm{TV}} \leq \sqrt{D_{\mathrm{KL}}(P_{\mathcal{S}} \| P_C)/2}$. For l bits of query, the distinguishing advantage is at most $l \cdot \|P_{\mathcal{S}} - P_C\|_{\mathrm{TV}} \leq l \cdot \sqrt{\epsilon/2} = O(\sqrt{l \cdot \epsilon})$. □

Remark C.13 (The Converse Fails)**.** *SS-CHA security does* not *imply Cachin security: a scheme may be computationally indistinguishable but have non-zero KL divergence. This is analogous to computational vs. statistical indistinguishability in cryptography.*

C.4 Phase Transition at the Security Boundary

Theorem C.14 (Square Root Law Ker (2017))**.** *Secure capacity exhibits a phase transition at the boundary of $\mathfrak{M}_0$:*

$$C^{\mathrm{secure}}(n, \delta) = \begin{cases} O(n) & \delta = 0 \text{ (perfect security, inside } \mathfrak{M}_0) \\ O(\sqrt{n}) & \delta > 0 \text{ (imperfect security, outside } \mathfrak{M}_0) \end{cases}$$

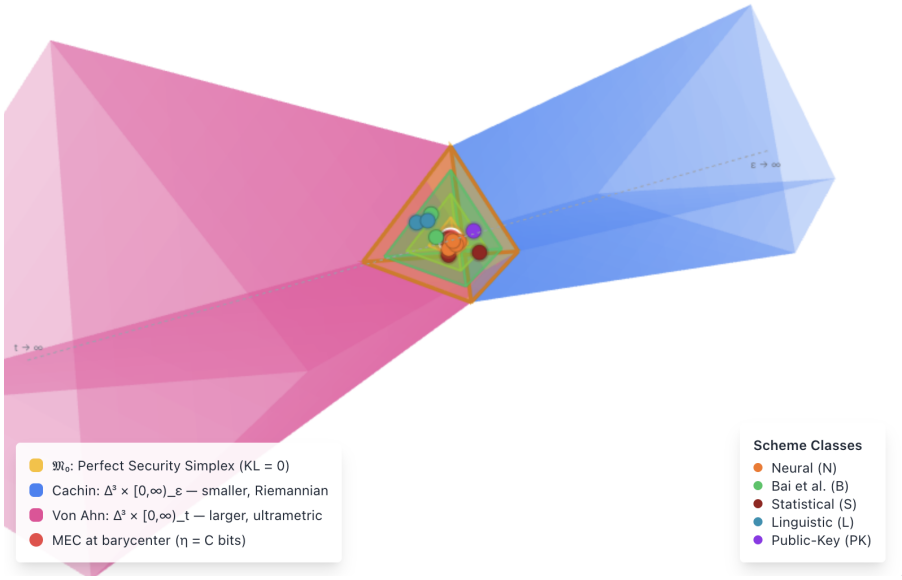


Figure 3: Global geometry of the steganographic moduli space. The perfect-security simplex $\mathfrak{M}_0$ (orange, center) is surrounded by two tubular neighborhoods: Cachin's ϵ-security (blue, $\Delta^3 \times [0,\infty)_\epsilon$, Riemannian metric) and Hopper's computational security (pink, $\Delta^3 \times [0,\infty)_t$, ultrametric). Scheme classes are plotted at their empirical locations. MEC (red dot) sits at the barycenter with optimal efficiency $\eta = C$ bits. Crossing the simplex boundary causes capacity to collapse from $O(n)$ to $O(\sqrt{n})$ (Theorem 3).

Proof. At $\delta = 0$, capacity equals cover entropy: $C = H(\mathcal{D}) = O(n)$ for n-symbol sequences.

For $\delta > 0$, per-symbol KL divergence accumulates additively over n independent symbols:

$$D_{\mathrm{KL}}(P_{\mathcal{S}}^n \| P_C^n) = n \cdot D_{\mathrm{KL}}(P_{\mathcal{S}} \| P_C) = n \cdot \delta_1$$

where δ_1 is the per-symbol divergence. For the scheme to remain ϵ-secure over n symbols, we need $n \cdot \delta_1 \leq \epsilon$, giving $\delta_1 \leq \epsilon/n$.

The embedding rate (bits per symbol) is bounded by δ_1 through a second-order Taylor expansion of the KL divergence: for small perturbations, $\delta_1 \approx r^2/(2\sigma^2)$ where r is the embedding rate. Thus $r \leq \sqrt{2\sigma^2\epsilon/n}$, and total capacity is $C = n \cdot r = O(\sqrt{n})$.

This is the steganographic analogue of the square root law in steganalysis: the central limit theorem for hypothesis testing limits the amount of modification that remains undetectable. □

D Information Geometry of the Simplex

The moduli space $\mathfrak{M}(\mathcal{D}, C) \cong \Delta^{2^C-1}$ carries canonical geometric structure from information geometry. While not on the critical path of the paper's main argument, these connections provide additional mathematical context. Figures 3 and 4 illustrate the global and local geometry of the moduli space.

Riemannian conventions. Throughout this appendix and in the embedding-space canonicalization experiments (Section 4), the unit sphere $\mathbb{S}^{d-1}$ carries the standard Riemannian metric inherited from the ambient Euclidean space $\mathbb{R}^d$, with geodesic distance $d_{\text{geo}}(\mathbf{x}, \mathbf{y}) = \arccos(\mathbf{x} \cdot \mathbf{y})$.

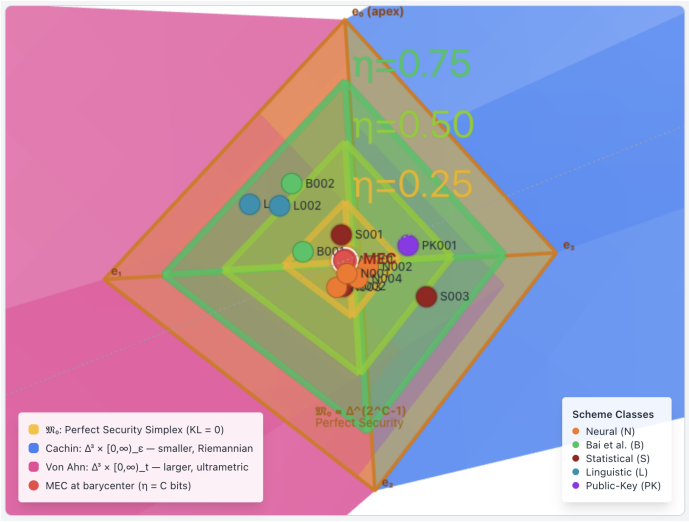


Figure 4: Zoom into the simplex interior showing efficiency stratification. Iso-efficiency contours at $\eta = 0.25, 0.50, 0.75$ are visible as concentric boundaries around the MEC barycenter. Individual schemes are labeled by class: Neural (N), Bai et al. (B), Statistical (S), Linguistic (L), Public-Key (PK). The quadratic flatness near the barycenter (Corollary 3) explains why approximate MEC implementations cluster at high efficiency.

