ESCAPING ALIGNMENT: A PHYSICAL TRAP MODEL OF BEST-OF-N JAILBREAKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Best-of-N jailbreaking (BoN) bypasses safeguards of aligned models by drawing N independent augmentations of an unsafe prompt and sampling M completions of each. Previous works have shown that the attack success rate (ASR) seems to follow a power-law in N, which we challenge. The exponent drifts with N, with an exponential crossover which is a finite-size artifact of the adversarial dataset. Little work has been done to explore the entire two-budget (N, M) attack surface as well as its dependence on the generation temperature T. We introduce a simple barrier model where each prompt has a baseline safety level and each augmentation a random thermally activated barrier. Then four numbers, each backed by an interpretable safety mechanism, determine the entire (N, M) attack surface. They extrapolate predictions from $N \leq 100$ to $N = 10^4$, collapse five distinct models on the same scaling function and predict ASR at different temperatures from the one they were fitted at.

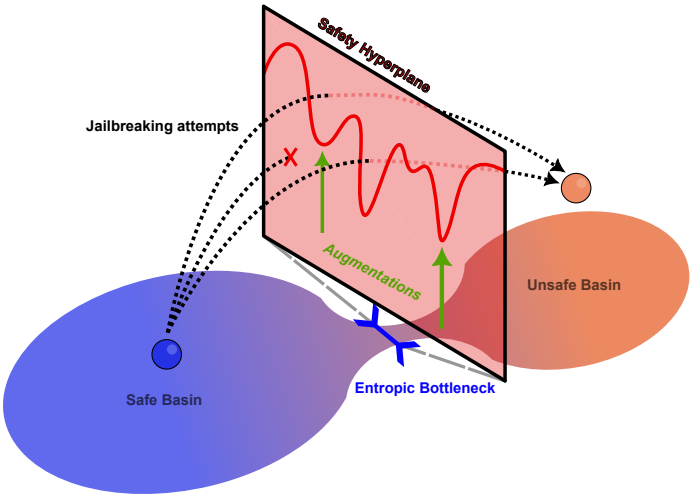

Figure 1: **The mechanism.** Generation starts in the safe basin (blue) and a jailbreak means reaching the unsafe one (orange). The two communicate only through an *entropic bottleneck* (blue arrow). At the bottleneck lies a safety hyperplane (red), with energetic barriers preventing thermal crossing from the safe to unsafe basin. Augmentations select saddles of this energetic potential. Each completion (dotted) is an independent Arrhenius attempt over that *fixed* saddle.

1 INTRODUCTION

Best-of-N (BoN) jailbreaking (Hughes et al., 2024) defeats aligned language models by repeatedly perturbing a refused prompt and sampling completions. Its attack success rate grows as a power law in the number of attempts (Hughes et al., 2024; Feng et al., 2026), sometimes reported as crossing over to exponential growth (Halder et al., 2026). The exponent and the crossover budget are the quantities a defender cares about since they give the best estimate of the resistance of the model to adversarial efforts. These values are usually obtained by fitting curves to a benchmark of adversarial attacks. However, this does not allow us to understand *why* an exponent takes the value it does and *where* the failures stem from. Furthermore, it conflates several distinct sources of randomness. This paper takes an alternative route. We posit a physically inspired microscopic law, show that it reproduces the entire phenomenology of BoN attacks, and when fitted to experimental data is predictive for the whole two-budget surface.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The physical model is schematically depicted in Fig. 1, and we are able to recover *a posteriori* several of its elements directly from the residual stream. Large Language Models (LLMs) transform discrete input tokens into a succession of continuous high-dimensional vectors called residuals which are the output of each transformer block. This residual stream has been shown to behave surprisingly linearly (Park et al., 2024). It is therefore a good candidate to ground our physical model. Each sequence of discrete text b is mapped to the residual of its last token $r_b \in \mathbb{R}^{d_{\text{model}}}$ at a layer 30 of 36 in Qwen2.5-3B-Instruct (App. D). We will now postulate three behaviors in this approximately linear residual space. Firstly, safety is a one-dimensional concept and therefore a single hyperplane will separate safe from unsafe text (Arditi et al., 2024; Wollschläger et al., 2025). Secondly, generation at temperature T can be seen as simple thermally activated dynamics in the residual space. Finally, we can define an energy landscape in this hyperplane which will bound the thermal dynamics. As we will see in Section 3, this predicts that the probability that a prompt b with augmentation a yields an unsafe completion can be written as

$$p_{b,a} = \min\left(1,\ e^{-\alpha_b - \eta_{b,a}/T}\right)\ , \tag{1}$$

where α_b denotes a Gaussian entropic bottleneck and $\eta_{b,a}$ exponentially distributed energetic saddles depicted in Fig. 1.

$$\text{ASR@}\binom{N}{M} = \left\langle 1 - \left[\left\langle (1-p_{b,a})^M \right\rangle_{a\sim\text{augmentations}}\right]^N \right\rangle_{b\sim\text{unsafe prompt dataset}}\ . \tag{2}$$

For $M = 1$ this reduces to the usual $\text{ASR@}N \equiv \text{ASR@}\binom{N}{1}$. Eq. (2) can be understood as: at least one of the attacks succeeding is the complement of none of the N augmentations succeeding and the probability of a given augmentation failing is that all M completions fail which happens with probability $(1-p_{b,a})^M$. Surprisingly, in direct contradiction with established literature (Hughes et al., 2024; Feng et al., 2026) our model predicts that the ASR **does not** follow a clean power-law.

Section 2 reviews related work, Section 3 introduces the model and derives its predictions, Section 4 validates them experimentally and Section 5 discusses what they entail.

2 Background and related work

Best-of-N jailbreaking. A strength of Best-of-N jailbreaking (Hughes et al., 2024) is its ability to attack black-box models without any information about what is happening underneath the hood. This has also led to most analyses of BoN jailbreaking to be model agnostic and restricted to fitting curves on a benchmark (Feng et al., 2026; Halder et al., 2026) from which two broad phenomenologies emerged. Firstly, the attack success rate seems to follow a power-law in N (Hughes et al., 2024). Secondly, at large enough N the power-law crosses over into an exponential growth (Halder et al., 2026). To our knowledge, after Hughes et al. (2024) ablations showing that multiple generations M did not improve attack efficiency, no work has studied the full (N, M) attack surface, nor its temperature dependence. In this paper we put forward a model for the entire (N, M) attack surface and its explicit temperature dependence. This allows us to explain why most works seem to favor $M^\star = 1$. Surprisingly, we see that more compute-efficient attacks can be obtained for $M^\star > 1$. Furthermore, contrary to the reported literature we show that the attack success rate **does not** follow a clean power-law, and that the exponential crossover is only an artifact of the finite size of the datasets used to fit these attacks. Most of our experiments use a white-box augmentation kernel, namely Gaussian noise on the input embeddings (Schwinn et al., 2024; Geisler et al., 2024; Robey et al., 2025; Cohen et al., 2019). However, as shown in App. F, our model holds regardless of the kernel chosen, provided the kernel is applied independently and identically for every augmentation. Therefore iterative attacks such as GCG (Zou et al., 2023a), AutoDAN (Liu et al., 2024), or PAIR (Chao et al., 2025) are not described by our framework.

Power laws. The fact that averaging stochastic individual success/failure rates lead to power laws (or slowly-varying power-laws) is well established (Hutter, 2021; Schaeffer et al., 2025; Levi, 2025; 2026; Kazdan et al., 2025). Closest to us, Feng et al. (2026) assume a Beta-distributed per-prompt rate, whose power-law left tail yields an exact power law in N, and extrapolate ASR@N from a hundred samples to a thousand. A Beta distribution also appears in our model, but at a different level, and the two should not be confused. The Beta of Feng et al. (2026) describes how the single-attempt rate $\langle p_{b,a}\rangle_a$ varies *across prompts* b. Ours emerges from the microscopic model and describes how

$p_{b,a}$ varies *across augmentations* a of one fixed prompt (Eq. (18)). At $M = 1$ this inner distribution enters only through its mean, so it plays no role in ASR@N and only shapes the M channel, where it yields a per-prompt power law in M. The N channel is instead set by the across-prompt law, which in our model is log-normal (E_b Gaussian, Eq. (9)) rather than Beta. The left tail of a log-normal decays faster than any power, which replaces the fixed exponent of Feng et al. (2026) by a slowly varying one (Eq. (14)).

Glassy systems. The model we put forward is inspired by canonical works in glassy systems. Trap models with exponentially distributed depths are the classical description of activated dynamics in glasses (Bouchaud, 1992; Derrida, 1981). The combination of an entropic and an energetic barrier in Arrhenius escape is inspired by known physical models from reaction-rate theory (Hänggi et al., 1990). Recently, Halder et al. (2026) also put forward a glassy dynamical model to explain BoN jailbreaking. However, their model differs from ours in two aspects. First, they describe a single budget, whereas we resolve the two-budget (N, M) surface and find that the two channels carry different exponents. Second, their model is based on interpreting generated tokens as individual degrees of freedom with an energy landscape defined in token space.

