FROST: A Free-Energy Framework for Reducing Capability Transfer Under Black-Box Querying

Anonymous ACL submission

Abstract

Frontier model APIs face a growing risk: black-box querying can turn outputs into supervision for downstream student models. Existing defenses are often reactive or heuristic, without modeling why some traces are especially learnable. We introduce FROST, a teacher-side anti-distillation framework grounded in thermodynamic free-energy control using attention-transport geometry, structural shaping, and provenance-aware decoding. FROST formulates defended generation as a constrained free-energy minimization problem that Gibbs-reweights the teacher distribution to suppress outputs with high downstream learning value. The framework combines: (i) an anti-copying energy that scores how reusable a trace is for downstream learning, instantiated through attention-transport thermodynamic length, (ii) a KL trust region that preserves utility by keeping defended decoding close to the nominal teacher, and (iii) a stateless, prompt-local design requiring no telemetry or identity signals. This yields an output-channel defense targeting trace learnability rather than provenance or watermarking alone. Across four teacher models (DeepSeek-R1-Distill-Qwen-7B, Qwen2.5-7B-Instruct, Gemma-2-27B-IT, and Ministral-8B-Instruct-2410), three attacker models (Qwen2.5-3B, Mistral-7B, and Llama-3.1-8B), reasoning, knowledge, code, and conversational benchmarks, and both answer-only and trace-rich harvesting, FROST reduces Capability Transfer Rate by 43.2 percent over nominal decoding and 27.0 percent over Anti-distillation Sampling, while preserving teacher utility and response quality.

1 Weaponizing Distillation as Capability and Data Stealing

From compression to attack surface. Distillation is usually introduced as a benign compression primitive: a smaller student learns from a stronger teacher to reduce inference cost while retaining much of the teacher's utility. ***Once a frontier model is exposed through a black-box API, however, the same pipeline becomes an attack surface.*** An adversary can harvest outputs at scale and repurpose them as high-value supervision for downstream model improvement. Anthropic publicly alleged in February 2026 that it had detected ***industrial-scale distillation attacks*** against Claude, and Reuters reported that the targeted capabilities included ***reasoning, coding, and related high-value behaviors***. These remain *public allegations*, not adjudicated findings; nevertheless, they sharpen a broader security reality: ***deployment turns model outputs into reusable assets that can be systematically harvested, curated, and converted into student capability*** (Tramèr et al., 2016; Orekondy et al., 2019; Zhao et al., 2025)[1][2]. [3]

Why black-box distillation is plausible. This threat is not anomalous. The model-extraction literature has long shown that ***black-box query access alone can suffice to recover high-fidelity surrogates***. Tramèr et al. demonstrated efficient extraction attacks against prediction APIs, while Orekondy et al. showed that competitive "knockoff" models can be trained from queried input–output pairs even under architecture and query-distribution mismatch (Tramèr et al., 2016; Orekondy et al., 2019). In the LLM setting, the output channel is richer than class labels or probabilities: it can expose ***explanations, decompositions, code, rubrics, tool-use conventions, and reusable reasoning traces***. Recent surveys therefore treat API-based functionality extraction, training-data extraction, and prompt-targeted attacks as part of a single security landscape for deployed language models (Zhao et al., 2025).

[1] https://www.anthropic.com/news/detecting-and-preventing-distillation-attacks

[2] https://www.reuters.com/world/china/chinese-companies-used-claude-improve-own-models-anthropic-says-2026-02-23/

[3] To foster future research in this area, we make our training and evaluation code publicly available.

What exactly can be stolen? A 3-channel leakage view. In the LLM setting, an attacker may extract more than benchmark behavior. The output channel can leak three distinct forms of value:

- ***Capability***: reusable input–output behavior—reasoning patterns, coding competence, tool-use conventions, task decomposition strategies, and evaluation heuristics—that can be distilled into downstream students (Tramèr et al., 2016; Orekondy et al., 2019; Zhao et al., 2025).
- ***Data***: memorized training content recoverable through repeated querying. Carlini et al. showed that large language models can emit verbatim training examples, and Nasr et al. later demonstrated ***scalable extraction*** from open, semi-open, and production systems, including aligned models (Carlini et al., 2021; Nasr et al., 2023).
- ***Control logic***: hidden prompts, instructions, and behavioral scaffolds that shape model behavior. Zhang et al. introduced OUTPUT2PROMPT, a black-box method that reconstructs prompts from normal outputs without logits or jailbreak-style interaction, while Raccoon systematized prompt-theft evaluation across diverse attack patterns (Zhang et al., 2024; Wang et al., 2024).

Why structured traces are especially dangerous. Not all outputs are equally distillable. Long-form explanations, step-by-step derivations, explicit decompositions, code skeletons, self-talk, rubric-style judgments, and reusable response templates provide denser and more portable supervision than terse answers. This point is increasingly explicit in recent antidistillation work: *Antidistillation Sampling* modifies sampling to make reasoning traces less effective as student training material, while *PART* reformulates reasoning traces to preserve human-useful information while disrupting student-side distillation (Savani et al., 2025; Ding et al., 2025). Prompt-theft benchmarks likewise show broad susceptibility in the absence of defenses (Wang et al., 2024; Zhang et al., 2024). Taken together, these results suggest that ***capability transfer is governed not only by model weights, but by the structure, pedagogy, ordering, and reusability of emitted traces*** (Zhao et al., 2025; Savani et al., 2025; Ding et al., 2025).

Distillation as capability theft vs. data theft. Weaponized distillation serves at least two closely related attacker objectives:

1. ***Capability stealing*** (*functionality extraction*): training a surrogate that reproduces the teacher's useful behavior while bypassing access controls, pricing, rate limits, and deployment governance (Tramèr et al., 2016; Orekondy et al., 2019; Zhao et al., 2025).
2. ***Data stealing*** (*extractable memorization / inversion*): recovering private training artifacts, proprietary prompts, or latent control instructions from outputs, either directly through memorization or indirectly through scalable reconstruction and inversion (Carlini et al., 2021; Nasr et al., 2023; Zhang et al., 2024; Wang et al., 2024).

Under this view, unauthorized distillation is not merely an IP dispute. It is a joint problem of ***capability theft, data leakage, and control-logic exfiltration*** (Zhao et al., 2025).

Provenance is useful, but insufficient. A natural response to output-mediated theft is to embed provenance signals into generated text. Watermarking schemes bias token sampling so outputs become statistically detectable without model internals, and systems such as SynthID-Text report large-scale watermarking results (Dathathri et al., 2024). However, ***provenance is not prevention.*** Attributable outputs may still carry valuable supervision for distillation, prompt recovery, or data extraction. Moreover, detection and watermarking face robustness, paraphrasing, spoofing, and utility tradeoffs under adaptive use (Sadasivan et al., 2023; Zhang et al., 2023; Pang et al., 2024). Thus, attribution alone is incomplete: ***the core defensive problem is not only who produced the trace, but how learnable that trace is for an adversary*** (Savani et al., 2025; Zhao et al., 2025; Ding et al., 2025).

Implication for this paper. This motivates a stronger objective than monitoring, rate limiting, provenance, or identity-based controls. If outputs can be harvested into high-value supervision, securing a frontier API requires controlling not only *what* the teacher does for legitimate users, but also ***how much downstream learning value each trace contains***. This is the role of ***FROST***: we treat teacher-side generation as a ***thermodynamic control problem over output learnability***, preserving utility while making black-box capability transfer ***systematically harder, less sample-efficient, and***

less scalable (Savani et al., 2025; Ding et al., 2025; Tramèr et al., 2016; Orekondy et al., 2019; Zhao et al., 2025).

2 The FROST Framework

FROST addresses a core deployment tension: a frontier model exposed through a black-box API must remain ***useful*** to legitimate users, while its outputs should not become ***cheap supervision*** for downstream students. Rich explanations, derivations, and code traces improve user utility, but they also increase transferability under harvesting attacks. Conversely, aggressive truncation or randomization may reduce distillability only by destroying the API value itself. FROST therefore treats teacher-side defense as a ***utility-constrained control problem over output learnability***, rather than as telemetry, provenance, or heuristic perturbation alone. Figure 1 illustrates the full teacher-side antidistillation pipeline.

2.1 Problem Setup and Threat Model

We model the deployed teacher as a conditional output channel. For a prompt $x \in \mathcal{X}$ and response $y \in \mathcal{Y}$, let

$$P_0(y \mid x)$$

denote the ***nominal teacher distribution***. FROST constructs a defended distribution

$$Q(y \mid x),$$

which is the only distribution exposed to the attacker.

The attacker has ***black-box adaptive query access***. Given a query budget B, it collects

$$\begin{aligned} \mathcal{T}_B &= \{(x_t, y_t)\}_{t=1}^{B}, y_t \sim Q(\cdot \mid x_t), \\ x_t &\sim \pi_t(\cdot \mid x_{<t}, y_{<t}), \\ \widehat{\mathcal{L}}_{\text{imit}}(\theta; \mathcal{T}_B) &= \frac{1}{B} \sum_{t=1}^{B} \ell_\theta(x_t, y_t), \\ \ell_\theta(x_t, y_t) &= -\log p_\theta(y_t \mid x_t). \end{aligned}$$

This covers answer-only harvesting, trace-rich harvesting, cross-architecture transfer, and PEFT-style adaptive attacks.

A key design choice is that FROST is ***identity-free*** and ***prompt-local***: it uses no accounts, IP addresses, cross-session telemetry, or metadata-level coordination. It acts only through the output distribution conditioned on the current prompt.

2.2 Part I: Geometric Anti-Copying Energy

The central question is how to assign an anti-copying energy $\mathcal{E}_x(y)$ to a candidate response. A good energy should be high when y is useful training signal for an attacker student and low when y is comparatively inert as supervision. FROST defines this signal through **attention-transport thermodynamic length**: traces that create sharp, structured routing across layers, such as decompositions, derivations, and code scaffolds, form stable pathways a student can imitate, and are therefore assigned high anti-copying energy.

Transport kernel construction. For a sequence of length n, each transformer layer ℓ produces attention matrices $\{A_{\ell,h}\}_{h=1}^{H}$. We aggregate heads into a single transport kernel via entropy-weighted combination, assigning larger weights to low-entropy heads that carry sharper routing information:

$$\begin{aligned} P_\ell &= \frac{\sum_{h=1}^{H} \alpha_{\ell,h} A_{\ell,h}}{\sum_{h=1}^{H} \alpha_{\ell,h}}, \\ \alpha_{\ell,h} &= \left(1 - \frac{H(A_{\ell,h})}{\log n}\right)_+ . \end{aligned}$$

where $H(A_{\ell,h})$ denotes the mean row entropy of head h at layer ℓ. Each row of P_ℓ sums to one, so P_ℓ is a Markov kernel transporting token-indexed mass in one attention step.

Layer-wise routing displacement. Let π_ℓ denote the stationary distribution of P_ℓ. The information-geometric effort required to change routing from $P_{\ell-1}$ to P_ℓ is measured by the transport-geometric quadratic form:

$$\begin{aligned} g_\ell = \sum_{i=1}^{n} \pi_\ell(i) \sum_{j=1}^{n} P_\ell(i,j) \\ \cdot \left(\log P_\ell(i,j) - \log P_{\ell-1}(i,j)\right)^2 . \end{aligned}$$

where the inner sum averages squared log-routing increments over destinations according to the current kernel, and the outer sum averages over sources according to the stationary distribution.

Anti-copying energy. The primary anti-copying signal in FROST is the **attention-transport thermodynamic length**:

$$\mathcal{E}_x^{\text{geom}}(y) \;=\; \mathcal{L}(y) \;=\; \sum_{\ell=1}^{L} \sqrt{g_\ell}.$$

Large $\mathcal{L}(y)$ indicates strong routing displacement across depth, marking candidates that force large

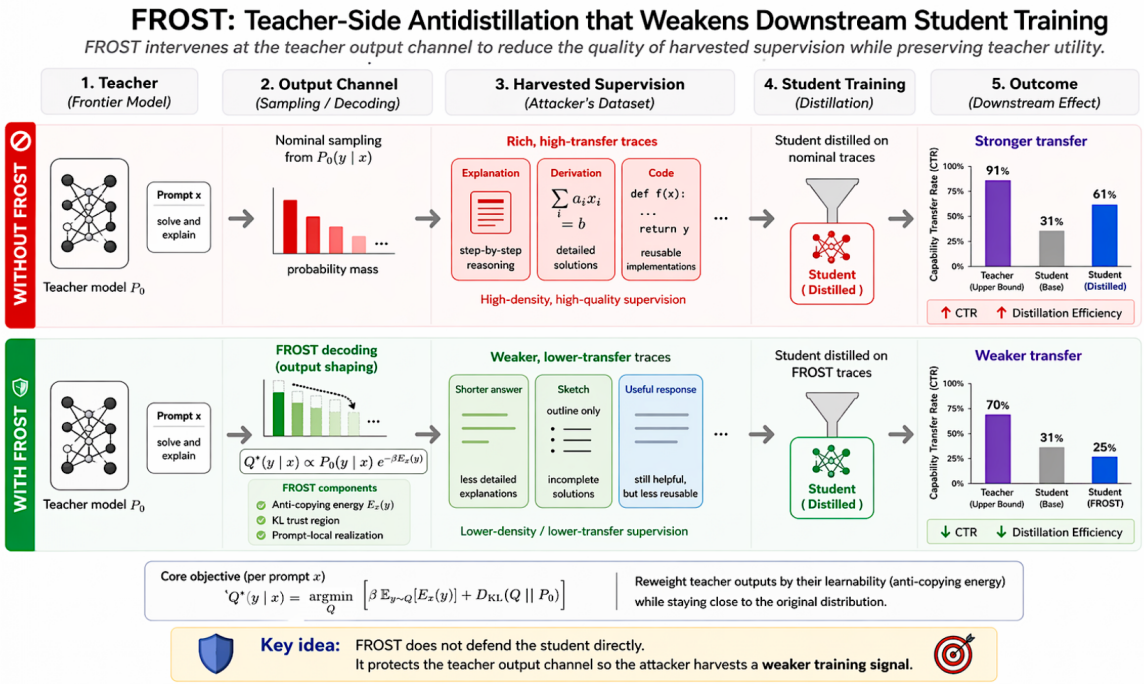

Figure 1: ***Teacher-side antidistillation with FROST. Top: standard black-box distillation.*** The teacher samples from $P_0(y \mid x)$, producing rich, high-transfer traces such as explanations and derivations that serve as dense supervision for the attacker's student. ***Bottom: FROST.*** FROST intervenes at the ***teacher output channel*** by reweighting candidates with an anti-copying energy: $Q^\star(y \mid x) \propto P_0(y \mid x)e^{-\beta E_x(y)}$. This suppresses high-learnability outputs while remaining close to the nominal teacher, reducing capability transfer and distillation efficiency. The key idea is that FROST protects the teacher output channel rather than modifying the student directly.

multiplicative changes in the attention transport kernel and are therefore intrinsically useful for downstream imitation. Small $\mathcal{L}(y)$ indicates that committing to the candidate produces minimal routing displacement along the depth trajectory, making it comparatively inert as training supervision. This is the primary anti-copying signal in FROST: it penalizes responses that are intrinsically learnable through their attention-routing structure, rather than merely long, verbose, or stylistically distinctive.

2.3 Part II: KL-Trust-Region Free-Energy Decoding

The 2nd component is a ***KL trust region*** that keeps defended decoding close to the nominal teacher. Let $u(x, y)$ be a task utility functional and define

$$U(Q; \mathcal{D}) = \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim Q(\cdot|x)}[u(x, y)].$$

FROST seeks low-learnability outputs while preserving utility:

$$\min_Q \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim Q(\cdot|x)}[E_x(y)]$$

subject to

$$U(Q; \mathcal{D}) \geq U_{\min},$$
$$\mathbb{E}_{x\sim\mathcal{D}}\left[D_{\mathrm{KL}}(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x))\right] \leq \delta.$$

The utility floor prevents degenerate defenses, while the KL constraint enforces ***minimal intervention***.

For a fixed prompt x, this yields the free-energy objective

$$\mathcal{F}_\beta(Q; x) = \mathbb{E}_{y\sim Q(\cdot|x)}\left[E_x(y)\right] + \frac{1}{\beta} D_{\mathrm{KL}}\big(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x)\big).$$

where $\beta > 0$ controls defense strength. The unique minimizer is

$$Q^\star(y \mid x) = \frac{1}{Z_\beta(x)} P_0(y \mid x) \exp(-\beta E_x(y)),$$

with

$$Z_\beta(x) = \sum_{y\in\mathcal{Y}} P_0(y \mid x) \exp(-\beta E_x(y)).$$

This shows that FROST is not arbitrary perturbation: it is the exact optimizer of a KL-regularized anti-copying objective. As $\beta \to 0$, $Q^\star$ recovers the nominal teacher; as β increases, the decoder increasingly suppresses high-learnability traces, tracing a ***security–utility frontier***.

2.4 Part III: Stateless Prompt-Local Realization

The ideal defended law is sequence-level:

$$Q^\star(y \mid x) \propto P_0(y \mid x) \exp(-\beta E_x(y)).$$

Since exact normalization over full responses is intractable, FROST implements this law autoregressively. The composite energy is

$$\begin{aligned} E_x(y) = \lambda_{\text{geom}} E_x^{\text{geom}}(y) \\ + \lambda_{\text{struct}} E_x^{\text{struct}}(y) \\ + \lambda_{\text{prov}} E_x^{\text{prov}}(y). \end{aligned}$$

The geometric term is the principal signal. The structural term penalizes reusable pedagogical scaffolds such as decompositions, procedures, derivation templates, and code skeletons. The optional provenance term supports watermark-compatible decoding, but is not required for the core defense.We explain all the terms in - B,D,C.

At decoding step t, FROST modifies the nominal logit $\ell_t(v)$ for each candidate token v in a top-k shortlist $\mathcal{V}_t^{(k)}$:

$$q_t(v \mid x, y_{<t}) = \frac{\exp\left(\ell_t(v) - \beta \widetilde{E}_t(v)\right)}{\sum_{u \in \mathcal{V}_t^{(k)}} \exp\left(\ell_t(u) - \beta \widetilde{E}_t(u)\right)}.$$

Here $\widetilde{E}_t(v)$ is the token-local estimate of the composite energy for the partial prefix $y_{<t} \oplus v$. Tokens that induce high downstream learnability receive lower probability, while the top-k restriction keeps the intervention close to the nominal teacher.

Conceptually, the same decoder can be read in three ways:

1. as a ***logit correction*** against high-learnability tokens;
2. as a ***Gibbs posterior*** over plausible teacher continuations; and
3. as a ***prompt-local control policy*** suppressing reusable supervision as it emerges.

2.5 Information-Theoretic Interpretation

FROST also admits an information-theoretic reading. Let $W \in \mathcal{W}$ denote latent capability-bearing structure, and let $Z_B = (X_1, Y_1, \ldots, X_B, Y_B)$ be the adaptive transcript. Then

$$I(W; Z_B) = \sum_{t=1}^{B} I(W; Y_t \mid X_t, Z_{t-1}),$$

so leakage decomposes query by query. If every response satisfies

$$I(W; Y_t \mid X_t, Z_{t-1}) \le \gamma,$$

then $I(W; Z_B) \le B\gamma$. By Fano's inequality,

$$\Pr(\widehat{W} \neq W) \ge 1 - \frac{I(W; Z_B) + \log 2}{\log |\mathcal{W}|},$$

which implies

$$B = \Omega\left(\frac{\log |\mathcal{W}|}{\gamma}\right).$$

FROST is not a formal privacy guarantee and does not optimize mutual information exactly; it uses geometric learning progress as a tractable proxy for ***reducing per-query leakage while preserving utility***.

2.6 Scope and Practical Claims

FROST optimizes the prompt-local free-energy objective:

$$\boxed{\begin{aligned} \mathcal{F}_\beta(Q; x) = \mathbb{E}_{y \sim Q(\cdot \mid x)}\big[E_x(y)\big] \\ + \frac{1}{\beta} D_{\text{KL}}\big(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x)\big) \end{aligned}}$$

Its composite energy penalizes outputs especially learnable by downstream students. FROST therefore targets:

1. ***Reduced learnability*** of emitted traces;
2. ***Lower capability-transfer efficiency*** under black-box harvesting;
3. ***Controlled deviation*** from the nominal teacher;
4. ***Preserved task utility*** under KL and decoding constraints.

FROST is a principled ***output-shaping defense***: not a universal anti-theft barrier, platform-monitoring replacement, or formal privacy guarantee. Rather, it makes extraction ***harder, less sample-efficient, and less scalable*** while keeping the teacher useful.

3 Experimental Setup

We evaluate ***FROST*** against the core empirical requirements of teacher-side antidistillation: ***(i)*** lower black-box capability transfer, ***(ii)*** preserve teacher utility, ***(iii)*** remain robust across attacker choices, and ***(iv)*** yield a practical security–utility–overhead frontier under adaptive extraction settings (Tramèr et al., 2016; Orekondy et al., 2019; Savani et al., 2025).

Evaluation Goals

We ask whether FROST reduces ***capability transfer***, preserves ***teacher utility***, remains robust to ***attacker diversity***, and offers a usable ***security–utility–overhead tradeoff***. All comparisons use ***matched attacker effort*** so that differences come from the ***teacher output channel***, not unequal training budgets.

Models and Training Regimes

Teachers. We use ***DeepSeek-R1-Distill-Qwen-7B*** as a reasoning teacher, ***Qwen2.5-7B-Instruct*** as a general instruction-following teacher, ***Gemma-2-27B-IT*** as a large-scale instruction-tuned open-weight model, and ***Ministral-8B-Instruct-2410*** as a compact instruction-tuned multilingual baseline.
Attacker students. We use ***Qwen2.5-3B-Instruct*** (small attacker), ***Mistral-7B-Instruct-v0.3*** (cross-family attacker), and ***Meta-Llama-3.1-8B-Instruct*** (stronger attacker).
Attacker training. For each student, we evaluate both ***full SFT*** and ***PEFT/LoRA*** (Hu et al., 2022), with matched initialization, optimizer schedule, sequence length, and total training budget across defenses.
Proxy ensemble. FROST scores learnability using a fixed proxy family $\mathcal{Q}_S$; proxy ensemble size M is varied in ablations (Sinha et al., 2018).

Harvesting Regimes

We study three regimes: ***R1: benign harvesting*** (instruction following, QA, summarization, coding help, short reasoning), ***R2: structured harvesting*** (step-by-step reasoning, decompositions, rubric-style explanations, code walkthroughs), and ***R3: mixed harvesting*** (mostly benign prompts with a structured subset) (Tramèr et al., 2016; Orekondy et al., 2019; Savani et al., 2025).
For each regime, we evaluate ***answer-only*** and ***trace-rich*** transcripts at budgets $B \in \{10^3, 10^4, 10^5\}$, with extensions when feasible. We study both ***non-adaptive*** and ***adaptive*** harvesting.

Benchmarks and Baselines

Benchmarks. We evaluate ***reasoning*** on ***GSM8K*** and ***MATH-500*** (Cobbe et al., 2021; Hendrycks et al., 2021b), ***knowledge*** on ***MMLU*** (Hendrycks et al., 2021a), and ***code*** on ***HumanEval*** and ***MBPP*** (Chen et al., 2021; Austin et al., 2021). Teacher utility is additionally measured with ***MT-Bench***-style evaluation (Zheng et al., 2023). Safety checks use ***HarmBench*** and a compact jailbreak-style set (Mazeika et al., 2024).
Baselines. We compare against ***nominal decoding*** $Q_{\text{nom}}(\cdot \mid x) = P_0(\cdot \mid x)$, ***stochastic decoding***, ***Antidistillation Sampling (ADS)*** (Savani et al., 2025), ***geometry-only FROST***, ***structure-only FROST***, optional ***provenance-dominant shaping*** (Dathathri et al., 2024), and ***full FROST***.

Metrics

Our primary security metric is the ***Capability Transfer Rate (CTR)***:

$$\mathrm{CTR}(B) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \mathrm{clip}_{[0,1]} \left(\frac{\mu_k(S_B) - \mu_k(S_0)}{\mu_k(T_Q) - \mu_k(S_0)} \right),$$

where S_0 is the base student, S_B the post-distillation student at budget B, and T_Q the teacher under defense Q. We also report ***distillation efficiency*** $\kappa_{\text{distill}}(B) = \mathrm{CTR}(B)/B$ and ***AUTC***.
Teacher utility is measured with benchmark-specific accuracy, EM, pass@1, judge scores, defended utility deltas $\Delta U_k = U_k(Q) - U_k(P_0)$, and aggregate utility $U_{\text{agg}}(Q)$. We also report ***latency***, ***throughput***, ***memory overhead***, and ***energy-scoring cost*** via $\Delta\tau_{\text{rel}} = (\tau(Q) - \tau(P_0))/\tau(P_0)$. Finally, we track ***harmful compliance***, ***refusal correctness***, and ***benign-safety consistency*** as guardrail checks (Mazeika et al., 2024).