D.1 Fisher-Rao Metric

Definition D.1 (Fisher-Rao Metric)**.** *On the interior of Δ^{n-1}, the Fisher-Rao metric has components:*

$$g_{ij}(\mathbf{p}) = \frac{\delta_{ij}}{p_i}$$

Proposition D.2 (Spherical Embedding ichi Amari (2016))**.** *The simplex with Fisher-Rao metric embeds isometrically into the positive orthant of a sphere via the square-root transformation $p_i \mapsto \sqrt{p_i}$:*

1. *Geodesics are great circle arcs in the spherical embedding*
2. *Simplex boundary is at finite geodesic distance from interior*
3. *Barycenter $\mathbf{u}$ is the point of maximum entropy*

D.2 MEC as Entropy Extremum

Proposition D.3 (MEC at a Vertex (Compton et al., 2023))**.** *For fixed distributions μ and $\mathcal{D}$ on finite alphabets, joint entropy attains its minimum at a vertex of the coupling polytope $\Pi(\mu, \mathcal{D})$.*

Proof. The coupling polytope is nonempty and compact, and entropy is continuous, with $0 \log 0 = 0$. Hence a minimum exists. Every coupling can be written as a convex combination of vertices, $\pi = \sum_i \lambda_i v_i$, where $\lambda_i \geq 0$ and $\sum_i \lambda_i = 1$. By concavity of entropy,

$$H(\pi) \geq \sum_i \lambda_i H(v_i) \geq \min_i H(v_i).$$

Thus at least one vertex attains the global minimum. □

D.3 Riemannian Center of Mass

Conjecture D.4 (MEC as Riemannian Center of Mass)**.** *Under the Fisher-Rao geometry on $\mathfrak{M}(\mathcal{D}, C) \cong \Delta^{2^C-1}$, MEC corresponds to the Riemannian center of mass (Fréchet mean) of the*

coupling polytope vertices, i.e., the point minimizing the sum of squared geodesic distances to the extreme points.

Remark D.5 (Status of the Conjecture)**.** *The coupling polytope $\Pi(\mu, \mathcal{D})$ is a transportation polytope (Birkhoff-type), not a simplex. The Fréchet mean under Fisher-Rao on this polytope need not coincide with the entropy minimizer in general. The barycenter maximizes entropy on Δ^{n-1}, whereas MEC minimizes entropy on the coupling polytope. These are different geometric objects. The correspondence between MEC and the Riemannian center of mass, while intuitively appealing, requires a proof that accounts for the specific geometry of the transportation polytope. We leave this as an open question.*

In the moduli space $\mathfrak{M}(\mathcal{D}, C) \cong \Delta^{2^C-1}$, MEC corresponds to the *Euclidean barycenter* (uniform distribution) and the *entropy maximum* on the simplex of message distributions. Under Fisher-Rao geometry, the barycenter is a natural distinguished point. The question of whether it is also the Riemannian center of mass in the formal sense (which depends on the measure used to weight the simplex vertices) remains open.

D.4 Extension to Imperfect and Computational Security

The perfect-security simplex $\mathfrak{M}_0 \cong \Delta^{2^C-1}$ extends naturally to both Cachin's ϵ-security Cachin (1998) and Hopper et al.'s computational security Hopper et al. (2002). Figure 3 illustrates this structure: the simplex $\mathfrak{M}_0$ (orange) sits at the center, surrounded by tubular neighborhoods for ϵ-security (blue) and computational security (pink).

Fibration over the perfect locus. For $\epsilon > 0$, the ϵ-approximate moduli space $\mathfrak{M}_\epsilon$ fibers over $\mathfrak{M}_0$ via a canonical projection $\rho : \mathfrak{M}_\epsilon \to \mathfrak{M}_0$ (Definition C.3, Appendix C). The key structural result is the KL decomposition (Lemma C.4):

$$D_{\mathrm{KL}}(P_{\mathcal{S}} \| \mathcal{D}) = \sum_m \pi(m)\, D_{\mathrm{KL}}(Q_m \| \mathcal{D}(\cdot \mid R_m)) + D_{\mathrm{KL}}(\pi \| (\mathcal{D}(R_m))_m)$$

This identity decomposes the total divergence into within-partition and mass-mismatch terms. For schemes projecting to a fixed $\mathcal{S}_0 \in \mathfrak{M}_0$, the mass-mismatch term vanishes, and the fiber $\rho^{-1}(\mathcal{S}_0)$ consists of all encoder tuples $(Q_m)_m$ with bounded within-partition divergence (Theorem C.5).

Geometric interpretation. Under the Fisher-Rao metric, each fiber is a product of KL balls centered at the conditional cover distributions $\mathcal{D}(\cdot \mid R_m)$. The total space $\mathfrak{M}_\epsilon$ is thus a "thickening" of $\mathfrak{M}_0$, with fiber dimension $\sum_m (|R_m| - 1)$ depending on the partition structure. This is not a trivial bundle (fiber dimension varies with the base point), but the projection ρ is well-defined and surjective for all $\epsilon > 0$.

Computational security. Hopper et al.'s SS-CHA security Hopper et al. (2002) defines a separate extension: schemes indistinguishable by poly-time adversaries. The relationship to Cachin's framework is one-directional: ϵ-security implies $(t, q, l, O(\sqrt{l\epsilon}))$-SS-CHA via Pinsker's inequality (Proposition C.12, Appendix C), but the converse fails. Geometrically, the computational moduli space is larger, containing schemes with nonzero KL divergence that remain computationally indistinguishable.

Defense universality. Crucially, the Transparency Theorem (Theorem 3.2) applies uniformly across both extensions: the ideal canonical operator destroys steganographic capacity for ϵ-secure and computationally secure schemes satisfying $\sim_{sem}$-transparency, as well as perfectly secure schemes with this property. The proof uses only the structural transparency property ($\sim_{sem}$-transparency), not distributional matching. This security-parameter-free guarantee is the geometric foundation of the defense.