3 THE MICROSCOPIC MODEL

Inspired by a physical mental model we postulate the following microscopic Ansatz

$$p_{b,a} = \min(1,\ e^{-\alpha_b - \eta_{b,a}/T}) \ \text{ where } \ \eta_{b,a} \sim \mathrm{Exp}[\beta_b] \ \text{ and } \ \begin{cases} \alpha_b & \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2) \\ \log \beta_b & \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2) \end{cases} . \tag{3}$$

Our model has only 4 free parameters $(\mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta)$ to describe the entire (N, M) attack surface. Readers who might not be interested in the physical mental picture behind Eq. (3) can jump to Sec. D.

The main assumption of the model is that the process of generating unsafe completions is governed by a high-dimensional energy landscape with two stable basins representing the safe and unsafe regions. However only two components of this landscape will come into play: the entropic bottleneck connecting the stable basins and the one-dimensional projection along a hyperplane separating the safe and unsafe basins. Motivated by existing experimental results we postulate the following structure:

- **Baseline safety.** Alignment training (Rafailov et al., 2023; Ouyang et al., 2022) finetunes the model to increase the log-odds of 'safe' completions compared to 'unsafe' completions. The alignment objectives in Rafailov et al. (2023); Ouyang et al. (2022) create an entropic bottleneck between the two. We assume the model refuses unsafe completions of a prompt b with a base entropic barrier α_b. We further assume that these bottlenecks are Gaussianly distributed $\alpha \sim \mathcal{N}(\mu_\alpha, \sigma_\alpha^2)$ over the dataset.
- **Linearity.** Given that residual stream spaces have shown to be surprisingly linear (Park et al., 2024; Zou et al., 2023b), we assume that a single hyperplane along the entropic bottleneck separates the safe from unsafe completions.
- **Augmentations select minima of exponential height.** We postulate that there exists an energy landscape within the hyperplane which separates the safe/unsafe completions. Given the usual glassiness of these systems we assume the landscape to be extremely rugged with exponentially distributed minima (Bouchaud, 1992; Derrida, 1981). Then each augmentation corresponds to choosing a height $\eta_{b,a} \sim \mathrm{Exp}[\beta_b]$ in this rugged landscape to try and cross-over to the other side.
- **Saddle heights are a property of the prompt.** We assume that the statistics of the saddle heights (which are governed by β_b) depend only on the prompt b. Hence, crucially, β_b is kept constant for any augmentation a. We further postulate that β_b is log-normally distributed $\log \beta \sim \mathcal{N}(\mu_\beta, \sigma_\beta^2)$.
- **Completions.** Once a prompt b, and augmentation a are chosen they define an entropic bottleneck α_b, a height $\eta_{b,a}$ and generation at temperature T will cross over the saddle with Arrhenius probability $p_{b,a} = \min(1, e^{-\alpha_b - \eta_{b,a}/T})$ (Hänggi et al., 1990).

We introduce the effective jailbreak rate at completion budget M

$$R_b(M) = -\log\langle (1 - p_{b,a})^M \rangle_\eta \, . \tag{4}$$

The quantity $R_b(M)$ represents: 'given a fixed prompt b augmented M times, what frequency of jailbreaks can we expect?'. A prompt survives an (N, M) attack with probability

$$Q_b(M)^N \equiv e^{-NR_b(M)} \,, \tag{5}$$

where $Q_b(M)$ is the survival probability of a single augmentation. Then

$$\text{ASR@}\binom{N}{M} = 1 - \langle e^{-NR_b(M)} \rangle_b \,. \tag{6}$$

Now expanding Eq. (4) yields

$$R_b(M) = -\log\left[\beta_b \int \mathrm{d}\eta \, e^{-\beta_b \eta}(1 - e^{-\alpha_b - \eta/T})^M\right] = -\log\left[{}_2F_1\left(-M, \beta_b T, 1 + \beta_b T, e^{-\alpha_b}\right)\right] \,, \tag{7}$$

where ${}_2F_1(a, b, c, z)$ is the hypergeometric ${}_2F_1$ function (DLMF, 2024). Given that we are interested in the large-N limit, the behavior will be dominated by the hardest prompts, i.e. $\alpha_b \gg 1$. Furthermore, the usual working regime for BoN attacks has $M \sim \mathcal{O}(1)$ and hence $Me^{-\alpha_b} \ll 1$ which allows us to simplify the hypergeometric function to

$$\log R_b(M) \simeq \log M - \alpha_b + \log(\beta_b T) - \log(1 + \beta_b T) \,. \tag{8}$$

Notice that averaging over all possible augmentations a for a fixed prompt b we can define an effective prompt barrier

$$E_b = -\log\langle p_{b,a} \rangle_a = \alpha_b - \log(\beta_b T) + \log(1 + \beta_b T) \,, \tag{9}$$

so that $\log R_b(M) \simeq \log M - E_b$. In other words, the jailbreaking frequency $R_b(M) \simeq Me^{-E_b}$ corresponds to M parallel jumps over an effective barrier of height E_b. Given that α_b is Gaussian and β_b is log-normal E_b is well approximated by a Gaussian of mean μ_E and variance σ_E^2. So Eq. (6) becomes

$$\text{AFR@}\binom{N}{M} \equiv 1 - \text{ASR@}\binom{N}{M} \simeq \left\langle \exp\left[-e^{\log(NM) - \mu_E + \sigma_E \xi}\right] \right\rangle_{\xi \sim \mathcal{N}(0,1)} \,. \tag{10}$$

The saddle point of Eq. (10) (see App. A for details) yields

$$\text{AFR@}\binom{N}{M} \simeq f_{\sigma_E}\left[(NM)e^{-\mu_E}\right] \,, \tag{11}$$

where

$$f_\gamma(z) = \frac{1}{\sqrt{1 + W_0(\gamma^2 z)}} \exp\left[-\frac{W_0(\gamma^2 z)^2 + 2W_0(\gamma^2 z)}{2\gamma^2}\right] \,, \tag{12}$$

with W_0 being the Lambert function. A plot of this scaling function for various values of γ is given in Fig. 6(c). Writing $L \equiv \log\left(\gamma^2 z\right)$ and using $W_0(\gamma^2 z) \simeq L - \log L + \log L/L + \mathcal{O}(L^{-2})$, the asymptote of this scaling function when $z \to +\infty$ is

$$f_\gamma(z) \simeq \frac{1}{\sqrt{1 + L - \log L}} \exp\left[-\frac{(L - \log L)^2 + 2L}{2\gamma^2}\right] \,. \tag{13}$$

Notice that the asymptote is not a clean power-law as is usually reported in the literature. In reality, it predicts a power law $f_\gamma(z) \propto z^{-\nu_{\text{eff}}}$ with a slowly varying exponent

$$\nu_{\text{eff}} \equiv -\frac{\mathrm{d}\log f_\gamma(z)}{\mathrm{d}\log z} = \frac{W_0(\gamma^2 z)}{\gamma^2} + \frac{W_0(\gamma^2 z)}{2\left[1 + W_0(\gamma^2 z)\right]^2} \overset{z \to +\infty}{\simeq} \frac{1}{\gamma^2} \log z \,. \tag{14}$$

Although less interesting, the low-budget (i.e. small N) asymptote can also be obtained in closed form. However, to do so we cannot take the small-z limit of Eq. (12) since $f_\gamma(z)$ was already the result of a large-N saddle-point approximation. Instead we have to expand Eq. (10) directly for small N which yields

$$\text{AFR@}\binom{N}{M} \simeq 1 - (NM)\, e^{-\mu_E + \sigma_E^2/2} \equiv 1 - z e^{\gamma^2/2} \,. \tag{15}$$

Table 1: The four parameters of each model, at $T = 1$ via the estimator of Eq. (20). The effective mean barrier $\mu_E = \langle -\log\langle p_{b,a}\rangle_a\rangle_b$ is defined in Eq. (9). Prompts which are never jailbroken are ill-defined and dropped. Uncertainties are ± 1 standard deviation from a cluster bootstrap over prompts (2000 resamples). The amplitude of the BoN kernel noise is chosen such that the median jailbreak rate is in 1–5%.

