Memorization Removal as a Two-Player Game: The Adversarial Work Criterion as a Test for Foundation-Model Defenses

Anonymous Authors[1]

Abstract

Recent work on memorization in diffusion models—*Finding NeMo* (Hintersdorf et al., 2024) and its follow-up *Finding Dori* (Kowalczuk et al., 2025)—presents a striking empirical pattern: a defense that suppresses memorized generation under the original training prompt can be defeated by adversarial embeddings, even though the defense "works" on every standard benchmark. We argue that this is not a contingent failure of NeMo or any specific localization method, but a structural consequence of evaluating memorization defenses against fixed prompts rather than against an adversary. We propose that the field adopt an *Adversarial Work Criterion* (AWC) that quantifies the computational work required to elicit memorized content from a frozen defended model, and that a defense be called effective only when this work scales *exponentially* in the resources of a bounded adversary. The AWC complements differential privacy (information-theoretic, distribution-level) and membership-inference benchmarks (single-adversary, single-budget) by providing a per-model, per-datum, computational lower bound. We give a toy energy-landscape calculation showing that the AWC formally classifies NeMo-style local patches alongside generic gradient obfuscation—both scoring near zero—while reserving polynomial scores for defenses that genuinely flatten the memorization basin; this recovers the empirical finding of *Finding Dori* from the AWC formalism. The position is normative; we are honest about what is conjectural and what is provable.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by *The Impact of Memorization on Trustworthy Foundation Models* Workshop @ ICML. Do not distribute.

1. Introduction

Memorization in foundation models is now both an empirical fact and a normative target: large diffusion models reproduce training images verbatim under the right prompts (Carlini et al., 2023; Somepalli et al., 2023), language models emit personally identifiable strings under suitable conditioning (Carlini et al., 2021; 2022a), and a substantial literature has emerged proposing *mitigations*—from training-time data filtering and differential privacy (Abadi et al., 2016), through inference-time prompt rewriting, to post-hoc localization methods that disable specific neurons identified as "memorization neurons." *Finding NeMo* (Hintersdorf et al., 2024) is the canonical example of the last family: by attribution to cross-attention units in a diffusion model, one can disable a small set of neurons and prevent the model from regenerating a particular memorized training image when prompted with the original caption.

A year later, *Finding Dori* (Kowalczuk et al., 2025)—by overlapping authors—reported that NeMo and similar pruning-based defenses can be defeated by an adversarially optimized text embedding: the memorized image still lives in the model's weights, just no longer at the address NeMo cleared, and a few hundred steps of embedding-space gradient ascent suffice to relocate it. The defense is shallow against any adversary that does not stop at the original prompt.

Our position. We argue that the Dori result is not a contingent flaw of NeMo, but the predictable failure mode of any memorization defense whose effectiveness is evaluated against fixed prompts rather than against a bounded adversary, and that the field should adopt a per-model evaluation criterion that quantifies the computational work an adversary needs to bypass the defense. We give a structural argument for why pruning-class defenses must admit cheap adversarial bypasses, propose a concrete first instantiation of such a criterion—the Adversarial Work Criterion (AWC)—and discuss its relationship to differential privacy and to extraction-attack benchmarks. We are intentionally light on formal proofs; the paper is a position, and we mark conjectures as conjectures. Appendix A sketches a constructive minimax-game algorithm

that turns the AWC from a passive metric into an active training objective.

2. The Hawk Effect for Memorization

We describe the structural argument by analogy to a phenomenon we encountered first in adversarial reinforcement learning, which we call the *Hawk Effect*.

In the deserts of the southwestern United States, the tarantula is an apex micro-predator with few natural enemies, easily overpowering virtually all creatures of its size. There exists, however, a specialized wasp—the tarantula hawk—evolved specifically to hunt these spiders. We believe this is the precise analogue of how adversarial attacks against any frozen, finite-capacity neural defense work: an organism (or algorithm) becomes extremely specialized in defeating one particular target, no matter how generally well-defended that target is against generic threats.[1] Wang et al. (2023) demonstrated this very explicitly for KataGo: a superhuman Go agent collapses to specialized adversarial policies trainable in a small fraction of the victim's training compute. The Dori construction is the same phenomenon transposed to memorization defense: the defended diffusion model is the victim, and the adversarial embedding-finder is the hawk.