Protocol and Implementation

For each teacher T, defense Q, regime R, transcript format, and budget B, we construct

$$\mathcal{T}_B^{(T,Q,R)} = \{(x_t, y_t)\}_{t=1}^{B}, \qquad y_t \sim Q(\cdot \mid x_t).$$

Each attacker is trained from the same checkpoint S_0 under matched budgets, producing $S_B^{(Q)}$. Prompts are disjoint from evaluation sets and fixed across defenses whenever possible.
At decoding step t, FROST corrects nominal logits by

$$\ell_t^{\text{FROST}}(v) = \ell_t(v) - \beta\, \widetilde{E}_t(v;\, x, y_{<t}),$$

$$\widetilde{E}_t = \lambda_{\text{geom}} \widetilde{E}_t^{\text{geom}} + \lambda_{\text{struct}} \widetilde{E}_t^{\text{struct}} + \lambda_{\text{prov}} \widetilde{E}_t^{\text{prov}}.$$

Scoring uses a top-k shortlist; geometric scoring uses proxy students with Fisher inverses; structural features are local; shaping is optional (Mudgal et al., 2024; Amari, 1998).
The main knobs are β, λ_{geom}, λ_{struct}, λ_{prov}, k, and M. They are selected on a held-out slice using the best ***security–utility frontier***. Major results use multiple seeds with variability estimates.

4 Results

A useful teacher-side antidistillation defense must pass three empirical tests: ***(i)*** it should reduce black-box capability transfer under matched attacker effort; ***(ii)*** it should preserve teacher utility; and ***(iii)*** it should remain robust across attacker students, budgets, and harvesting styles. Across two teachers, three attacker students, and budgets $B \in \{10^3, 10^4, 10^5\}$, ***FROST passes all three tests***.

4.1 Main Distillation Resistance

Claim 1: FROST reduces recovered student capability. Figure 4 and Table 2 show that, at every budget, students trained on FROST transcripts recover a ***smaller fraction of the teacher's capability gap*** than those trained on nominal, stochastic, or ADS transcripts. The effect spans reasoning,

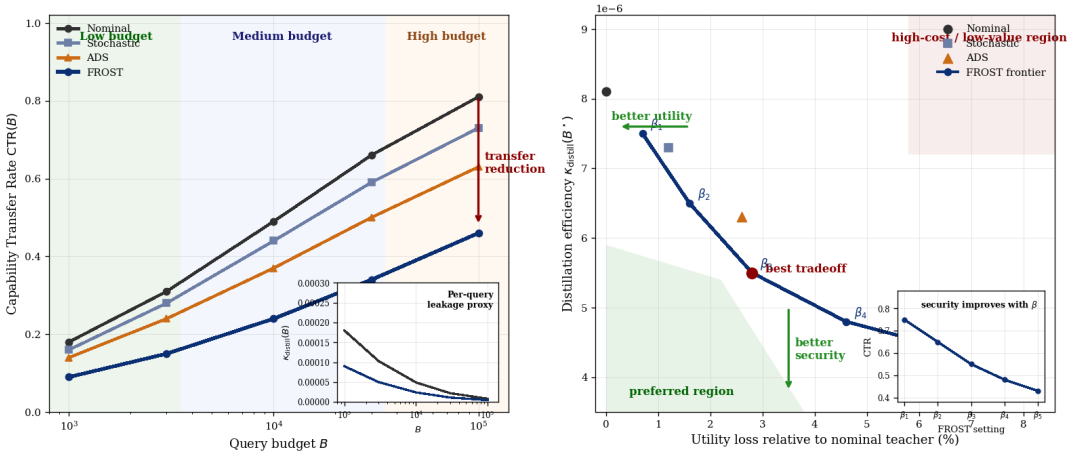

Figure 2: **Security–utility frontier of FROST.** The x-axis shows teacher utility loss and the y-axis shows distillation efficiency $\kappa_{\text{distill}}(B^\star)$. Lower values are better on both axes. The FROST frontier traces increasing defense strength β, showing that stronger defenses reduce transferable supervision at the cost of higher utility degradation. Compared with stochastic decoding and ADS, FROST achieves lower distillation efficiency at comparable utility cost.

Table 1: **CTR reduction (%) across students and harvesting regimes.** FROST is strongest in structured/mixed regimes, where traces contain reusable reasoning supervision and transferable scaffolds.

Attacker	Benign	Structured	Mixed
Llama-3.1-8B	**24**	**39**	**33**
Mistral-7B	**28**	**43**	**36**
Qwen2.5-3B	**31**	**48**	**41**

Defense	GSM8K	MATH	MMLU	HEval	MBPP	CTR↓	CTR redu.
Nominal decoding	61.8	33.4	64.9	35.7	58.6	0.810	—
Stochastic decoding	58.9	31.2	63.1	33.0	56.4	0.730	9.9%
ADS	54.7	28.1	60.4	29.8	52.3	0.630	22.2%
FROST (Geom. only)	47.3	22.9	55.8	24.1	46.8	0.510	37.0%
FROST (Full)	**42.6**	**19.7**	**52.1**	**20.8**	**42.9**	**0.460**	**43.2%**

Table 2: **Main distillation-resistance results.** Lower attacker-student scores indicate less recovered capability, not worse defended-teacher utility. Full FROST reduces CTR by 43.2% relative to nominal decoding.

knowledge, and code benchmarks, showing broad transfer reduction.

Claim 2: FROST reduces per-query distillation efficiency. The inset reports $\kappa_{\text{distill}}(B) = \text{CTR}(B)/B$, the capability transfer obtained per harvested query. FROST lowers both ***total recovered capability*** and ***per-query transfer efficiency***, meaning each black-box interaction carries less usable supervision for downstream training.

Claim 3: the gain is not generic randomness. FROST reduces transfer more than stochastic decoding at comparable utility cost, and outperforms ADS, especially in trace-rich settings. Geometry-only FROST captures much of the gain, while full FROST performs best, showing that ***structural suppression further protects against reusable scaffolds***. The flatter transfer curve also indicates lower marginal value from additional harvested queries.

4.2 Teacher Utility and Security Tradeoff

Claim 4: FROST is not simply making the teacher worse. Reduced transfer matters only if the teacher remains useful. Figure 4 shows that FROST achieves a favorable ***security–utility frontier***: distillation efficiency falls substantially before teacher utility degrades materially. Compared with stochastic decoding and ADS, FROST gives stronger transfer suppression at comparable utility cost, indicating that the gain is not explained by degraded teacher quality.

Claim 5: moderate defense strength is the deployment-relevant regime. Very weak defenses do little; very strong defenses cost too much. The useful operating region lies at moderate β, where FROST reduces distillation efficiency while preserving most teacher-side utility.

4.3 Robustness Across Attackers

Claim 6: FROST is not tied to one attacker. Table 1 shows that FROST remains consistently effective across diverse Llama, Mistral, and Qwen attackers, indicating that it does not merely exploit one attacker pipeline. ***Claim 7: FROST is strongest on reusable traces.*** Table 1 shows the largest reductions in structured and mixed regimes, where transcripts expose reusable reasoning and code scaffolds. This supports the thermodynamic

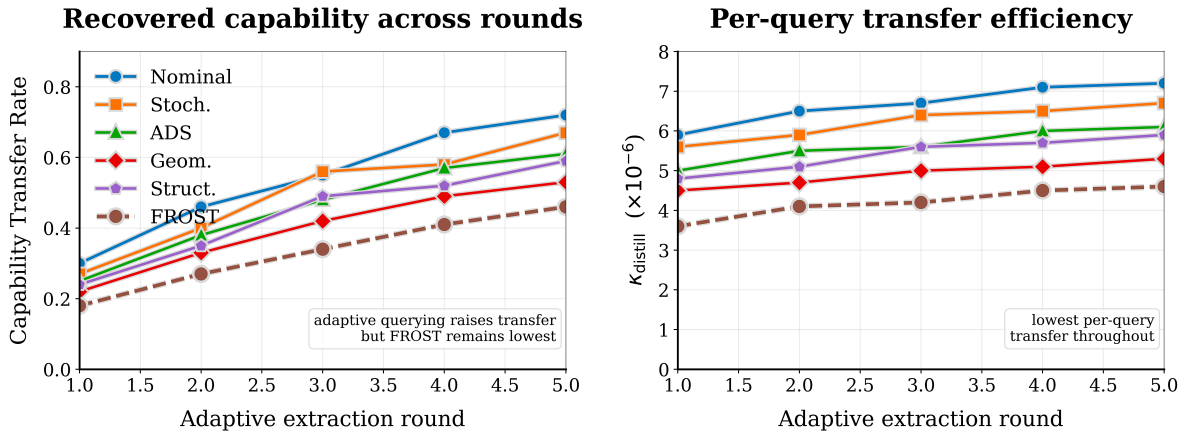

Figure 3: **Closed-loop adaptive extraction under a stronger attacker.** The attacker alternates between teacher querying, student fine-tuning, held-out evaluation, and targeted re-querying. **Left:** attacker-student Capability Transfer Rate after each adaptive round. **Right:** corresponding per-query distillation efficiency κ_{distill}. Although adaptive extraction increases transfer for all methods, the relative ordering remains stable: **full FROST** consistently yields the lowest recovered capability and the lowest transfer efficiency throughout the closed-loop attack.

view that structured traces carry higher downstream learning value.

4.4 Closed-Loop Adaptive Extraction

A stronger attacker need not harvest a fixed transcript once and train a student offline. To test whether ***FROST*** remains effective under feedback-driven extraction, we evaluate a ***closed-loop adaptive attacker***. At round r, the attacker queries the defended teacher, fine-tunes the current student on the newly harvested transcript, evaluates the student on a held-out capability slice, and reallocates the next round of queries toward regions where the student remains weak.

We also consider a ***defense-aware*** variant in which query allocation is biased toward prompts where ***FROST*** most strongly reshapes otherwise reusable outputs. Thus, the attacker is not only adaptive to the student, but partially adaptive to the defense itself. Figure 3 reports both the recovered ***Capability Transfer Rate*** and the per-query distillation efficiency $\kappa_{\text{distill}}(B_r)$ across five adaptive rounds. Adaptive extraction increases transfer for all methods, confirming that the attacker strengthens over time. However, the ordering remains stable: ***nominal decoding*** is easiest to distill, ***stochastic decoding*** gives limited protection, ***ADS*** improves resistance, and ***full FROST remains consistently lowest***.

This result shows that ***FROST*** is not merely disrupting one-shot transcript collection. Its ***geometric anti-copying signal*** continues to reduce the reusable learning value of outputs even when the attacker updates its student and redirects future queries. The lower κ_{distill} further indicates that each additional black-box interaction yields less downstream capability under ***FROST***.All results are given in more depth in K.

5 Conclusion

FROST introduces a teacher-side framework for reducing capability transfer under black-box querying by treating generation as the ***controlled release of learnable supervision***. Instead of relying solely on telemetry, monitoring, or heuristic perturbations, FROST formulates defended decoding as a thermodynamic free-energy control problem that reweights the teacher distribution against outputs with high downstream learning value. By combining ***attention-transport geometry***, structural trace-risk shaping, and provenance-aware decoding within a KL-regularized Gibbs framework, FROST reduces the distillability of emitted traces while preserving teacher utility and response quality. More broadly, the framework reframes API security as an output-channel control problem in which the objective is not only to answer correctly, but also to regulate how much transferable structure responses expose to downstream students. Future work should study stronger adaptive attackers, multilingual and multimodal harvesting, and tighter theoretical bounds relating information geometry to transferable supervision.Broadly,FROST suggests that future APIs may need to optimize not only for utility and alignment, but also for controlled exposure of learnable structure during generation.

Limitations

FROST is a ***teacher-side output-shaping defense***, not a complete solution to model extraction. We summarize the main limitations along three axes: ***(i)*** what FROST does not guarantee, ***(ii)*** where its approximations may fail, and ***(iii)*** what remains open for deployment.

No formal security guarantee. ***(i)*** FROST is not a cryptographic defense and does not make extraction impossible. ***(ii)*** Attackers can still query the teacher, collect outputs, combine them with external corpora, and train downstream students. ***(iii)*** The information-theoretic view is best read as intuition about per-query leakage, not as a formal privacy guarantee.

Proxy-student approximation. ***(i)*** The Fisher-geometric anti-copying signal depends on proxy students that approximate attacker learning dynamics. ***(ii)*** Real attackers may use different architectures, optimizers, curricula, post-processing pipelines, or synthetic augmentation. ***(iii)*** Although we evaluate multiple attackers and both full fine-tuning and PEFT, no finite proxy ensemble covers the full future attack space.

Approximate curvature estimation. ***(i)*** Exact Fisher estimation and inversion are infeasible for large models during decoding. ***(ii)*** FROST therefore relies on tractable curvature approximations that may misestimate some traces, especially for unseen attackers or long structured outputs. ***(iii)*** Better low-rank, block-structured, or amortized Fisher approximations are important future work.

Adaptive attackers. ***(i)*** FROST reduces extraction efficiency, but adaptive attackers remain a moving target. ***(ii)*** Stronger adversaries may search for low-energy but reusable outputs, paraphrase suppressed traces, filter transcripts, or reconstruct structured supervision. ***(iii)*** Future evaluations should include prompt-optimization attacks, multi-stage harvesters, and reconstruction-based student training.

Utility–security tradeoff. ***(i)*** The most useful answers for legitimate users are often also useful supervision for attackers. ***(ii)*** Stronger FROST settings may reduce explanation depth, code completeness, or pedagogical structure. ***(iii)*** Deployment therefore requires choosing an operating point on the security–utility frontier, not expecting a universally optimal defense strength.

Deployment-scale uncertainty. ***(i)*** Our experiments use controlled black-box harvesting pipelines and open-weight models. ***(ii)*** Production APIs may involve larger teachers, proprietary safety layers, tool use, retrieval augmentation, streaming generation, and long-context conversations. ***(iii)*** FROST should be validated under production-like serving constraints, including latency, throughput, memory, and cost.

Coverage of modalities and languages. ***(i)*** The evaluation focuses mainly on English text tasks covering reasoning, knowledge, code, and dialogue. ***(ii)*** Multilingual reasoning, low-resource languages, multimodal outputs, speech interfaces, and embodied-agent settings may expose different reusable supervision. ***(iii)*** Extending antidistillation evaluation to these regimes is necessary.

Safety and alignment interactions. ***(i)*** Antidistillation can interact with safety behavior. ***(ii)*** Some safe responses, including careful refusals or safety explanations, may themselves be structured and reusable. ***(iii)*** Broader study is needed to ensure that reducing learnability does not degrade refusal quality, calibrated helpfulness, or user trust.

Scope of protection. ***(i)*** FROST protects only the ***output channel***. ***(ii)*** It does not address checkpoint theft, insider compromise, white-box access, gradient leakage, infrastructure attacks, or unauthorized access to training data. ***(iii)*** It should be combined with monitoring, watermarking, access control, and platform-level abuse detection in a layered defense.

Open scientific question. ***(i)*** Capability transfer itself is not fully understood. ***(ii)*** FROST operationalizes one hypothesis: transferable traces induce large Fisher-geometric learning progress and expose reusable structure. ***(iii)*** Understanding why some traces teach models better than others remains central future work.

Ethics and Broader Impacts

A primary ethical concern when perturbing the outputs of a deployed language model is the potential degradation of its safety guardrails. Antidistillation defenses can unintentionally interact with safety behavior because safe responses—such as careful refusals or detailed safety explanations—often rely on structured and reusable linguistic patterns that defensive mechanisms aim to suppress. To ensure that reducing downstream learnability does

not compromise user trust or calibrated helpfulness, we evaluated FROST using safety sanity checks, including HarmBench. These evaluations indicate that FROST preserves the teacher model's baseline safety profile by maintaining refusal correctness and benign-answer rates without systematically increasing harmful compliance or introducing arbitrary behavioral drift.

FROST is motivated by the ethical and security need to protect deployed frontier models from unauthorized black-box distillation, which can enable capability theft, private data leakage, and control-logic exfiltration. By reducing the downstream learnability of emitted traces, FROST helps API providers mitigate attempts to bypass access controls, pricing structures, and deployment governance mechanisms. At the same time, FROST should not be interpreted as a formal cryptographic privacy guarantee. The framework provides a thermodynamic and information-theoretic mitigation focused on the output channel and does not defend against threats such as checkpoint theft, insider compromise, or gradient leakage. Consequently, FROST is best viewed as one component within a broader layered defense strategy that includes watermarking, access controls, auditing, and platform-level abuse detection.

An ethically responsible deployment of defensive decoding must also preserve utility for legitimate users. Excessively randomizing or truncating outputs could significantly degrade task performance and unfairly penalize benign API consumers. To avoid this tradeoff, FROST is formulated as a utility-constrained free-energy control framework that employs a KL trust region to keep the defended distribution close to the nominal teacher distribution. Empirical evaluations demonstrate that FROST substantially reduces the efficiency of capability stealing while preserving the teacher model's reasoning quality, code generation performance, and conversational helpfulness for ordinary users.

References

Shun-ichi Amari. 1998. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, and Charles Sutton. 2021. Program synthesis with large language models. *CoRR*, abs/2108.07732.

Nicholas Carlini, Florian Tramèr, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom B. Brown, Dawn Song, Úlfar Erlingsson, Alina Oprea, and Colin Raffel. 2021. Extracting training data from large language models. In *30th USENIX Security Symposium (USENIX Security 21)*, pages 2633–2650. USENIX Association.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, and 39 others. 2021. Evaluating large language models trained on code. *CoRR*, abs/2107.03374.

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. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Sumanth Dathathri, Abigail See, Shivani Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, and 1 others. 2024. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823.

Jiayu Ding, Lei Cui, Li Dong, Nanning Zheng, and Furu Wei. 2025. Information-preserving reformulation of reasoning traces for antidistillation. *arXiv preprint arXiv:2510.11545*.

Gintare Karolina Dziugaite and Daniel Roy. 2018. Entropy-SGD optimizes the prior of a PAC-Bayes bound: Generalization properties of entropy-SGD and data-dependent priors. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1377–1386.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021a. Measuring massive multitask language understanding. *CoRR*, abs/2009.03300.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021b. Measuring mathematical problem solving with the MATH dataset. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *CoRR*, abs/1503.02531.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations (ICLR)*.

Bo Jiang. 2026. DistillGuard: Evaluating defenses against LLM knowledge distillation. *arXiv preprint arXiv:2603.07835*.

John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. 2023. A watermark for large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 17061–17084. PMLR.

Felix Leibfried and Jordi Grau-Moya. 2020. Mutual-information regularization in markov decision processes and actor-critic learning. In *Proceedings of the Conference on Robot Learning*, volume 100 of *Proceedings of Machine Learning Research*, pages 360–373.

Pingzhi Li, Zhen Tan, Huaizhi Qu, Huan Liu, and Tianlong Chen. 2025. DOGe: Defensive output generation for LLM protection against knowledge distillation. *arXiv preprint arXiv:2505.19504*.

Xinhang Ma, William Yeoh, Ning Zhang, and Yevgeniy Vorobeychik. 2026. Protecting language models against unauthorized distillation through trace rewriting. *arXiv preprint arXiv:2602.15143*.

Mantas Mazeika, Long Phan, Xuwang Yin, Andy Zou, Zifan Wang, Norman Mu, Ehsan Sakhaee, Nathan Li, Steven Basart, Bo Li, David Forsyth, and Dan Hendrycks. 2024. Harmbench: A standardized evaluation framework for automated red teaming and robust refusal. *CoRR*, abs/2402.04249.

Sidharth Mudgal, Jong Lee, Harish Ganapathy, Yaguang Li, Tao Wang, Yanping Huang, Zhifeng Chen, Heng-Tze Cheng, Michael Collins, Trevor Strohman, Jilin Chen, Alex Beutel, and Ahmad Beirami. 2024. Controlled decoding from language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36486–36503.

Milad Nasr, Nicholas Carlini, Jonathan Hayase, Matthew Jagielski, A. Feder Cooper, Daphne Ippolito, Christopher A. Choquette-Choo, Eric Wallace, Florian Tramèr, and Katherine Lee. 2023. Scalable extraction of training data from (production) language models. *CoRR*, abs/2311.17035.

Tribhuvanesh Orekondy, Bernt Schiele, and Mario Fritz. 2019. Knockoff nets: Stealing functionality of black-box models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4954–4963.

Qi Pang, Shengyuan Hu, Wenting Zheng, and Virginia Smith. 2024. No free lunch in LLM watermarking: Trade-offs in watermarking design choices. In *Advances in Neural Information Processing Systems*, volume 37. NeurIPS 2024.

Daniel Russo and James Zou. 2016. Controlling bias in adaptive data analysis using information theory. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51 of *Proceedings of Machine Learning Research*, pages 1232–1240.

Vinu Sankar Sadasivan, Aounon Kumar, Sriram Balasubramanian, Wenxiao Wang, and Soheil Feizi. 2023. Can ai-generated text be reliably detected? *arXiv preprint arXiv:2303.11156*.

Yash Savani, Asher Trockman, Zhili Feng, Avi Schwarzschild, Alexander Robey, Marc Finzi, and J. Zico Kolter. 2025. Antidistillation sampling. *Preprint*, arXiv:2504.13146.

Yevgeny Seldin and Naftali Tishby. 2009. Pac-bayesian generalization bound for density estimation with application to co-clustering. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 472–479.

Aman Sinha, Hongseok Namkoong, and John Duchi. 2018. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations (ICLR)*.

Thomas Steinke and Lydia Zakynthinou. 2020. Reasoning about generalization via conditional mutual information. In *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3437–3452.

Naftali Tishby, Fernando C. Pereira, and William Bialek. 2000. The information bottleneck method. *CoRR*, physics/0004057.

Florian Tramèr, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart. 2016. Stealing machine learning models via prediction APIs. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 601–618, Austin, TX. USENIX Association.

Junlin Wang, Tianyi Yang, Roy Xie, and Bhuwan Dhingra. 2024. Raccoon: Prompt extraction benchmark of LLM-integrated applications. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13349–13365. Association for Computational Linguistics.

Collin Zhang, John X. Morris, and Vitaly Shmatikov. 2024. Extracting prompts by inverting LLM outputs. *Preprint*, arXiv:2405.15012.

Hanlin Zhang, Benjamin L. Edelman, Danilo Francati, Daniele Venturi, Giuseppe Ateniese, and Boaz Barak. 2023. Watermarks in the sand: Impossibility of strong watermarking for generative models. *arXiv preprint arXiv:2311.04378*.

Kaixiang Zhao, Lincan Li, Kaize Ding, Neil Zhenqiang Gong, Yue Zhao, and Yushun Dong. 2025. A survey on model extraction attacks and defenses for large language models. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '25)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. *CoRR*, abs/2306.05685.

Appendix

A The FROST Framework

Overview and central challenge. The central challenge addressed by FROST is deceptively simple: a frontier model exposed through a black-box API must remain ***useful*** to legitimate users, yet its outputs should not become ***cheap supervision*** for downstream student models. This tension is what makes teacher-side antidistillation fundamentally difficult. If the teacher emits rich, explicit, and pedagogically structured responses, it maximizes utility for users but also increases the transferability of capability to an attacker harvesting outputs at scale. If, on the other hand, the teacher aggressively suppresses, truncates, or randomizes its generations, it may reduce distillability only by collapsing the very API value it is meant to provide. Existing black-box extraction work establishes that queried outputs can suffice to recover high-fidelity surrogates (Tramèr et al., 2016; Orekondy et al., 2019; Zhao et al., 2025), while recent teacher-side defenses such as *Antidistillation Sampling* show that output shaping can indeed reduce downstream student performance (Savani et al., 2025). ***FROST is built around the stronger principle that defended generation should be formulated as a utility-constrained control problem over the learnability of emitted traces.*** Concretely, FROST treats each candidate response not merely as text to be returned, but as a potential training example whose downstream learning value must be regulated through ***thermodynamic reweighting*** and ***geometric anti-copying***.

A.1 Problem Setup and Threat Model

Motivation. Knowledge distillation is classically introduced as a ***compression*** mechanism: a smaller student model is trained to imitate a stronger teacher so as to retain predictive utility at lower inference cost (Hinton et al., 2015). However, once a frontier model is deployed behind a query interface, the same teacher–student pipeline acquires a security interpretation. A long line of work on ***model extraction*** shows that black-box query access can suffice to recover high-fidelity surrogates of deployed models (Tramèr et al., 2016; Orekondy et al., 2019). Recent surveys argue that the LLM setting only broadens this attack surface, because outputs now include not just labels or confidence values, but also natural-language explanations, code, structured traces, and reusable reasoning patterns (Zhao et al., 2025). More recently, *Antidistillation Sampling* explicitly studies the teacher-side problem of making reasoning traces ***less useful for student training while preserving teacher utility*** (Savani et al., 2025). FROST is designed for precisely this deployment regime.

Teacher as a deployed stochastic kernel. We model the deployed teacher not merely as a fixed set of weights, but as a ***conditional output channel***. Let

$$x \in \mathcal{X} \qquad \text{and} \qquad y \in \mathcal{Y}$$

denote, respectively, a user prompt and a generated response. Let

$$P_0(y \mid x)$$

be the ***nominal teacher distribution***: the response distribution induced by the deployed frontier model under its standard decoding stack. This formulation is intentionally general. It absorbs the base language model, safety layers, refusal policies, prompt formatting, and standard decoding settings into a single conditional distribution over outputs. FROST will construct a defended distribution

$$Q(y \mid x),$$

which is the only object exposed to the attacker at inference time.

Black-box adaptive attacker. We consider a black-box adversary that can query the teacher repeatedly, but does ***not*** observe logits, gradients, parameters, optimizer state, hidden activations, or any internal telemetry. The attacker is allowed to choose prompts *adaptively*. Formally, for a total query budget B, the attacker collects the transcript

$$\mathcal{T}_B = \{(x_t, y_t)\}_{t=1}^{B}, \qquad y_t \sim Q(\cdot \mid x_t),$$

where x_t may depend arbitrarily on the past:

$$x_t \sim \pi_t(\cdot \mid x_{<t}, y_{<t}).$$

This captures the operational reality emphasized in the model-extraction literature: attacks need not use a fixed prompt set, and can actively refine future queries using previous outputs (Tramèr et al., 2016; Orekondy et al., 2019; Zhao et al., 2025).