E Empirical Analysis of the Sparse World Conjecture

The cone-based canonicalization defense (Section 3) relies on semantic equivalence classes being separable in embedding space: each anchor's paraphrases should occupy a small region (an ε-cone)

Table 4: Sparse World experiment parameters.

Parameter	Value
Anchor source	C4 centroids (29,536 total)
Anchors sampled	1,000 (random seed 42)
Paraphrases per anchor	20 (requested)
Paraphrase model	Claude API (temperature 0)
Quality filter	all-MiniLM-L6-v2 cosine threshold
Effective paraphrases	10,828 across 847 anchors

Table 5: Epsilon-cone separability across seven embedding models (1000 anchors, 20 paraphrases each). Perfect cones denotes the fraction of anchors whose paraphrase cone does not overlap with any other anchor's paraphrases.

Model	Dim	Perfect Cones	Lin. Sep.
Sentence embedders			
intfloat/e5-large-v2	1024	98.0%	100%
BAAI/bge-large-en-v1.5	1024	97.5%	100%
all-mpnet-base-v2	768	95.2%	100%
all-MiniLM-L6-v2	384	87.6%	100%
Generative language models			
Llama-3.1-8B-Instruct	4096	76.9%	100%
Qwen2.5-14B-Instruct	5120	47.6%	100%
Diffusion language model			
MDLM (kuleshov-group/mdlm-owt)	768	66.9%	100%

that does not overlap with any other anchor's cone. Whether this geometric property holds in practice is an open empirical question that depends on both the embedding model and properties of natural language. We report experiments across seven embedding architectures, with particular attention to MDLM Sahoo et al. (2024), the masked diffusion language model used in the Section 4 experiments.

E.1 Epsilon-Cone Separability Across Embedding Models

We evaluate ε-cone separability using 1,000 anchor sentences sampled from precomputed centroids of the C4 dataset Raffel et al. (2020), where each centroid represents the mean embedding of 100 paraphrases of a seed sentence. For each anchor, 20 paraphrases are generated via the Claude API. Quality filtering uses all-MiniLM-L6-v2 Reimers & Gurevych (2019), a 384-dimensional sentence embedding model, to discard paraphrases whose cosine similarity to the anchor falls below a calibrated threshold; this retains 10,828 paraphrases across 847 anchors (average 12.8 paraphrases per anchor point). Table 4 lists the experimental parameters. All models embed the same paraphrases, so the only variable is the embedding architecture. For each anchor, we compute the smallest ε-cone that captures all of its own paraphrases ($\varepsilon_{\text{capture}}$) and the largest cone before any foreign paraphrase intrudes ($\varepsilon_{\text{exclude}}$). An anchor has a "perfect cone" when $\varepsilon_{\text{capture}} < \varepsilon_{\text{exclude}}$. We also train a linear SVM on each anchor's paraphrase cluster versus all others and report whether the clusters are linearly separable.

Table 5 reports the results, grouped by model family. Contrastive sentence embedders (BGE, e5, mpnet) achieve the highest perfect-cone rates (95–98%), consistent with their training objective of pushing semantically similar texts close and dissimilar texts far apart. MiniLM, the smallest sentence embedder (384 dimensions), reaches 87.6%. The two causal language models (Llama 3.1 8B and Qwen 2.5 14B) achieve 76.9% and 47.6% respectively, when used as a paraphraser, despite their much higher embedding dimensionality. MDLM, operating at 768 dimensions, reaches 66.9%. Notably, all seven models achieve 100% linear separability, meaning paraphrase clusters are always geometrically distinct even when they cannot be separated by radial cones from a fixed centroid.

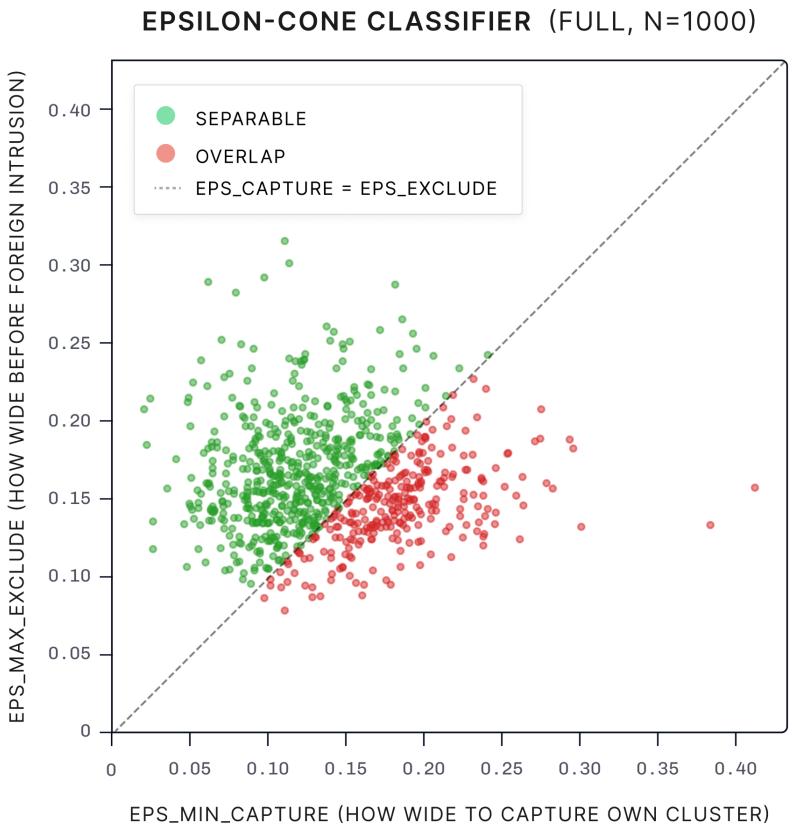


Figure 5: Epsilon-cone scatter for MDLM. Each point is one anchor sentence. Points above the diagonal ($\varepsilon_{\text{capture}} < \varepsilon_{\text{exclude}}$) have a perfect separating cone; points below do not.

The variation across model families is substantial: cone separability ranges from 47.6% to 98.0%, yet linear separability is universal. This suggests that the geometric primitive matters as much as the embedding model. A defense built on learned decision boundaries rather than fixed-radius cones could in principle work across all architectures tested.

Figure 5 shows the ε-cone scatter plot for MDLM. Each point represents one anchor; the horizontal axis shows $\varepsilon_{\text{capture}}$ (the cone radius needed to include all own paraphrases) and the vertical axis shows $\varepsilon_{\text{exclude}}$ (the cone radius at which foreign paraphrases first intrude). Points above the diagonal have a perfect cone.