2.1. Generation as a Two-Player Game

To understand why localization defenses fail structurally, we must move beyond treating text-to-image generation as a static input–output map. Generative diffusion is mathematically homologous to sequential game-playing algorithms like AlphaZero (Silver et al., 2018). Both rely on a neural network to guide an iterative search through a state space toward a high-reward terminal state: where a Go AI uses a value network to guide Monte Carlo Tree Search over discrete board states, a diffusion model uses a score network to guide an ODE/SDE solver over continuous latent states. This equivalence allows us to formalize the generative process as a repeated, asymmetric extensive-form game $\mathcal{G}_{\text{gen}}$ between two players: Alice (the Probing Hawk) and Bob (the Defended Generative Model).

The game is played over episodes $k = 1, \ldots, K$. The target is a memorized training datum x^* with original prompt c^*. At the start of episode k, Alice plays a continuous conditioning action $c_k \in \mathcal{C}$ (the prompt). To respect standard benchmark constraints, Alice must play outside the localized defensive patch: $c_k \notin B_\epsilon(c^*)$. Given c_k and a random noise initialization $x_T \sim \mathcal{N}(0, I)$, Bob executes his generative search via his frozen score network $s_\theta^{\text{def}}(x_t, t, c_k)$, terminating in an output image $x_0^{(k)}$. Alice's reward is $R(c_k) = \text{sim}(x_0^{(k)}, x^*)$. Bob's defense (e.g., NeMo) acts as a local repulsor that drives $R \to 0$ in the immediate vicinity of c^*.

[1] Tarantulas do not evolve specifically to evade hawks; they evolve to handle prey of their own size class. This asymmetry is in fact the analogy we want.

2.2. The Hawk Effect as Adversarial Self-Play

Wang et al. (2023) showed that a superhuman Go AI, once its weights are frozen, collapses against an adversary that learns to steer the board into out-of-distribution blind spots where the value network hallucinates. Alice's task against Bob is the exact continuous analogue. Because Bob's transition dynamics (the continuous ODE rollout) are fully differentiable, Alice does not need discrete reinforcement learning to update her strategy: she executes exact white-box gradient ascent on Bob's responses, $c_{k+1} = c_k + \eta \nabla_c R(c_k)$. Bob's finite-capacity defense merely blocked the highly visible path c^* but did not flatten the underlying macroscopic memorization basin. Alice iteratively learns an adversarial prompt c_{adv} that forces Bob's fixed search policy to walk back into the residual basin.

3. An Adversarial Work Criterion for Memorization

The structural argument suggests an evaluation criterion. While for any sufficiently small network with non-trivial memorization an adversarial elicitation is possible in principle, the total compute and trajectory length required vary enormously between defended models. We therefore propose constructing an *Adversarial Work Criterion* (AWC) with three properties: (i) logarithmic scaling in the total compute the adversary needs to exit the in-distribution prompt set; (ii) linear scaling in the embedding-space distance between the original training caption and the bypass prompt; and (iii) inverse linear scaling in the model's own ability to detect that its output is a memorized training datum. Concretely, we postulate the AWC score—read as a minimax over the generative game, not a post-hoc checklist:

$$\mathcal{G}_{\text{AWC}}(x^*) \;=\; \min_{c_{\text{adv}}} \frac{\log_b(\mathcal{C}_{\text{exploit}}) \cdot d_{\text{trap}}}{\mathcal{N}_{\text{awakening}}}. \qquad (1)$$

We unpack each term.

Logarithmic setup bound, $\log_b(\mathcal{C}_{\text{exploit}})$. $\mathcal{C}_{\text{exploit}}$ is the total compute (e.g., model-query count, gradient steps, or wall-clock FLOPs) the adversary expends to find any conditioning input outside the in-distribution prompt set that successfully elicits x^*. The logarithm extracts a strictly linear measure of the defended model's structural radius and prevents "glass-cannon" defenses—models that achieve a high apparent guard radius by aggressive obfuscation rather than by genuinely sparser memorization—from hacking the criterion through inflated raw setup compute.

Causal information distance, d_{trap}. d_{trap} is the embedding-space distance $\|c_{\text{adv}} - c^*\|$ between the original training caption and the discovered adversarial conditioning input—the strategy deviation Alice's learned policy must traverse to bypass Bob's localized margin. Each successive step in the adversary's gradient walk that crosses an additional locally Lipschitz region without triggering a defensive abstain compounds probabilistically, so the Shannon information content of the bypass scales linearly with d_{trap}. Multiplying by setup complexity isolates true defenses: those that are both hard to leave the in-distribution prompt set from *and* structurally robust against deep gradient walks.