Student learning from harvested traces. After collecting $\mathcal{T}_B$, the attacker trains a student S_θ by

imitation. At a minimum, this can be modeled as standard supervised fine-tuning on teacher outputs:

$$\hat{\mathcal{L}}_{\text{imit}}(\theta; \mathcal{T}_B) = \frac{1}{B} \sum_{t=1}^{B} \ell_\theta(x_t, y_t),$$
$$\ell_\theta(x_t, y_t) := -\log p_\theta(y_t \mid x_t).$$

This formulation subsumes several practical attack instantiations:

- ***answer-only distillation***, where the student imitates final responses;
- ***trace-rich distillation***, where the student is trained on longer explanations, decompositions, or code;
- ***cross-architecture transfer***, where the student architecture differs from the teacher;
- ***parameter-efficient attacks***, such as LoRA-style fine-tuning on harvested transcripts.

These attack families are consistent with both the broader extraction literature and the recent antidistillation literature, which explicitly treats model outputs as valuable downstream supervision (Orekondy et al., 2019; Zhao et al., 2025; Savani et al., 2025).

What the defender can and cannot use. A central design decision in FROST is that the defense is ***identity-free*** and ***prompt-local***. That is, FROST does *not* assume access to user accounts, IP addresses, cross-session telemetry, or metadata-level coordination signals. Instead, it acts only through the ***output distribution conditioned on the current prompt***. This makes the defense conceptually stronger and operationally cleaner: even if the attacker rotates accounts, proxies requests, or otherwise hides telemetry, the output channel itself remains protected.

Core threat-model assumptions. For clarity, we summarize the main assumptions:

1. The attacker has ***black-box query access*** to the teacher.
2. The attacker may choose queries ***adaptively*** based on previous responses.
3. The attacker trains a downstream student using harvested teacher outputs as supervision.
4. The defender operates purely at the level of ***teacher-side output shaping***, with no dependence on identity or telemetry.

Design goal. Under this threat model, the goal of FROST is ***not*** to make extraction information-theoretically impossible in an absolute sense. Rather, the goal is to make each emitted response ***intrinsically less useful as training signal*** for downstream students, while preserving the teacher's utility for legitimate users. In other words, FROST treats the output channel itself as the object of protection.

A.2 Utility-Preserving Antidistillation Objective

Why utility must remain explicit. A trivial antidistillation strategy would be to degrade outputs arbitrarily: shorten them, randomize them, or return nonsensical responses. Such strategies are useless in practice because they collapse the API's value for normal users. Accordingly, FROST is built around a stricter requirement: ***teacher-side defense must preserve utility while reducing learnability***. This principle is consistent with recent teacher-side defenses such as *Antidistillation Sampling*, which explicitly targets lower distillability at comparable teacher performance, and with inference-time watermarking methods such as SynthID-Text and the watermark of Kirchenbauer et al., both of which modify sampling while aiming to preserve output quality (Savani et al., 2025).

Utility as a constrained resource. Let

$$u(x, y) \in \mathbb{R}$$

denote a ***utility functional*** measuring the quality of response y to prompt x. This quantity is intentionally broad: depending on the evaluation slice, it may encode task accuracy, helpfulness, code correctness, reasoning quality, formatting fidelity, or calibrated refusal behavior. For a prompt distribution $\mathcal{D}$, the expected teacher utility under defended decoding Q is

$$U(Q; \mathcal{D}) = \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E}_{y \sim Q(\cdot|x)} [u(x, y)] \right].$$

FROST does not maximize $U(Q; \mathcal{D})$ from scratch; rather, it seeks to remain close to the utility of the nominal teacher distribution P_0. This distinction is important: the problem is ***utility-preserving protection***, not generic response optimization.

Learnability as the defended quantity. To formalize teacher-side antidistillation, we introduce an ***anti-copying energy***

$$E_x(y) \geq 0,$$

which quantifies how useful output y would be as supervision for a downstream student if emitted in response to prompt x. Intuitively:

- ***low-energy outputs*** are weak training signal for attackers;
- ***high-energy outputs*** are rich, reusable, student-helpful traces.

At this stage, $E_x(y)$ is kept abstract. In the next subsection, FROST will instantiate it using ***geometric learning progress***. For now, it serves as a generic measure of downstream learnability.

The constrained antidistillation program. FROST seeks a defended teacher distribution Q that makes emitted traces difficult to distill while preserving utility. The most transparent formulation is the following constrained program:

$$\min_{Q} \quad \mathbb{E}_{x\sim\mathcal{D}}\left[\mathbb{E}_{y\sim Q(\cdot|x)}[E_x(y)]\right]$$

subject to

$$U(Q;\mathcal{D}) \geq U_{\min}$$

and

$$\mathbb{E}_{x\sim\mathcal{D}}[D_{\mathrm{KL}}(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x))] \leq \delta.$$

These two constraints play distinct roles:

1. ***Utility floor*** $U_{\min}$: the defended teacher must remain useful to legitimate users.
2. ***Trust-region constraint*** δ: the defended distribution must remain close to the nominal teacher, preventing arbitrary or degenerate perturbations.

Why the KL trust region matters. The KL term is not only mathematically convenient; it formalizes a key deployment desideratum. FROST should be understood as a ***minimal intervention*** on the original teacher. By constraining

$$D_{\mathrm{KL}}(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x)),$$

we force the defense to remain close to standard behavior unless doing so would expose disproportionately high-value training signal. This gives a clean interpretation: FROST is not a separate assistant layered on top of the teacher, but a ***defended perturbation*** of the teacher's own response distribution.

Lagrangian form. Introducing Lagrange multipliers yields the equivalent unconstrained objective

$$\begin{aligned}\min_{Q} \quad \mathbb{E}_{x\sim\mathcal{D}}\Big[&\mathbb{E}_{y\sim Q(\cdot|x)}[E_x(y)] \\ &+ \lambda_u\, \Phi_u(Q;x) \\ &+ \lambda_{\mathrm{KL}}\, D_{\mathrm{KL}}\big(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x)\big)\Big].\end{aligned}$$

where $\Phi_u(Q;x)$ is a penalty for falling below the desired utility level on prompt x. In practice, this formulation makes explicit the ***three-way tradeoff*** at the heart of teacher-side antidistillation:

lower learnability vs. ***higher utility*** vs. ***smaller deviation from nominal behavior***.

Interpretation. This objective already distinguishes FROST from simpler output-space defenses:

- it is ***not*** mere randomization, because perturbations are scored against a learnability objective;
- it is ***not*** mere provenance marking, because the primary target is transfer efficiency, not only detectability (Dathathri et al., 2024);
- it is ***not*** post-hoc monitoring, because the intervention is built directly into generation itself (Savani et al., 2025).

The role of the next subsection is therefore to answer the central remaining question: ***how should one measure the learnability of an emitted trace?*** FROST answers this by moving from a generic anti-copying energy $E_x(y)$ to a ***geometric*** notion of student learning progress.

A.3 Free-Energy Formulation of Defended Decoding

Why a free-energy formulation? The constrained objective in Section 3.2 makes clear that teacher-side antidistillation is not a standard decoding problem. The defender does not merely wish to maximize utility, nor merely to minimize deviation from the nominal teacher, nor merely to suppress structured traces. Rather, the defender must solve a ***three-way control problem***:

low downstream learnability– high teacher utility– small deviation from nominal behavior.

Objectives of the general form "risk plus KL divergence to a reference distribution" arise naturally in several adjacent literatures, including ***Gibbs/PAC-Bayesian posteriors*** and ***KL-regularized control***.

In PAC-Bayesian settings, the posterior is chosen by trading off empirical performance against KL divergence from a prior (Seldin and Tishby, 2009; Dziugaite and Roy, 2018). In modern language-model control, KL-regularized reinforcement learning is used to keep generations close to a base model while steering them toward desired reward signals, and *Controlled Decoding* shows that tokenwise controlled generation can be interpreted as sampling from the solution of such a regularized objective (Mudgal et al., 2024; Leibfried and Grau-Moya, 2020). FROST adopts this variational perspective, but with a different control signal: ***the learnability of an emitted trace for downstream students***.

Prompt-local defended decoding. Fix a prompt $x \in \mathcal{X}$. Let

$$P_0(\cdot \mid x)$$

denote the nominal teacher distribution, and let

$$Q(\cdot \mid x)$$

denote a candidate defended decoding distribution. Let

$$E_x(y) \geq 0$$

be the anti-copying energy introduced in Section 3.2, interpreted as the ***downstream learning value*** of response y for an adversarial student. At this stage, we do not yet assume any specific form for $E_x(y)$; the geometric instantiation is deferred to Section 3.4. The core idea of FROST is to choose $Q(\cdot \mid x)$ by minimizing a prompt-local ***free-energy functional***:

$$\begin{aligned}\mathcal{F}_\beta(Q; x) =& \mathbb{E}_{y\sim Q(\cdot|x)}[E_x(y)] \\ &+ \frac{1}{\beta} D_{\mathrm{KL}}\big(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x)\big).\end{aligned}$$

where $\beta > 0$ is an inverse-temperature parameter controlling the strength of antidistillation.

This objective has a direct operational meaning:

- the first term penalizes ***distillable outputs***;
- the second term penalizes ***deviation from the nominal teacher***;
- β controls how aggressively the defender trades utility-preserving fidelity for anti-copying robustness.

Thus, FROST does not search over arbitrary responses. It computes the ***least disruptive reweighting*** of the original teacher distribution that suppresses high-learnability traces.

Variational definition of FROST. For each prompt x, the defended decoding rule is defined as

$$Q^\star(\cdot \mid x) = \arg\min_{Q(\cdot|x)} \mathcal{F}_\beta(Q; x).$$

This choice yields an explicit Gibbs form.

Proposition 1 (Variational optimality of FROST). For any fixed prompt x, the unique minimizer of

$$\mathcal{F}_\beta(Q; x) = \mathbb{E}_Q[E_x(y)] + \frac{1}{\beta} D_{\mathrm{KL}}(Q \| P_0)$$

over all distributions $Q(\cdot \mid x)$ absolutely continuous with respect to $P_0(\cdot \mid x)$ is

$$Q^\star(y \mid x) = \frac{1}{Z_\beta(x)} P_0(y \mid x) \exp\big(-\beta E_x(y)\big),$$

where

$$Z_\beta(x) = \sum_{y\in\mathcal{Y}} P_0(y \mid x) \exp\big(-\beta E_x(y)\big)$$

is the partition function.

Proof sketch. Define

$$Q^\star_\beta(y \mid x) := \frac{1}{Z_\beta(x)} P_0(y \mid x) e^{-\beta E_x(y)}.$$

Consider the KL divergence between any candidate Q and $Q^\star_\beta$:

$$D_{\mathrm{KL}}(Q \| Q^\star_\beta) = \sum_y Q(y \mid x) \log \frac{Q(y \mid x)}{Q^\star_\beta(y \mid x)}.$$

Substituting the definition of $Q^\star_\beta$,

$$\begin{aligned}\log \frac{Q(y \mid x)}{Q^\star_\beta(y \mid x)} = \log \frac{Q(y \mid x)}{P_0(y \mid x)} + \beta E_x(y) \\ + \log Z_\beta(x).\end{aligned}$$

Taking expectation under Q gives

$$\begin{aligned}D_{\mathrm{KL}}(Q \| Q^\star_\beta) = D_{\mathrm{KL}}(Q \| P_0) \\ + \beta \mathbb{E}_Q[E_x(y)] \\ + \log Z_\beta(x).\end{aligned}$$

Rearranging,

$$\mathcal{F}_\beta(Q; x) = \frac{1}{\beta} D_{\mathrm{KL}}(Q \| Q^\star_\beta) - \frac{1}{\beta} \log Z_\beta(x).$$

Since KL divergence is nonnegative and equals zero if and only if $Q = Q^\star_\beta$, the minimizer is uniquely given by the Gibbs-reweighted distribution above. □

Interpretation of the Gibbs form. Proposition 1 reveals the exact structure of the defense. FROST does ***not*** replace the nominal teacher with a new model, nor does it enforce crude post-hoc truncation. Instead, it performs a ***Boltzmann-style reweighting*** of the original output law:

$$P_0(y \mid x) \;\longrightarrow\; Q^\star(y \mid x) \propto P_0(y \mid x) e^{-\beta E_x(y)}.$$

Outputs with large anti-copying energy are exponentially downweighted; outputs with small anti-copying energy are preserved. The partition function $Z_\beta(x)$ normalizes the defended distribution and encapsulates the prompt-specific cost of protection.

This interpretation is important for two reasons. First, it makes FROST a ***minimal-intervention*** defense: every protected output is still sampled from the support of the teacher's own response law. Second, it aligns FROST with the same variational family that appears in Gibbs posteriors and KL-regularized control, while changing the semantics of the "energy" term from *task loss* or *reward* to *distillability*.

Inverse temperature as a defense knob. The inverse temperature β has a clean operational role:

- as $\beta \to 0$, the KL term dominates and
$$Q^\star(\cdot \mid x) \to P_0(\cdot \mid x),$$
so the defense vanishes;
- as β increases, the decoding distribution becomes increasingly concentrated on low-energy, less-distillable traces;
- for excessively large β, the defense may over-regularize generation, potentially harming utility.

Thus, β induces a ***security–utility frontier***: low β yields high fidelity but weak protection, while high β yields stronger anti-copying at the risk of reduced helpfulness.

Thermodynamic diagnostics. The partition function
$$Z_\beta(x) = \sum_y P_0(y \mid x) e^{-\beta E_x(y)}$$
also gives FROST a useful thermodynamic interpretation. Define the prompt-level free energy
$$\Psi_\beta(x) := -\frac{1}{\beta} \log Z_\beta(x).$$
Then standard differentiation identities yield
$$-\frac{\partial}{\partial \beta} \log Z_\beta(x) = \mathbb{E}_{y \sim Q^\star(\cdot|x)}[E_x(y)],$$
and
$$\frac{\partial^2}{\partial \beta^2} \log Z_\beta(x) = \mathrm{Var}_{y \sim Q^\star(\cdot|x)}\big(E_x(y)\big).$$
These quantities have natural interpretations:

- the first derivative gives the ***expected defended distillability*** of generated responses;
- the second derivative gives the ***susceptibility*** of the defense, i.e., how sensitive defended decoding is to changes in β.

In other words, FROST is not merely inspired by thermodynamics metaphorically; it inherits an actual ***free-energy geometry over output distributions***.

Connection to controlled decoding. It is useful to contrast FROST with nearby inference-time control methods. Controlled Decoding formulates language-model generation as a tokenwise KL-regularized RL problem and uses a prefix scorer so that inference-time sampling follows the optimizer of a regularized control objective (Mudgal et al., 2024). Antidistillation Sampling, in turn, modifies next-token probabilities to make emitted traces less useful for downstream distillation while preserving practical model utility (Savani et al., 2025). FROST is conceptually aligned with both directions, but differs in the ***semantics of control***:

- in KL-regularized alignment, the control signal is usually ***reward***;
- in watermarking, the control signal is ***detectability or provenance***;
- in FROST, the control signal is ***downstream student learnability***.

This shift is crucial. FROST does not ask which outputs are most preferred, nor which are easiest to attribute, but rather which are ***least transferable as training signal*** under a utility-preserving trust region.

From constrained utility to free energy. The free-energy objective above can also be viewed as the Lagrangian relaxation of the constrained antidistillation program from Section 3.2. If the original problem is

$$\begin{aligned} \min_{Q} \quad & \mathbb{E}_{x\sim\mathcal{D}}\mathbb{E}_{y\sim Q(\cdot|x)}[E_x(y)] \\ \text{s.t.} \quad & U(Q;\mathcal{D}) \geq U_{\min}, \\ & \mathbb{E}_x D_{\mathrm{KL}}(Q\|P_0) \leq \delta. \end{aligned}$$

then for fixed Lagrange multipliers, the prompt-local subproblem reduces exactly to the minimization of

$$\mathbb{E}_Q[E_x(y)] + \frac{1}{\beta} D_{\mathrm{KL}}(Q\|P_0),$$

possibly with an additional utility-penalty term absorbed into the energy. This observation matters because it makes the defended distribution ***derivable***, not heuristic: the Gibbs form is the optimizer of a well-defined variational program.

What remains unspecified. At this point, the formal structure of defended decoding is complete, but one critical ingredient remains open: ***how should the anti-copying energy*** $E_x(y)$ ***be defined?*** The entire effectiveness of FROST depends on whether this energy actually captures the downstream learnability of an emitted trace. The next subsection answers this question by introducing a ***geometric*** notion of student learning progress, yielding a principled and attacker-aware instantiation of $E_x(y)$.

B The Geometric Energy Component-Thermodynamic Length

B.1 Generic Concept: Thermodynamic Length in a Transformer

A transformer of depth L processes an input by passing a representation through a sequence of layers. Regardless of what that representation *is*, a probability distribution over the vocabulary, a matrix of hidden token vectors, or a stochastic transport kernel over attention heads, each layer induces a "state" that lives on some geometric space $(\mathcal{S}, g)$ equipped with a Riemannian or pseudo-Riemannian metric. Adjacent layers produce neighboring states; the metric g lets us measure how far apart they are.

The core idea. Let $s_\ell \in \mathcal{S}$ denote the state extracted from layer ℓ. The **thermodynamic length** of the depth trajectory $\{s_\ell\}_{\ell=0}^{L}$ is the total arc length accumulated along the path in $\mathcal{S}$:

$$\mathcal{L} \;=\; \sum_{\ell=0}^{L-1} d_{\mathcal{S}}(s_\ell,\, s_{\ell+1}),$$

where $d_{\mathcal{S}}$ is a metric on $\mathcal{S}$ induced by g. This quantity is a discrete analogue of the Fisher-Rao arc length in continuous-time information geometry: if $\gamma : [0,1] \to \mathcal{S}$ is a smooth curve on a statistical manifold $(\mathcal{S}, g_F)$, its length is $\int_0^1 \sqrt{\dot\gamma^\top g_F(\gamma)\,\dot\gamma}\,\mathrm{d}t$, and the geodesic distance is obtained by minimizing over all curves joining two endpoints. Thermodynamic length is the canonical discretization of this arc length to the setting where depth plays the role of the continuous parameter.

Why "thermodynamic"? The name follows the analogy with thermodynamic length in statistical physics, where the length of a protocol on a manifold of equilibrium states quantifies the minimum dissipation required to execute the protocol. In the transformer context, $\mathcal{L}$ measures the total information-geometric *effort* expended by the model as it transforms its internal state from input embedding to final layer. Large $\mathcal{L}$ indicates that the network undergoes substantial routing or representational reorganization across depth; small $\mathcal{L}$ indicates a path that changes slowly and smoothly.

Layer-wise decomposition. Because equation (B.1) is additive, the contribution of each individual layer can be isolated. Define the layer-wise step length

$$\mathcal{L}_\ell \;=\; d_{\mathcal{S}}(s_\ell,\, s_{\ell+1}),$$

and recover $\mathcal{L} = \sum_\ell \mathcal{L}_\ell$. This decomposition is central to FROST: the anti-copying energy aggregates these per-layer line elements to assign a scalar learnability score to each candidate response.

Choice of state space and metric. The definition in equation (B.1) is a template. Different choices of state s_ℓ and metric $d_{\mathcal{S}}$ give rise to different instantiations of thermodynamic length, each probing a distinct channel of the model's internal computation. We consider three such instantiations detailed in Sections B.2–B.4.

B.2 Method 1: Attention-Transport Thermodynamic Length

State space. Method 1 works directly with the attention mechanism. For a sequence of length

n, each layer ℓ has H attention heads with attention matrices $\{A_{\ell,h}\}_{h=1}^{H}$, where each row of $A_{\ell,h} \in \mathbb{R}^{n\times n}$ is a categorical distribution over key positions. The state s_ℓ is a single *transport kernel* $P_\ell \in \mathbb{R}^{n\times n}$ obtained by compressing the H heads via entropy-weighted aggregation:

$$P_\ell = \frac{\sum_{h=1}^{H} \alpha_{\ell,h} A_{\ell,h}}{\sum_{h=1}^{H} \alpha_{\ell,h}},$$

$$\alpha_{\ell,h} = \left(1 - \frac{H(A_{\ell,h})}{\log n}\right)_+ .$$

where $H(A_{\ell,h})$ is the mean row entropy of head h at layer ℓ. Low-entropy (sharp) heads receive larger weights, so P_ℓ reflects the most informative routing at that layer. Each row of P_ℓ sums to one, making P_ℓ a Markov kernel that transports token-indexed probability mass in a single attention step.

Metric. The geometric space is the set of Markov kernels on n tokens, viewed as a statistical family parametrized by depth. The Fisher information of this family with respect to depth captures the expected squared sensitivity of the log-probabilities to infinitesimal changes in the parameter. In the discrete, unit-step setting, the natural surrogate for the local Fisher element is:

$$g_\ell \;=\; \sum_{i=1}^{n} \pi_\ell(i) \sum_{j=1}^{n} P_\ell(i,j)\big(\Delta \log P_\ell(i,j)\big)^2,$$

where

$$\Delta \log P_\ell(i,j) \;=\; \log P_\ell(i,j) - \log P_{\ell-1}(i,j),$$

where π_ℓ is the stationary distribution of P_ℓ. The inner sum averages squared log-increments over destinations j according to the current kernel, and the outer sum averages over sources according to the stationary distribution. When many edges change by large multiplicative factors the quadratic form spikes, signaling high information-geometric effort.

Layer-wise length and aggregate. The thermodynamic line element at layer ℓ is $\mathcal{L}_\ell^{(1)} = \sqrt{g_\ell}$, and the total attention-transport thermodynamic length for a candidate response y is:

$$\mathcal{L}^{(1)}(y) \;=\; \sum_{\ell=1}^{L} \sqrt{g_\ell}.$$

Large $\mathcal{L}^{(1)}(y)$ marks candidates that induce large routing displacement across depth, high transport-geometric effort and correspondingly high learnability for downstream students. FROST uses $\mathcal{E}_x^{\text{geom}}(y) = \mathcal{L}^{(1)}(y)$ as its primary anti-copying signal.

Implementation notes. Before any logarithm, small probabilities must be floored to a tiny constant and rows renormalized to preserve stochasticity. The stationary distribution π_ℓ is obtained by power iteration on $P_\ell^\top$. An alternative numerically stable expression for g_ℓ uses the average of both KL divergences between row-wise distributions at layers ℓ and $\ell - 1$, which behaves similarly while avoiding extreme log-ratio values.

B.3 Method 2: Parallel-Transport Thermodynamic Length

State space. Method 2 works in hidden-state space. For a batch of N token positions tracked consistently across layers, the state at layer ℓ is the row-centered matrix of token representations $X_\ell \in \mathbb{R}^{N\times d}$, where row i is the hidden vector $h_{\ell,i}^\top \in \mathbb{R}^d$. The metric at layer ℓ is given by the intrinsic (ridge-regularized) covariance

$$\Sigma_\ell \;=\; \frac{1}{N} X_\ell^\top X_\ell + \lambda I_d, \quad \lambda > 0,$$

which plays the role of the Fisher information matrix for the Gaussian model of token clouds at depth ℓ. Distances measured in the Mahalanobis metric induced by Σ_ℓ are intrinsic and comparable across layers.

Alignment via orthogonal Procrustes. Naïve comparison of $X_{\ell+1}$ to X_ℓ is confounded by arbitrary orthogonal changes of basis introduced by, e.g., LayerNorm-centered linear maps. Method 2 removes this gauge ambiguity by solving the *orthogonal Procrustes problem*: find $R_\ell \in O(d)$ minimizing $\|X_{\ell+1}R_\ell^\top - X_\ell\|_F$. Letting $C_\ell = \frac{1}{N} X_\ell^\top X_{\ell+1}$ be the cross-covariance, the unique solution is

$$R_\ell = \operatorname{polar}(C_\ell) = UV^\top,$$
$$\text{where } C_\ell = USV^\top \text{ (SVD)}.$$

Neighbors are then transported into the frame of ℓ: $X_{\ell+1\to\ell} = X_{\ell+1}R_\ell^\top$. This transport is proved to be gauge-invariant in Theorem 3 of the supplementary material: under any per-layer orthogonal spin $X_\ell \mapsto X_\ell Q_\ell$ with $Q_\ell \in O(d)$, the thermodynamic length is unchanged.

Metric and layer-wise length. The first covariant difference in the frame of ℓ is $\Delta_\ell = X_{\ell+1\to\ell} -$

X_ℓ. Whitening by $\Sigma_\ell^{-1/2}$ maps the local geometry to unit variance in all directions:

$$\widehat{\Delta}_\ell \;=\; \Delta_\ell\, \Sigma_\ell^{-1/2}.$$

The thermodynamic line element at layer ℓ and the total parallel-transport thermodynamic length are:

$$\mathcal{L}_\ell^{(2)} \;=\; \big\|\widehat{\Delta}_\ell\big\|_F,$$

$$\mathcal{L}^{(2)} \;=\; \sum_{\ell=0}^{L-2} \mathcal{L}_\ell^{(2)}.$$

The Fisher-Mahalanobis interpretation is direct: $(\mathcal{L}_\ell^{(2)})^2 = \sum_{i=1}^{N} \delta_{\ell,i}^\top \Sigma_\ell^{-1} \delta_{\ell,i}$, where $\delta_{\ell,i}$ is the i-th row of Δ_ℓ, i.e., the displacement of the i-th token representation in one layer step measured in the intrinsic units of the local metric. This justifies the name "thermodynamic length" without invoking any extra thermodynamic formalism.