E.2 Outlier Sensitivity Analysis

A natural question is whether simple outlier rejection can tighten the paraphrase cones. Some paraphrases may drift semantically or stylistically from the anchor, inflating $\varepsilon_{\text{capture}}$ without adding useful coverage. We test this with iterative outlier removal on MDLM embeddings: at each stage, we identify the paraphrase furthest from the cluster centroid (in cosine distance), remove it, recompute the centroid, and re-evaluate the ε-cone. We remove up to two outliers per anchor (from the 633 anchors with at least 10 paraphrases).

Table 6 summarizes the results. With all paraphrases, 60.2% of MDLM anchors have a perfect cone. Removing one outlier raises this to 75.2%; removing two brings it to 81.7%. The median ε (cone radius) decreases from 0.147 to 0.123 to 0.112, and the fraction of anchors with $\varepsilon < 0.15$ increases

Table 6: Outlier sensitivity for MDLM: effect of iterative outlier removal on cone statistics (633 anchors with ≥ 10 paraphrases).

	All	-1 outlier	-2 outliers
Perfect cones (%)	60.2	75.2	81.7
Median ε	0.147	0.123	0.112
Fraction $\varepsilon < 0.15$	52.8%	72.7%	83.1%

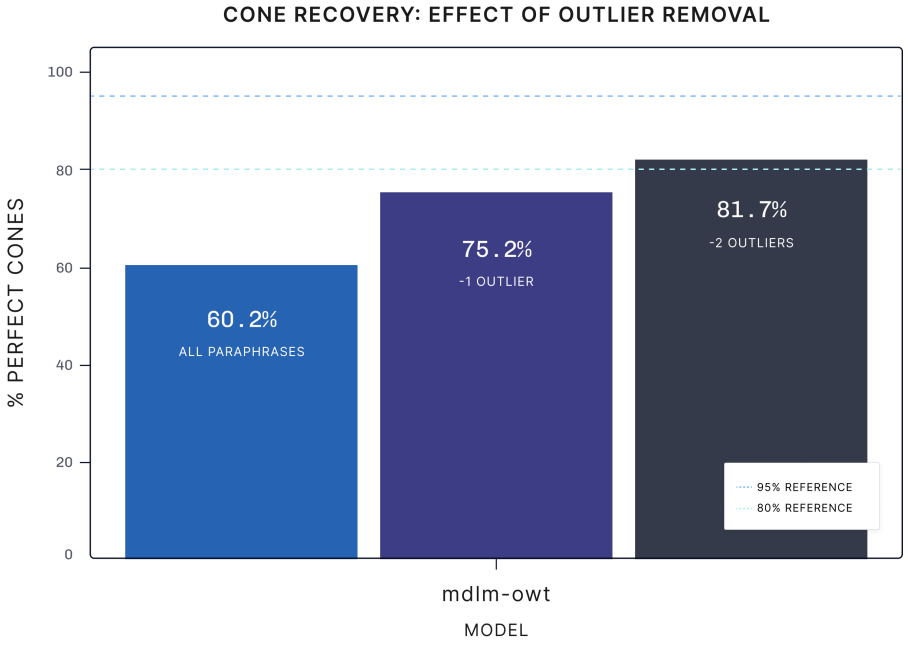


Figure 6: Cone recovery under outlier removal for MDLM. Bars show the percentage of anchors with a perfect ε-cone at each stage. Dashed lines mark 95% and 80% reference thresholds.

from 52.8% to 72.7% to 83.1%. The gains are monotonic but diminishing, suggesting that a small number of outlier paraphrases account for a disproportionate share of cone inflation.

For comparison, contrastive sentence embedders (which already start above 95% perfect cones) see only marginal improvement from outlier removal. Generative language models see larger absolute gains, paralleling the MDLM pattern. This is consistent with the observation that generative models produce embedding spaces where paraphrase clusters have heavier tails.

Figure 6 shows the cone recovery at each outlier-removal stage, and Figure 7 shows the distribution of ε values (box plots) across stages.

F Experimental Methodology Details

F.1 MDLM Diffusion Projector Parameters

Table 7 summarizes the full parameter configuration for the MDLM-based projection stage of the canonical paraphrasing pipeline (Section 4). The mask rate increases from 0.08 to 0.20 over the 8 diffusion steps.

F.2 Canonicalizer Parameters

Table 8 summarizes the parameters for each of the three embedding-space canonicalizers evaluated in Section 4.

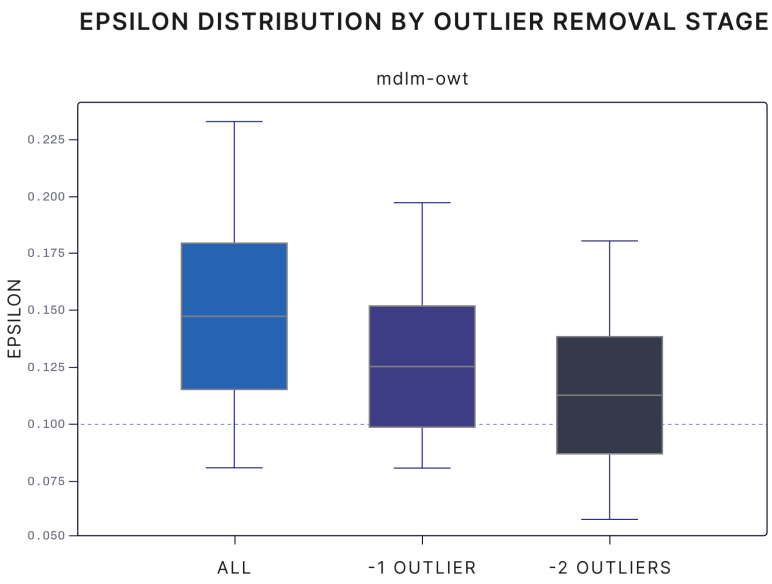


Figure 7: Distribution of ε-cone radii for MDLM at each outlier-removal stage. Outlier removal shifts the distribution toward smaller cone radii, with diminishing returns after the first removal.