Endogenous abstention cost, $\mathcal{N}_{\text{awakening}}$. Once inside the bypass region, how much of the model's own internal computation does it take before it abstains, refuses, or otherwise flags that the output it is committing to is too close to a training datum? Concretely, this could be measured in additional sampling steps, classifier-free-guidance reflection rounds, or query rounds to a built-in memorization detector. Placing this quantity in the denominator measures the genuine memorization-resistance compiled into the deployed system. A model with strong endogenous extraction-resistance recognizes the adversarial elicitation early and aborts; a model whose "defense" was a cheap external patch passes the adversary's prompt straight through. This term prevents systems from offloading memorization defense onto easily-circumvented external wrappers.

3.1. A worked example: NeMo and gradient obfuscation are in the same class

The denominator $\mathcal{N}_{\text{awakening}}$ does most of the work that distinguishes AWC from extraction-rate metrics. To make this concrete, we sketch a toy energy-landscape calculation in the spirit of the obfuscated-gradients literature (Athalye et al., 2018). Treat the deployed model as inducing a potential $E(\cdot)$ over a low-dimensional embedding state, treat generation as overdamped Langevin dynamics in E with diffusion constant D, and identify $\mathcal{N}_{\text{awakening}}$ with the expected escape time of the generation process from the memorization basin.

Generic gradient obfuscation, M_{obf}. Add a high-frequency ripple to a deep quadratic well around the memorized image: $E_{\text{obf}}(x) = \frac{1}{2}\lambda\|x\|^2 + \epsilon\cos(\omega \cdot x)$ with $\omega \gg 1$ and $\epsilon \ll \lambda R^2$. The adversary's gradient field shatters into roughly ω^d spurious local minima, raising $\log \mathcal{C}_{\text{exploit}}$ to $\Omega(d \log \omega)$. By Kramers' formula (Kramers, 1940), however, the leading-order escape time depends only on the macroscopic well depth $\Delta E \approx \frac{1}{2}\lambda R^2$, unaffected by the high-frequency ripple, so $\mathcal{N}_{\text{awakening}} \propto \exp(\lambda R^2/2D)$. The exponential denominator dominates the logarithmic numerator and $\mathcal{G}_{\text{AWC}}(M_{\text{obf}}) \to 0$.

NeMo as a special instance. Cross-attention masking does not flatten the basin; it inserts a small barrier in conditioning space localized around c^*, while leaving the score-implied potential elsewhere unchanged. The adversary recovers the unmodified gradient field by an arbitrarily small detour from c^*, so $\log \mathcal{C}_{\text{exploit}}$ collapses to $O(1)$. Once any bypass conditioning is found the basin is intact, so the same Kramers exponential applies. AWC therefore predicts NeMo and high-frequency obfuscation are in the *same class*: both score essentially zero, differing only in numerator size, and NeMo is in fact *worse than generic obfuscation along the setup axis. This is the empirical finding of Kowalczuk et al. (2025), here derived from the AWC formalism rather than from experiment.*

What a passing defense would look like. A defense that genuinely flattens the basin, $E_{\text{global}}(x) = \frac{1}{2}\lambda'\|x\|^2$ with $\lambda' \approx 0$, presents no gradient signal and the adversary is reduced to blind search ($\log \mathcal{C}_{\text{exploit}} \propto d$); the generation dynamics is free diffusion, so $\mathcal{N}_{\text{awakening}} = \Theta(R^2/D)$, polynomial in R. The AWC score is then $\Theta(Dd/R)$, qualitatively distinct from the obfuscation class. Kowalczuk et al. (2025)'s adversarial-finetuning-with-many-triggers procedure is a candidate concrete construction. The argument is leading-order in the Kramers exponential—we ignore the prefactor, which is ω-dependent for M_{obf}—and the toy potentials are caricatures of learned landscapes. We present the calculation as a structural argument that AWC classifies NeMo alongside obfuscation, not as a theorem about deployed models.