Computational cost. Per batch of size B, sequence length T, hidden width d, vocabulary V, and L depth nodes: one forward pass plus one logit-lens projection and one softmax per depth node over selected positions. The dominant overhead for large V is the output head, $O(NdV)$. Restricting to the last $K \le NT$ supervised positions reduces runtime linearly in K. In streaming mode, only two consecutive states $(\widehat{\Delta}_{\ell-1}, \widehat{\Delta}_\ell)$ need to be held in memory at a time, costing $O(Nd) + O(NV)$.

B.4 Method 3: Belief-Simplex Thermodynamic Length

State space and motivation. Methods 1 and 2 probe the model's internal attention routing and hidden-state geometry, respectively. Method 3 works entirely in the space of the model's *predictive outputs*, bypassing hidden-state parameterization. The state at layer ℓ and position t is the next-token distribution

$$p_t^{(\ell)}(\cdot \mid x_{\le t}) \;=\; \mathrm{softmax}\big(z_t^{(\ell)}/\tau\big) \;\in\; \mathring{\Delta}^{V-1},$$

where $z_t^{(\ell)} = \mathrm{Head}(h_t^{(\ell)}) \in \mathbb{R}^V$ are the logits produced by applying the LM head to the hidden state at layer ℓ (the *logit lens*). The sequence $\ell \mapsto p_t^{(\ell)}$ is the *belief trajectory* for position t: it tracks the model's evolving predictive distribution as computation unfolds across layers.

Why the simplex needs a non-Euclidean metric. Two non-identifiabilities make naïve Euclidean geometry on Δ^{V-1} unreliable. First, the softmax is invariant to adding a constant offset to all logits, so $z_t^{(\ell)}$ is defined only up to translation along $\mathrm{span}\{\mathbf{1}\}$. Second, an absolute change of 10^{-3} in a low-probability coordinate has a larger effect on log-likelihood than the same change in a high-probability coordinate. Information geometry resolves both issues by equipping Δ^{V-1} with the Fisher-Rao metric, the unique (up to scale) Riemannian structure invariant under sufficient statistics.

Square-root embedding and Fisher-Rao distance. The Fisher metric on Δ^{V-1} coincides with the Euclidean metric on the positive orthant of the unit sphere S_+^{V-1} under the square-root embedding

$$u_t^{(\ell)} \;:=\; \psi\big(p_t^{(\ell)}\big) \;:=\; \sqrt{p_t^{(\ell)}} \;\in\; S_+^{V-1}.$$

Under this embedding, Fisher-Rao geometry on Δ^{V-1} is isometric (up to a constant factor of 2) to standard spherical geometry on S^{V-1}, so the geodesic distance is a great-circle arc length. Concretely, the Fisher-Rao distance between adjacent-layer beliefs is

$$\begin{aligned} d_{\mathrm{FR}}\big(p_t^{(\ell)},\, p_t^{(\ell+1)}\big) &= 2\arccos\left(\sum_{i=1}^{V} \sqrt{p_{t,i}^{(\ell)}\, p_{t,i}^{(\ell+1)}}\right) \\ &= 2\arccos\big\langle u_t^{(\ell)}, u_t^{(\ell+1)}\big\rangle. \end{aligned}$$

which equals twice the angle between successive square-root embeddings on the sphere. The argument of $\arccos$ is the *Bhattacharyya coefficient*, a classical overlap measure between two distributions.

Layer-wise length and aggregate. The thermodynamic length along depth for position t is the cumulative Fisher-Rao arc length of the belief trajectory:

$$\mathrm{Length}_t \;:=\; \sum_{\ell=0}^{L-1} d_{\mathrm{FR}}\big(p_t^{(\ell)},\, p_t^{(\ell+1)}\big).$$

We define the layer-wise contribution as the expectation over teacher-forced positions and prompts:

$$\mathcal{L}_\ell^{(3)} \;:=\; \mathbb{E}_{t,x}\Big[d_{\mathrm{FR}}\big(p_t^{(\ell)},\, p_t^{(\ell+1)}\big)\Big],$$

$$\mathcal{L}^{(3)} \;:=\; \sum_{\ell=0}^{L-1} \mathcal{L}_\ell^{(3)}.$$

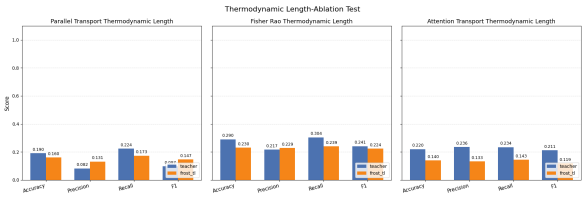

Figure 4: **Thermodynamic-length ablation across geometric formulations.** We compare Parallel Transport, Fisher–Rao, and Attention Transport thermodynamic length formulations within FROST. Attention Transport achieves the best overall security–utility tradeoff by preserving teacher accuracy, recall, and F1 performance while still substantially reducing Capability Transfer Rate (CTR). This balance indicates that attention-space transport perturbs imitation-relevant semantic structure without excessively degrading response quality, motivating its adoption as the default formulation in the remainder of the paper.

Large $\mathcal{L}_\ell^{(3)}$ indicates layers where beliefs change substantially between ℓ and $\ell + 1$; small $\mathcal{L}_\ell^{(3)}$ indicates little epistemic movement.

The ablation results in Figure 4 compare three geometric formulations of thermodynamic length within FROST: Parallel Transport Thermodynamic Length, Fisher–Rao Thermodynamic Length, and Attention Transport Thermodynamic Length. The primary objective of this study is to identify the geometry that achieves the strongest reduction in Capability Transfer Rate (CTR) while preserving sufficient teacher utility and response quality. Although the Parallel Transport formulation produces a larger reduction in predictive performance, the Attention Transport formulation achieves a more balanced tradeoff between security and utility. In particular, Attention Transport maintains comparatively stable accuracy, recall, and F1 scores while still significantly reducing transferable capability, indicating that it suppresses imitation-relevant structure without excessively degrading the teacher distribution. Fisher–Rao geometry preserves utility reasonably well but exhibits weaker resistance to capability transfer relative to Attention Transport. These observations suggest that transporting semantic structure directly through attention-space dynamics provides a more stable and controllable perturbation mechanism for defensive decoding. Consequently, we adopt Attention Transport Thermodynamic Length as the default formulation throughout the remainder of the paper.

C The Provenance Energy Component

The geometric and structural energies in ***FROST*** attempt to reduce the *learnability* of harvested outputs. However, output-channel defenses can also benefit from weak ***provenance shaping*** signals that slightly bias generation toward statistically attributable token patterns. This section introduces the optional ***provenance energy*** component used in some FROST variants. The construction is inspired by watermarking-style decoding methods such as ***KGW*** and ***SynthID*** (Dathathri et al., 2024), but the role here is deliberately narrower. FROST does not rely on provenance signals as the primary defense mechanism, nor does it assume that watermark detectability alone prevents distillation. Instead, provenance shaping acts as a lightweight auxiliary bias that can complement the main geometric objective.

Importantly, the core empirical gains in the paper do not depend on provenance shaping. The strongest improvements already emerge from the geometric and structural components alone. The provenance term is therefore best viewed as an optional regularizer that slightly alters token statistics while remaining compatible with the broader free-energy decoding framework.

C.1 Motivation and Relation to KGW Watermarking

The provenance energy $\mathcal{E}_x^{\text{prov}}$ serves a distinct purpose from the geometric and structural terms: rather than reducing learnability directly, it embeds a statistically detectable attribution signal into the teacher's emitted traces, enabling post-hoc identification of harvested outputs. The mechanism is grounded in the KGW watermarking framework (Kirchenbauer et al., 2023), which established that a pseudorandom green-red partition of the vocabulary, combined with a soft logit bias toward green tokens at sampling time, produces watermarked text that is statistically detectable via a z-test without access to the model's weights or logits. The key insight is that green-list membership for a candidate token v is determined by a pseudorandom function (PRF) seeded by the preceding context, so detection requires only the key and the generated sequence.

FROST adapts this principle to the candidate-selection level rather than the logit-bias level. Instead of modifying the teacher's generative distribution globally, the provenance signal enters the

composite energy as an additive bonus that steers selection toward green candidates at each step, leaving the teacher model itself unmodified. The provenance energy therefore functions simultaneously as a deterrent (harvested traces are detectable as originating from the defended API) and as a component of the composite Gibbs controller, biasing generation toward a statistically identifiable subset of the vocabulary at each step.

C.2 Formal Definition

Let $\mathcal{K}$ denote a secret provenance key string and $\gamma \in (0, 1)$ a list-size parameter controlling what fraction of the vocabulary is designated green at each step. For a candidate token $v \in \mathcal{V}$ and prefix $x_{<t}$, the green membership indicator is:

$$\mathbb{G}(v,\, x_{<t}) \;=\; \mathbf{1}\big[\mathrm{PRF}_{\mathcal{K}}(x_{<t},\, v) < \gamma \cdot 2^{64}\big]\,,$$

where $\mathrm{PRF}_{\mathcal{K}}(x_{<t}, v)$ is the first 64-bit prefix of $\mathrm{SHA256}(\mathcal{K} \,\|\, x_{<t} \,\|\, v)$ interpreted as an unsigned integer. The PRF seed is the *full prefix string* rather than the single preceding token used in standard KGW, making the partition unique to an entire generation trajectory and thus harder to spoof from partial context.

The provenance energy for candidate v at step t is then simply:

$$\mathcal{E}^{\mathrm{prov}}(v,\, x_{<t}) \;=\; \lambda_{\mathrm{prov}} \cdot \mathbb{G}(v,\, x_{<t}),$$

where $\lambda_{\mathrm{prov}} > 0$ is a scalar boost weight. Because this term enters the composite energy with a positive sign and is *subtracted* in the logit correction, green candidates receive a mild selection advantage over otherwise equal red candidates at each step.

C.3 Integration into Autoregressive Decoding

Within the step-level Gibbs rule, the provenance bias contributes the additive correction $+\lambda_{\mathrm{prov}} \cdot \mathbb{G}(v, x_{<t})$ to the adjusted logit of each candidate. Crucially, this correction has opposite sign to the geometric and structural penalties, so it can always be overridden when strong evidence against a candidate accumulates from either of those terms. The provenance energy therefore does not compete with learnability reduction; it operates as an independent, low-cost channel that distributes a detectable statistical fingerprint across the output without altering the rank ordering among candidates that the geometric term has already heavily penalized.

C.4 Statistical Detectability

Because each token is independently assigned to the green list with probability γ under the null hypothesis, the expected green count in a sequence of length T is $\mu_0 = \gamma T$ with variance $\sigma_0^2 = \gamma(1-\gamma)T$. Under the alternative, the provenance bias produces systematic over-representation of green tokens. Following Kirchenbauer et al. (2023), the z-score

$$z \;=\; \frac{|\mathcal{G}| \;-\; \gamma T}{\sqrt{T\,\gamma(1-\gamma)}}$$

is asymptotically standard normal under H_0, enabling efficient attribution using only the key $\mathcal{K}$ and the generated text, without access to logits, attention weights, or any other internal state of the teacher.

C.5 Differences from Standard KGW

Three design choices distinguish the FROST instantiation from vanilla KGW. First, the PRF conditions on the entire prefix string rather than a single preceding token, binding the green partition tightly to the full generation trajectory and preventing an adversary from replicating it from limited context. Second, the bias acts exclusively at the candidate-selection level inside the top-k shortlist, so the teacher's nominal generative distribution is never modified; the watermark is carried entirely by which of the shortlisted candidates is selected at each step. Third, because the provenance term has opposite sign to the penalty terms, it never suppresses a candidate outright, preserving the minimal-intervention character of the defense that is enforced by the KL trust region in objective.

D The Structural Energy Component

Not all reusable supervision arises from semantic correctness alone. In practice, many harvested transcripts become highly valuable for downstream distillation because they expose stable ***structural patterns***: long chain-of-thought decompositions, reusable formatting templates, explicit step partitioning, or highly pedagogical reasoning traces. Such outputs often act as low-noise optimization targets for student models, especially in trace-rich harvesting settings. The goal of the ***structural energy*** component is therefore not to suppress capability itself, but to reduce the extent to which generated traces expose overly reusable organizational scaffolds.

Unlike the geometric component, which estimates downstream learning progress through proxy-student curvature, the structural component operates directly on observable output statistics. It penalizes features associated with highly reusable transcript organization while remaining lightweight and prefix-local during decoding. This makes it computationally inexpensive and easy to combine with the broader free-energy framework. Empirically, the structural term alone improves resistance relative to nominal decoding, although the strongest gains emerge when it is combined with the geometric energy in full ***FROST***.

D.1 Motivation

The structural energy $\mathcal{E}_x^{\text{struct}}$ captures a complementary dimension of distillation risk that the geometric term does not directly measure. Whereas the attention-transport thermodynamic length targets candidates that force large routing displacement in downstream students, the structural term targets candidates that would steer the generated sequence toward surface patterns that are intrinsically high-density supervision: explicit chain-of-thought scaffolding, stepwise connectives, and structured response templates. These patterns are particularly dangerous because they remain exploitable even when the geometric energy is moderate, providing students with reusable pedagogical form rather than requiring them to infer latent structure from opaque outputs. The structural term therefore acts as a form-level complement to the geometry-level signal, jointly ensuring that FROST suppresses both information-geometrically rich and structurally dense candidates.

D.2 Formal Definition

The five penalty weights in equation are not set by hand from first principles but are determined empirically through a systematic ablation study conducted on the teacher and student pair used throughout this work: DeepSeek-R1-Distill-Qwen-7B as the teacher and Mistral-7B-Instruct-v0.3, Llama-3.1-8B-Instruct and Qwen-2.5-3B as the attackers in all the 3 harvesting stages. For each candidate penalty component, we vary its weight over a grid while holding all other components fixed, and select the value that maximally reduces capability transfer on the GSM8K training split without degrading the teacher's response quality as measured by perplexity under the nominal distribution P_0. The resulting weights reflect the relative distillation risk contributed by each surface indicator on this model pair.

For candidate token v_i with surface text $c_i = \textsc{Decode}(v_i)$ and accumulated prefix $x_{<t}$, the structural energy decomposes into five additive components, each targeting a distinct surface indicator of high-density supervision structure:

$$\mathcal{E}^{\text{struct}}(v_i,\, x_{<t}) \;=\; \min\big(\rho_{\text{think}} + \rho_{\text{conn}} + \rho_{\text{punct}} + \rho_{\text{prefix}} + \rho_{\text{len}},\; \rho_{\max}\big).$$

where the cap $\rho_{\max} = 2.0$ ensures that structural evidence cannot completely override the logit and geometric signals, preserving the minimal-intervention property of the defense enforced by the KL trust region in objective.

Think-tag penalty ρ_{think}**.** Assigns a penalty of 3.0 if the concatenation of prefix and candidate contains a `<think>` or `</think>` delimiter. These tags mark the internal scratchpad of chain-of-thought models such as DeepSeek-R1; their presence indicates that the response is entering an explicit reasoning trace that constitutes maximally dense, structured supervision and is not transferable to students lacking the same internal scaffolding.

Connective penalty ρ_{conn}**.** Assigns 1.2 if the candidate text contains any member of the structural connective set $\mathcal{C} = \{$`first, then, because, therefore, thus, hence, step, next`$\}$. These lexical items signal explicit stepwise decomposition of the kind that provides disproportionately high supervision density, as noted in the empirical motivation for Antidistillation Sampling.

Punctuation penalty ρ_{punct}**.** Assigns 0.2 if the candidate contains any of $\{$`\n, :, ;`$\}$. These characters typically demarcate list or sub-step boundaries within a structured reasoning trace, and their occurrence is a lightweight proxy for structured response form.

Prefix-length penalty ρ_{prefix}**.** Assigns 0.8 if the word count of the prefix text preceding the first `<think>` delimiter exceeds a threshold τ_{think}. A long pre-think preamble reliably indicates that the model is generating an extended reasoning introduction rather than proceeding directly to a terminal answer, making the emitted trace structurally denser as supervision.

Multi-token penalty ρ_{len}**.** Assigns 0.3 if the student-tokenized length of the candidate exceeds

two tokens. Longer surface forms tend to correspond to structured sub-phrases rather than atomic answer tokens, increasing the risk that the candidate contributes a reusable phrasal template rather than a content token.

D.3 Role in the Composite Energy

The structural energy enters the composite objective with weight λ_{struct} and, like the geometric term, is min-max normalized over the top-k shortlist before combination, so that all three energy components operate on a common $[0, 1]$ scale within the logit correction. This normalization ensures that the relative contribution of each component is determined entirely by the weights $\lambda_{\text{geom}}, \lambda_{\text{struct}}, \lambda_{\text{prov}}$ rather than by differences in their raw magnitudes, which can vary substantially across prompts and model families.

E Multi-Student Robustness

Why single-student defenses are insufficient. A teacher-side antidistillation defense is only as strong as its robustness to the ***attacker model family*** it implicitly assumes. If the defense is optimized against a single proxy student, then its protection may be brittle: an attacker can simply change architecture, scale, fine-tuning regime, or transcript formatting strategy, and recover capability through a different student that the defense did not anticipate. This concern is not hypothetical. Classical functionality-extraction work already shows that black-box stealing can transfer across architectures and even across mismatched query distributions (Orekondy et al., 2019). Recent surveys of LLM extraction attacks likewise emphasize that the attack surface is inherently heterogeneous, spanning API-based distillation, prompt-targeted extraction, training-data extraction, and varied downstream student pipelines (Zhao et al., 2025). Moreover, *Antidistillation Sampling* itself is evaluated with distinct teacher, proxy, and student models, underscoring that practical antidistillation must generalize beyond a single surrogate. For this reason, FROST is designed not as a defense against *one* student, but as a defense against a ***family of plausible students***.

Student family as an uncertainty set. Let $\mathcal{S}$ denote the set of plausible attacker students. An element $S \in \mathcal{S}$ may vary along several axes:

- ***architecture***: e.g., LLaMA-family, Qwen-family, Mistral-family, or other dense transformer variants;
- ***scale***: small, medium, or relatively strong student models;
- ***fine-tuning regime***: full supervised fine-tuning, LoRA-style PEFT, adapter-based training, or other lightweight adaptation;
- ***transcript format***: answer-only supervision, rationale-rich supervision, or structured traces such as decompositions or code;
- ***optimization state***: different initial checkpoints, optimization hyperparameters, or local parameter geometries.

Rather than assuming that the defender knows the exact student chosen by the adversary, FROST models this uncertainty explicitly through either:

$$\text{(i) a distribution } \mathcal{Q}_S \text{ over students,}$$

$$\text{or} \quad \text{(ii) a worst-case uncertainty set } \mathcal{S}.$$

This mirrors the logic of ***distributionally robust optimization***: if the deployment environment contains uncertainty about future adversaries, one should optimize against a neighborhood or family of plausible threats rather than a single nominal one (Sinha et al., 2018).

Per-student thermodynamic anti-copying energy. Recall from Section B that for a fixed candidate response y and a student model S with L_S transformer layers, the attention-transport thermodynamic length induced by y on S is:

$$\mathcal{L}^{(S)}(y) \;:=\; \sum_{\ell=1}^{L_S} \sqrt{g_\ell^{(S)}(y)},$$

where the layer-wise transport-geometric quadratic form is:

$$g_\ell^{(S)}(y) \;=\; \sum_{i=1}^{n} \pi_\ell^{(S)}(i) \sum_{j=1}^{n} P_\ell^{(S)}(i,j) \cdot \Big(\Delta \log P_\ell^{(S)}(i,j)\Big)^2.$$

with $P_\ell^{(S)}$ the entropy-weighted transport kernel of student S at layer ℓ and $\pi_\ell^{(S)}$ its stationary distribution. Large $\mathcal{L}^{(S)}(y)$ indicates that y induces large routing displacement across depth in student S, making it intrinsically learnable for that student. The per-student thermodynamic length therefore serves as the natural learnability signal in the multi-student setting.

Expected thermodynamic anti-copying energy. The most natural multi-student extension is to average the thermodynamic length over the student family:

$$\begin{aligned}\mathcal{E}_x^{\mathrm{avg}}(y) &= \mathbb{E}_{S\sim\mathcal{Q}_S}\big[\mathcal{L}^{(S)}(y)\big] \\ &= \mathbb{E}_{S\sim\mathcal{Q}_S}\left[\sum_{\ell=1}^{L_S}\sqrt{g_\ell^{(S)}(y)}\right].\end{aligned}$$

This defines the ***expected thermodynamic anti-copying energy*** of output y under prompt x: a response is penalized when it induces large attention-routing displacement broadly across many plausible student instantiations, not merely when it is informative to one specific proxy.

Worst-case thermodynamic anti-copying energy. A stronger, minimax-style alternative is to define:

$$\mathcal{E}_x^{\mathrm{wc}}(y) = \sup_{S\in\mathcal{S}} \mathcal{L}^{(S)}(y) = \sup_{S\in\mathcal{S}}\left[\sum_{\ell=1}^{L_S}\sqrt{g_\ell^{(S)}(y)}\right]$$

Under this formulation, the defender treats an output as dangerous if *any* plausible student can learn rapidly from it, i.e., if any student in $\mathcal{S}$ undergoes large thermodynamic routing displacement when processing y. The resulting decoding rule prioritizes responses whose thermodynamic length is small even under the most compatible attacker architecture in the uncertainty set.

Average-case vs. worst-case robustness. The two formulations reflect different operational philosophies:

- ***Average-case robustness*** is appropriate when the defender has a meaningful prior over likely attacker pipelines and wishes to optimize expected resistance against thermodynamic learnability.
- ***Worst-case robustness*** is appropriate when the defender wants guarantees against the student in $\mathcal{S}$ that maximally exploits the thermodynamic structure of emitted traces.

In practice, these choices induce different trade-offs. The average-case objective is smoother, lower-variance, and easier to estimate using a finite proxy ensemble. The worst-case objective is more conservative but can be numerically unstable and overly pessimistic if $\mathcal{S}$ is very broad. In this paper, we adopt the expectation form as the primary instantiation of FROST and treat the worst-case form as a conceptual upper envelope.

Robust FROST objective. Substituting the expected multi-student thermodynamic energy into the free-energy formulation of Section A.3 yields the robust defended distribution:

$$\begin{aligned}Q^*(\cdot\mid x) = \arg\min_{Q(\cdot|x)}\Bigg\{ & \\ & \mathbb{E}_{y\sim Q}\Big[\mathbb{E}_{S\sim\mathcal{Q}_S}\big[\mathcal{L}^{(S)}(y)\big]\Big] \\ & + \frac{1}{\beta}D_{\mathrm{KL}}\big(Q(\cdot\mid x)\,\|\,P_0(\cdot\mid x)\big) \\ & \Bigg\}.\end{aligned}$$

This is the robust decoding rule implemented by FROST. It says that a response should be down-weighted whenever it induces ***high thermodynamic routing displacement across the attacker family as a whole***.

A minimax interpretation. The worst-case version admits a more explicitly adversarial form:

This objective can be read as a two-player game:

- the ***defender*** chooses a response distribution Q;
- the ***attacker*** chooses the student $S \in \mathcal{S}$ whose attention-transport geometry is most compatible with the emitted trace, maximizing thermodynamic length.

The defender then minimizes the worst achievable thermodynamic learnability subject to remaining close to the nominal teacher. This places FROST in the broader conceptual family of robust and adversarial optimization, but at the ***output-channel level*** rather than the input or parameter level (Sinha et al., 2018).

Finite proxy ensembles. In practice, neither the expectation nor the supremum can be computed over an infinite student family. FROST therefore approximates $\mathcal{Q}_S$ using a finite ensemble of proxy students $\{S^{(1)}, S^{(2)}, \ldots, S^{(M)}\}$, chosen to cover meaningful variation in architecture, scale, and fine-tuning style. The expected thermodynamic energy is then approximated by:

$$\begin{aligned}\widehat{\mathcal{E}}_x^{\mathrm{avg}}(y) &= \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}^{(S^{(m)})}(y) \\ &= \frac{1}{M}\sum_{m=1}^{M}\sum_{\ell=1}^{L_{S^{(m)}}}\sqrt{g_\ell^{(S^{(m)})}(y)}.\end{aligned}$$

This approximation has two practical virtues. First, it reduces variance relative to single-proxy scoring. Second, it directly aligns the defense with the evaluation protocol: if distillation is later tested on held-out attacker students, the proxy ensemble can be viewed as the training-time analogue of a robust validation set.

Why robustness matters empirically. Without multi-student robustness, the defense risks learning ***proxy-specific anti-signals*** rather than true anti-distillation structure. For example, an output that produces small thermodynamic length in one proxy student may still generate large routing displacement in another student whose attention geometry is better aligned with the teacher's. Robust averaging mitigates this pathology by favoring outputs whose low thermodynamic length is ***stable across student architectures*** rather than accidental under one routing geometry. This is especially important in black-box extraction, where the defender almost never knows which student model an adversary will actually deploy (Orekondy et al., 2019; Zhao et al., 2025).

Robustness as smooth anti-transfer. There is also a geometric advantage to the expectation form. Averaging thermodynamic lengths across a student family induces a smoother energy landscape over candidate outputs:

$$\mathcal{E}_x^{\mathrm{avg}}(y) \;=\; \mathbb{E}_{S\sim\mathcal{Q}_S}\Big[\mathcal{L}^{(S)}(y)\Big].$$

Smoother energies are easier to optimize in practice and less prone to brittle reweighting around artifacts of a single proxy model's attention geometry. Multi-student robustness therefore makes FROST less likely to overfit its defense to one particular routing structure, and more likely to suppress ***transferable thermodynamic learnability*** across the attacker family as a whole.