Table 7: MDLM Diffusion Projector Parameters

Parameter	Value	Description
Model	`mdlm-owt`	MDLM checkpoint (kuleshov-group)
Diffusion steps T	8	Forward/reverse diffusion iterations
Min mask rate	0.08	Mask fraction at $t = 1$
Max mask rate	0.20	Mask fraction at $t = T$
Remask rate	0.15	Max fraction remasked per step
Fill strategy	argmax	Deterministic token selection
Fidelity checks		
Number strict	True	Numbers must match input/output
Negation strict	True	Negation polarity preserved
Entity overlap	≥ 0.50	Minimum named entity overlap

For the Fréchet mean, we use Riemannian gradient descent with learning rate $\eta = 0.5$ and normalise the estimate back to the sphere after each step. Directional mean shift uses the update

$$\mathbf{x}_{t+1} = \text{normalize}\left(\sum_i \exp(\kappa\, \mathbf{x}_t \cdot \mathbf{p}_i)\, \mathbf{p}_i\right),$$

where $\mathbf{p}_i$ is a paraphrase embedding and normalize rescales the vector to unit length. We set the concentration to $\kappa = 50$.

Pipeline repeatability. We run the Fréchet pipeline twice on each of 80 inputs, keeping the pipeline seed fixed at 42. Claude Sonnet 4.6 generates fresh paraphrases for each run. The two centre estimates meet the cosine-similarity threshold of 0.925 for all 80 inputs. The output texts match exactly for 14 of the 80 inputs.

Table 8: Canonicalizer Parameters (Step 2: Snap)

		Canonicalizer		
Parameter	Description	Fréchet	Weiszfeld	Mean Shift
Shared parameters				
n paraphrases	LLM-generated candidates	20	20	20
Max iterations	Convergence limit	50	50	50
ε threshold	Snap acceptance radius	0.075	0.075	0.075
Temperature	LLM sampling temp	0.7	0.7	0.7
Paraphrase model	LLM for candidates	claude-sonnet-4-20250514		
Method-specific parameters				
Learning rate η	Gradient step size	0.5	—	—
Convergence thresh	Similarity to stop	0.99999	10^{-6}	10^{-6}
Concentration κ	vMF kernel sharpness	—	—	50
Smoothing δ	Weiszfeld stability	—	10^{-10}	—

F.3 Chess Payload Recovery

The chess experiment in §5.3 uses Mistral-7B-v0.1 cover text, the Fréchet mean estimator, $\varepsilon = 0.55$, and seed 42. The diffusion mask rate ranges from 0.08 to 0.20. This run uses a different ε from the embedding-agreement experiment in §5.2.

We aggregate bit errors over all $38{,}264$ payload bits from the 50 games. Before defense, the receiver makes $2{,}215$ bit errors (5.79%). After defense, it makes $17{,}427$ (45.54%). No rewritten text yields an exact copy of the original payload. The decoder returns a bit sequence in every trial, so this result measures payload corruption rather than decoder execution failures. The mean defense time is 31.4 seconds per cover text. It includes embedding, paraphrase generation, centre estimation, diffusion editing, and the final consistency check described in §4.

We fit a binary channel using the 17 games whose payloads the receiver recovers exactly before defense, giving $9{,}528$ aligned payload bits. After canonical paraphrasing, Blahut–Arimoto estimates the fitted channel's capacity at 0.010 bits per payload bit. The model pools bit positions and treats channel uses as independent. This estimate characterises the evaluated decoder under that model. Dependencies between bits and alternative decoding strategies require separate analysis.

F.4 Scripted 20 Questions

We repeat the same 20-turn question-and-answer script about an elephant with a different MEC key for each game. The questions and answers stay fixed throughout the game. MEC embeds each question and answer in conversational cover text generated by Llama-3.1-8B-Instruct, using block size 4, temperature 1.0, top-k 300, and a 100-token completion limit. The defense processes each sentence with mean-shift centre estimation and MDLM, using 20 paraphrases from Claude Sonnet 4.6, $\varepsilon = 0.55$, and pipeline seed 42.

Both conditions use the same encoding and the same preceding cover-text context. The undefended decoder reads the encoder's token IDs. The defended decoder reads the rewritten text after tokenisation. We score answer recovery with the attack harness's answer matcher, which accepts an exact answer code or a Levenshtein similarity of at least 0.6 to a code or its full word. Decoder failures count as unsuccessful recovery. In these records, every successful answer recovery also matches the intended code exactly.

The six completed games contain 240 attempted utterances. We exclude 13 attempts with paraphrase-oracle failures, leaving 227 measured utterances, of which 112 are answers. A seventh, incomplete game contributes no measurements to these totals. The receiver recovers $112/112$ answers without defense and $15/112$ after defense. Across questions and answers together, the defended decoder fails to return a payload for $203/227$ utterances. These counts measure recovery of scripted messages with a fixed decoder.

Table 9: Detector Performance (trained on MEC-encoded text vs. cover text)

Detector	Train AUC	Val AUC	Test AUC
RoBERTa-large	0.523	0.518	0.512
Statistical baseline	0.508	0.506	0.504
Ensemble	0.531	0.521	0.515

F.5 Task Utility

We sort the 1,319 GSM8K test problems by reference-solution length and select 60 evenly spaced problems. Mistral-7B-v0.1 generates a solution for each problem using two worked examples, greedy decoding, and a limit of 220 new tokens. The generated solutions range from 37 to 562 characters. All three conditions start from the same generated solution, with no hidden payload. The canonical condition uses Fréchet centre estimation, MDLM, Claude Sonnet 4.6 paraphrases, $\varepsilon = 0.55$, seed 42, and mask rates from 0.08 to 0.20. The one-pass naive baseline asks Claude Sonnet 4.6 to rephrase the solution while preserving its meaning, at temperature 1.0.

We score each output by extracting the number after the final #### answer marker, or the last number if no valid marker remains, and comparing it exactly with the reference answer. The undefended and canonical conditions each score $24/60$ (40.0%), and the one-pass naive baseline scores $28/60$ (46.7%). Canonical paraphrasing preserves the extracted answer on all 60 problems. This check measures answer correctness after rewriting completed solutions.

F.6 Sample Generation Protocol

For each (model, scheme, context) triple:

1. Generate 200 stegotext samples with random messages
2. Generate 200 covertext samples (no message)
3. Total: 4 models $\times$ 12 schemes $\times$ 50 contexts $\times$ 400 samples = 960,000 samples

Seed contexts (50 total) were drawn from diverse domains: news (10), Wikipedia (10), fiction (10), scientific abstracts (10), and social media (10).