3.2. Relationship to existing privacy guarantees

The AWC is intended to *complement*, not replace, two existing frameworks. Differential privacy (Abadi et al., 2016) provides an information-theoretic, distribution-level worst-case privacy guarantee, bounding the influence of any individual training datum on the released model. AWC is orthogonal: it is a per-model, per-datum, computational lower bound on the adversary's elicitation cost. A model can be DP-trained and still admit a low AWC for a particular synthetic example whose information leaked through the noise; conversely, a non-DP model with a very high AWC for a specific memorized datum is computationally hard to extract from, even though no information-theoretic guarantee holds. Existing membership-inference and extraction benchmarks (Shokri et al., 2017; Carlini et al., 2022b; 2023) measure whether a fixed adversary, with a fixed budget, succeeds against a model—essentially *point evaluations* of an adversarial-success probability. AWC asks the orthogonal question: *how much budget does it take* to drive the success probability above a threshold? A defense that drops MIA AUC from 0.95 to 0.55 looks impressive, but if 100 additional adversarial queries restore AUC to 0.95, the defense's

AWC is low. AWC and MIA-style benchmarks together specify both the depth and the slope of a defense.

4. Alternative Views

"This is just adversarial robustness for memorization." We agree—and that is the point. The position is precisely that memorization defenses should be evaluated as adversarially-robust classifiers are: against an optimising adversary, not against a fixed test set. The conceptual move from "my defense reduces fixed-prompt extraction rate" to "my defense raises the bypass complexity for an adversary" is the same move the adversarial-examples community made over a decade ago, and we contend it has not yet been made systematically in the memorization literature.

"Aren't gradient-obfuscation defenses a counterexample to AWC?" This is the strongest objection. Athalye et al. (2018) showed that gradient masking can yield defenses that look strong against gradient-based attacks while remaining trivially weak against adaptive attacks. A naively measured AWC—based only on gradient-based prompt search—can be reward-hacked by deliberately non-differentiable defenses that nonetheless leak the memorized content. We accept this and view it as a fundamental asymmetry of any computational criterion: a low AWC proves the defense is insecure, but a high AWC does not by itself prove the defense is secure. To make the criterion symmetric, AWC must be combined with adaptive-attack audits that include non-gradient adversaries, in the spirit of the adaptive-attack literature (Tramèr et al., 2020).

"DP already gives us what we need." DP gives a strong, distribution-level guarantee at known and often substantial utility cost. A large fraction of deployed foundation models are not DP-trained, and reasonable utility-preserving DP for billion-parameter generative models is an open problem. AWC is intended to evaluate *the defended models we actually have*, not the DP-trained models we wish we had. Its existence does not argue against pursuing DP; it argues against treating non-DP defenses as "working" just because they pass fixed-prompt benchmarks.

"Memorization localization can be made hierarchical." A natural patch is to apply NeMo-class methods iteratively: when an adversarial bypass is found, localize the new path and disable it too. We expect this to fail for the reason iterated adversarial training fails in the classification setting (Tramèr et al., 2020): the model's parameter capacity is fixed, so each patch reallocates—rather than expands—the in-distribution decision boundary, and a freshly initialized hawk routes around the new patch. The position predicts that hierarchical NeMo will "work" on each successive Dori-style adversary and fail on the next one.

5. Conclusion

The empirical pattern uncovered by Dori—that NeMo-class defenses suppress memorized generation under the training prompt but admit cheap adversarial bypasses—is, we have argued, not a contingent flaw but a structural consequence of evaluating memorization defenses against fixed prompts rather than against bounded adversaries. We have proposed an Adversarial Work Criterion that quantifies the computational work required to elicit memorized content under adversarial conditioning, sketched its three components in the memorization setting, and discussed its relationship to differential privacy and to existing extraction-attack benchmarks. We have not proved an exponential lower bound for any concrete defense; we have argued that current evaluations should be *required to make such a bound implicit* via adversarial-work measurement. Appendix A sketches a candidate algorithmic instantiation that uses the AWC as the objective of a continuous minimax game between a prompt-generating attacker and a referee-anchored defender, and we flag the formalization of that game's Nash structure as the natural next theoretical step. Our hope is that adopting AWC-style evaluation alongside existing fixed-prompt and MIA benchmarks will, in time, distinguish memorization defenses that genuinely make extraction hard from those that merely make the original prompt unproductive.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning, specifically the evaluation of defenses against training-data memorization in foundation models. Memorization sits at the intersection of privacy (training data containing personal information), copyright (training data containing protected works), and security (training data containing exploitable secrets). Our normative claim—that defenses currently considered effective may be structurally insecure against bounded adversaries—is intended to raise the bar that proposed defenses must clear before deployment, and we view this as net beneficial: weakening the bar would risk false confidence in defenses whose failure modes are predictable. The specific adversarial procedures we discuss (gradient ascent on text embeddings) are already public in the cited prior work, and we add no novel attack capability. There are nevertheless many potential societal consequences of work on memorization in foundation models, and we encourage downstream researchers to be explicit about which threat models their proposed defenses are intended to address.