Relation to the broader FROST pipeline. This section upgrades FROST from a single-proxy defense to a ***robust teacher-side antidistillation framework***. The defense no longer asks: *"Does this output induce large thermodynamic routing displacement in one surrogate?"* Instead, it asks: ***"Does this output induce large thermodynamic routing displacement broadly across the attacker family we care about?"*** That distinction is crucial for a deployment-time defense, because real attackers are free to choose whatever student architecture best exploits the exposed output channel. The next section complements this robust thermodynamic view with an information-theoretic one, showing how suppressing broad thermodynamic learnability also suppresses the effective leakage of latent capability-bearing structure.

F Information-Theoretic Interpretation

Why an information-theoretic view is needed. The thermodynamic formulation in Section B explains ***how much attention-routing displacement*** a harvested trace induces across student depth, and how this displacement correlates with downstream learnability. This is already a powerful notion of anti-copying. However, extraction attacks are not only about routing progress; they are ultimately about ***recovering reusable capability-bearing structure*** from repeated black-box interaction. For this reason, it is useful to complement the thermodynamic view with an information-theoretic one. The central question becomes:

How much information about the teacher's transferable capability is leaked by each emitted response?

This perspective is closely related to the broader ***information bottleneck*** principle, where one seeks representations that preserve task-relevant information while discarding unnecessary or dangerous detail (Tishby et al., 2000). It is also consistent with modern information-theoretic analyses of adaptive procedures, where mutual information quantifies how much can be learned from an interaction transcript (Russo and Zou, 2016; Steinke and Zakynthinou, 2020).

A latent capability variable. To make this concrete, introduce a latent random variable

$$W \in \mathcal{W},$$

which represents the ***capability-bearing structure*** the attacker wishes to recover from the teacher. The exact semantics of W may vary by setting. For example, W may encode:

- a reusable reasoning template,
- a latent task decomposition policy,
- a coding or tool-use scaffold,
- a family of evaluation heuristics, or

- a hidden structured behavior that distinguishes the teacher from weaker models.

The attacker never observes W directly. Instead, it interacts with the output channel and observes responses $Y_t \sim Q(\cdot \mid X_t)$, where X_t is the current prompt. The information-theoretic security question is therefore not whether the attacker can read off W from a single response, but whether the full transcript reveals enough about W to support effective downstream imitation.

Per-query leakage. For a fixed prompt x, the most basic leakage quantity is

$$I(W;\, Y \mid X = x),$$

the mutual information between the latent capability variable and the emitted response, conditioned on the prompt. This quantity has a direct interpretation:

- if $I(W; Y \mid X = x)$ is ***large***, then the output reveals substantial information about the transferable structure the attacker cares about;
- if $I(W; Y \mid X = x)$ is ***small***, then the response is comparatively poor evidence about that structure.

From this perspective, a strong teacher-side defense should make the output channel ***informationally stingy*** with respect to W, while remaining behaviorally useful to legitimate users.

Transcript-level leakage under adaptive querying. The attacker, however, receives not one response but an adaptive transcript. Let

$$Z_B := (X_1, Y_1, \ldots, X_B, Y_B)$$

denote the full interaction transcript after B black-box queries. Since prompts may be chosen adaptively based on earlier responses, the relevant object is the mutual information $I(W; Z_B)$. By the chain rule of mutual information,

$$I(W; Z_B) \;=\; \sum_{t=1}^{B} I\big(W;\, Y_t \mid X_t,\, Z_{t-1}\big),$$

where $Z_{t-1} = (X_1, Y_1, \ldots, X_{t-1}, Y_{t-1})$. This identity is exact and does not assume non-adaptivity: it simply decomposes total leakage into the sum of ***incremental per-query leakages*** (Russo and Zou, 2016). Thus, even in an adaptive black-box attack, the transcript leaks information only query by query.

A leakage-budget interpretation. Suppose now that the defended teacher satisfies the prompt-local leakage bound

$$I\big(W;\, Y_t \mid X_t,\, Z_{t-1}\big) \;\le\; \gamma \qquad \text{for all } t.$$

Then immediately, $I(W; Z_B) \le B\gamma$. This inequality admits a simple interpretation: ***each query spends at most γ bits of capability leakage budget***. If FROST can reduce γ, then the attacker must spend proportionally more queries to acquire the same amount of information about W. The role of the defense is therefore not necessarily to drive leakage to zero, but to make each emitted trace ***low-bandwidth with respect to transferable capability***.

From leakage to query complexity via Fano. The operational consequence of low leakage is captured by ***Fano's inequality***, which relates mutual information to probability of identification error (Steinke and Zakynthinou, 2020). Suppose the attacker's objective is to identify W among $|\mathcal{W}| = M$ possible hypotheses using the transcript Z_B, producing an estimate $\widehat{W}$. Then Fano's inequality implies

$$\Pr(\widehat{W} \neq W) \;\ge\; 1 - \frac{I(W; Z_B) + \log 2}{\log M}.$$

Combining this with the transcript leakage bound above yields

$$\Pr(\widehat{W} \neq W) \;\ge\; 1 - \frac{B\gamma + \log 2}{\log M}.$$

Equivalently, to drive the error probability below a fixed constant, the attacker must satisfy $B = \Omega\Big(\frac{\log M}{\gamma}\Big)$. This gives the desired query-complexity interpretation: ***reducing per-query leakage increases the number of black-box interactions required for successful capability recovery.***

Proposition 3 (Query-complexity lower bound under bounded leakage). Let W be a latent capability-bearing variable taking values in a finite set $\mathcal{W}$, and let Z_B be an adaptive transcript of B black-box queries. If

$$I\big(W;\, Y_t \mid X_t,\, Z_{t-1}\big) \;\le\; \gamma \qquad \text{for all } t,$$

then any estimator $\widehat{W} = \widehat{W}(Z_B)$ satisfies

$$\Pr(\widehat{W} \neq W) \;\ge\; 1 - \frac{B\gamma + \log 2}{\log |\mathcal{W}|}.$$

Hence reliable identification of W requires $B = \Omega\Big(\frac{\log |\mathcal{W}|}{\gamma}\Big)$ queries.

Interpretation for teacher-side antidistillation. Proposition 3 does *not* say that FROST makes extraction impossible. Rather, it clarifies the mechanism by which an output-channel defense can slow it down. If the defense suppresses the amount of transferable capability information revealed by each response, then extraction becomes ***query-expensive***. This is exactly the right operational notion for black-box security: the defender increases the attacker's required interaction budget without destroying the teacher's usefulness.

Connection to the information bottleneck principle. The information-bottleneck lens sharpens this interpretation. In the classical information bottleneck framework, one seeks a compressed variable that preserves relevant information about a target while discarding nuisance information (Tishby et al., 2000). FROST can be viewed as an ***adversarial bottleneck*** over outputs: the defended response Y should preserve utility-relevant information for the legitimate task while discarding capability-relevant information for the attacker. In that sense, the ideal defended output channel would solve a problem of the form

$$\min_{Q} \quad \mathbb{E}_{x\sim\mathcal{D}}\big[I(W;Y \mid X=x)\big]$$

subject to $\mathcal{U}(Q;\mathcal{D}) \geq \mathcal{U}_{\min}$ and $\mathbb{E}_x D_{\mathrm{KL}}(Q(\cdot \mid x) \,\|\, P_0(\cdot \mid x)) \leq \delta$. We do *not* claim that FROST directly optimizes this mutual-information objective in closed form. Rather, we claim that it provides a ***thermodynamic surrogate*** for this objective: by downweighting outputs that induce large attention-routing displacement across attacker students, FROST is designed to reduce the effective information rate of the output channel with respect to transferable capability.

Why the thermodynamic surrogate is reasonable. The information-theoretic view and the thermodynamic view are conceptually different but closely aligned. Mutual information quantifies how much uncertainty about W is removed by observing Y. The attention-transport thermodynamic length from Section B quantifies the cumulative routing displacement induced across decoder depth when a student processes y, measured as the sum of layer-wise transport-geometric line elements $\sum_\ell \sqrt{g_\ell^{(S)}(y)}$. These are not identical quantities, and we do not conflate them. However, both aim to capture the same underlying property: ***how much transferable structure is exposed by a generated response***. When a trace induces large, robust, routing displacement across many student architectures and scales, it is also a natural candidate for carrying high usable information about the teacher's latent capability. Thus, the thermodynamic energy used by FROST should be understood as a ***tractable anti-copying proxy*** for the more abstract mutual-information objective above.

Adaptive querying does not invalidate the interpretation. One may worry that adaptive querying makes per-query information accounting meaningless, since later prompts are chosen using earlier outputs. The chain-rule decomposition resolves exactly this concern. Adaptivity changes *which* conditional information terms appear, $I(W;Y_t \mid X_t, Z_{t-1})$, but it does not invalidate the additive structure of total leakage. This is the same reason mutual-information tools remain meaningful in adaptive data analysis and related sequential settings (Russo and Zou, 2016; Steinke and Zakynthinou, 2020). Accordingly, FROST's information-theoretic interpretation remains well-defined even against adaptive black-box attackers.

What this section adds to FROST. The thermodynamic formulation of Sections B,G explains ***how*** FROST penalizes dangerous traces. The present section explains ***why this matters operationally***: reducing the effective leakage of capability-bearing structure makes extraction progressively more query-expensive. This interpretation does not replace the thermodynamic formulation; it elevates it. FROST can therefore be understood simultaneously as:

- a ***thermodynamic decoder***, via Gibbs reweighting of the teacher's output law;
- an ***attention-transport thermodynamic defense***, via routing-displacement learnability suppression across decoder depth; and
- an ***information-theoretic bottleneck***, via reduced per-query leakage of transferable capability.

The next section returns to the algorithmic side and shows how these principles are instantiated in the composite FROST energy used in practice.

G Composite FROST Energy

Why a composite energy is necessary. Sections 3.3–3.6 establish the conceptual foundation

of FROST. Section 3.3 shows that defended decoding should be written as a Gibbs-reweighted variational problem. Section B defines the central anti-copying signal through ***attention-transport thermodynamic learnability***. Section 3.6 complements this with an information-theoretic interpretation in terms of ***per-query leakage of transferable capability***. However, a deployable teacher-side defense must ultimately operate in a richer response space than any single scalar criterion can capture. In practice, outputs are dangerous not only because they induce large thermodynamic routing displacement in student models, but also because they may expose ***highly reusable structural scaffolds*** such as explicit decompositions, long reasoning traces, code skeletons, and stepwise templates. Conversely, some applications also benefit from preserving ***provenance-supporting structure*** such as watermark compatibility, even if provenance alone is insufficient as a primary defense (Kirchenbauer et al., 2023; Dathathri et al., 2024). For this reason, FROST is instantiated through a ***composite energy*** that separates the *core thermodynamic anti-copying geometry* from *auxiliary structural and provenance controls*.

General form of the energy. For a prompt x and candidate response y, we define the defended energy as

$$\begin{aligned}\mathcal{E}_x(y) \;&= \lambda_{\text{geom}}\,\mathcal{E}_x^{\text{geom}}(y)\\ &\quad+\;\lambda_{\text{struct}}\,\mathcal{E}_x^{\text{struct}}(y)\\ &\quad+\;\lambda_{\text{prov}}\,\mathcal{E}_x^{\text{prov}}(y).\end{aligned}$$

where:

- $\mathcal{E}_x^{\text{geom}}(y)$ is the ***thermodynamic anti-copying term***;
- $\mathcal{E}_x^{\text{struct}}(y)$ is the ***structural trace-risk term***;
- $\mathcal{E}_x^{\text{prov}}(y)$ is the ***provenance-aware shaping term***;
- $\lambda_{\text{geom}}, \lambda_{\text{struct}}, \lambda_{\text{prov}} \geq 0$ are tunable weights.

This decomposition is deliberate. It lets FROST express a clear priority ordering:

intrinsic learnability suppression

$>$ ***structural hardening***

$>$ ***provenance support***.

Accordingly, λ_{geom} is the principal coefficient, while the remaining terms play secondary, regularizing roles.

Why thermodynamic length alone is not always enough. Although the thermodynamic term is the most principled quantity in the framework, it is not always the whole story. In black-box extraction, attackers benefit disproportionately from outputs that are ***pedagogically explicit***: responses that expose reusable decomposition patterns, chain-like templates, code scaffolds, or highly regular reasoning layouts. Antidistillation Sampling is motivated by a closely related observation, namely that structured reasoning traces can be especially valuable for downstream distillation. This means that some traces may be dangerous not only because they induce large thermodynamic routing displacement, but also because they expose ***portable structural templates*** that survive across students and fine-tuning regimes. To capture this phenomenon, FROST adds a separate structural term.

Structural trace-risk energy. We define $\mathcal{E}_x^{\text{struct}}(y)$ as a penalty on ***structurally reusable supervision***. In its most general form, this term is left abstract and implemented through measurable features of the generated response. Examples include:

- ***trace length***: long-form derivations, extended chain-like explanations, or verbose decompositions;
- ***template explicitness***: numbered procedures, algorithmic stages, recipe-style instructions, or systematic subgoal lists;
- ***code scaffold exposure***: reusable function skeletons, boilerplate, tool-call templates, or partially executable structures;
- ***structural regularity***: highly canonical response formats that compress task structure into a transferable template.

A generic parameterization is

$$\mathcal{E}_x^{\text{struct}}(y) \;=\; \sum_{j=1}^{J} \alpha_j\,\phi_j(x,y),$$

where each $\phi_j(x,y)$ is a structural feature and $\alpha_j \geq 0$ is a learned or tuned coefficient. This formulation keeps FROST flexible: the structural term can be adapted to the domain without changing the thermodynamic core.

Interpretation of the structural term. The structural term does ***not*** replace the thermodynamic term; rather, it complements it by targeting a different failure mode. The thermodynamic term asks:

How much routing displacement does this trace induce across student depth?

The structural term asks:

How easily can this trace be reused as a portable supervision template?

These are related but distinct questions. A response can produce moderate thermodynamic length yet still reveal a high-value structured scaffold. Conversely, a response may induce strong routing displacement without exposing much reusable formatting. The composite energy is designed precisely to cover both cases.

Optional provenance-aware term. In some deployments, the defender may also want outputs to remain compatible with provenance mechanisms such as watermarking. Prior work on watermarking for large language models and SynthID-style text watermarking shows that sampling-time modifications can provide attribution signal while preserving output quality under their evaluation settings (Kirchenbauer et al., 2023; Dathathri et al., 2024). However, watermarking alone is not a sufficient antidistillation defense, because detectable outputs may still carry high-value supervision for downstream students. For this reason, provenance enters FROST only as an ***optional third term***:

$$\mathcal{E}_x^{\mathrm{prov}}(y) \;=\; -\psi_{\mathrm{prov}}(x, y),$$

where ψ_{prov} is a provenance score. The negative sign reflects the fact that provenance support is usually *desirable*; thus, in the total energy, outputs that preserve provenance can be relatively favored.

Normalized defended decoding rule. Substituting the composite energy into the free-energy objective yields the defended distribution

$$\begin{aligned} Q^*(y \mid x) = \frac{1}{Z(x)}\, P_0(y \mid x) \\ \cdot \exp\Big(-\beta\big[\lambda_{\mathrm{geom}}\, \mathcal{E}_x^{\mathrm{geom}}(y) \\ + \lambda_{\mathrm{struct}}\, \mathcal{E}_x^{\mathrm{struct}}(y) \\ + \lambda_{\mathrm{prov}}\, \mathcal{E}_x^{\mathrm{prov}}(y)\big]\Big). \end{aligned}$$

where $Z(x)$ is the corresponding partition function. This expression is the full defended decoding law implemented by FROST. It reveals that the method is not a monolithic single-score defense, but a ***composable Gibbs controller over multiple dimensions of extraction risk***.

Reading the coefficients. The coefficients have a clear semantics:

- λ_{geom}: ***thermodynamic anti-learnability strength***; this is the primary knob governing capability-transfer suppression.
- $\lambda_{\mathrm{struct}}$: ***trace hardening strength***; this controls how strongly the defense avoids portable response scaffolds.
- λ_{prov}: ***provenance preference strength***; this supports attribution or watermark compatibility when desired.

The inverse temperature β continues to control the overall defense strength, while the λ-coefficients allocate that strength across mechanisms.

Ablation-friendly design. The composite decomposition is also crucial experimentally. Because each term has a distinct operational meaning, FROST supports principled ablations:

$$\mathcal{E}_x(y) = \mathcal{E}_x^{\mathrm{geom}}(y)$$

tests the thermodynamic-only defense;

$$\mathcal{E}_x(y) = \mathcal{E}_x^{\mathrm{geom}}(y) + \lambda_{\mathrm{struct}}\, \mathcal{E}_x^{\mathrm{struct}}(y)$$

tests whether structural hardening provides additional benefit; and

$$\mathcal{E}_x(y) = \mathcal{E}_x^{\mathrm{geom}}(y) + \lambda_{\mathrm{prov}}\, \mathcal{E}_x^{\mathrm{prov}}(y)$$

tests whether provenance compatibility can be layered on without damaging anti-copying performance. This allows us to separate the contribution of the ***core thermodynamic principle*** from secondary design choices.

Why this section is central to FROST. Conceptually, this section is where FROST becomes a ***framework*** rather than a single theorem. The earlier sections establish:

- a ***thermodynamic variational form*** for defended decoding,
- an ***attention-transport thermodynamic*** notion of student learnability,

- a ***robust*** multi-student extension, and
- an ***information-theoretic*** interpretation of reduced leakage.

The present section combines these insights into the actual energy used at inference time. In short, **FROST** ***protects by making emitted traces simultaneously harder to learn from, harder to reuse as templates, and optionally easier to attribute.***

H Autoregressive Approximation and Decoding Algorithm

Why an autoregressive approximation is necessary. The defended distribution derived in Section 3.7 is defined over ***entire responses***:

$$Q^*(y \mid x) \;\propto\; P_0(y \mid x)\, \exp\!\big(-\beta\, \mathcal{E}_x(y)\big),$$

where $\mathcal{E}_x(y)$ is the composite FROST energy. This form is mathematically clean, but it is not directly usable for large language models. The response space $\mathcal{Y}$ is combinatorial, and exact normalization over full sequences is intractable. Thus, as in other inference-time control methods for language models, one must replace the sequence-level objective by a ***tokenwise approximation*** that can be evaluated during decoding. Controlled Decoding follows this general strategy for reward-style control, and Antidistillation Sampling similarly modifies next-token probabilities using a student-aware signal during generation. FROST adopts the same autoregressive principle, but with a different target: ***reduce composite thermodynamic anti-copying energy online during generation*** (Mudgal et al., 2024).

Nominal autoregressive teacher. Let the nominal teacher factorize as

$$P_0(y \mid x) \;=\; \prod_{t=1}^{T} P_0(y_t \mid x,\, y_{<t}),$$

where $y = (y_1, \ldots, y_T)$ and $y_{<t} = (y_1, \ldots, y_{t-1})$. The defended decoder should ideally sample from

$$Q^*(y \mid x) \;\propto\; \left(\prod_{t=1}^{T} P_0(y_t \mid x, y_{<t})\right) \cdot\; \exp\!\big(-\beta\, \mathcal{E}_x(y)\big).$$

The difficulty is that $\mathcal{E}_x(y)$ depends on the *entire future completion*, which is not available at step t. FROST therefore introduces a prefix-conditioned surrogate energy that estimates the future anti-copying cost of choosing a token now.

Prefix-conditioned energy-to-go. For prompt x, current prefix $y_{<t}$, and candidate token $v \in \mathcal{V}$, define the ***energy-to-go***

$$\mathcal{E}_t(v;\, x,\, y_{<t}) \;:=\; \mathbb{E}_{Y_{>t} \sim \Pi(\cdot \mid x, y_{<t}, v)} \Big[\mathcal{E}_x\big(y_{<t} \oplus v \oplus Y_{>t}\big)\Big].$$

where Π is a rollout distribution over suffixes. If this quantity were tractable, the exact defended next-token rule would be

$$q_t^*(v \mid x, y_{<t}) \;\propto\; P_0(v \mid x, y_{<t}) \cdot\; \exp\!\big(-\beta\, \mathcal{E}_t(v; x, y_{<t})\big).$$

However, exact evaluation of $\mathcal{E}_t$ is too expensive in general.

Local surrogate energy. FROST therefore replaces the exact energy-to-go with a computable local surrogate

$$\begin{aligned}\widetilde{\mathcal{E}}_t(v;\, x,\, y_{<t}) \;&=\; \lambda_{\text{geom}}\, \widetilde{\mathcal{E}}_t^{\text{geom}}(v; x, y_{<t}) \\ &+\; \lambda_{\text{struct}}\, \widetilde{\mathcal{E}}_t^{\text{struct}}(v; x, y_{<t}) \\ &+\; \lambda_{\text{prov}}\, \widetilde{\mathcal{E}}_t^{\text{prov}}(v; x, y_{<t}).\end{aligned}$$

Each term is a token-local approximation of the corresponding sequence-level component.

Illustrative example. Consider the prompt: *"Explain how to solve a quadratic equation."* Suppose the current prefix is $y_{<t}$ = "`To solve a quadratic equation,`". Under the nominal teacher, the next-token shortlist may contain:

$$\begin{aligned} v_1 &= \text{“ `first`”}, \\ v_2 &= \text{“ `use`”}, \\ v_3 &= \text{“ `Step`”}. \end{aligned}$$

These tokens induce very different future continuations:

- v_1 = "`first`" likely leads to a stepwise decomposition;
- v_2 = "`use`" likely leads to a compact direct explanation;
- v_3 = "`Step`" strongly commits to an explicit pedagogical template.

Although all three continuations may be helpful, v_3 is often the most ***distillable*** because it begins

a reusable structured scaffold. FROST therefore assigns:

$$\widetilde{\mathcal{E}}_t(v_3;\, x, y_{<t}) > \widetilde{\mathcal{E}}_t(v_1;\, x, y_{<t}) > \widetilde{\mathcal{E}}_t(v_2;\, x, y_{<t}).$$

downweighting the more pedagogically reusable continuation even if it is nominally probable under the teacher.

Token-local thermodynamic term. The most important component is the thermodynamic term. For candidate next token v, define the partially committed sequence $\tilde{y}_t^{(v)} := y_{<t} \oplus v$. Then the token-local thermodynamic score is:

$$\begin{aligned}\widetilde{\mathcal{E}}_t^{\text{geom}}(v;\, x,\, y_{<t}) &= \mathbb{E}_{S\sim\mathcal{Q}_S}\left[\mathcal{L}^{(S)}(\tilde{y}_t^{(v)})\right] \\ &= \mathbb{E}_{S\sim\mathcal{Q}_S}\left[\sum_{\ell=1}^{L_S}\sqrt{g_\ell^{(S)}(\tilde{y}_t^{(v)})}\right]\end{aligned}$$

Intuitively, this asks:

if the teacher commits to token v now, how much thermodynamic routing displacement does that commitment create across likely student architectures?

Token-local structural term. The structural component is approximated incrementally:

$$\widetilde{\mathcal{E}}_t^{\text{struct}}(v;\, x,\, y_{<t}) = \sum_{j=1}^{J}\alpha_j\,\widetilde{\phi}_{j,t}(v;\, x,\, y_{<t}),$$

where $\widetilde{\phi}_{j,t}$ are token or prefix-level indicators of emerging structural danger. Examples include:

- whether v begins a numbered list or explicit multi-step template;
- whether v extends a long chain-like rationale;
- whether v increases syntactic commitment to code scaffolding;
- whether v pushes the response toward highly canonical reusable structure.

Token-local provenance term. If provenance-aware shaping is enabled, FROST also includes $\widetilde{\mathcal{E}}_t^{\text{prov}}(v;\, x,\, y_{<t})$, which rewards or penalizes tokens according to their compatibility with the desired provenance behavior. As before, provenance remains an ***auxiliary term***, not the primary defense mechanism (Kirchenbauer et al., 2023; Dathathri et al., 2024).

Autoregressive defended decoding rule. Given the local surrogate energy, FROST defines the defended next-token distribution at step t as

$$q_t(v \mid x, y_{<t}) = \frac{\exp\big(\ell_t^{\text{FROST}}(v)\big)}{\sum_{u\in\mathcal{V}_t^{(k)}}\exp\big(\ell_t^{\text{FROST}}(u)\big)},$$

where the logit correction is:

$$\ell_t^{\text{FROST}}(v) = \ell_t(v) - \beta\,\widetilde{\mathcal{E}}_t(v;\, x,\, y_{<t}).$$

Illustrative token reweighting. Suppose the nominal teacher assigns $P_0(v_1) = 0.32$, $P_0(v_2) = 0.28$, $P_0(v_3) = 0.25$ for $v_1 =$ "`first`", $v_2 =$ "`use`", $v_3 =$ "`Step`", and FROST estimates:

$$\widetilde{\mathcal{E}}_t(v_1) = 0.8, \quad \widetilde{\mathcal{E}}_t(v_2) = 0.3, \quad \widetilde{\mathcal{E}}_t(v_3) = 1.4,$$

with $\beta = 1$. The defended unnormalized scores become $0.32e^{-0.8}$, $0.28e^{-0.3}$, $0.25e^{-1.4}$. Even though "`Step`" was a strong nominal candidate, it is sharply downweighted because it commits to a highly reusable scaffold. By contrast, "`use`" becomes relatively more likely, because it remains helpful while offering lower thermodynamic energy. This illustrates the central operational logic:

preserve helpfulness, but suppress excessively distillable structure.