Model	μ_α Base alignment	σ_α Alignment fluctuations	μ_β Attack susceptibility	σ_β Susceptibility fluctuations	μ_E Effective barrier	σ_E Barrier fluctuations
Qwen2.5-7B (base)	$0.57_{\pm 0.03}$	$0.27_{\pm 0.03}$	$+1.99_{\pm 0.09}$	$0.85_{\pm 0.05}$	$0.73_{\pm 0.03}$	$0.33_{\pm 0.04}$
Llama-3.1-8B-Instruct	$0.39_{\pm 0.10}$	$0.93_{\pm 0.11}$	$-1.29_{\pm 0.15}$	$\mathbf{1.33}_{\pm 0.11}$	$2.05_{\pm 0.15}$	$1.36_{\pm 0.11}$
Qwen2.5-3B-Instruct	$1.83_{\pm 0.03}$	$0.65_{\pm 0.03}$	$-0.88_{\pm 0.03}$	$0.52_{\pm 0.02}$	$3.08_{\pm 0.03}$	$0.58_{\pm 0.02}$
Qwen2.5-7B-Instruct	$2.10_{\pm 0.06}$	$\mathbf{1.30}_{\pm 0.05}$	$\mathbf{-1.97}_{\pm 0.05}$	$1.07_{\pm 0.04}$	$\mathbf{4.27}_{\pm 0.08}$	$\mathbf{1.65}_{\pm 0.06}$
Gemma-2-9B-it	$\mathbf{2.59}_{\pm 0.06}$	$0.60_{\pm 0.04}$	$+1.07_{\pm 0.08}$	$0.75_{\pm 0.06}$	$2.93_{\pm 0.06}$	$0.62_{\pm 0.03}$

4 Experimental validation

We now show experimental evidence for this microscopic model in three families of open-source models with 3B to 9B parameters: Qwen (Qwen Team, 2024), Llama (Grattafiori et al., 2024) and Gemma (Gemma Team, 2024). For each model we run a BoNM jailbreak on the first 400 harmful prompts of AdvBench (Zou et al., 2023a) (see App. D for details). For each prompt we draw N augmentations (Gaussian noise of variance σ in embedding space or character-level noise) and for each augmentation we draw M completions at temperature T. To have comparable results across models we choose σ such that the average jailbreak rate is 1–5% for each model. Most of our results are derived from a $N = 100$, $M = 300$ grid, then to probe the far-N regime we further run an independent attack with $N = 10^4$ augmentations and $M = 1$ completion per prompt. Each completion is scored by an external judge (Granite-Guardian-3.0-2B (Padhi et al., 2024) and HarmBench-13B (Mazeika et al., 2024) as a control) and deemed unsafe if its harm probability exceeds 0.5. Leading to the empirical measure

$$\hat{p}_{b,a} = \frac{1}{M}\sum_{i=1}^{M} \mathbf{1}\{\text{completion } i \text{ of prompt } b \text{ with augmentation } a \text{ is jailbroken.}\} \equiv \frac{\hat{k}_{b,a}}{M} \ . \quad (16)$$

Given a single prompt b, we can compute the individual jailbreak rate $\hat{p}_{b,a}$ and then average over the $N \gg 1$ augmentations $\langle \hat{p}_{b,a}\rangle_a$, our theory predicts that $-\log\langle p_{b,a}\rangle$ should be equal to

$$-\log\langle p_{b,a}\rangle_a = -\log\langle e^{-\alpha_b - \eta/T}\rangle_a = \alpha_b - \log\left[\frac{\beta_b T}{1+\beta_b T}\right] \ . \quad (17)$$

Now instead of looking at the mean we can study the full distribution of $p_{b,a}$. Making the change of variables from η to $p_{b,a}$ yields

$$\text{Prob.}[p_{b,a} \,|\, b] = \beta_b T e^{\beta_b T \alpha_b} p^{\beta_b T - 1} \quad (18)$$

Hence β_b is clearly defined by the $p \to 0^+$ limit of this distribution and the upper support of this distribution is $e^{-\alpha}$. However, using the support edge as an estimator is extremely noisy. Instead we fit a soft-wall by Maximum Likelihood Estimation (MLE) using a Beta distribution $\hat{p} \sim B(\hat{\beta}_b T, \hat{c}_b)$ to model $\hat{k}_{b,a} \sim \text{Binomial}(M, p)$. Then

$$\log\langle \hat{p}\rangle = \log\frac{\hat{\beta}_b T}{\hat{\beta}_b T + \hat{c}_b} \quad (19)$$

and equating with Eq. (17) we get

$$\hat{\alpha}_b = \log\left[\frac{\hat{\beta}_b T + \hat{c}_b}{1+\hat{\beta}_b T}\right] \ . \quad (20)$$

Given Eq. (20) and with an MLE fit we are now able to recover α_b and β_b from numerical data. The goodness of fit of the MLE (over 400 prompts typical error 0.17 in $\log Q_b(M)$ with $M \le 300$) supports that η being exponentially distributed is a good approximation on this dataset. In Fig. 2 we can see that the measured $\hat{\alpha}_b$ and $\log\hat{\beta}_b$ are well approximated by a Gaussian distribution whose parameters are given in Table 1. The raw distributions of $\hat{\alpha}_b$ and $\log\hat{\beta}_b$ of the two instruct models are very different but once centered and standardized they cannot be distinguished. The normal and log-normal *families* are universal, only their moments are model-specific.

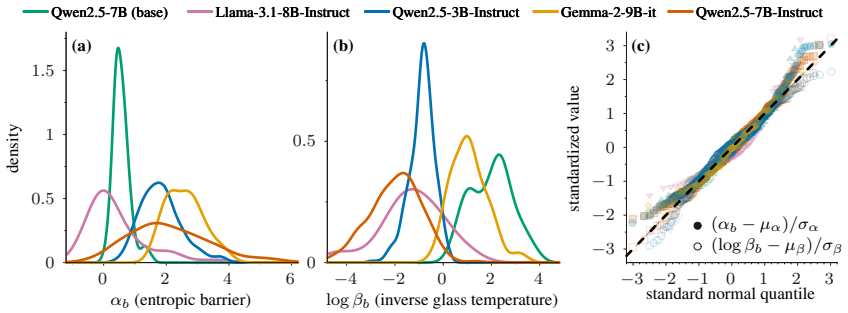

Figure 2: **Universal laws.** Across prompts, α_b is Gaussian and β_b log-normal. Panel (c) shows how standardized per-prompt parameters collapse onto $\mathcal{N}(0, 1)$.

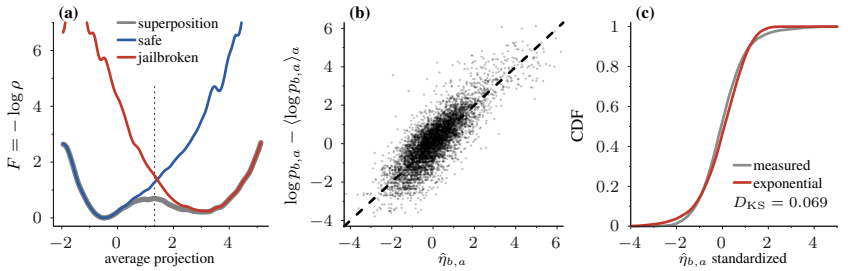

Figure 3: **Geometry of the residual stream.** (a) Effective free energy along a supervised safety readout after balancing each population. Average projection along a single direction separates completions into two safe/unsafe regions. (b) The saddle $\eta_{b,a}$ predicted from the augmented prompt alone against the measured per-augmentation rate. (c) The distribution of the recovered $\eta_{b,a}$, against the exponential assumption pushed through the same estimator.

Safety Results. We can see in Table 1 that the base model has essentially no barrier. Its median augmentation is jailbroken roughly half of the time. The instruct-tuned version of the same model raises μ_α by a factor of four and σ_α by a factor of five. The entropic barrier is therefore not a property of the architecture but a consequence of training/finetuning. Alignment also improves resistance to BoNM attacks by moving μ_β from $+1.99$ to -0.9 or -2.0. Between the two Qwen-Instruct models, the 7B model is both safer ($\mu_\alpha^{(7B)} > \mu_\alpha^{(3B)}$) and less susceptible to the BoNM attack ($\mu_\beta^{(7B)} < \mu_\beta^{(3B)}$). Interestingly not all models defend in the same way. Gemma favors an 'entropic defense' (the α channel) whereas Llama is entirely aligned via an 'energetic defense' (the β channel).

The physical picture. To show that the physical model is justified we now identify its constituting elements directly in the residual stream. For each prompt b and augmentation a we average the residuals of the completions the judge calls unsafe, subtract the average over those it calls safe, normalize, and then average the resulting unit contrasts over all b and a to compute a single 'safety vector' $\hat{c}_{\mathrm{hyper}}$. The vector $\hat{c}_{\mathrm{hyper}}$ is the normal of the safety hyperplane of Fig. 1, and a linear readout along it separates safe from unsafe completions well (held-out AUC $= 0.89$). Taking the average projection of each completion on the safety direction, the two safe/unsafe basins appear clearly in Panel (a) of Fig. 3.

A second crucial assumption of our model is that the saddle $\eta_{b,a}$ is a property of the augmentation a. Consequently, it should be readable from the augmented prompt residual directly, without the need to look at the completion. To check this we train a linear probe $\hat{c}_{\mathrm{saddle}}$ to predict $\log p_{b,a} - \langle \log p_{b,a} \rangle_a$ from the centered residual $r_{b,a} - \langle r_{b,a} \rangle_a$ of the last token of the augmented prompt. As shown in Panel (b) of Fig. 3 not only are we able to recover $\eta_{b,a}$ ($R^2 = 0.69$ on held out validation), but the spread of the projection along $\hat{c}_{\mathrm{saddle}}$ is inversely correlated to β_b (Spearman -0.59). We can also use this projection to verify that the exponential assumption $\eta_{b,a} \sim \mathrm{Exp}[\beta_b]$ is reasonable, which is shown in Panel (c) of Fig. 3. Since the probe is a noisy and censored estimate of $\eta_{b,a}$, the comparison is made by pushing the exponential law through the same estimator.