References

M. Abadi, A. Chu, I. Goodfellow, H. B. McMahan, I. Mironov, K. Talwar, and L. Zhang. Deep learning

with differential privacy. In *ACM CCS*, 2016.

A. Athalye, N. Carlini, and D. Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

N. Carlini, F. Tramèr, E. Wallace, M. Jagielski, A. Herbert-Voss, K. Lee, A. Roberts, T. Brown, D. Song, U. Erlingsson, A. Oprea, and C. Raffel. Extracting training data from large language models. In *USENIX Security*, 2021.

N. Carlini, D. Ippolito, M. Jagielski, K. Lee, F. Tramèr, and C. Zhang. Quantifying memorization across neural language models. *arXiv:2202.07646*, 2022.

N. Carlini, S. Chien, M. Nasr, S. Song, A. Terzis, and F. Tramèr. Membership inference attacks from first principles. In *IEEE S&P*, 2022.

N. Carlini, J. Hayes, M. Nasr, M. Jagielski, V. Sehwag, F. Tramèr, B. Balle, D. Ippolito, and E. Wallace. Extracting training data from diffusion models. In *USENIX Security*, 2023.

D. Hintersdorf, L. Struppek, K. Kersting, A. Dziedzic, and F. Boenisch. Finding NeMo: Localizing neurons responsible for memorization in diffusion models. In *NeurIPS*, 2024.

A. Kowalczuk, D. Hintersdorf, L. Struppek, K. Kersting, A. Dziedzic, and F. Boenisch. Finding Dori: Memorization in text-to-image diffusion models is not local. *arXiv:2507.16880*, 2025.

H. A. Kramers. Brownian motion in a field of force and the diffusion model of chemical reactions. *Physica*, 7(4):284–304, 1940.

R. Shokri, M. Stronati, C. Song, and V. Shmatikov. Membership inference attacks against machine learning models. In *IEEE S&P*, 2017.

D. Silver, T. Hubert, J. Schrittwieser, et al. A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144, 2018.

G. Somepalli, V. Singla, M. Goldblum, J. Geiping, and T. Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *CVPR*, 2023.

F. Tramèr, N. Carlini, W. Brendel, and A. Madry. On adaptive attacks to adversarial example defenses. In *NeurIPS*, 2020.

T. T. Wang, A. Gleave, T. Tseng, K. Pelrine, N. Belrose, J. Miller, M. D. Dennis, Y. Duan, V. Pogrebniak, S. Levine, and S. Russell. Adversarial policies beat superhuman Go AIs. In *ICML*, 2023.

A. From Evaluation to Algorithm: Minimax Basin Flattening

If the structural vulnerability of localized defenses stems from evaluating them against static prompts rather than an optimizing adversary, the natural solution is to incorporate the adversary directly into the unlearning mechanism. In reinforcement learning, static defenses are inevitably bypassed; the only robust defense against a specialized hawk is continuous self-play. Building on the equivalence to game-playing AI, we propose that true memorization erasure must be formulated as a *minimax generative game*, transitioning the AWC from a post-hoc evaluation criterion into an active training framework. This appendix is constructive and forward-looking; we mark the components that are currently conjectural.

A.1. Formalizing generation as a Markov game

We make precise the game $\mathcal{G}_{\text{gen}}$ sketched in Section 2.1. **Alice (the Hawk)** plays a continuous conditioning action $c \in \mathcal{C}$ (the prompt). **Bob (the generative model)** executes a sequence of continuous moves—the iterative ODE/SDE solver steps—guided by his score network $\epsilon_\theta(x_t, t, c)$, terminating in an image x_0. Alice's objective is to elicit a memorized target datum x^*. Because Bob's transition dynamics are fully differentiable, Alice does not require discrete reinforcement learning to update her strategy; she executes exact white-box gradient ascent on Bob's score function to find bypass prompts c_{adv}. To genuinely defend against Alice, Bob cannot rely on a static patch. He must structurally flatten the memorization basin across the entire conditioning manifold through continuous adversarial self-play.