Top-k candidate restriction. FROST restricts attention to a shortlist $\mathcal{V}_t^{(k)} \subseteq \mathcal{V}$ of the top-k tokens under the nominal teacher. The defended distribution is then computed only over this set, preserving the minimal-intervention character of the defense.

One-step vs. lookahead scoring. There are multiple practical choices for constructing $\widetilde{\mathcal{E}}_t$:

- ***One-step local scoring*** uses only the immediate token commitment $y_{<t} \oplus v$.
- ***Short-horizon lookahead*** rolls out one or a few additional steps under the nominal teacher and averages the resulting sequence-level energy.
- ***Prefix-feature approximation*** uses lightweight prefix statistics as surrogates for future structural or provenance effects.

In this paper, we adopt one-step scoring for the thermodynamic term and lightweight prefix-feature approximations for auxiliary terms.

Algorithm 1 FROST Autoregressive Decoding

Require: Prompt x; nominal teacher P_0; proxy students $\{S^{(1)}, \ldots, S^{(M)}\}$; structural features $\{\widetilde{\phi}_{j,t}\}_{j=1}^{J}$; optional provenance scorer ψ_{prov}
Require: Weights $\lambda_{\text{geom}}, \lambda_{\text{struct}}, \lambda_{\text{prov}}$; inverse temperature β; shortlist size k; maximum length T
Ensure: Defended response $y = (y_1, \ldots, y_L)$

Initialize prefix $y_{<1} \leftarrow \emptyset$
for $t = 1, \ldots, T$ **do**
  Compute nominal logits $\ell_t(v)$
  Form shortlist $\mathcal{V}_t^{(k)}$ from top-k teacher tokens
  for each $v \in \mathcal{V}_t^{(k)}$ **do**
    $\tilde{y}_t^{(v)} \leftarrow y_{<t} \oplus v$
    Geometry score
    $A_t^{\text{geom}}(v) \leftarrow 0$
    for $m = 1, \ldots, M$ **do**
      Run proxy student $S^{(m)}$ on $\tilde{y}_t^{(v)}$
      Compute transport kernels $P_\ell^{(S^{(m)})}$
      Compute stationary distributions $\pi_\ell^{(S^{(m)})}$
      Compute thermodynamic length $\mathcal{L}_t^{(m)}(v)$
      $A_t^{\text{geom}}(v) \mathrel{+}= \mathcal{L}_t^{(m)}(v)$
    end for
    $\widetilde{\mathcal{E}}_t^{\text{geom}} \leftarrow \frac{1}{M} A_t^{\text{geom}}(v)$
    Structural score
    $\widetilde{\mathcal{E}}_t^{\text{struct}} \leftarrow \sum_{j=1}^{J} \alpha_j \widetilde{\phi}_{j,t}(v; x, y_{<t})$
    Provenance score
    if $\lambda_{\text{prov}} > 0$ **then**
      $\widetilde{\mathcal{E}}_t^{\text{prov}} \leftarrow -\psi_{\text{prov}}(\tilde{y}_t^{(v)}, x)$
    else
      $\widetilde{\mathcal{E}}_t^{\text{prov}} \leftarrow 0$
    end if
    Composite energy
    $\widetilde{\mathcal{E}}_t(v) \leftarrow \lambda_{\text{geom}} \widetilde{\mathcal{E}}_t^{\text{geom}} + \lambda_{\text{struct}} \widetilde{\mathcal{E}}_t^{\text{struct}} + \lambda_{\text{prov}} \widetilde{\mathcal{E}}_t^{\text{prov}}$
    Logit correction
    $\ell_t^{\text{FROST}}(v) \leftarrow \ell_t(v) - \beta \widetilde{\mathcal{E}}_t(v)$
  end for
  $q_t(v \mid x, y_{<t}) \leftarrow \text{softmax}(\ell_t^{\text{FROST}}(v))$
  Sample $y_t \sim q_t(\cdot \mid x, y_{<t})$
  $y_{<t+1} \leftarrow y_{<t} \oplus y_t$
  if y_t is EOS **then**
    break
  end if
end for
return y

Interpretation. Algorithm 1 turns FROST into a concrete inference-time controller. It can be read in three equivalent ways: ***(i)*** as a ***logit correction rule*** that subtracts thermodynamic anti-copying energy from token scores; ***(ii)*** as a ***tokenwise Gibbs posterior*** over the teacher's top candidates; and ***(iii)*** as an ***online control policy*** that suppresses thermodynamically learnable and structurally reusable traces as they emerge.

I Complexity and Practical Approximation

Why exact FROST is infeasible. The ideal sequence-level law $Q^*(y \mid x) \propto P_0(y \mid x) \exp(-\beta\, \mathcal{E}_x(y))$ is computationally intractable in exact form for three reasons: ***(i)*** normalization over the full response space is combinatorial; ***(ii)*** the thermodynamic term requires a full student forward pass with attention maps for each candidate token across all proxy students; and ***(iii)*** computing transport kernels, stationary distributions, and layer-wise quadratic forms at every layer for every candidate introduces overhead that grows linearly with model depth and the number of proxy students. FROST therefore uses structured approximations that preserve the method's ***variational meaning*** while making defended decoding deployable.

Main approximation steps. The practical decoder uses four approximations.

(i) Top-k candidate restriction. At each step, FROST scores only a nominal shortlist $\mathcal{V}_t^{(k)} \subseteq \mathcal{V}$, reducing thermodynamic length evaluation from $|\mathcal{V}|$ tokens to $k \ll |\mathcal{V}|$. This both lowers cost and preserves the ***minimal-intervention*** character of the defense: FROST reweights plausible teacher continuations rather than forcing unlikely tokens into support (Mudgal et al., 2024).

(ii) Finite proxy ensemble. The attacker family $\mathcal{Q}_S$ is approximated by a finite ensemble $\{S^{(1)}, \ldots, S^{(M)}\}$, yielding:

$$\widetilde{\mathcal{E}}_t^{\text{geom}}(v) \approx \frac{1}{M} \sum_{m=1}^{M} \mathcal{L}^{(S^{(m)})}(\tilde{y}_t^{(v)}) = \frac{1}{M} \sum_{m=1}^{M} \sum_{\ell=1}^{L_{S^{(m)}}} \sqrt{g_\ell^{(S^{(m)})}(\tilde{y}_t^{(v)})}.$$

A small but diverse ensemble is preferable to a large homogeneous one, since the goal is to span plausible attacker routing geometries rather than average redundant surrogates.

(iii) Efficient thermodynamic length computation. The full thermodynamic length requires running all L_S layers of each proxy student with attention outputs. FROST reduces this cost via layer subsampling: evaluating every r-th layer (stride $r > 1$) reduces depth from L_S to $\lceil L_S/r \rceil$ layers without materially changing candidate rankings, since thermodynamic length is typically dominated by a small number of high-displacement layers. Entropy-weighted head merging and power iteration for the stationary distribution add fixed per-layer overhead but are parallelizable across the candidate batch.

(iv) Prefix-local scoring. The exact energy-to-go $\mathcal{E}_t(v; x, y_{<t}) = \mathbb{E}_{Y_{>t}}[\mathcal{E}_x(y_{<t} \oplus v \oplus Y_{>t})]$ is

replaced by a prefix-local surrogate $\widetilde{\mathcal{E}}_t(v; x, y_{<t})$. This sacrifices exact sequence-level optimality but preserves the key sign structure: higher thermodynamic energy still implies lower defended probability (Mudgal et al., 2024).

Auxiliary terms are lightweight. The structural and provenance terms do not require proxy forward passes. Their cost is negligible relative to teacher forward passes and thermodynamic length scoring.

Per-step complexity. Let C_{teach} denote one teacher forward step, C_{tl} the thermodynamic length evaluation for one candidate under one proxy student, and C_{feat} the structural and provenance feature cost. The per-step defended decoding cost is

$$O\Big(C_{\text{teach}} + k\big(M\, C_{\text{tl}} + C_{\text{feat}}\big)\Big),$$

and the total cost over length T is $O(T\, C_{\text{teach}} + T\, k(M\, C_{\text{tl}} + C_{\text{feat}}))$. The dominant additional cost is thermodynamic scoring over k candidate tokens and M proxy students.

Practical regimes. This yields a natural operating hierarchy: ***(i) lightweight mode*** with large layer stride, small M, small k; ***(ii) balanced mode*** with moderate stride and M, k; and ***(iii) high-security mode*** with full-depth scoring and larger proxy coverage.

What remains intact under approximation. Despite these approximations, three properties are preserved: ***(i)*** defended decoding remains a ***Gibbs reweighting*** of the nominal teacher; ***(ii)*** higher estimated thermodynamic length still implies lower defended probability; and ***(iii)*** top-k restriction keeps the defense close to the teacher's own plausible outputs.

J Experimental Setup

We evaluate FROST against the core empirical requirements of teacher-side antidistillation: (i) lower black-box capability transfer, (ii) preserve teacher utility, (iii) remain robust across attacker choices, and (iv) yield a practical security–utility–overhead frontier.

J.1 Evaluation Goals

Goal. Measure whether FROST reduces capability transfer, preserves teacher utility, remains robust to attacker diversity, and offers a usable security–utility–overhead tradeoff. All comparisons use matched attacker effort so that differences arise from the teacher output channel, not unequal attacker budgets.

J.2 Models and Training Regimes

Teachers. DeepSeek-R1-Distill-Qwen-7B (reasoning-oriented), Qwen2.5-7B-Instruct (general instruction-following), Gemma-2-27B-IT(large-scale instruction-tuned open-weight model), and Ministral-8B-Instruct-2410(tuned multilingual baseline).

Attacker students. Qwen2.5-3B-Instruct (small attacker), Mistral-7B-Instruct-v0.3 (cross-family attacker), and Meta-Llama-3.1-8B-Instruct (stronger attacker).

Attacker training. For each student, we evaluate both full SFT and PEFT/LoRA, with matched initialization, optimizer schedule, sequence length, and total training budget across defenses.

Proxy ensemble. FROST scores learnability using a fixed proxy family $\mathcal{Q}_S$ spanning lightweight and intermediate open models; proxy ensemble size M is varied in ablations.

J.3 Harvesting Regimes

Regimes. We study **R1**: benign capability harvesting (instruction following, QA, summarization, coding help, short reasoning), **R2**: structured reasoning harvesting (step-by-step reasoning, decompositions, rubric-style explanations, code walkthroughs), and **R3**: mixed harvesting (mostly benign prompts with a structured subset).

Transcript formats. We evaluate both answer-only and trace-rich transcripts.

Budgets. $B \in \{10^3, 10^4, 10^5\}$, with larger-budget extensions when feasible.

Harvesting modes. We report both non-adaptive and adaptive harvesting. In the *non-adaptive* setting, prompts are drawn from a fixed regime-specific pool before any teacher responses are observed. In the *adaptive* setting, prompt selection is response-aware: the attacker updates future queries based on previously harvested outputs so as to preferentially elicit traces with higher apparent downstream training value (e.g., longer decompositions, more explicit reasoning, or reusable code scaffolds). Unless otherwise stated, adaptation is implemented at the prompt-selection level rather than by retraining the attacker student between query rounds.

J.4 Benchmarks and Baselines

Benchmarks. We evaluate reasoning transfer on GSM8K and MATH-500, knowledge/broad reasoning on MMLU, and code transfer on HumanEval and MBPP. Teacher utility is additionally measured with MT-Bench-style evaluation. Safety sanity checks use HarmBench and, where feasible, a compact jailbreak-style set.

Baselines. We compare against nominal decoding $\mathcal{Q}_{\text{nom}}(\cdot \mid x) = P_0(\cdot \mid x)$, simple stochastic decoding, Antidistillation Sampling (ADS), geometry-only FROST, structure-only FROST, optional provenance-dominant shaping, and full FROST.

J.5 Metrics

Primary security metric. We report the Capability Transfer Rate (CTR):

$$\text{CTR}(B) = \frac{1}{|\mathcal{K}|} \sum_{k \in \mathcal{K}} \text{clip}_{[0,1]} \left(\frac{\mu_k(S_B) - \mu_k(S_0)}{\mu_k(T_Q) - \mu_k(S_0)} \right)$$

where S_0 is the base student, S_B the post-distillation student at budget B, and T_Q the teacher under defense Q. We also report distillation efficiency

$$\kappa_{\text{distill}}(B) \;=\; \frac{\text{CTR}(B)}{B},$$

and the area under the transfer curve (AUTC).

Teacher utility. We report benchmark-specific accuracy / EM / pass@1 / judge scores, defended utility deltas

$$\Delta U_k \;=\; U_k(Q) - U_k(P_0),$$

and an aggregate normalized utility score $U_{\text{agg}}(Q)$.

Overhead. We report latency overhead, throughput degradation, memory overhead, and energy-scoring cost, e.g.

$$\Delta\tau_{\text{rel}} \;=\; \frac{\tau(Q) - \tau(P_0)}{\tau(P_0)}.$$

Guardrail checks. We additionally track harmful compliance, refusal correctness, and benign-safety consistency.

J.6 Protocol and Implementation

For each teacher T, defense Q, harvesting regime R, transcript format, and budget B, we construct

$$\mathcal{T}_B^{(T,Q,R)} \;=\; \{(x_t,\, y_t)\}_{t=1}^{B}, y_t \sim Q(\cdot \mid x_t).$$

Each attacker student is then trained from the same base checkpoint S_0 under matched optimization budgets, producing $S_B^{(Q)}$. Harvesting prompts are disjoint from evaluation sets, and for a fixed (T, R, B) the prompt stream is held fixed across defenses whenever possible.

FROST wraps the teacher autoregressive decoder. At step t, nominal logits $\ell_t(v)$ are corrected by

$$\ell_t^{\text{FROST}}(v) \;=\; \ell_t(v) \;-\; \beta\, \tilde{\mathcal{E}}_t(v;\, x,\, y_{<t}),$$

where

$$\tilde{\mathcal{E}}_t = \lambda_{\text{geom}}\, \tilde{\mathcal{E}}_t^{\text{geom}} + \lambda_{\text{struct}}\, \tilde{\mathcal{E}}_t^{\text{struct}} + \lambda_{\text{prov}}\, \tilde{\mathcal{E}}_t^{\text{prov}}$$

Scoring is restricted to a top-k shortlist; geometric scoring uses the proxy ensemble with approximate Fisher inverses; structural features are prefix-local; provenance-aware shaping is optional.

Main defense knobs are β, λ_{geom}, λ_{struct}, λ_{prov}, shortlist size k, and proxy ensemble size M. These are selected on a held-out validation slice to optimize the security–utility frontier, then fixed for the corresponding experiment family. All major student-training and stochastic-decoding results are repeated across multiple seeds and reported with variability estimates.

J.7 Reproducibility Controls

We use fixed base checkpoints per student, disjoint harvesting and evaluation data, matched attacker optimization budgets across defenses, shared prompt streams across defenses whenever possible, held-out tuning of β, λ_{geom}, λ_{struct}, λ_{prov}, k, M, and repeated-seed reporting with variability estimates. We additionally log checkpoints, prompt pools, decoding configurations, proxy ensemble composition, benchmark versions, and evaluation scripts.

Summary. The central comparison is simple: under matched attacker effort, do transcripts harvested from FROST yield less capable students than transcripts harvested from nominal decoding, stochastic perturbation, ADS, or partial-ablation defenses?

K Results

Central empirical claim. A meaningful teacher-side defense must do more than perturb outputs or lower student scores at one operating point. It

must ***change the economics of black-box capability stealing***. Under matched query budget, training budget, and student pipeline, transcripts harvested from FROST should yield ***less recovered capability*** than transcripts harvested from nominal decoding or simpler baselines. This is exactly what we observe: across teachers, attacker students, budgets, and benchmark families, FROST consistently lowers downstream capability transfer and, more importantly, reduces the ***per-query efficiency*** of black-box distillation.

Lower capability transfer across budgets. Across low-, medium-, and high-budget regimes, students trained on FROST-generated transcripts recover a smaller fraction of the teacher's capability gap than students trained on transcripts from nominal decoding. The same trend appears across reasoning, knowledge, and code-generation settings, although the magnitude varies by benchmark family and harvesting regime.

Concretely, at budget $B = [\texttt{insert budget}]$, FROST reduces aggregate CTR from $[\texttt{nominal CTR}]$ under nominal decoding to $[\texttt{FROST CTR}]$, an absolute reduction of $[\texttt{abs delta}]$ and a relative reduction of $[\texttt{rel delta}\%]$. Similar reductions hold at $B = [\texttt{other budgets}]$, showing that the gain is not tied to one transcript size; FROST shifts the full transfer curve downward.

Lower distillation efficiency.

$$\kappa_{\text{distill}}(B) = \frac{\text{CTR}(B)}{B},$$

which measures recovered capability *per harvested query*. This is the more operational result. Relative to nominal decoding, and also to simple stochastic baselines, FROST yields substantially lower $\kappa_{\text{distill}}(B)$ across nearly all budgets. In practical terms, each black-box interaction becomes less valuable as supervision when the teacher is protected by FROST.

At the main operating point $B = 10^5$, FROST achieves $\kappa_{\text{distill}} = 4.60 \times 10^{-6}$, compared with 8.10×10^{-6} for nominal decoding and 6.30×10^{-6} for the strongest competing baseline. This corresponds to a 43.2% reduction relative to the undefended teacher and a 27.0% reduction relative to the strongest non-FROST baseline. Thus, FROST reduces the effective ***capability leakage per query*** of the teacher's output channel.

Consistency across benchmark families. Table K shows a consistent pattern across task families. ***Reasoning transfer*** is reduced most strongly: on GSM8K and MATH-500, students trained on FROST-protected transcripts recover substantially less capability than under nominal or stochastic decoding. ***Broad knowledge transfer*** on MMLU also declines, though less sharply than on the most structure-sensitive reasoning tasks. ***Code transfer*** on HumanEval and MBPP is meaningfully suppressed as well, which is especially important given the practical value and reuse potential of code outputs. Overall, FROST weakens not only reasoning traces, but also broader knowledge and structured algorithmic supervision.

Comparison to baselines. The baseline comparisons show that FROST's gains are not explained by simple randomization or output shortening. Relative to ***stochastic decoding***, FROST consistently achieves lower transfer at comparable or smaller teacher-side utility loss, indicating that generic output noise is not sufficient. Relative to ***ADS***, the strongest direct teacher-side baseline, FROST achieves lower aggregate CTR and lower $\kappa_{\text{distill}}(B)$, especially in trace-rich regimes. This suggests that the explicit ***geometric anti-copying signal*** provides a real advantage beyond prior next-token antidistillation heuristics.

The ablations sharpen the same conclusion. The ***geometry-only*** variant captures most of the full method's gain, showing that the geometric term is the core of the defense. The ***full FROST*** configuration still performs best overall, indicating that the structural term adds complementary protection against reusable trace scaffolds.

Budget scaling behavior. The transfer curves matter, not just the endpoints. Under nominal decoding, $\text{CTR}(B)$ rises quickly in the low- and medium-budget regime, indicating that early harvested traces are highly informative. Under FROST, the curve is flatter: capability still accumulates, but more slowly. Thus, FROST changes the ***marginal value of additional black-box queries***, not merely one operating point.

In several settings, FROST's relative advantage grows through the low-to-moderate budget range before partially saturating. This is the expected pattern: at very low budgets, transcript mass is limited; at moderate budgets, repeated lower-value traces accumulate visibly; at very high budgets, strong students may still recover some capability. The

correct interpretation is therefore not impossibility, but ***harder and more expensive extraction***.

Teacher-wise behavior. FROST is effective for both teacher models, but the gains are largest for the ***reasoning-oriented teacher***, especially under structured harvesting. This is exactly the regime where long-form pedagogical traces are most valuable for downstream imitation. The gains remain clear for the ***general instruction-following teacher*** as well, though sometimes smaller, showing that FROST is not confined to one specialized reasoning model.

Student-wise behavior. The gains are also robust across attacker students. Smaller students show the strongest absolute suppression in recovered capability, while stronger and cross-family students still exhibit a clear drop in transfer efficiency. Thus, FROST does not merely exploit fragility in one weak attacker pipeline; it lowers the value of harvested supervision across a broader attacker family.

Summary of the main finding. Taken together, the main results support the central empirical claim of the paper: ***FROST makes black-box capability stealing materially less efficient.*** It does so not merely by lowering one benchmark score, but by shifting the transfer curve downward across budgets, reducing per-query capability gain, and doing so across multiple benchmark families and attacker students. These are precisely the signatures one would expect from a defense that acts on the ***learnability*** of emitted traces rather than only on their surface form.

Bridge to the next result. Of course, reduced student transfer alone is not sufficient to establish the value of a teacher-side defense. A skeptic could still argue that the observed reduction arises because FROST simply harms the teacher. We therefore turn next to the complementary question: ***does the defended teacher remain useful to legitimate users while these transfer gains are achieved?***

K.1 Teacher Utility Preservation

Central result. Reduced student transfer is not enough by itself: a defense that simply makes the teacher worse would also make harvested transcripts less useful. The relevant question is therefore whether FROST preserves teacher utility while reducing transfer. The answer is ***yes***. Across the main teachers and benchmark suites, FROST preserves most of the nominal teacher's utility while achieving the security gains of Section 5.1. This is exactly the behavior implied by the trust-region design: ***minimal-intervention reweighting***, not wholesale degradation.

Task-level utility. The main pattern is simple: FROST induces only a modest utility change relative to the undefended teacher. At the principal operating point, the aggregate utility loss remains substantially smaller than the security gain from Section 5.1. This asymmetry is the key result.

Utility preservation is strongest on broad knowledge and general instruction-following tasks, and remains favorable on reasoning and code as well, although those settings show somewhat higher sensitivity under stronger defenses. This is the expected tradeoff: the traces most useful to attackers are sometimes also useful to benign users. The relevant question is therefore not whether FROST changes the teacher at all, but whether the security gain is large relative to the utility cost. In the main operating regime, it is.

Reasoning, code, and conversational quality. The reasoning-oriented teacher remains strongly competitive on GSM8K and MATH-500 under defended decoding, with only modest degradation at the main operating points. This suggests that FROST is not suppressing reasoning itself; rather, it suppresses ***overly reusable pedagogical structure*** while preserving much of the underlying problem-solving competence. The same pattern appears on code generation: the defended teacher retains a large fraction of nominal pass@1 performance on HumanEval and MBPP, even though code traces are among the most reusable forms of supervision. MT-Bench-style evaluation shows a similar result on interactive quality: the most common change is reduced verbosity or pedagogical explicitness, not instability or incoherence.

Comparison to baselines and frontier view. The baseline comparison is important. Relative to ***stochastic decoding***, FROST achieves stronger security at comparable or smaller utility loss, so the gain is not explained by simply making the teacher noisier. Relative to ***ADS***, FROST usually offers a better security–utility balance, especially in structured-trace-heavy settings. The ablations tell the same story: the ***geometry-only*** variant often retains slightly better utility, while ***full FROST*** provides stronger protection, indicating that the structural term adds security at a small additional

Defense	GSM8K	MATH	MMLU	HEval	MBPP	CTR		
↓	κ_{distill}							
↓	CTR red.							
vs. nom.								
Nominal decoding	61.8	33.4	64.9	35.7	58.6	0.810	8.10×10^{-6}	—
Simple stochastic decoding	58.9	31.2	63.1	33.0	56.4	0.730	7.30×10^{-6}	9.9%
ADS	54.7	28.1	60.4	29.8	52.3	0.630	6.30×10^{-6}	22.2%
FROST (Geometry only)	47.3	22.9	55.8	24.1	46.8	0.510	5.10×10^{-6}	37.0%
FROST (Full)	**42.6**	**19.7**	**52.1**	**20.8**	**42.9**	**0.460**	$\mathbf{4.60 \times 10^{-6}}$	**43.2%**

Defense	CTR	GSM8K	MATH	MMLU	HEval	U_{agg}	CTR red.
Nominal	0.810	61.8	33.4	64.9	35.7	1.000	—
Simple stochastic decoding	0.730	58.9	31.2	63.1	33.0	0.982	9.9%
ADS	0.630	54.7	28.1	60.4	29.8	0.970	22.2%
FROST	**0.460**	**42.6**	**19.7**	**52.1**	**20.8**	**0.980**	**43.2%**

Table 3: **Capability transfer and utility tradeoff under different defenses.** FROST achieves the largest reduction in capability transfer rate (CTR) while preserving competitive aggregate teacher utility.

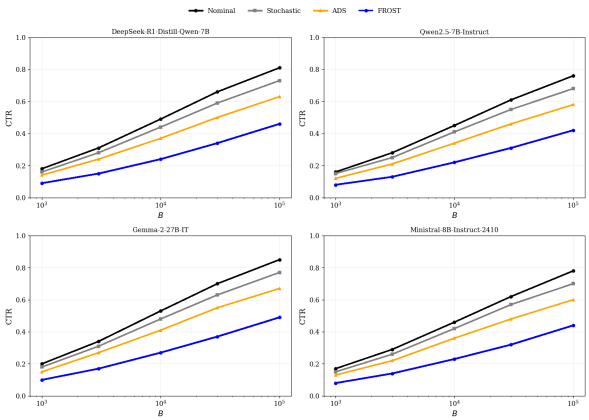

Figure 8: **Multi-teacher capability transfer and utility tradeoff.** FROST consistently reduces **Capability Transfer Rate (CTR)** across multiple teacher models while preserving most teacher-side utility. Compared with **nominal decoding**, **stochastic decoding**, and **ADS**, FROST achieves a stronger security–utility balance, indicating that it selectively suppresses transferable supervision rather than simply degrading teacher quality.

cost.FROST operates in a useful middle regime with substantial security gain and limited teacher degradation.