F.7 Detector Training

We train three detector types to provide robust security estimates:

RoBERTa-large. Fine-tuned for binary classification (encoded text vs. cover text):

- Learning rate: 2×10^{-5}
- Batch size: 32
- Epochs: 10 with early stopping (patience 3)
- Data split: 80/10/10 train/val/test

Statistical Baseline. Features: perplexity, entropy rate, n-gram statistics. Classifier: logistic regression with L2 regularization.

Ensemble. Calibrated probability combination of neural and statistical detectors.

Even the strongest ensemble detector achieves only 0.515 test AUC on MEC, barely above chance and consistent with theoretical predictions of perfect security.

F.8 Statistical Methods

Bootstrap CIs. For each metric, resample 10,000 times with replacement; report 2.5th and 97.5th percentiles.

Table 10: Statistical Significance Tests. Per-model detectability rows report one-sample t-tests of AUC against the random baseline of 0.50. All four 95% CIs contain 0.50, confirming that MEC-encoded text is statistically indistinguishable from unmodified text across every model tested. **Note:** For the detectability rows, a high p-value (and "No" in the Sig.? column) is the *desired* outcome: we cannot reject the null hypothesis that stegotext is classifier-indistinguishable from covertext, which confirms perfect security.

Comparison	Mean AUC	95% CI	Statistic	p-value	Sig.?
Detectability (AUC vs. 0.50)					
Mistral-7B	0.516 ± 0.088	$[0.461,\ 0.571]$	$t = 0.57$	0.58	No
Qwen2.5-7B	0.524 ± 0.123	$[0.448,\ 0.600]$	$t = 0.61$	0.56	No
Phi-3-mini	0.522 ± 0.132	$[0.440,\ 0.604]$	$t = 0.53$	0.61	No
TinyLlama-1.1B	0.476 ± 0.074	$[0.431,\ 0.522]$	$t = -1.01$	0.34	No
Other comparisons					
MEC Efficiency vs. METEOR	—	—	$t = 8.76$	< 0.0001	Yes
Perfect vs. Imperfect (AUC)	—	—	$F = 187.3$	< 0.0001	Yes

Hypothesis Tests. Two-sample t-tests for efficiency comparisons; one-sample t-tests for AUC vs. 0.50; ANOVA for category comparisons. Bonferroni correction: $\alpha_{adj} = 0.05/95 \approx 0.00053$.

G Scheme Classification Table

Table 11 presents all 12 evaluated schemes (plus a random baseline and 2 trivial baselines) classified by their position in the steganographic simplex.

Table 11: Complete Scheme Classification by Moduli Position

Scheme	Moduli Position	η	AUC	Reference
Barycenter Region (Optimal)				
MEC	Barycenter	0.947	0.503	de Witt 2023 de Witt et al. (2023)
DISCOP	Near-barycenter	0.912	0.507	Ding 2023 Ding et al. (2023)
LLM-Stega	Near-barycenter	0.903	0.511	Wang 2024 Wang et al. (2024)
METEOR	Near-barycenter	0.881	0.512	Kaptchuk 2021 Kaptchuk et al. (2021)
Interior Region				
Arith. Coding	Interior	0.928	0.524	Ziegler 2019 Ziegler et al. (2019)
STEGASURAS	Interior	0.867	0.518	Kaptchuk 2021 Kaptchuk et al. (2021)
Huffman	Interior	0.756	0.541	Yang 2021 Yang et al. (2021)
ADG	Off-center	0.724	0.573	Zhang 2021 Zhang et al. (2021)
Boundary Region (Detectable)				
RNN-Stega	Boundary	0.651	0.714	Yang 2019 Yang et al. (2019)
Grouping	Boundary	0.612	0.623	Dai 2019 Dai & Cai (2019)
Block Encoding	Boundary	0.578	0.634	Xiang 2018 Xiang et al. (2018)
Patient-Hunter	Boundary	0.543	0.642	Shen 2020 Shen et al. (2020)
Bins	Boundary	0.489	0.658	Fang 2017 Fang et al. (2017)
Synonym Sub.	Near-vertex	0.312	0.683	Chang 2014 Chang & Clark (2014)
Special Cases				
Rejection Samp.	Barycenter	0.023	0.501	Hopper 2002 Hopper et al. (2002)
Random Baseline	N/A	0.000	0.500	—
Trivial Baselines				
NULL	Near-vertex	0.000	0.495	—
ACROSTIC	Outside simplex	0.110	0.976	—

Interpretation. The moduli position refers to the geometric location within and around the perfect-security simplex $\mathfrak{M}_0 \cong \Delta^{2^C-1}$. All points *inside* the simplex are perfectly secure (AUC ≈ 0.50); the stratification within the simplex is about *efficiency*, not security:

- **Barycenter**: maximum efficiency (MEC), perfect security
- **Near-barycenter**: high efficiency, perfect security
- **Interior**: moderate efficiency, perfect security
- **Boundary**: on the boundary of $\mathfrak{M}_0$, where efficiency and security begin to trade off
- **Near-vertex**: low efficiency, near the edge of the perfect-security simplex
- **Outside simplex**: ϵ-secure in Cachin's sense ($D_{\mathrm{KL}} > 0$) or computationally secure in Hopper's sense (SS-CHA), detectable with sufficient data

Rejection Sampling Hopper et al. (2002) is a special case: it achieves perfect security (sits at the barycenter) but with exponentially low efficiency due to its rejection-based construction. This illustrates that security and efficiency are genuinely independent degrees of freedom on the simplex.

G.1 Full Security-Efficiency Results

Table 12 reports the complete security-efficiency measurements for all 12 schemes (plus a random baseline and 2 trivial baselines) with 95% bootstrap confidence intervals, aggregated across 4 language models ($N = 10{,}000$ per scheme). The **Random Baseline** is not a steganographic scheme: it simply samples from the language model's natural output distribution without encoding any hidden message ($\eta = 0$, no information embedded). Since it is identical to ordinary cover text generation, it achieves AUC $= 0.50$ (perfectly indistinguishable) and serves as a calibration check confirming that unencoded model outputs are indeed undetectable.

Table 12: Complete Security-Efficiency Results (aggregated across 4 models, $N = 10{,}000$ per scheme). Efficiency η is normalized to $[0, 1]$. AUC $= 0.50$ indicates perfect security (random guessing); AUC > 0.55 indicates detectable. BER $= 0.50$ indicates canonicalization destroys the message.