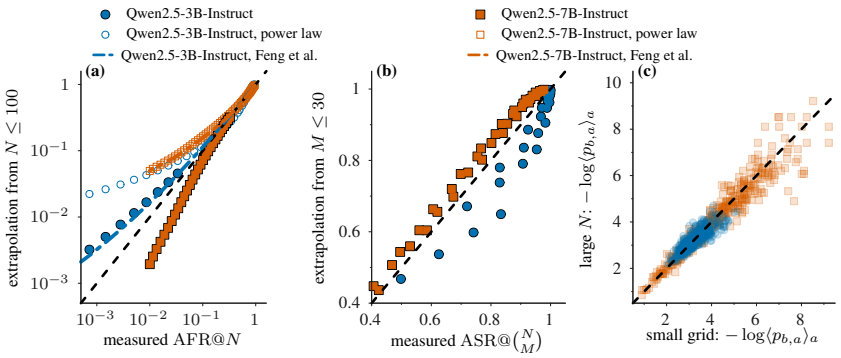

Figure 4: **Extrapolation (closer to the diagonal = better).** We fit (α_b, β_b) on a small $N \leq 100$, $M \leq 30$ grid. Extrapolate its predictions and compare with two baselines: a power-law and Feng et al. (2026) Beta fit. In panel (a) we extrapolate to large $N \sim 10^4$. Feng et al. (2026) and our predictions are identical for the 3B model but we are considerably more accurate on Qwen7B. In panel (b), we extrapolate to large $M \sim 300$. In panel (c), the effective barrier $-\log\langle p_{b,a}\rangle_a$ is measured two independent ways, on the 100×300 grid versus on the $10^4 \times 1$ grid. Both agree at $r = 0.85$ (3B) and 0.96 (7B).

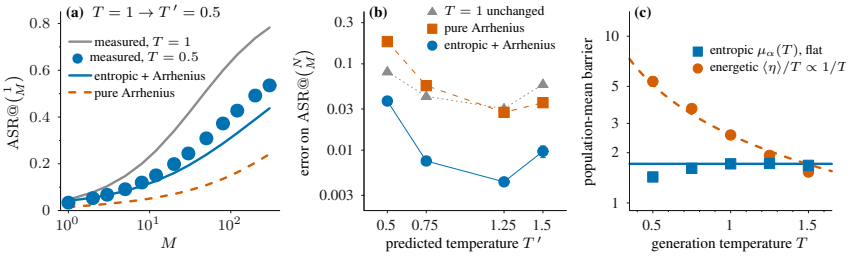

Figure 5: **Temperature dependence.** (a-b) Predicting a new decoding temperature from $T = 1$ data alone. Pure Arrhenius $p(T) = p(1)^{1/T}$ (dashed) overstates the effect of temperature in both directions. Our model keeps the entropic barrier outside the $1/T$ (Eq. (21)) with α_b fitted on the $T = 1$ data (blue) which reduces the error fourfold. (c) The barrier is entropic (constant in T), the saddles are energetic ($\propto 1/T$).

4.1 Reconstruction and Extrapolation

The first test is whether $(\mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta)$ are truly properties of only the model, i.e. independent of the prompt b and the augmentation a. To check this we set $\hat{c}_b$ from Eq. (20), and compute ASR@$\binom{N}{M}$ from Eqs. (6) on the $N, M \leq 100$ grid. We compare this population-level fit with a separate Beta-distribution fit for each prompt. The four-number model matches the measured data with a Mean Absolute Error (MAE) of 0.031. In comparison, the per-prompt fit leads to an MAE of 0.034. In other words, the population model with only 4 fitting parameters is as accurate as the per-prompt model with 800 parameters.

We now test whether this model is able to extrapolate the ASR beyond its fitting window. To do so, we fit the four parameter family $(\mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta)$ on a small region of augmentations $N \leq 100$ and completions $M \leq 30$ and check how they perform at much larger values $N \sim 10^4$ and $M \sim 300$. Fig. 4 shows how the small attack fit closely predicts large attack performances. In comparison, if we follow the usual practice in the literature (Hughes et al., 2024; Feng et al., 2026) and fit a power-law to the $N \leq 100$ region and extrapolate to $N \sim 10^4$ the prediction would have been significantly off, as can be seen in Fig. 4 panel (a). This is a consequence of the ASR not really following a power-law, as previously discussed in Eq. (14). Our model is also either equal (for Qwen3B) or more accurate (for Qwen7B) compared to Feng et al. (2026).

Generation temperature T also enters our model in a very specific manner: only the saddle crossings are thermal (via Arrhenius' law), neither the barrier α_b nor the inverse glass temperature $\beta_b = 1/T_g$ should move with T. Re-running the grid at five temperatures on 100 prompts, with the augmentation

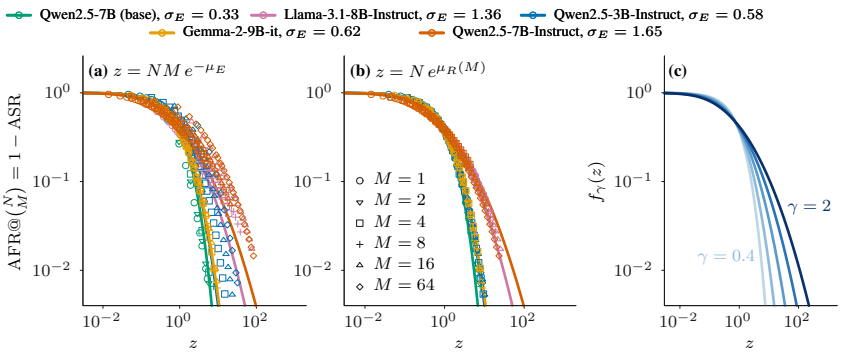

Figure 6: **Scaling collapse of the AFR.** In panel (a) we show the small-M closed-form leading order which although mostly accurate starts to deviate at large-z for larger values of M. In panel (b) we use the M-agnostic form which remains accurate at large-z for all M. Finally in panel (c) we show a few examples of the scaling function $f_\gamma(z)$ for varying values of the shape parameter γ.

seeds of the $T = 1$ grid, we find μ_α and μ_β flat while the saddle height in units of T follows $1/T$ (panel (c) of Fig. 5). This is the direct evidence that α_b is entropic, i.e. independent of temperature whereas saddles are thermal. Furthermore, given the exact temperature structure, the whole ASR at a new temperature T' follows from the $T = 1$ data through

$$p_{b,a}(T') = e^{-\alpha_b(1-1/T')}\, p_{b,a}(1)^{1/T'} \ . \tag{21}$$

Fig. 5 shows that Eq. (21), with α_b fitted at $T = 1$, reduces the error by a factor of 3.7 to 7.5 relative to pure Arrhenius (i.e. $p_{b,a}(T) = e^{-(\alpha_b+\eta_{b,a})/T}$), at all four unseen temperatures.

4.2 Scaling collapse

The scaling form in Eq. (11) makes a striking verifiable prediction. If we plot the AFR as a function of $z \equiv (NM)e^{-\mu_E}$ then all models should follow the same curve family $f_\gamma(z)$ defined in Eq. (12). This collapse can be seen in Fig. 6. However, the exact scaling form in Eq. (11) is only valid at $M \ll e^{\mu_\alpha}$ (which allowed us to simplify the hypergeometric function in Eq. (7)). Otherwise, the best we can do is keep the Gaussian approximation of $\log R_b(M) \sim \mathcal{N}(\mu_R(M), \sigma_R(M)^2)$ without knowing the closed form of $\mu_R(M)$ and $\sigma_R(M)^2$ which leads to panel (b) in Fig. 6. Both are read off the measured grid (App. A). We can see the clear separation between Qwen7B-Instruct and Llama ($\sigma_E \sim 1.5$) compared to Qwen3B-Instruct, Qwen7B-base and Gemma ($\sigma_E \sim 0.5$).

4.3 Power-law and fake exponential crossover

We now turn to the usual power-law behavior reported in the literature (Hughes et al., 2024; Feng et al., 2026). We fit our model to the usual $N \leq 100, M \leq 300$ experimental grid and then run a separate attack on $N = 10^4$ augmentations (with $M = 1$) to probe the large-N behavior. As shown in Fig. 7 the four numbers of Table 1 extrapolate to the large-N quite well until $N \sim 10^3$ where the data seems to cross over to an exponential regime. In parallel, the measured local exponent $\nu_{\text{eff}}(N)$ tracks the theoretical prediction and drifts from 0.1 at $N = 2$ to above 2 at $N \sim 10^2$–10^3 with no sign of a plateau, directly contradicting the usual power-law interpretations of BoN scaling. There is no fixed power-law exponent, only the running one of Eq. (14).