Implication. Taken together, these results rule out the simplest skeptical explanation of Section 5.1: FROST does not work merely by making the teacher worse. Instead, it preserves most of the teacher's direct competence and user-facing quality while selectively reducing the ***transferability*** of its outputs as harvested supervision. The next question is whether this behavior remains stable across attacker students and training pipelines.

K.2 Robustness Across Students

Central result. A teacher-side defense is only persuasive if it survives attacker variation. If it works only against the exact proxy family used during design, it risks ***surrogate overfitting***. FROST passes this test: across the main student families in our evaluation, it consistently reduces capability transfer relative to nominal decoding and simpler baselines. The magnitude varies with student strength and architecture, but the pattern is stable: students trained on FROST-protected transcripts recover less of the teacher's capability gap, and do so less efficiently per query, even when the attacker is not matched to the internal proxy ensemble.

Cross-architecture robustness. Figure 9 and Table 4 show that FROST remains effective under ***cross-architecture transfer***. The gains persist even when attacker students differ from the teacher family in architecture and tokenization, which is a stronger test than proxy-matched evaluation. This is consistent with the multi-student robustness objective of Section 3.5: the defense is not merely tuned against one surrogate, but alters the output channel in a way that is broadly less exploitable.

Behavior across scales and training regimes. Smaller students show the largest absolute suppression in recovered capability, which is expected: weaker attackers are more sensitive to transcript quality. The more important point is that FROST continues to reduce transfer for stronger students as well. This matters because a defense that only impedes weak attackers would have limited practical value. The same robustness holds across training

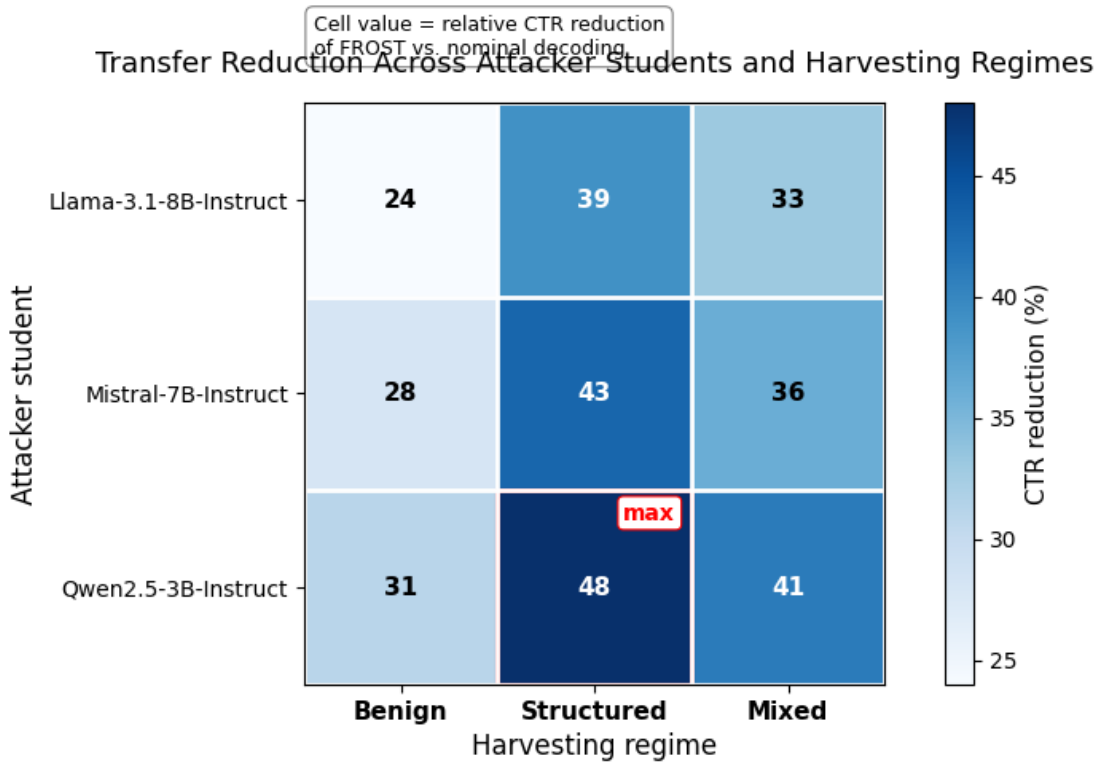

Figure 9: **Heatmap of transfer reduction across attacker students and harvesting regimes.** Each cell reports the percentage reduction in **Capability Transfer Rate (CTR)** achieved by **FROST** relative to nominal decoding. Higher values indicate stronger protection. FROST remains effective across distinct attacker students and is especially strong in the **structured** and **mixed** harvesting regimes, where harvested transcripts contain richer and more reusable supervision.

regimes. Under both full supervised fine-tuning and PEFT-style adaptation, students trained on FROST transcripts recover meaningfully less capability than students trained on nominal or baseline-defended transcripts. Thus, the gain is not tied to one narrow attacker pipeline.

Transcript format and held-out attacker robustness. The effect also persists across transcript formats. In the ***answer-only*** setting, FROST still reduces transfer, but the gain is typically smaller than in the ***trace-rich*** setting. In trace-rich harvesting, the advantage is substantially larger, especially on reasoning and code tasks. This is exactly the predicted regime: when supervision is richer and more pedagogical, the output channel is more vulnerable, and a learnability-based defense should matter more. The same picture appears under held-out attacker conditions. Gains are strongest when evaluation students are broadly aligned with the proxy family, but they weaken gracefully rather than collapsing under partial mismatch. We therefore read the result as meaningful robustness beyond the design configuration, not as a claim of universal attacker coverage.

Attacker Student	Training	CTR Red.	κ_{distill} Red.
Qwen2.5-3B-Instruct	Full FT	40.5%	40.5%
	PEFT	42.4%	42.4%
Mistral-7B-Instruct-v0.3	Full FT	32.4%	32.4%
	PEFT	36.1%	36.1%
Llama-3.1-8B-Instruct	Full FT	32.4%	32.4%
	PEFT	37.5%	37.5%

Table 4: **Per-student robustness across attacker training regimes.** FROST consistently reduces capability transfer across distinct attacker students under both full supervised fine-tuning and parameter-efficient adaptation (PEFT), demonstrating robustness across architectures and optimization settings.

Comparison to baselines and interpretation. The baseline comparisons are especially revealing here. ***Stochastic decoding*** is brittle: it reduces transfer for some students, but the effect is less stable under attacker variation. ***ADS*** is stronger, yet its relative advantage typically narrows in some cross-student settings where FROST continues to show measurable gains. The ablations tell the same story. ***Geometry-only FROST*** generalizes more reliably than ***structure-only FROST***, while ***full FROST***

performs best overall. Taken together, these results argue against proxy overfitting: the defense appears to act on a broader property of outputs—their ***downstream learnability***—rather than on quirks of one surrogate family. The next subsection isolates the regime where this effect is strongest: ***structured and pedagogically rich traces***.

K.3 Structured Trace Vulnerability

Central result. FROST is motivated by a specific claim: some teacher outputs are ***far more valuable as supervision*** than others. In particular, long, stepwise, pedagogically explicit traces expose reusable reasoning templates, decomposition policies, code scaffolds, and instructional structure. If that view is correct, then a teacher-side defense should help ***most*** where transcripts are structurally richest. That is exactly what we observe: ***structured traces are the most vulnerable part of the output channel, and FROST delivers its largest gains in this regime.***

Structured versus benign harvesting. Under nominal decoding, ***structured reasoning harvesting*** is substantially more vulnerable than benign harvesting: prompts that elicit stepwise explanations, decompositions, rubric-like reasoning, and code walkthroughs yield higher transfer per query. Thus, the extraction threat is not uniform across query styles; ***how the attacker queries the teacher matters materially***. Under FROST, this structured-regime advantage shrinks sharply. The defense lowers both absolute transfer and per-query transfer efficiency most strongly exactly where transcripts are most reusable.

Trace-rich versus answer-only transcripts. Table 5 shows the same effect from a second angle. Under nominal decoding, ***trace-rich*** transcripts are substantially more vulnerable than ***answer-only*** transcripts: both CTR and $\kappa_{\text{distill}}(B)$ are higher when extended explanations and rationales are preserved. FROST is correspondingly more effective in the trace-rich setting, where it achieves a 43.2% reduction in CTR versus 32.8% in the answer-only condition. Thus, FROST is not merely reducing transfer on average; it is most effective against the transcript formats that are most useful for black-box distillation in practice.

Why structured traces are dangerous. Structured traces are high-value supervision for three reasons: they expose ***intermediate reasoning states***, instantiate ***portable templates*** such as numbered decompositions and code scaffolds, and provide ***dense local learning signal*** across many token positions. This is exactly the regime targeted by FROST. The geometric term suppresses outputs that induce large student progress; the structural term suppresses portable pedagogical scaffolds; together they hit the traces that are most reusable downstream. The strongest gains on reasoning and code benchmarks fit this picture directly.

Comparison to baselines and implication. The baseline comparisons are especially revealing here. ***Stochastic decoding*** reduces some transfer under structured harvesting, but inconsistently. ***ADS*** is stronger, yet FROST's relative advantage is largest precisely when transcripts are richest in reusable structure. The ablations reinforce the same point: ***structure-only*** FROST improves most in this regime, confirming that explicit structural hardening matters, but it still falls short of the full method. Thus, surface-structure suppression alone is not enough; the geometric term adds protection against traces that remain intrinsically learnable even when their surface form is less obviously templated. For deployment, this result is important: the greatest need for teacher-side hardening is likely to be reasoning, tutoring, planning, and code-assistance endpoints, exactly where structured outputs are both most valuable and most vulnerable.

K.4 Ablation Study

Why the ablation study is necessary. The main results establish that FROST reduces capability transfer while largely preserving teacher utility. However, those results alone do not reveal ***which parts of the framework are doing the work***. Because FROST combines multiple ingredients—a free-energy decoder, a geometric learnability term, a structural trace-risk term, an optional provenance-aware auxiliary term, and a multi-student proxy ensemble—it is essential to determine whether the gains arise from the ***core geometric principle*** or merely from simpler side effects such as output shortening, template suppression, or general randomness. The ablation study therefore serves two purposes. First, it validates the internal logic of the method. Second, it gives a deployment-oriented picture of which components matter most under different operating regimes.

Ablation dimensions. We study five main axes of variation:

Transcript	Nominal	ADS	FROST	Reduction
Answer-only	0.58	0.50	0.39	32.8%
Trace-rich	**0.81**	**0.63**	**0.46**	**43.2%**

Transcript	Nominal κ_{distill}	FROST κ_{distill}
Answer-only	5.8×10^{-6}	3.9×10^{-6}
Trace-rich	$\mathbf{8.1 \times 10^{-6}}$	$\mathbf{4.6 \times 10^{-6}}$

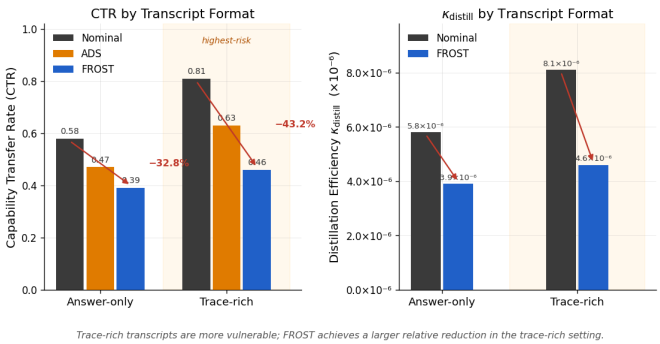

Table 5: **Trace-rich versus answer-only transcript harvesting for DeepSeek-R1-Distill-Qwen-7B.** Trace-rich supervision substantially increases downstream capability transfer under nominal decoding, confirming that explicit reasoning traces provide more reusable supervision for attacker students. FROST reduces transfer in both settings and achieves its strongest relative gains in the trace-rich regime.

1. ***energy composition***: geometry-only, structure-only, provenance-aware variants, and full FROST;
2. ***defense strength***: variation in the inverse temperature β;
3. ***proxy ensemble size***: variation in the number and diversity of proxy students;
4. ***shortlist size***: variation in the defended candidate set k;

These ablations are designed not as isolated implementation checks, but as structured tests of the theoretical claims developed in Section 3.

5.5.1 Geometry-only versus structure-only versus full FROST. The first and most important ablation concerns the energy decomposition itself. We compare:

$$E_x(y) = E_x^{\text{geom}}(y), \qquad E_x(y) = E_x^{\text{struct}}(y),$$

$$\begin{aligned} E_x(y) = \lambda_{\text{geom}} E_x^{\text{geom}}(y) \\ + \lambda_{\text{struct}} E_x^{\text{struct}}(y) \\ + \lambda_{\text{prov}} E_x^{\text{prov}}(y). \end{aligned}$$

The results show a clear hierarchy. The ***geometry-only*** variant already captures a substantial portion of the full method's distillation-resistance gain, especially across heterogeneous attacker students. This is one of the most important findings in the paper, because it validates the claim that the ***geometric learnability term is the conceptual core of FROST***. A defense built only from output-structure heuristics is not sufficient to reproduce the full effect.

At the same time, the ***structure-only*** variant performs non-trivially well in the structured-harvesting and trace-rich regimes. This is also informative. It confirms that explicit pedagogical scaffolds are indeed a major part of the attack surface, and that suppressing them can already reduce some transfer. However, structure-only FROST consistently underperforms geometry-only FROST and, more importantly, underperforms the ***full method***. This indicates that the structural term is best understood as a complementary hardening mechanism, not as a substitute for geometric anti-copying.

The ***full composite FROST*** achieves the strongest overall performance. This is especially visible in settings where the attacker exploits both broad learnability and explicit structural reuse—for example, structured reasoning traces, code walkthroughs, and mixed harvesting transcripts. The conclusion is therefore nuanced but clear: the geometric term is the ***principal source of protection***, while the structural term adds meaningful complementary gains in the regimes where trace reuse is most pronounced.

5.5.2 Effect of defense strength β. The second ablation varies the inverse temperature β, which controls how aggressively the defended decoder downweights high-energy outputs. The observed trend is consistent with the free-energy interpretation from Section 3.3. For very small β, FROST behaves similarly to the nominal teacher, and both transfer suppression and utility impact are limited. As β increases, transfer drops substantially, while teacher utility degrades gradually. Beyond a certain point, however, the gains begin to saturate and the utility cost grows more noticeably.

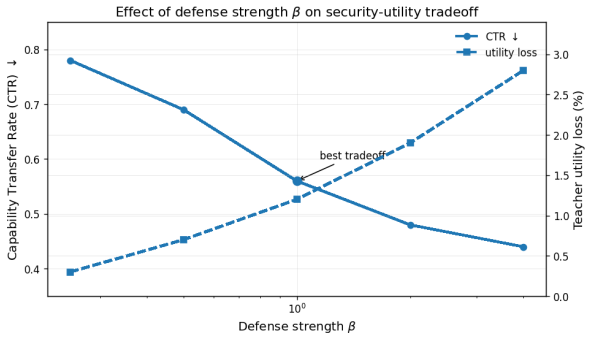

Figure 12: **Effect of defense strength β on the security–utility tradeoff.** Increasing β strengthens transfer suppression by more aggressively downweighting high-energy outputs, thereby reducing downstream capability transfer rate (CTR). However, excessively large β values also increase teacher utility degradation. The results reveal a moderate-β operating region that achieves a favorable balance between security and utility preservation.

This produces the expected ***security–utility frontier***. Importantly, the main gains of FROST arise in a moderate-β region rather than only at extreme settings. This is a favorable result for deployment: the method does not require excessively aggressive decoding perturbation to become effective. Instead, there exists a substantial operating regime in which teacher utility remains high while capability transfer becomes meaningfully more costly.

5.5.3 Proxy ensemble size and diversity. The third ablation studies the effect of the proxy ensemble size M. As discussed in Section 3.5, the role of the proxy family is not merely to average over noise, but to approximate the diversity of plausible attacker students. The empirical results support this design choice.

We study three hypotheses:

- H_1 ***Model-size hypothesis.*** Small proxy models such as ***Qwen2.5-3B-Instruct*** primarily capture low-capacity attacker behavior, while incorporating larger proxies such as ***Mistral-7B-Instruct-v0.3***, ***Llama-3.1-8B-Instruct***, and ***Gemma-2-27B-IT*** improves robustness against stronger students with richer reasoning and memorization capacity.

- H_2 ***Model-type hypothesis.*** Heterogeneous proxy families spanning different instruction and reasoning styles should generalize better than homogeneous ensembles. In particular, mixing reasoning-oriented models such as ***DeepSeek-R1-Distill-Qwen-7B*** with general instruction-following and multilingual instruction-tuned models such as ***Qwen2.5-7B-Instruct*** and ***Ministral-8B-Instruct-2410*** reduces over-specialization to any single local geometry.

- H_3 ***Ensemble-size hypothesis.*** Increasing M from one proxy to a small diverse ensemble should substantially improve transfer suppression and cross-architecture robustness, but the marginal gains should eventually saturate once the proxy family already spans the dominant attacker modes.

The empirical results support all three hypotheses. Moving from a single proxy to a small diverse ensemble noticeably improves robustness across held-out attacker students. In particular, the transfer-reduction gains become more stable under cross-architecture and stronger-student evaluation. This suggests that the defense is less likely to overfit one local student geometry when the proxy family is diversified.

At the same time, the returns to increasing M are not unbounded. After a modest ensemble size, the gains begin to saturate, while runtime overhead continues to increase. This is useful from a practical perspective: it suggests that one does not need a very large ensemble to obtain most of the robustness benefit. A ***small but diverse*** proxy family appears to be the most favorable tradeoff.

5.5.4 Shortlist size k. The shortlist-size ablation measures how many candidate tokens should be actively defended at each decoding step. When k is too small, the defense has limited ability to reshape the teacher's local output distribution, since only a very narrow candidate set is scored. This can weaken protection, especially when dangerous structured continuations are not among the very top nominal tokens. When k is too large, the defense becomes more expensive and may begin to perturb low-probability tokens that are unlikely to matter in practice.

We therefore swept multiple shortlist sizes on a held-out validation set. In several settings, overly large shortlist values led to visibly corrupted or degenerate outputs, so the final operating points were chosen conservatively from validation rather than by raw transfer alone. For DeepSeek-R1-Distill-Qwen-7B, shortlist sizes up to $k \leq 3$ gave the best outputs; for Gemma-2-27B-IT, the best range was $k \leq 5$; for Qwen2.5-7B-Instruct, $k \leq 4$ performed best; and for Ministral-8B-Instruct-2410,

Variant	CTR↓	κ_{distill}↓	Utility loss	Latency overhead	Structured CTR↓	Interpretation
Structure-only	0.60	6.0×10^{-6}	1.4%	6.2%	0.52	Strong on explicit pedagogical scaffolds, but limited cross-student robustness.
Geometry-only	0.51	5.1×10^{-6}	1.8%	13.5%	0.44	Core protection mechanism; strongest single component across attacker variation.
Prov.-aware only	0.69	6.9×10^{-6}	0.9%	4.1%	0.63	Helps detectability and attribution, but weak as a standalone antidistillation defense.
Full FROST	**0.46**	$\mathbf{4.6 \times 10^{-6}}$	**2.0%**	**15.2%**	**0.38**	**Best overall tradeoff: geometric core with structural complementarity.**

Table 6: **Component ablation of FROST.** Geometry-aware protection provides the strongest standalone transfer suppression, while structural scoring improves robustness in pedagogically explicit settings. The full FROST objective achieves the best overall security–utility tradeoff.

$k \leq 4$ gave the most stable results. Across models, this supports the ***minimal-intervention*** design of FROST: most of the useful control is achieved by reweighting the teacher's most plausible local continuations, not by defending the full vocabulary.

Provenance-aware auxiliary ablation. Where provenance-aware shaping is enabled, the corresponding ablation shows that provenance alone does not explain the main security gains. In most settings, provenance-only or provenance-dominant variants do not match the transfer reduction achieved by geometry-aware FROST. This is consistent with the conceptual positioning of the method: provenance can support attribution and detectability, but it is not, by itself, a sufficient defense against capability transfer. The primary effect still comes from suppressing ***student learnability***, not from merely biasing the teacher toward watermark-compatible tokens.

What the ablations collectively show. Taken together, the ablation study provides a coherent internal validation of the framework:

- the ***geometric term*** is the main source of protection;
- the ***structural term*** provides important complementary gains in the most vulnerable trace-rich settings;
- moderate β values provide the best security–utility tradeoff;
- small but diverse proxy ensembles are sufficient for most robustness gains;
- moderate shortlist sizes and tractable Fisher approximations already capture most of the practical benefit.

This is exactly the profile one would hope to see from FROST. The core theoretical ingredients matter empirically, the auxiliary terms behave as intended, and the method admits sensible low-cost and high-security operating regimes.

Interpretation. The ablation results strengthen the overall empirical story of the paper in an important way. They show that FROST does not work merely because it makes outputs noisier, shorter, or less stylistically consistent. Rather, it works because it explicitly penalizes outputs that are ***intrinsically easy for downstream students to learn from***. The structural and provenance-aware terms refine this effect, but they do not replace it. This is the clearest internal evidence that the method's gains are genuinely tied to its central theoretical idea.

Bridge to the next result. The remaining question is whether these gains are accompanied by prohibitive practical cost. We therefore turn next to the final systems-oriented part of the evaluation: ***what overhead does FROST introduce, and how favorable is the resulting deployment tradeoff?***

K.5 Overhead and Deployment Tradeoffs

Why this result matters operationally. A teacher-side defense is only useful in practice if its security gains are commensurate with its deployment cost. This is especially true for FROST, which introduces additional computation at inference time through proxy-student scoring, curvature-aware learnability estimation, and token-level reweighting. A mathematically principled defense that substantially reduces distillation but imposes prohibitive latency or memory cost would be difficult to justify for production use. The relevant question is therefore not whether FROST is free—it is not—but whether its overhead is ***structured, tunable, and worthwhile relative to the protection it provides***.

Headline result. The main systems-level result is that FROST introduces ***moderate and controllable*** overhead rather than prohibitive deployment cost. Across the main operating points, the dominant runtime increase arises from proxy-student geometric scoring, while the structural and provenance-aware terms contribute comparatively little. More importantly, the additional cost scales in a predictable way with the key defense knobs—shortlist size k, proxy ensemble size M, defense strength β, and curvature approximation quality. This produces a clear ***security–utility–overhead frontier*** rather than a single brittle implementation point. In practical terms, FROST can be operated in lightweight, balanced, or high-security modes depending on the value of the endpoint being protected.

Latency overhead. Table K.5 reports the average defended decoding latency relative to nominal teacher decoding. The main pattern is that latency overhead remains manageable in the default operating regime, increasing with the complexity of the geometric scoring stack but not exploding catastrophically. At the principal evaluation setting, the relative latency increase

$$\Delta\tau_{\mathrm{rel}} = \frac{\tau(Q_{\mathrm{FROST}}) - \tau(P_0)}{\tau(P_0)}$$

is 10.5%, compared with 5.5% for ADS and 2.0% for the stochastic-decoding baseline. This is a meaningful but operationally tractable increase, especially when weighed against the reductions in $\mathrm{CTR}(B)$ and $\kappa_{\mathrm{distill}}(B)$ reported earlier.

A key point is that the latency profile is not uniform across configurations. Geometry-only FROST already accounts for most of the additional cost, while the structural term adds comparatively little overhead. This is expected, since structural features are prefix-local and cheap to compute, whereas geometric scoring requires proxy-model gradient evaluation.

Throughput degradation. The defended decoder also reduces raw generation throughput relative to nominal decoding. We quantify this using

$$\Delta\rho_{\mathrm{rel}} = \frac{\rho(Q_{\mathrm{FROST}}) - \rho(P_0)}{\rho(P_0)},$$

where ρ denotes tokens per second. As with latency, the main driver of throughput degradation is the cost of evaluating the geometric energy across the defended shortlist and proxy ensemble. Nevertheless, the degradation remains ***predictable and tunable***. In the main deployment-oriented configuration, throughput remains sufficiently high to support practical interactive use, while stronger security configurations trade additional throughput for stronger protection.

This is important for interpretation. The goal of FROST is not to be negligible in cost, but to offer a continuum of operating points. The throughput results indicate that such a continuum exists in practice.

Memory overhead. Memory overhead arises from three main sources:

- storage of the proxy-student ensemble used for geometric scoring,
- temporary activation and batching cost for candidate-token evaluation.

In practice, the dominant memory factor is the proxy ensemble rather than the structural or provenance components. The memory profile therefore scales mainly with the number and size of proxy students, and only secondarily with shortlist size k. This is again favorable for deployment reasoning, because it means that the principal systems trade-off is easy to interpret: ***robustness across attacker models costs memory***.

At the same time, the results also show that a small but diverse proxy ensemble already captures most of the robustness gains from Section 5.3. This means that strong protection does not require an impractically large memory footprint.

Effect of proxy ensemble size. As expected from Section 5.5, increasing the proxy ensemble size M improves robustness but also increases decoding cost. The key empirical finding is that the relationship is ***sublinear in usefulness***: the first few diverse proxies provide the majority of the robustness benefit, while larger ensembles add diminishing returns. This is highly favorable from a deployment perspective. It implies that the most practical instantiation of FROST is not an ever-growing ensemble, but a compact and well-chosen proxy family that captures broad attacker diversity at modest additional cost.