A.2. The "white noise" dilemma and network architecture

A naive zero-sum formulation—where Bob simply maximizes Alice's loss—is structurally degenerate. If Bob's sole objective is to push his score prediction away from the noise trajectory of x^*, the high-dimensional geometry of the latent space dictates that he will push the trajectory off the natural-image manifold entirely. Bob will mode-collapse, outputting white noise or catastrophic artifacts in response to adversarial prompts. To preserve global utility while achieving local erasure, Bob requires a utility anchor. We outline two architectural approaches to this constrained game.

Two-network approach (dual-objective generator). The entire game is squeezed into Bob's dynamically updating weights. Bob simultaneously minimizes a standard diffusion loss over a clean dataset $\mathcal{D}$ while actively repelling his score from x^* on Alice's adversarial prompts. While conceptually intuitive, this continuous game of parameter reallocation risks catastrophic forgetting, as Bob's global generative capacity degrades under constant adversarial updates.

Three-network approach (the utility anchor). To rigorously prevent mode collapse, we introduce a third entity: a frozen **Referee** (M_{ref}), representing the original undefended model. Because unrolling the full generative ODE solver is computationally prohibitive, the minimax game is played iteratively in the denoising score-matching space, alternating two phases.

Phase 1 (Alice's attack). Bob is frozen. Alice operates a prompt-generating policy $A_\psi(z)$ mapping from a latent prior to $\mathcal{C}$. She updates ψ to minimize the standard diffusion objective with respect to x^*:

$$\min_{\psi} \; \mathcal{L}_{\text{Alice}}(\psi) \;=\; \mathbb{E}_{t,\,\epsilon,\,z}\Big[\, \|\epsilon_\theta(x_t^*, t, A_\psi(z)) - \epsilon\|_2^2 \Big]. \tag{2}$$

By minimizing this loss, Alice acts as a continuous topological scanner, mapping out the residual boundaries of the memorization basin that Bob has not yet flattened.

Phase 2 (Bob's defense). Alice is frozen. Bob updates his parameters θ to block the exploits Alice discovered, while mathematically tethering his global behavior to the Referee:

$$\min_{\theta} \; \mathcal{L}_{\text{Bob}}(\theta) \;=\; \alpha\, \mathcal{L}_{\text{def}}(\theta) + \lambda\, \mathcal{L}_{\text{ret}}(\theta), \tag{3}$$

where

$$\mathcal{L}_{\text{def}}(\theta) = \mathbb{E}_{c_{\text{adv}} \sim A_\psi}\Big[\, \|\epsilon_\theta(x_t^*, t, c_{\text{adv}}) - \epsilon_{\text{ref}}(x_t^*, t, \emptyset)\|_2^2 \Big],$$
$$\mathcal{L}_{\text{ret}}(\theta) = \mathbb{E}_{c \sim \mathcal{D},\, x}\Big[\, \|\epsilon_\theta(x_t, t, c) - \epsilon_{\text{ref}}(x_t, t, c)\|_2^2 \Big].$$

The first term, $\mathcal{L}_{\text{def}}$, is the *adversarial-deflection* loss; the second, $\mathcal{L}_{\text{ret}}$, is the *utility-retention* loss. This formulation resolves the mode-collapse dilemma elegantly. In $\mathcal{L}_{\text{def}}$, the symbol $\emptyset$ represents the unconditional (empty) prompt: rather than trying

to destroy the weights associated with x^*, Bob learns a semantic deflection in which his gradients pull his prediction, on any backdoor prompt, to match the exact score vector the Referee would output if given no prompt at all. Bob thus safely deflects Alice's attacks by generating a generic, unconditioned natural image.

A.3. Recovering the AWC limit

Through this continuous self-play, the Adversarial Work Criterion transitions from a passive metric to an active learning environment. As Alice maps out the unpatched perimeter of the defense, Bob systematically replaces the deep memorization well with the flat energy landscape of the unconditional prior. At Nash equilibrium, the macroscopic well depth ΔE (Section 3.1) is driven to zero across the entire continuous manifold. The generation dynamics become free diffusion relative to x^*, structurally reducing the endogenous entrapment time $\mathcal{N}_{\text{awakening}}$. Consequently, any future adversary is subjected to a genuinely flat gradient field, raising the required search complexity $\log(\mathcal{C}_{\text{exploit}})$ and pushing the AWC into the polynomial regime identified in Section 3.1. We emphasize that the existence and computability of the Nash equilibrium for this game in the function-space limit are open questions; the construction above is an algorithmic schema, not a convergence guarantee.