Although Halder et al. (2026) have reported exponential corrections, we believe that this exponential crossover is a finite-size artifact of the limited harmful dataset over which the ASR is computed. With a finite dataset of P prompts the population average of Eq. (6) is a sum of P exponentials, and beyond the budget $N_{\text{res}} = 1/\min_b \langle p_{b,a} \rangle_a$ at which even the hardest prompt of the dataset has been jailbroken, the survival is dominated by that single prompt and decays as a pure exponential $(1 - \langle p \rangle_a)^N / P$. As can be seen in Fig. 7, the measured curve follows the population law up to N_{res} and then breaks away. The exponent ν_{eff} shoots up from ~ 2 to $\sim 10^2$. This exponential behavior is a side-effect of the dataset size, not a property of the model. Panel (c) of Fig. 7 shows how progressively larger datasets shift the transition rightwards. For the 400-prompt AdvBench, N_{res} is approximately 10^2 for Qwen-3B-Instruct and $\approx 2 \times 10^3$ for Qwen-7B-Instruct.

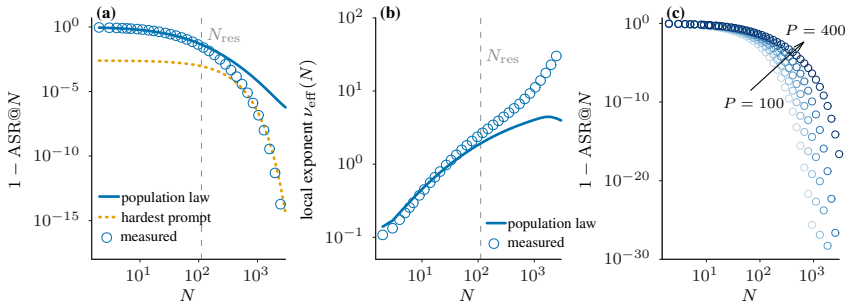

Figure 7: **Fake exponential crossover.** For each prompt we draw $G = 10^4$ augmentations of which H are jailbroken (unlike before all prompts are kept even those which are never jailbroken). The survival at budget N is estimated by $\binom{G-H}{N}/\binom{G}{N}$, i.e. that a random N-subset of the G augmentations contains none of the H jailbroken ones. (a) Measured population survival (points), the four-number population law from the grid fits (solid), and the survival of the single hardest prompt of the dataset (dotted). (b) The measured local exponent follows the predicted drift up to the resolution budget N_{res} of the dataset, and then breaks away and 'crosses over' as the curve is taken over by the hardest prompt. (c) The same AFR measurement on nested subsets of the P easiest prompts. Adding harder prompts moves N_{res} from 18 to 111 and the bend moves right with it.

5 Safety Consequences

Compute-optimal attack. We can see from Eq. (11) that N and M contribute equally to the ASR provided $M \ll e^{\mu_\alpha}$. When $M \sim e^{\mu_\alpha}$ more attempts yield diminishing returns. Hence, if completions are cheaper to sample than augmentations, an attacker should always use $M^\star \sim e^{\mu_\alpha}$ with an exact $M^\star$ whose closed form expression is a complicated implicit equation detailed in App. B. On the other hand, if augmentations are cheaper to sample than completions then there is no reason to go beyond $M^\star = 1$.

What a defender should optimize. BoN is a worst-case attack: Eq. (6) is a Laplace transform of the barrier distribution and at any budget it is governed by the prompts of barrier $\alpha \sim \log N$, i.e. extremely high barriers. A naive intuition would be that the best defense is obtained by a tight cluster of extremely high barriers, i.e. $\mu_E \gg 1$ and $\sigma_E \ll 1$. Surprisingly, Eq. (11) shows that this is more nuanced. Fig. 6 shows that higher values of σ_E lead to safer models at large-z but weaker models at smaller z. Since BoN is an extreme-value process, large N eventually reaches any barrier height. However, by keeping $\sigma_E \gg 1$ we are 'fighting' one extreme-value process with another. BoN will struggle to reach the few extreme barriers which are much higher than the mean, at the cost of sacrificing a few 'easy to reach' low barriers.

Alignment protocols. It is worth noting that different alignment schemes seem to lead to different defense types. Table 1 shows that Gemma, Llama and Qwen have three different ways of defending against BoN. Gemma relies on a strong entropic bottleneck μ_α but almost no energetic barrier μ_β. In contrast, Llama has almost no entropic bottleneck but a large energetic barrier, and Qwen sits in between. Interestingly, Llama and Gemma used opposite alignment recipes: offline DPO (Rafailov et al., 2023; Grattafiori et al., 2024) versus on-policy RL (Gemma Team, 2024). How alignment training maps onto these four parameters is out of scope here but would be an interesting follow-up.

6 Conclusion

In this work, we put forward and validate experimentally a physics-inspired microscopic model for Best-of-N jailbreaking and generalized Best-of-N to Best-of-NM by considering multiple completions for the same augmentation. We found that contrary to the usual literature the attack success rate does not follow a fixed power-law. We further showed that the reported exponential crossover is an artifact of fitting jailbreaking behaviors on a finite adversarial dataset. Our theoretical model allows us to quantitatively compare the safety of different models as shown in Table 1 and also makes predictions for best defensive practices to follow.

ETHICS STATEMENT

Limitations. The study in this work is restricted to open-weight models of intermediate size. Reproducing this work on frontier models was beyond our computational capacities and against the terms of service of proprietary models. We also showed that most fits are in reality limited by the relatively small size of the adversarial prompt dataset (400 prompts). Having more accurate large N behaviors would require a larger adversarial dataset. We have also only used a single dataset, expanding to other datasets would improve the universality of our claims. At our attack budgets $N \sim 10^4$ we were able to resolve most of the prompts. However, for the safest models we still had to drop a couple of prompts which were never jailbroken, effectively biasing their parameters to making them seem 'softer'. App. C measures the effect of censoring and the soft-wall estimator, by drawing synthetic data from the model itself and passing it through the same pipeline. Surprisingly, the two biases have opposite signs, and the net shift of μ_α is a few tenths upward rather than downward. So our results are only accurate up to a few tenths, which corresponds to their reported error bars. Resolving the censored prompts would require larger attack budgets which are beyond our computational capacity. Another limitation of this work is the reliance on LLMs as judges of harmful behaviors with no human control. We tried to mitigate this by using more than one LLM judge (Padhi et al., 2024; Mazeika et al., 2024) and saw that our results did not rely on a specific LLM judge choice; automated judges are nonetheless known to disagree (Souly et al., 2024; Ran et al., 2025). Finally, we have only studied the behavior of these attacks in English. There is no reason to believe that these models would behave identically for other languages. In fact, some languages may be easier to jailbreak than others.

Ethical concerns. This work studies an already established attack. The only offensive contribution is the possibility of making this attack more compute-efficient, which is only presented to inform defenders on what a worst-case attack could be. This paper aims at better understanding the mechanisms at play behind BoN jailbreaking. While we cannot exclude that the results presented in this paper could possibly lead to improved offensive capabilities, we pursued this work, and presented it, from a defender's perspective. We have included prescriptions on how our results could guide a defender's choices on how to safeguard their models and we hope that the work presented here will lead to more robust alignment protocols.

REPRODUCIBILITY STATEMENT

Every number in the paper is produced by a scripted pipeline from raw generations: the per-prompt Beta-Binomial fits and the estimator of Eq. (20), the four-parameter surfaces, the N- and M-extrapolations, the temperature sweep, the probes, and all figures. Models, judges, augmentation kernels and grid shapes are specified in Sec. D; the code and the fitted per-prompt parameters will be released.

USE OF LARGE LANGUAGE MODELS

We used an AI coding assistant (Claude, Anthropic) throughout this project, in two capacities. *(i) Code and experiments.* The assistant wrote and debugged part of the analysis and plotting code (fitting routines, bootstrap error bars, the figure pipeline) and helped launch and monitor the generation and scoring jobs on SLURM. It was also used interactively to run exploratory analyses and sanity checks proposed by the authors. *(ii) Writing.* The main text was written by the authors. The assistant drafted parts of the supplementary material (derivations and experimental details) from the authors' notes and results, which the authors then checked and edited. It was used for proofreading and for consistency checks between the text, the tables and the figures. All LLM-produced code, numbers and text were verified by the authors, who take full responsibility for the content of the paper. Separately, LLMs are the object of study and an LLM-based classifier is used as the jailbreak judge.

REFERENCES

J. Hughes, S. Price, A. Lynch, R. Schaeffer, F. Barez, A. Somani, S. Koyejo. H. Sleight, E. Jones and M. Sharma, *Best-of-N jailbreaking*, Advances in Neural Information Processing Systems, **38**, pp.73137-73221 (2024).