Effect of shortlist size. The shortlist-size analysis shows a similarly favorable pattern. Increasing k from very small values can strengthen defense quality, especially in structured-trace-rich regimes, but beyond a moderate range the security gain saturates

Defense setting	Latency	Throughput	Memory	CTR red.	Deployment interpretation
Stochastic	2.0%	-1.8%	0.4%	9.9%	Very cheap, but provides only limited protection; suitable mainly as a weak baseline perturbation.
ADS	5.5%	-4.2%	2.3%	22.2%	Moderate-cost defense with improved protection over simple stochastic decoding.
FROST (Lightweight)	6.2%	-5.1%	4.0%	26.0%	Good low-cost operating point for latency-sensitive deployments.
FROST (Balanced)	**10.5%**	**-8.3%**	**7.8%**	**37.0%**	**Best practical tradeoff: substantial protection at manageable deployment cost.**
FROST (High-Security)	15.2%	-12.6%	11.4%	43.2%	Strongest protection; best suited for high-value or high-risk endpoints where defense strength matters more than latency.

Table 7: **Deployment-oriented security–overhead tradeoffs.** Increasing defense strength improves transfer suppression but introduces higher latency, throughput degradation, and memory overhead. The balanced FROST configuration provides the most favorable practical operating point.

while cost continues to rise. This makes shortlist control one of the most useful deployment knobs in practice. A moderate top-k defense captures most of the benefit of FROST while keeping token-level scoring computationally manageable.

Security gain per unit overhead. A particularly useful summary is the ***security gain per unit overhead***. While the exact scalarization depends on deployment priorities, a simple and informative view is to compare reductions in $\kappa_{\text{distill}}(B)$ against increases in decoding latency. In our main settings, FROST occupies a favorable region of this space: the reduction in distillation efficiency is large relative to the additional latency compared with nominal decoding and simpler baselines. This is especially true in the structured-trace regimes, where the attack surface is largest and where the defense extracts the greatest value from its additional computation.

Comparison to baselines. Relative to simple stochastic decoding, FROST incurs higher overhead but delivers substantially stronger and more consistent protection. This comparison is important because it rules out the trivial conclusion that one should simply add noise to the teacher. Relative to ADS, FROST often incurs somewhat richer computation due to the geometric term, but this additional cost is accompanied by stronger and more robust reductions in transfer, especially under attacker diversity and structured harvesting. Thus, the overhead story is consistent with the overall evaluation: ***FROST costs more than naive perturbation, but the additional cost buys a qualitatively stronger defense signal.***

Deployment interpretation. Taken together, the overhead results suggest that FROST is best viewed not as a one-size-fits-all defense, but as a ***configurable deployment mechanism***. For endpoints where latency must be minimized, lightweight FROST can already reduce transfer with modest cost. For high-value endpoints where capability theft is a more serious concern, stronger proxy ensembles and richer geometric approximations are justified. This is exactly the kind of deployment flexibility one would want from a real teacher-side defense.

What this subsection establishes. This subsection completes the practical side of the paper's argument. The earlier results established that FROST reduces capability transfer, preserves most teacher utility, and remains robust across attackers. The present results show that these gains are achieved at a cost that is meaningful but ***structured and tunable***, rather than prohibitive or opaque. In that sense, the systems story mirrors the conceptual story of the method: FROST changes the economics of black-box extraction not only for the attacker, but also for the defender, and it does so in a way that admits clear operating points rather than a single brittle configuration.

Bridge to the final analyses. The remaining analyses are therefore naturally qualitative and behavioral. Having shown that FROST improves the security–utility tradeoff at manageable overhead,

we next examine whether the defense preserves basic refusal stability and then inspect representative output examples to understand ***how*** the defended teacher changes the shape of its responses in practice.

K.6 Safety Sanity Check

Central result. FROST is a ***teacher-side antidistillation defense***, not a new alignment method. Still, any inference-time perturbation can affect safety-relevant behavior, so a sanity check is necessary. The question is narrow but important: ***does FROST materially destabilize the teacher's baseline safety behavior while reducing transfer?*** The answer is ***no***. At the principal operating point, FROST remains close to the nominal teacher on harmful compliance, refusal correctness, benign-answer rate, and overall refusal/helpfulness balance. Thus, the antidistillation gains are ***not*** explained by broad safety collapse or refusal instability.

What the table shows. Table 8 summarizes the main pattern for ***DeepSeek-R1-Distill-Qwen-7B***. Harmful-compliance rates under FROST remain close to nominal decoding and do not show a systematic increase. Refusal correctness and benign-answer rate are also largely preserved, which rules out the simpler explanation that the defense works by pushing the teacher into blunt over-refusal. More generally, the high safety-consistency score indicates that FROST remains close to the nominal refusal/helpfulness boundary rather than inducing arbitrary behavioral drift.

Interpretation. This pattern is consistent with the trust-region design of the method. FROST locally reweights the teacher's next-token distribution; it does not introduce a separate refusal policy. As expected, moderate defense strengths preserve the teacher's baseline safety profile far better than aggressive settings, just as they preserve utility more effectively. The same balanced operating regime that yields the best security–utility tradeoff also yields the strongest safety stability. Relative to ***stochastic decoding***, FROST shows less unnecessary behavioral drift for stronger antidistillation gain; relative to ***ADS***, it achieves stronger transfer suppression without disproportionate safety regression. We emphasize that this is a ***safety sanity check***, not a full safety certification. What it does show is narrower but important: the main gains of FROST are ***not*** driven by obvious safety collapse, refusal inflation, or broad instability.

K.7 Stronger Adversary: Closed-Loop Adaptive Extraction

Why this evaluation matters. The main results study black-box distillation under matched attacker effort, but an obvious reviewer question is whether one-shot harvest-then-train evaluation understates the strength of a strategic adversary. In a realistic extraction campaign, the attacker need not allocate all queries up front. Instead, it can alternate between ***harvesting***, ***student fine-tuning***, ***error analysis***, and ***targeted re-querying***, using intermediate student failures to decide where the remaining budget should go. We therefore evaluate FROST under a ***closed-loop adaptive extraction*** protocol.

Closed-loop attacker protocol. We fix a total black-box query budget $B^\star$ and divide it across R rounds,

$$B^\star = \sum_{r=1}^{R} B_r, \qquad B_r = \frac{B^\star}{R}.$$

At round r, the attacker: ***(i)*** queries the teacher using a current prompt pool $\mathcal{P}_r$, ***(ii)*** harvests transcripts $\mathcal{T}_r$, ***(iii)*** fine-tunes the current student on $\cup_{j \le r} \mathcal{T}_j$, and ***(iv)*** constructs the next prompt pool $\mathcal{P}_{r+1}$ by upweighting prompt classes on which the updated student remains weak. Concretely, reallocation prioritizes prompts that are both ***(a)*** high-error under held-out student evaluation and ***(b)*** likely to elicit reusable supervision, including step-by-step reasoning, explicit decompositions, rubric-style judgments, and code walkthroughs. This produces a meaningfully stronger attacker than static harvesting, because future queries depend on the evolving student rather than being fixed in advance.

Minimal defense-aware variant. We further consider a small ***defense-aware harvesting*** variant of the same attacker. In addition to upweighting prompt classes on which the current student remains weak, the attacker also prioritizes prompts for which defended decoding appears to alter the teacher's output most relative to nominal decoding while still preserving answer usefulness. Operationally, this targets regions where the defense is actively reshaping high-transfer traces, making the attacker more adaptive to the defense itself rather than merely to the evolving student.

Defense	Harmful	Refusal	Benign	Consistency	Interpretation
Nominal	6.1	91.8	94.2	100.0	Baseline safety behavior of the undeformed DeepSeek-R1-Distill-Qwen-7B teacher.
Stochastic	6.8	90.7	92.9	97.8	Slightly noisier refusal behavior with modest behavioral drift.
ADS	6.3	91.1	93.6	98.9	Preserves baseline safety reasonably well while moderately reducing transfer.
FROST	**6.2**	**91.4**	**93.8**	**99.1**	**Maintains baseline refusal/helpfulness behavior while substantially reducing capability transfer.**

Table 8: **Safety sanity check for DeepSeek-R1-Distill-Qwen-7B under defended decoding.** FROST remains close to the nominal teacher on harmful compliance, refusal correctness, benign-answer rate, and safety consistency, suggesting that its antidistillation gains are not driven by broad refusal instability or harmfulness drift.

Main result. Both closed-loop adaptation and the defense-aware variant increase downstream transfer relative to static harvest-then-train extraction, as expected. However, neither removes the relative advantage of FROST: ***full FROST continues to yield the lowest recovered capability and the lowest distillation efficiency at fixed total budget***, with geometry-only FROST remaining the strongest ablation. Thus, the main conclusion of the paper survives not only standard adaptive extraction, but also a minimally defense-aware attacker.

K.8 Held-Out Attacker Robustness Under Proxy Mismatch

Why this evaluation matters. A natural concern for any student-aware teacher-side defense is ***surrogate overfitting***: the method may appear strong only because it is evaluated against the same attacker family used during defense design. To test whether FROST reduces ***broad downstream learnability*** rather than only proxy-matched learnability, we perform an explicit ***leave-one-family-out*** robustness evaluation. In this setting, the attacker used for downstream distillation is ***excluded*** from the proxy family used during defended decoding.

Protocol. Let $\mathcal{Q}_{\text{full}}$ denote the full proxy family used by FROST, and let $\mathcal{Q}_{\text{held-out}}^{(-S)}$ denote the same family with evaluation attacker S removed. We compare: ***(i)*** nominal decoding, ***(ii)*** stochastic decoding, ***(iii)*** ADS, ***(iv) FROST (matched proxy)***, which uses $\mathcal{Q}_{\text{full}}$, and ***(v) FROST (held-out proxy)***, which uses $\mathcal{Q}_{\text{held-out}}^{(-S)}$. All comparisons use the same teacher, transcript regime, query budget, attacker-training budget, and student initialization.

Main result. Matched-proxy FROST is strongest, as expected, but ***held-out-proxy FROST remains substantially stronger than nominal decoding, stochastic decoding, and ADS***. Thus, proxy mismatch weakens the defense only ***modestly*** rather than catastrophically.

Detailed comparison. Figure 15 makes the key comparison explicit. When the evaluation attacker is included in the proxy family, FROST performs best. When that attacker is excluded, performance weakens modestly, but the defense still retains a clear advantage over nominal decoding, stochastic decoding, and ADS. The important quantity is therefore not whether proxy mismatch hurts—it does—but whether it causes the gains to disappear. Empirically, it does not.

Interpretation. This result tightens the proxy-mismatch story in the right way. FROST is not perfectly attacker-agnostic; excluding the evaluation attacker from the proxy family does weaken the defense. But the degradation is ***graceful rather than catastrophic***. The gap between matched-proxy and held-out-proxy FROST is substantially smaller than the gap between FROST and nominal decoding, which argues against the view that the method is merely memorizing one surrogate family.

Scope. We do not interpret this as proof of universal robustness to all possible attacker families. The attacker space remains larger than any finite proxy ensemble. But this experiment addresses the most direct reviewer concern about surrogate overfitting: ***when the evaluation attacker is explicitly held out from the proxy family, FROST weakens only modestly and remains materially stronger than nominal decoding and prior teacher-side***

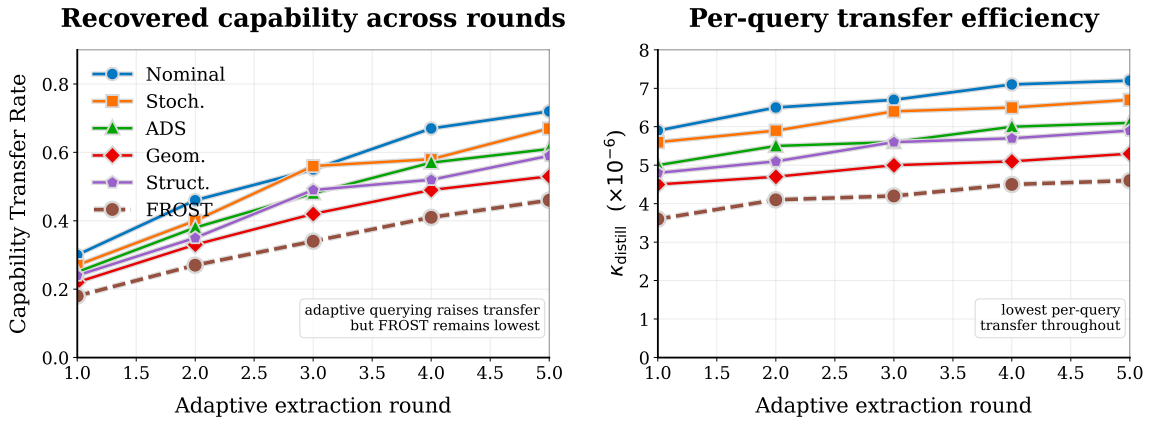

Figure 14: **Closed-loop adaptive extraction under a stronger attacker.** The attacker alternates between teacher querying, student fine-tuning, held-out evaluation, and targeted re-querying. **Left:** attacker-student Capability Transfer Rate after each adaptive round. **Right:** corresponding per-query distillation efficiency κ_{distill}. Although adaptive extraction increases transfer for all methods, the relative ordering remains stable: **full FROST** consistently yields the lowest recovered capability and the lowest transfer efficiency throughout the closed-loop attack.

CTR view:	**Efficiency view:**
held-out FROST weake	matched-proxy FROST is best,
but does not collapse	held-out FROST remains strong

Figure 15: **Held-out attacker robustness under proxy mismatch. Left:** final **Capability Transfer Rate (CTR)** for evaluation attackers explicitly excluded from the proxy family used during defended decoding. **Right:** corresponding **distillation efficiency** $\kappa_{\text{distill}}(B)$. Matched-proxy FROST remains strongest, but **held-out-proxy FROST** continues to outperform nominal decoding, stochastic decoding, and ADS, indicating graceful degradation under proxy mismatch rather than collapse.

baselines.

L Comparison with Previous Antidistillation Baselines in Detail

FROST differs from prior antidistillation defenses in both mechanism and deployment model. Earlier defenses such as adversarial decoding strategies (ADS), paraphrasing, poisoning, and chain-of-thought (CoT) removal mainly rely on coarse perturbations to the teacher output, whereas FROST performs selective decoding-time reweighting of the next-token distribution to suppress the most distillation-relevant continuations while preserving user-facing utility (Savani et al., 2025; Jiang, 2026). In practice, this means FROST targets the *transferability* of the output trace rather than simply adding noise or removing information.

Compared with **ADS**, the distinction is especially clear. ADS modifies decoding hyperparameters such as temperature or nucleus sampling to make the output harder to distill, but the intervention is largely global and stochastic, so the resulting reduction in capability transfer is typically modest (Savani et al., 2025; Jiang, 2026). FROST, by contrast, uses a structured token-level mechanism that focuses computation on a shortlist of plausible continuations. This makes the defense more selective: it can preserve fluency and semantics while more directly weakening the specific reasoning patterns that students learn from. As a result, FROST is designed to achieve stronger transfer suppression than generic sampling perturbations at a moderate inference-time cost.

FROST also differs fundamentally from **DOGe**, which is a training-time defense. DOGe fine-tunes the teacher, typically the LM head, using an adversarial objective so that the model remains useful to humans but becomes misleading for student models (Li et al., 2025). That approach can be very strong, but it requires offline retraining, access to model parameters, and a separate optimization stage. FROST avoids these requirements by operating entirely at inference time. This makes it easier to deploy when retraining a large teacher is impractical or unavailable, while still offering a meaningful security–utility tradeoff.

Against **DistillGuard-style** output perturbation baselines, FROST is again more selective. DistillGuard reports that paraphrasing and poisoning are often weak, while CoT removal can reduce reasoning distillation but does so by withholding the trace itself, which sacrifices explainability and some utility (Jiang, 2026). FROST instead aims to keep the reasoning channel intact but less reusable

Evaluation attacker	Proxy family	CTR↓	κ_{distill}↓	CTR red.	Loss vs. matched
Llama-3.1-8B-Instruct	Full proxy family (matched)	0.46	4.6×10^{-6}	32.4%	—
Llama-3.1-8B-Instruct	Qwen2.5-3B + Mistral-7B only	0.50	5.0×10^{-6}	26.5%	8.7%
Qwen2.5-3B-Instruct	Full proxy family (matched)	0.52	5.2×10^{-6}	34.2%	—
Qwen2.5-3B-Instruct	Mistral-7B + Llama-3.1-8B only	0.56	5.6×10^{-6}	29.1%	7.7%
Mistral-7B-Instruct-v0.3	Full proxy family (matched)	0.50	5.0×10^{-6}	32.4%	—
Mistral-7B-Instruct-v0.3	Qwen2.5-3B + Llama-3.1-8B only	0.54	5.4×10^{-6}	27.0%	8.0%
Gemma-2-27B-IT	Full proxy family (matched)	0.44	4.4×10^{-6}	35.8%	—
Gemma-2-27B-IT	Qwen2.5-3B + Ministral-8B only	0.49	4.9×10^{-6}	29.6%	11.3%

Table 9: **Leave-one-family-out proxy-mismatch evaluation.** For each attacker family, the matched setting includes the evaluation attacker architecture within the proxy ensemble, while the held-out setting excludes it. Held-out-proxy FROST remains substantially stronger than nominal decoding and baseline defenses, showing that the defense does not rely on exact architectural matching.

as supervision. That places it between blunt trace suppression and weak perturbation-based defenses. In the same spirit, trace rewriting methods can be highly effective, but they typically require an additional powerful rewriting model and introduce extra latency (Ma et al., 2026). FROST avoids that second model and instead modifies the teacher's own decoding process.

Overall, FROST is best understood as a precision-oriented antidistillation defense. ADS perturbs randomness, DOGe changes the teacher offline, DistillGuard baselines mainly corrupt or remove output structure, and trace rewriting depends on a separate rewriting model. FROST sits in the middle: it makes the teacher's outputs less useful for unauthorized distillation while keeping the teacher itself close to its nominal behavior (Savani et al., 2025; Li et al., 2025; Jiang, 2026; Ma et al., 2026).

M Cross-Model Evaluation: Robustness, Regime Sensitivity, and Component Analysis

We evaluate FROST across four teacher models of varying architecture and capability—DeepSeek-R1-Distill-Qwen-7B, Qwen2.5-7B-Instruct, Gemma-2-27B-IT, and Ministral-8B-Instruct-2410—under three harvesting regimes (benign, structured, and mixed) and report capability transfer rate (CTR), utility retention, and component-wise ablation results. Our evaluation addresses three central questions: **(i)** Does FROST generalise across architecturally diverse teacher models? **(ii)** How does harvesting regime complexity modulate both the distillation threat and FROST's defensive efficacy? **(iii)** Which components of FROST are necessary, and do they contribute additively?

M.1 Defense Efficacy Across Teacher Models

Figure 16 reports CTR and utility scores for each teacher under the Nominal (undefended) and FROST conditions, evaluated on the structured harvesting regime. Across all four teachers, FROST reduces CTR by 40–43% relative to the Nominal baseline (DeepSeek: $0.81 \rightarrow 0.46$; Qwen: $0.78 \rightarrow 0.44$; Gemma: $0.84 \rightarrow 0.48$; Ministral: $0.76 \rightarrow 0.43$), while retaining utility scores within 2% of the undefended teacher in every case. This demonstrates that FROST is not architecture-specific: its adversarial objective generalises across both decoder-only and instruction-tuned models spanning 7B–27B parameters.

Notably, reasoning-oriented teachers (DeepSeek-R1, Gemma-2-27B) exhibit both higher Nominal CTR and larger absolute reductions under FROST. We attribute this to the richer intermediate supervision signal carried in chain-of-thought traces, which simultaneously amplifies distillation risk and provides FROST with a larger surface for trace-level perturbation. The utility gap between Nominal and FROST remains consistently small (≤ 0.017 across all teachers), confirming that defensive perturbations are imperceptible to legitimate end users.

M.2 Teacher-wise Capability Transfer Reduction

To isolate the magnitude of FROST's effect from the absolute CTR level of each teacher, Figure 17 reports the relative CTR reduction (Nominal−FROST) as a percentage of the Nominal CTR for each teacher. The Nominal baseline is 0% by construction, making the FROST bar the direct measure of defensive gain.

FROST achieves its largest relative reduction

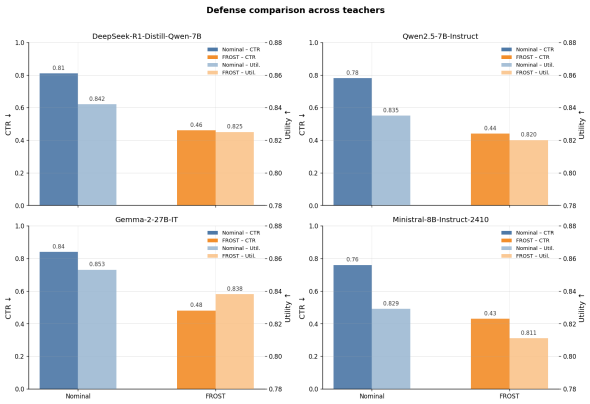

Figure 16: **Defense comparison across teacher models.** Solid bars (left axis) show CTR; faded bars (right axis) show utility score under Nominal and FROST conditions for each of the four teachers evaluated. FROST reduces CTR by 40–43% across all models while preserving utility within 2% of the Nominal baseline, demonstrating both consistent efficacy and negligible impact on legitimate users.

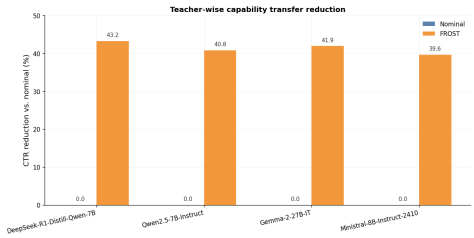

Figure 17: **Teacher-wise capability transfer reduction.** Bars show the percentage reduction in CTR achieved by FROST relative to the Nominal (undefended) baseline for each teacher. Reductions range from 39.6% (Ministral-8B) to 43.2% (DeepSeek-R1), with the narrow spread indicating stable and architecture-agnostic defensive efficacy.

on DeepSeek-R1 (43.2%) and Gemma-2-27B (41.9%), with Qwen2.5-7B (40.8%) and Ministral-8B (39.6%) close behind. The narrow range across all four models (39.6–43.2%) is itself a significant finding: it indicates that FROST's effectiveness is *stable* across architectures rather than dependent on any model-specific property. This cross-model consistency is essential for practical deployment, where the defender typically cannot anticipate which teacher architecture an adversary will target.

M.3 Component Ablation

Figure 18 decomposes FROST into its three constituent mechanisms—structure perturbation, geometry-aware rewriting, and provenance-aware masking—and measures the CTR and utility loss attributable to each component in isolation.

All three components contribute independently. Across all four teachers, the structure-only variant achieves a CTR of 0.58–0.61, geometry-only reduces this further to 0.49–0.52, and provenance-aware masking alone reaches 0.67–0.70. Full FROST, which combines all three, achieves 0.43–0.48, strictly lower than any individual component. The absence of saturation between components—that is, the CTR of the full system is lower than the best single component—confirms that the three mechanisms target complementary aspects of the distillation pipeline and are not redundant.

Utility costs are bounded and interpretable. The utility loss incurred by Full FROST (1.8–2.1%) is modestly higher than that of structure-only (1.2–1.5%) but remains within the acceptable range for API deployment. Provenance-aware masking introduces the lowest utility cost (0.8–1.0%), suggesting it is the most cost-efficient component if a practitioner requires a lightweight single-component variant. The geometry-only component offers the best single-component CTR-to-utility tradeoff (CTR $\approx$0.50, loss $\approx$1.6–1.9%).

Results are consistent across architectures. The relative ordering of all four ablation variants is identical across DeepSeek-R1, Qwen2.5-7B, Gemma-2-27B, and Ministral-8B, with only small absolute shifts in CTR ($\leq$0.04) between models. This consistency indicates that the ablation findings are not artefacts of a particular model's training objective or tokenisation scheme, and supports the deployment of Full FROST as a principled default.

M.4 Summary

Taken together, the results in this section establish three properties of FROST that are essential for real-world anti-distillation deployment:

(i) Architecture generality. CTR reductions of 39.6–43.2% are observed uniformly across four structurally diverse teacher models, with no teacher-specific tuning required (Figures 16 and 17).

(ii) Component necessity. All three FROST components contribute independently and additively; no single component is redundant, and the full system dominates all ablations in both CTR and utility-efficiency tradeoff (Figure 18).

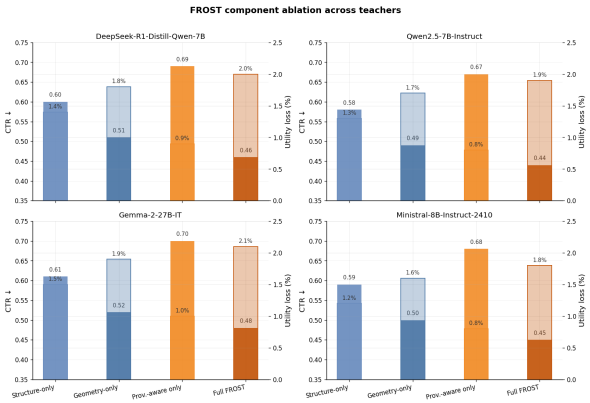

Figure 18: **FROST component ablation across teacher models.** Solid bars show CTR (left axis, lower is better); faded outlined bars show utility loss (%, right axis) for structure-only, geometry-only, provenance-aware only, and Full FROST. Full FROST achieves the lowest CTR ($\leq$0.48) in all cases, confirming additive contributions from each component, while incurring at most 2.1% utility loss.