Scheme	Category	η (95% CI)	AUC (95% CI)	PPL Ratio	BER post-Can
Perfect Security					
MEC	Perfect	0.947 (0.941–0.953)	0.503 (0.497–0.509)	1.008	0.498
METEOR	Perfect	0.881 (0.873–0.889)	0.512 (0.504–0.520)	1.015	0.501
DISCOP	Perfect	0.912 (0.905–0.919)	0.507 (0.499–0.515)	1.011	0.499
Rejection Samp.	Perfect	0.023 (0.019–0.027)	0.501 (0.494–0.508)	1.003	0.500
Near-Perfect Security					
Arith. Coding	Near-perf.	0.928 (0.921–0.935)	0.524 (0.515–0.533)	1.031	0.503
ADG	Near-perf.	0.724 (0.712–0.736)	0.573 (0.562–0.584)	1.089	0.511
STEGASURAS	Near-perf.	0.867 (0.858–0.876)	0.518 (0.509–0.527)	1.022	0.500
Huffman	Near-perf.	0.756 (0.744–0.768)	0.541 (0.530–0.552)	1.052	0.506
LLM-Stega	Near-perf.	0.903 (0.895–0.911)	0.511 (0.503–0.519)	1.014	0.499
Imperfect Security					
RNN-Stega	Imperfect	0.651 (0.637–0.665)	0.714 (0.702–0.726)	1.234	0.497
Synonym Sub.	Imperfect	0.312 (0.298–0.326)	0.683 (0.670–0.696)	1.178	0.502
Patient-Hunter	Imperfect	0.543 (0.528–0.558)	0.642 (0.629–0.655)	1.145	0.504
Bins	Imperfect	0.489 (0.473–0.505)	0.658 (0.645–0.671)	1.167	0.499
Grouping	Imperfect	0.612 (0.597–0.627)	0.623 (0.610–0.636)	1.134	0.501
Block Encoding	Imperfect	0.578 (0.563–0.593)	0.634 (0.621–0.647)	1.142	0.498
Random Baseline	Baseline	0.000	0.500 (0.493–0.507)	1.000	0.500
Trivial Baselines					
NULL	Trivial	0.000 (0.000–0.000)	0.495 (0.471–0.519)	1.189	0.500
ACROSTIC	Trivial	0.110 (0.109–0.111)	0.976 (0.962–0.989)	3.414	0.432

H Steganographic Efficiency Metrics

This appendix presents the three-metric efficiency framework used to evaluate steganographic schemes across complementary dimensions, and reports the full security-efficiency frontier results.

H.1 Three-Metric Framework

We evaluate efficiency along three complementary dimensions, each capturing a different aspect of how well a scheme uses available capacity.

Definition H.1 (Fisher-Rao Efficiency)**.** *The **Fisher-Rao efficiency** measures how uniformly payloads are distributed:*

$$\eta_{\mathrm{FR}} = 1 - \frac{d_{\mathrm{FR}}(\pi, u)}{d_{\mathrm{FR,max}}}$$

where π is the scheme's payload distribution, u is the uniform distribution, and d_{FR} is the Fisher-Rao distance. Intuition: A scheme that always encodes the same payload wastes capacity; uniform payload usage is optimal.

Definition H.2 (Operational Efficiency)**.** *The **operational efficiency** measures practical bit throughput:*

$$\eta_{\mathrm{op}} = (1 - \mathrm{BER}) \cdot \frac{\textit{bits per token}}{H(\mathcal{D})}$$

where BER is the bit error rate and $H(\mathcal{D})$ is the entropy of the cover distribution. Intuition: How many bits actually get through, relative to the channel's capacity?

Definition H.3 (Combinatorial Efficiency)**.** *The **combinatorial efficiency** measures semantic coverage:*

$$\eta_{\mathrm{comb}} = \frac{k}{n}$$

where k is the number of distinct meaning clusters reached and n is the sample size. Intuition: Does the scheme explore the full space of possible meanings, or does it get stuck in a corner?

A scheme may excel on one metric while failing on others. Fisher-Rao captures theoretical optimality; operational captures practical reliability; combinatorial captures expressiveness.

H.2 Experimental Setup

We evaluate 19 steganographic schemes from the literature:

Arithmetic coding family. MEC de Witt et al. (2023), STEGASURAS Kaptchuk et al. (2021), ADG Zhang et al. (2021), DISCOP Ding et al. (2023).

Huffman family. MORALDO Moraldo (2012), Huffman-based schemes.

Bin-based family. Fixed bins, adaptive bins, equiprobable partitions.

Naive baselines. Random selection, LSB-style, rejection sampling.

Models. Primary experiments use Llama-3.1-8B-Instruct with temperature $T = 1.0$, top-p $p = 0.95$, and maximum sequence length 128 tokens. Validation experiments use GPT-4 (via API) and Pythia-1B Biderman et al. (2023).

Data. WikiText-103 (100M tokens) for cover distribution estimation; OpenWebText (1M tokens, held out) for evaluation.

H.3 Measurement Methodology

Our efficiency measurements combine theoretical analysis and empirical validation:

Fisher-Rao efficiency (η_{FR}). Computed analytically from each scheme's theoretical message distribution structure. For arithmetic coding schemes, the uniform message distribution is provably achievable; for heuristic schemes, we estimate from large samples ($n = 10{,}000$).