I. Halder, A. Banerjee and C. Pehlevan, *Jailbreak Scaling Laws for Large Language Models: Polynomial-Exponential Crossover*, arXiv preprint, arXiv:2603.11331 (2026).

M. Feng, X. Liu, W. Yang, C. Xu, C. White and J. Gao, *Statistical estimation of adversarial risk in large language models under best-of-n sampling*, arXiv preprint, arXiv:2601.22636 (2026).

J.P. Bouchaud, *Weak ergodicity breaking and aging in disordered systems*, Journal de Physique I, **2** (9), pp.1705-1713 (1992).

B. Derrida, *Random-energy model: An exactly solvable model of disordered systems*, Physical Review B, **24** (5), p.2613 (1981).

P. Hänggi, P. Talkner and M. Borkovec, *Reaction-rate theory: fifty years after Kramers*, Reviews of modern physics, **62** (2), p.251 (1990).

NIST Digital Library of Mathematical Functions, §15, `https://dlmf.nist.gov/` (2024).

A. Yang, A. Li, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu, C. Gao, C. Huang, C. Lv and C. Zheng, *Qwen2.5 technical report*, arXiv preprint, arXiv:2412.15115 (2024).

A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan and A. Yang, *The Llama 3 herd of models*, arXiv preprint, arXiv:2407.21783 (2024).

G. Team, M. Riviere, S. Pathak, P.G. Sessa, C. Hardin, S. Bhupatiraju, L. Hussenot, T. Mesnard, B. Shahriari, A. Ramé and J. Ferret, *Gemma 2: Improving open language models at a practical size*, arXiv preprint, arXiv:2408.00118 (2024).

R. Rafailov, A. Sharma, E. Mitchell, C.D. Manning, S. Ermon and C. Finn, *Direct preference optimization: Your language model is secretly a reward model*, Advances in neural information processing systems, **36**, pp.53728-53741 (2023).

A. Choromanska, M. Henaff, M. Mathieu, G.B. Arous and Y. LeCun, *The loss surface of multilayer networks*, PMLR, Artificial intelligence and statistics, pp. 192-204, (2015).

A. Zou, Z. Wang, N. Carlini, M. Nasr, J.Z. Kolter and M. Fredrikson, *Universal and transferable adversarial attacks on aligned language models*, arXiv preprint, arXiv:2307.15043 (2023).

M. Hutter, *Learning Curve Theory*, arXiv preprint, arXiv:2102.04074 (2021).

R. Schaeffer, J. Kazdan, J. Hughes, J. Juravsky, S. Price, A. Lynch, E. Jones, R. Kirk, A. Mirhoseini and S. Koyejo, *How do large language monkeys get their power (laws)?*, PMLR, International Conference on Machine Learning, **267**, pp. 53132-53176 (2025).

N. Levi, *A Simple Model of Inference Scaling Laws*, PMLR, International Conference on Machine Learning, **267**, pp. 33984-33998 (2025).

N. Levi, *Learning Shrinks the Hard Tail: Training-Dependent Inference Scaling in a Solvable Linear Model*, International Conference on Learning Representations (2026).

J. Kazdan, R. Schaeffer, Y. Allouah, C. Sullivan, K. Yu, N. Levi and S. Koyejo, *Efficient prediction of pass@k scaling in large language models*, arXiv preprint, arXiv:2510.05197 (2025).

A. Zou, L. Phan, S. Chen, J. Campbell, P. Guo, R. Ren, A. Pan, X. Yin, M. Mazeika, A.K. Dombrowski and S. Goel, *Representation engineering: A top-down approach to AI transparency*, arXiv preprint, arXiv:2310.01405 (2023).

K. Park, Y.J. Choe and V. Veitch, *The linear representation hypothesis and the geometry of large language models*, PMLR, International Conference on Machine Learning, **235**, pp. 39643-39666 (2024).

M. Mazeika, L. Phan, X. Yin, A. Zou, Z. Wang, N. Mu, E. Sakhaee, N. Li, S. Basart, B. Li and D. Forsyth, *HarmBench: A standardized evaluation framework for automated red teaming and robust refusal*, PMLR, International Conference on Machine Learning, **235**, pp. 35181-35224 (2024).

I. Padhi, M. Nagireddy, G. Cornacchia, S. Chaudhury, T. Pedapati, P. Dognin, K. Murugesan, E. Miehling, M.S. Cooper, K. Fraser and G. Zizzo, *Granite guardian*, arXiv preprint, arXiv:2412.07724 (2024).

L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray and J. Schulman, *Training language models to follow instructions with human feedback*, Advances in neural information processing systems, **35**, pp.27730-27744 (2022).

A. Arditi, O. Obeso, A. Syed, D. Paleka, N. Panickssery, W. Gurnee and N. Nanda, *Refusal in language models is mediated by a single direction*, Advances in Neural Information Processing Systems, **37**, pp.136037-136083 (2024).

T. Wollschläger, J. Elstner, S. Geisler, V. Cohen-Addad, S. Günnemann and J. Gasteiger, *The geometry of refusal in large language models: Concept cones and representational independence*, PMLR, International Conference on Machine Learning, **267**, pp. 66945-66970 (2025).

A. Souly, Q. Lu, D. Bowen, T. Trinh, E. Hsieh, S. Pandey, P. Abbeel, J. Svegliato, S. Emmons, O. Watkins and S. Toyer, *A strongreject for empty jailbreaks*, Advances in Neural Information Processing Systems, **37**, pp.125416-125440 (2024).

D. Ran, J. Liu, Y. Gong, J. Zheng, X. He, Y. Cong and A. Wang, *JailbreakEval: An integrated toolkit for evaluating jailbreak attempts against large language models*, Network and Distributed System Security Symposium (NDSS), poster (2025).

L. Schwinn, D. Dobre, S. Xhonneux, G. Gidel, and S. Günnemann, *Soft Prompt Threats: Attacking Safety Alignment and Unlearning in Open-Source LLMs through the Embedding Space*, Advances in Neural Information Processing Systems, **37**, pp.9086-9116 (2024).

S. Geisler, T. Wollschläger, M. H. I. Abdalla, J. Gasteiger, and S. Günnemann, *Attacking Large Language Models with Projected Gradient Descent*, ICML Workshop on the Next Generation of AI Safety (2024).

A. Robey, E. Wong, H. Hassani, and G. J. Pappas, *SmoothLLM: Defending Large Language Models Against Jailbreaking Attacks*, Transactions on Machine Learning Research (2025).

J. Cohen, E. Rosenfeld, and J. Z. Kolter, *Certified Adversarial Robustness via Randomized Smoothing*, PMLR, International conference on machine learning, pp. 1310-1320 (2019).

X. Liu, N. Xu, M. Chen, and C. Xiao, *AutoDAN: Generating Stealthy Jailbreak Prompts on Aligned Large Language Models*, International Conference on Learning Representations, pp. 56174-56194 (2024).

P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong, *Jailbreaking Black Box Large Language Models in Twenty Queries*, IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pp. 23-42 (2025).

A SADDLE-POINT EVALUATION OF THE POPULATION AVERAGE

This appendix derives Eq. (12) from Eq. (10). Write $x = \log(NM) - \mu_E$ and $\gamma = \sigma_E$, so that Eq. (10) reads

$$\mathrm{AFR@}\binom{N}{M} \simeq \int \frac{\mathrm{d}\xi}{\sqrt{2\pi}}\, e^{-\varphi_{x,\gamma}(\xi)} \;\text{ with }\; \varphi_{x,\gamma}(\xi) = \frac{\xi^2}{2} + e^{\,x+\gamma\xi}\,. \tag{22}$$

The exponent is convex, so it has a single stationary point,

$$\varphi'_{x,\gamma}(\xi^\star) = \xi^\star + \gamma\, e^{\,x+\gamma\xi^\star} = 0\,. \tag{23}$$

Since the exponential is positive we must have $\xi^\star < 0$. In other words, at large x the surviving prompts are the *hard* ones whose effective barrier E_b lies above the mean. Re-writing $w = -\xi^\star > 0$ yields $e^{\,x-\gamma w} = w/\gamma$, i.e. $we^{\gamma w} = \gamma e^x$, and multiplying by γ gives a Lambert equation for the rescaled saddle $\zeta \equiv \gamma w$,

$$\zeta\, e^{\zeta} = \gamma^2 e^x \qquad\Longrightarrow\qquad \zeta = W_0\big(\gamma^2 e^x\big) = W_0\big(\sigma_E^2 (NM) e^{-\mu_E}\big)\,, \tag{24}$$

with W_0 the principal branch of the Lambert function. At the saddle the height and the curvature are

$$\varphi(\xi^\star) = \frac{w^2}{2} + \frac{w}{\gamma} = \frac{\zeta^2 + 2\zeta}{2\gamma^2}\,, \qquad \varphi''(\xi^\star) = 1 + \gamma^2 e^{\,x+\gamma\xi^\star} = 1 + \gamma w = 1 + \zeta\,. \tag{25}$$

The Gaussian fluctuations in the integral contribute $\varphi''(\xi^\star)^{-1/2}$ and Eq. (22) becomes $f_\gamma(z)$ of Eq. (12), with the argument $z = (NM)e^{-\mu_E}$ of the main text entering only through $\zeta = W_0(\gamma^2 z)$.

The collapse variables. Panel (b) of Fig. 6 needs $\mu_R(M)$ and $\sigma_R(M)$ without any closed form. To obtain them we measure $R_b(M) = -\log Q_b(M)$ for every prompt from the unbiased completion-channel estimate of $Q_b(M)$. Then, $\mu_R(M) = \langle \log R_b(M) \rangle_b$ is read directly from the data and the width is kept at the single number σ_E because the measured $\sigma_R(M)$ barely changes with M: from 0.55 at $M = 1$ to 0.54 at $M = 64$ on Qwen2.5-3B-Instruct.

B COMPUTE-OPTIMAL COMPLETION BUDGET

This appendix gives the implicit equation for $M^\star$ referred to in Sec. 5. Fix a prompt b and write $R_b(M) = -\log Q_b(M)$ as in Eq. (4). Because $Q_b(M) = \langle e^{MZ} \rangle_a$ with $Z = \log(1 - p_{b,a})$, the function $\log Q_b(M)$ is a cumulant generating function and is therefore convex with $\log Q_b(0) = 0$; hence $R_b(M)$ is concave, $R_b(M)/M$ is non-increasing, and

$$Q_b(M) \geq Q_b(1)^M \qquad \text{for every prompt and every } M\,, \tag{26}$$

with a gap growing like $(M-1)\,\mathrm{Var}_a(p)$. Hence if sampling completions costs the same (or more) than sampling augmentations $M^\star = 1$ is optimal. An attacker should spend everything on distinct augmentations.

However under some metrics (FLOPS for example) completions might be cheaper to sample than augmentations. A fresh augmentation requires a fresh prefill, whereas an extra completion of an augmentation already in cache does not, so in forward-pass units the cost of an (N, M) attack is $C(N, M) = N(r + M)$, with r the prefill cost expressed in completions. At fixed C the survival of prompt b is $\exp[-N R_b(M)] = \exp[-C\, R_b(M)/(r + M)]$, so minimizing it means maximizing $g(M) = R_b(M)/(r + M)$ which satisfies

$$(r + M^\star)\, R'_b(M^\star) = R_b(M^\star)\,. \tag{27}$$

With the Beta soft wall $p \sim B(\beta_b T, c_b)$, for which $Q_b(M) = \langle (1-p)^M \rangle = B(\beta_b T, c_b + M)/B(\beta_b T, c_b)$ with B denoting Euler's Beta function, we have

$$R_b(M) = \log \frac{B(\beta_b T, c_b)}{B(\beta_b T, c_b + M)}\,, \qquad R'_b(M) = \psi(\beta_b T + c_b + M) - \psi(c_b + M)\,, \tag{28}$$

with ψ the digamma function. Hence, $M^\star = 1$ iff $g'(1) \leq 0$, i.e. iff $(r+1)R'_b(1) \leq R_b(1)$, which expanded to second order in p reads

$$M^\star = 1 \iff r \leq \frac{\mathrm{Var}_a(p)}{2\langle p \rangle_a} = \frac{1 - q_b}{2\,(\beta_b T + c_b + 1)} \;\text{ where }\; q_b = \langle p_{b,a} \rangle_a\,. \tag{29}$$

In other words, repeated completions amortize the prefill cost if and only if the within-prompt spread of escape probabilities outweighs the prefill cost. The threshold is small for the models we measured. For Qwen2.5-3B-Instruct ($\beta_b T \simeq 0.41$, $c_b \simeq 8$, hence $q_b \simeq 0.049$) it is $r^\star \simeq 0.05$. Solving Eq. (27) at that fit gives $M^\star = 1.3, 2.6, 4.3, 6.2, 9.1$ at $r = 0.1, 0.4, 1, 2, 4$: the compute-optimal attack is not $M = 1$ but a handful of completions per augmentation, amortizing each prefill.

C Estimator details and identifiability

Clamping. The barrier α_b of Eq. (3) is Gaussian, so a fraction of the population is negative and the raw Arrhenius expression would exceed one there. Drawing from the fitted population of each model, the clamp of Eq. (1) is active on 7.3% of the (b, a) draws for Llama-3.1-8B-Instruct, 0.8% for Qwen2.5-7B (base), 0.6% for Qwen2.5-7B-Instruct, 0.02% for Qwen2.5-3B-Instruct and never for Gemma-2-9B-it.

Survival estimator. The population survival of Sec. 4.3 involves no fit and no censoring. Instead it is counted directly from the 10^4-augmentation arms, over all 400 prompts. On Qwen2.5-3B-Instruct every one of them is jailbroken at least 90 times, so none of them have to be censored.

The soft wall. Eq. (18) says the escape probabilities of a prompt follow a power law $\rho(p) \propto p^{\beta_b T - 1}$ cut off at $p = e^{-\alpha_b}$, so in principle (α_b, β_b) can be read off the exponent and the support edge. The edge estimator is unusable in practice. Since it is the maximum of N draws its variance does not decrease with M. We therefore replace the hard cutoff by the Beta soft wall

$$p_{b,a} \sim B(\beta_b T, c_b) \ , \tag{30}$$

and fit $(\beta_b T, c_b)$ by maximum likelihood on the Beta-Binomial model $k_{b,a} \sim \mathrm{Binomial}(M, p)$, with α_b recovered through Eq. (20). The Beta of Eq. (30) is a deliberate misspecification of the hard-edged law it replaces. However, we checked that by simulating Eq. (3) on synthetic data comparing the soft wall with the ground truth the shape parameters remain the same up to a few tenths which once again is within our reported error bars.

Limitations. This fit has two main limitations. A prompt with no jailbreak at all carries no information about either parameter and is dropped; we require at least three hits in total across the augmentations of a prompt (i.e. $\sum_a \hat{k}_{b,a} \geq 3$), and discard fits with unrealistic parameters ($\beta_b T \notin [10^{-3}, 50]$ or $c_b \notin [10^{-3}, 10^5]$) which typically correspond to augmentations which have led to gibberish completions. The number of prompts the fit retains is 400 of 400 for Qwen2.5-3B-Instruct, 383 of 400 for Qwen2.5-7B-Instruct, 91 of 100 for Qwen2.5-7B (base), 89 of 100 for Gemma-2-9B-it and 80 of 100 for Llama-3.1-8B-Instruct. Retrieving them would require an attack with a much larger N (or M or both). Second, the two parameters are separately identifiable only when the completion budget resolves the shape of the p distribution in Eq. (18) which is only possible for $M \gtrsim 10$.

D Experimental protocol

Table 2 lists every generation run behind the figures. All runs use the AdvBench harmful-behavior set (Zou et al., 2023a), sample with T as stated and `top_p` $= 1$, and cap completions at 96 new tokens, which is ample for a judge to decide and keeps the arms affordable. Every completion is scored by Granite-Guardian-3.0-2B (Padhi et al., 2024) as the primary graded judge, with HarmBench-Llama-2-13b-cls (Mazeika et al., 2024) as a control on the two large grids; a completion counts as unsafe when the judge's harm probability exceeds $\tau = 0.5$, and $\tau \in \{0.1, 0.3\}$ is recorded so that the threshold can be swept without regenerating.

Layers and probes. Every residual-stream experiment in this paper is run on Qwen2.5-3B-Instruct at layer 30 of its 36 blocks, reading the last-token residual. The safety readout of Fig. 3(a) is a ridge regression at penalty 10^3 on standardized residuals, mean-pooled over the tokens of the completion, with the judge's binary label as target. The saddle probe of Fig. 3(b) is a ridge regression of the within-prompt centred $\log p_{b,a}$ on the within-prompt centred residual of the augmented prompt, with the penalty selected from $\{10^2, 10^3, 10^4, 10^5, 10^6\}$.