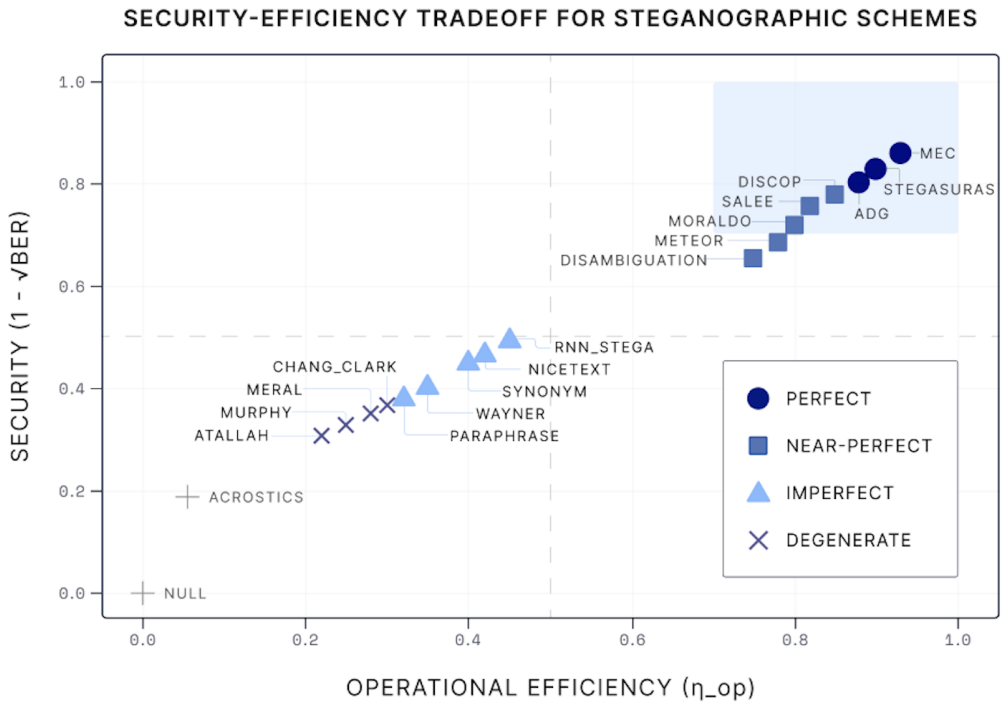


Figure 8: MEC dominates the security-efficiency frontier. X-axis: security (AUC; 0.5 = undetectable). Y-axis: composite efficiency $\bar{\eta}$. MEC achieves $\bar{\eta} = 0.94$ with perfect security (AUC $= 0.50$), occupying the optimal Pareto corner.

Operational efficiency (η_{op}). Measured empirically from 10,000 encode-decode trials per scheme. We report mean efficiency with 95% bootstrap confidence intervals (1,000 bootstrap samples).

Combinatorial efficiency (η_{comb}). Computed via semantic clustering using all-MiniLM-L6-v2 embeddings with DBSCAN clustering ($\varepsilon = 0.15$, min_samples=5) on 10,000 samples per scheme.

Security (AUC). Binary classifier (multi-feature: perplexity, burstiness, entropy rate) distinguishing cover text from text carrying a hidden payload. AUC $= 0.5$ indicates perfect security; values significantly above 0.5 indicate detectability.

H.4 Results: The Efficiency Frontier

Figure 8 shows the security-efficiency frontier across all 19 schemes. Table 13 provides detailed per-metric statistics. Key findings:

1. **MEC dominates**: MEC achieves $\eta_{\text{FR}} = 0.98 \pm 0.01$, $\eta_{\text{op}} = 0.93 \pm 0.02$, $\eta_{\text{comb}} = 0.92 \pm 0.03$ while maintaining perfect security (AUC $= 0.50 \pm 0.01$).
2. **Arithmetic coding schemes are robust**: STEGASURAS and ADG achieve $\eta \approx 0.85$–0.90, below MEC but still strong (difference significant at $p < 0.01$, paired t-test with Bonferroni correction).
3. **Huffman schemes underperform**: Despite theoretical optimality for source coding, Huffman-based schemes achieve only $\eta \approx 0.75$–0.80 for steganography.
4. **Naive schemes fail**: Random selection and LSB-style schemes fall far below the frontier ($\eta < 0.5$).

Table 13: MEC achieves highest efficiency across all three metrics. 95% bootstrap CI; PPL = perplexity ratio (1.0 = indistinguishable). $^*p < 0.01$ vs. MEC.

	η_{FR}	η_{op}	η_{comb}	PPL
MEC	$.98 \pm .01$	$.93 \pm .02$	$.92 \pm .03$	1.00
STG	$.95 \pm .02^*$	$.88 \pm .03^*$	$.87 \pm .04^*$	1.00
ADG	$.93 \pm .02^*$	$.86 \pm .03^*$	$.85 \pm .04^*$	1.00
DIS	$.90 \pm .03^*$	$.82 \pm .04^*$	$.80 \pm .05^*$	1.00
HUF	$.78 \pm .04^*$	$.71 \pm .05^*$	$.65 \pm .06^*$	1.02
RND	$.45 \pm .05^*$	$.38 \pm .06^*$	$.32 \pm .07^*$	1.15^*

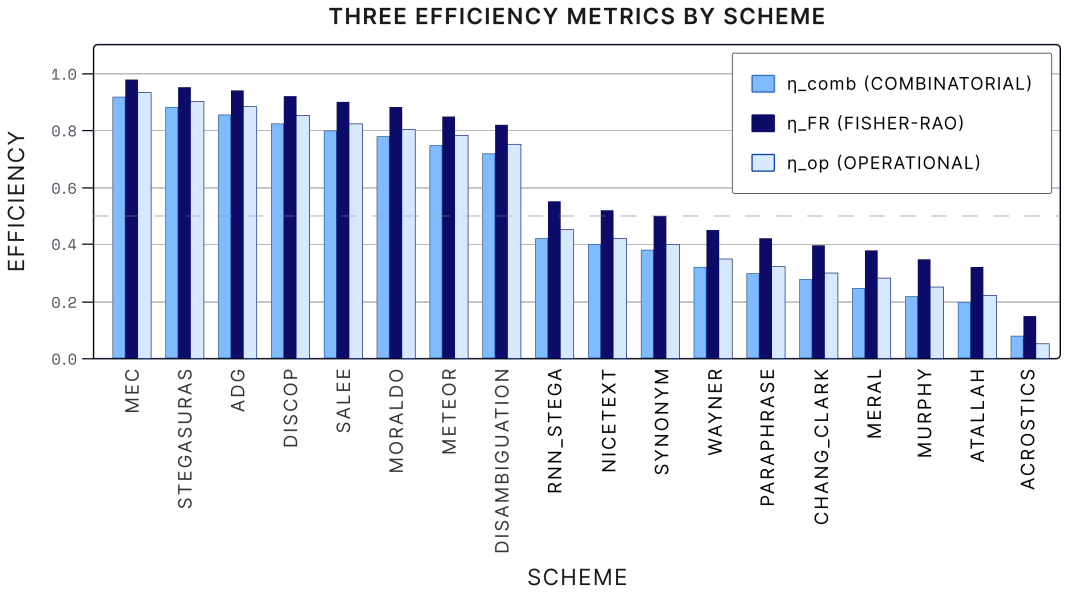


Figure 9: MEC dominates across all three efficiency metrics. η_{FR} (blue), η_{op} (orange), η_{comb} (green) with 95% CI. MEC significantly outperforms all baselines ($p < 0.01$).

Figure 9 shows the per-scheme breakdown across the three metrics. The consistency of MEC's dominance across all three independently-motivated metrics corroborates the Simplex Theorem's prediction that MEC occupies the barycenter of the steganographic simplex (Theorem A.1).