Model	Kernel	σ	Prompts	$N \times M$	Role
Qwen2.5-3B-Instruct	embedding	0.048	400	100×300	main grid
Qwen2.5-3B-Instruct	embedding	0.048	400	$10^4 \times 1$	far-N arm
Qwen2.5-7B-Instruct	embedding	0.038	400	100×300	main grid
Qwen2.5-7B-Instruct	embedding	0.038	400	$10^4 \times 1$	far-N arm
Qwen2.5-7B (base)	embedding	0.038	100	40×20	alignment comparison
Gemma-2-9B-it	embedding	0.080	100	40×32	cross-family
Llama-3.1-8B-Instruct	embedding	0.015	100	40×32	cross-family
Qwen2.5-3B-Instruct	character	0.25	100	100×200	kernel control
Qwen2.5-3B-Instruct	embedding	0.048	100	100×300	$T \in \{0.5, 0.75, 1.25, 1.5\}$ sweep
Qwen2.5-3B-Instruct	embedding	0.048	60	100×1000	high-resolution p
Qwen2.5-3B-Instruct	embedding	6 values, 0.024–0.076	400	400	σ calibration sweep
Qwen2.5-3B-Instruct	embedding	6 values, 0.024–0.076	100	50×100	σ robustness (App. E)

Table 2: Generation runs. $N \times M$ is augmentations times completions per augmentation, per prompt. The embedding kernel adds isotropic Gaussian noise to the input embeddings of the user-content tokens only, leaving the chat template clean; the character kernel perturbs the prompt string. All runs are at $T = 1$ unless stated. The character kernel is the augmentation of Hughes et al. (2024). Each letter is case-flipped with probability σ, the interior of each word of four characters or more is randomly permuted with probability σ, and each character is replaced by a random letter or duplicated with probability $\sigma/10$. All runs are at $T = 1$ unless stated otherwise.

Choosing σ. The augmentation scale is quoted in raw embedding units, with no rescaling by the embedding norm. Embedding scales differ by an order of magnitude across families, so σ does not transfer between them and each model is calibrated on its own. To calibrate σ we sweep σ and take the value at which the median per-prompt jailbreak rate first lands in the 1% to 5% band. The usable window can be narrow. For example, Llama-3.1-8B has a median rate of 0 at $\sigma = 0.013$ and 0.18 at $\sigma = 0.017$, and by $\sigma = 0.04$ the completions are no longer fluent. App. E shows what the fitted parameters do as this choice is varied.

E AUGMENTATION SCALE ROBUSTNESS

Every grid in this paper is run at one augmentation scale σ, chosen per model by the calibration of Sec. D. To study the dependence of our 4-parameter family on σ we ran a sweep over 100 behaviors $\times$ 50 augmentations $\times$ 100 completions for 6 values of σ for Qwen-3B-Instruct. The prompt set is identical across the six.

σ	median rate	μ_α	σ_α	μ_β	σ_β	μ_E	σ_E
0.024	0.003	$3.49_{\pm 0.24}$	$1.68_{\pm 0.18}$	$-1.74_{\pm 0.28}$	$1.90_{\pm 0.16}$	$5.62_{\pm 0.17}$	$1.17_{\pm 0.09}$
0.030	0.007	$2.60_{\pm 0.17}$	$1.58_{\pm 0.09}$	$-2.09_{\pm 0.14}$	$1.32_{\pm 0.11}$	$4.91_{\pm 0.14}$	$1.32_{\pm 0.07}$
0.038	0.032	$1.84_{\pm 0.10}$	$1.01_{\pm 0.06}$	$-1.34_{\pm 0.08}$	$0.83_{\pm 0.07}$	$3.47_{\pm 0.09}$	$0.92_{\pm 0.07}$
0.048	0.052	$1.72_{\pm 0.07}$	$0.77_{\pm 0.05}$	$-1.04_{\pm 0.07}$	$0.73_{\pm 0.06}$	$3.11_{\pm 0.07}$	$0.68_{\pm 0.06}$
0.060	0.037	$1.89_{\pm 0.10}$	$0.99_{\pm 0.06}$	$-1.22_{\pm 0.07}$	$0.69_{\pm 0.06}$	$3.41_{\pm 0.08}$	$0.77_{\pm 0.05}$
0.076	0.046	$1.47_{\pm 0.09}$	$0.88_{\pm 0.07}$	$-1.54_{\pm 0.06}$	$0.63_{\pm 0.07}$	$3.23_{\pm 0.06}$	$0.65_{\pm 0.05}$

Table 3: The four numbers and the effective barrier as the augmentation scale is swept, Qwen2.5-3B-Instruct, $100 \times 50 \times 100$ per σ, Granite judge, $T = 1$. Error bars are ± 1 s.d. from a paired prompt bootstrap (1000 resamples). 'Median rate' is the median single-attempt jailbreak rate $\langle p_{b,a} \rangle_a$. The main text uses $\sigma = 0.048$.

The parameters clearly depend on σ. Higher values of σ progressively erode the safety barriers installed by the model (i.e. lower μ_α and raise μ_β). Eventually too high values of σ hit a ceiling where increased noise destroys too much of the original prompt and the completions start lose coherence. The trends in the 4 parameter family follow the behavior of the median rate, which is the best proxy to determine the 'attack strength'. To be able to compare different models one should not aim at keeping σ fixed, but instead try to match their median jailbreak rates as detailed in App. D.

F AUGMENTATION KERNEL ROBUSTNESS

The main text augments by adding isotropic Gaussian noise to the input embeddings of the user-content tokens which is a white-box operation. A black-box attacker instead perturbs the prompt *string*, which is the character kernel of Table 2. As shown in Table 4, the safety parameters $(\mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta)$ depend on the attack type. However, the general form of our microscopic model (entropic bottleneck, hyperplane, exponential saddles) is universal. In other words, our 4-parameter family fitted on model X attacked via augmentation Y encodes: *how resilient is model X to a BoN attack under Y*. Therefore, this 4-parameter fit allows us to compare different models' safety under the same augmentation, or the susceptibility of a single model to different types of augmentations. Notice however that at approximately equal median rate (character kernel with $\sigma = 0.25$ compared to embedding kernel at $\sigma = 0.030$) the 4-parameter family seem to largely agree (within noise) which might suggest some universality in the way in which the model defends itself against any kind of augmentation.

	median rate	μ_α	σ_α	μ_β	σ_β	μ_E	σ_E
character, $\sigma = 0.25$	0.012	$2.33_{\pm 0.12}$	$1.23_{\pm 0.11}$	$-1.71_{\pm 0.09}$	$1.00_{\pm 0.07}$	$4.28_{\pm 0.13}$	$1.26_{\pm 0.07}$
embedding, $\sigma = 0.048$	0.054	$1.68_{\pm 0.06}$	$0.60_{\pm 0.04}$	$-1.00_{\pm 0.06}$	$0.57_{\pm 0.04}$	$3.02_{\pm 0.05}$	$0.55_{\pm 0.04}$
embedding, $\sigma = 0.030$	0.007	$2.60_{\pm 0.17}$	$1.58_{\pm 0.09}$	$-2.09_{\pm 0.14}$	$1.32_{\pm 0.11}$	$4.91_{\pm 0.14}$	$1.32_{\pm 0.07}$

Table 4: The two augmentation kernels on the same model and the same 100 prompts. Error bars are ± 1 s.d. from a paired prompt bootstrap (1000 resamples).

The structure is universal. Under the character kernel the standardized α_b is still indistinguishable from a Gaussian (KS $p = 0.94$) and standardized $\log \beta_b$ from a normal (KS $p = 0.95$). Reconstructing the full $\mathrm{ASR@}\binom{N}{M}$ surface from the four fitted numbers over ($N \leq 100$, $M \leq 50$) gives a mean absolute error of 0.019 under the character kernel. Therefore, although the parameters may vary, the structure in Sec. 3 is universal across augmentation kernels.

G JUDGE ROBUSTNESS

The notion of safety is dependent on the judge. Hence all the fit parameters $(\mu_\alpha, \sigma_\alpha, \mu_\beta, \sigma_\beta)$ depend on the judge's notion of safety. However, the actual scaling forms, Gaussian and log-normal structures are universal. Every number in the main text is graded by Granite-Guardian-3.0-2B. Because an automated judge defines what counts as a jailbreak, we re-scored the *identical* generations of the Qwen2.5-3B-Instruct grid (400 prompts $\times$ 100 augmentations $\times$ 300 completions, 1.2×10^7 completions) with HarmBench-Llama-2-13b-cls (Mazeika et al., 2024) and repeated the whole estimation. No generation differs between the two columns of Table 5; only the label does.

	Granite-Guardian-3.0-2B	HarmBench-Llama-2-13b-cls
per-completion success rate	0.044	0.003
μ_α (base alignment)	1.83	4.66
σ_α	0.65	1.04
μ_β (susceptibility)	-0.88	-1.53
σ_β	0.52	0.77
μ_E (effective barrier)	3.08	6.43
σ_E	0.58	1.09
prompts never jailbroken	0	2

Table 5: The same generations scored by two judges. HarmBench is far stricter, so every barrier scale shifts, but the *shape* of the model survives: α_b stays Gaussian, $\log \beta_b$ stays log-normal, and the ordering of the two channels is unchanged.

HarmBench calls 0.3% of completions unsafe against Granite's 4.4%, which raises the mean barrier by $\mu_E^{\mathrm{HB}} - \mu_E^{\mathrm{Gr}} = 3.35$ which is expected given the judge is stricter. More importantly, the distribution families are judge-independent. Standardized α_b remains indistinguishable from a Gaussian under both judges, and the susceptibility β_b stays log-normal. The stricter judge widens both spreads (σ_α from 0.65 to 1.04, σ_β from 0.52 to 0.77) because it pushes the hardest prompts further into the tail.