EFFICIENT CALIBRATION VIA REJECTION SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

Exponential Tilting (ET) is a classical technique for model calibration, alignment and reward maximization. For a given statistic (reward) function f, ET transforms a base model distribiution Q via the multiplicative update rule $Q^\alpha \propto Q \cdot \exp(\alpha \cdot f)$. Braverman et al. (ICML 2020) showed that fitting α to minimize cross-entropy loss against some reference distribution, ensures the tilted model does not lose quality, while maximizing the *average* value of f. Unfortunately, ET has a prohibitive computational cost of computing the normalizing partition function, requiring multiple foward passes to evaluate f on the full support of tokens, which renders this technique impreactical in the autoregressive LLM setup.

We propose a natural, highly efficient algorithm for implicitly sampling from the tilted distribution via rejection-sampling, which dramatically reduces the query access to the tilted model and avoids computing the normalizing partition function altogether. We demonstrate the efficiency of our technique in theory and practice, by analyzing the expected query count and upper bounding it in terms of the cross-entropy distance to the base model and the "variance" of f. We then corroborate our bounds on small-size LLMs (Qwen2.5, Qwen3 and SmolLM3), applying our approach to the substantive case study of *entropy calibration* of LLMs (Cao et al., NIPS 2025). Our experiments show improvement in entropy miscalibration in base models, while incurring 15–25% additional sampling attempts due to rejections. We expect this overhead to improve when moving to more capable and larger models.

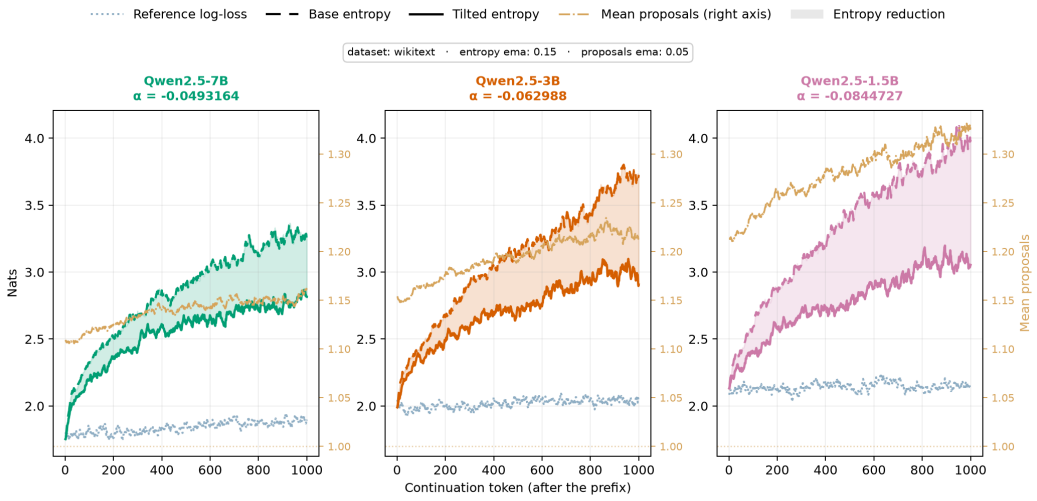

Figure 1: Entropy miscalibration in Qwen2.5 models. The x-axis shows the generation step. The left y-axis shows the average next-token entropy of the model for each step, over responses generated by the model. The dashed line shows the base model, while the full line shows the tilted model. The right y-axis shows the average number of proposals per token. The reference log-loss shows the average surprisal of the model over real responses, for each step.

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.

1 INTRODUCTION

A basic notion in machine learning is statistical calibration: if Q is our model and P is the reference latent distribution, a calibrated model should satisfy $\Pr_{x\sim Q}[f(x) = v] = \Pr_{x\sim P}[f(x) = v]$ for every statistic f and value $v \in \mathbb{R}$. A more modest (but tractable) objective is *moment-matching*, i.e., $\mathbb{E}_Q[f] = \mathbb{E}_P[f]$, which we henceforth call f*-calibration*. Braverman et al. (2020) have shown that a simple–yet intractable–solution exists, given by an *exponential tilt* (ET) of the base model. More precisely, define the α-ET of a model Q as $Q^\alpha(x) = Q(x)\exp(\alpha f(x))/Z(\alpha)$ where $Z(\alpha) = \sum_y Q(y)\exp(\alpha f(y))$ is the partition function (the sum is over the domain of Q). By minimizing the cross-entropy (CE) of Q^α against the reference distribution P, a value α_* is chosen so that Q^{α_*} is also a better model (in the sense of better CE error) and satisfies moment-matching.

Thus, calibration (in the sense of moment-matching) offers a way to overcome one of the prominent trade-offs that appear in the literature: the *quality–diversity* tradeoff. In sampling methods, such as temperature scaling (Ackley et al., 1985), top-k (Fan et al., 2018a; Noarov et al., 2026), top/min-p (Holtzman et al., 2020; Nguyen et al., 2025), η-sampling (Hewitt et al., 2022), selective (Zhang et al., 2020) or top-H (Potraghloo et al., 2026) to mention a few – the probability mass is truncated or sharpened, thus reducing diversity to achieve better quality responses. Similarly, in many post-training reward-based alignment techniques, such as RLHF (Kirk et al., 2024; Ziegler et al., 2020), KL-regularized RL (GX-Chen et al., 2025) or instruction tuning (Zhang et al., 2025), to mention a few – the diversity of results is compromised in order to increase performance in certain tasks. See further discussion in the related work section.

If quality is measured by the statistic f and diversity is measured by the CE error (measuring how well the model covers the reference distribution), then calibration is a clear method to increase quality without compromising diversity. We stress that f can be viewed as a reward function too, and under certain conditions which we discuss in Appendix D, calibration can be viewed as reward maximization. Thus, making calibration tractable might be useful in reward maximization tasks as well.

The problem that remains open is the **computational efficiency** of ET, and specifically computing the partition function. For example, in the auto-regressive language model setup, the model's domain is the vocabulary $\mathcal{V}$, usually consisting of tens of thousands of tokens. Computing f for all tokens might be prohibitive, if it cannot computed at once in parallel. For example, consider the problem of entropy calibration, where $f(w) = H(Q(\cdot \mid w))$ is the next-token-entropy of the model Q after appending the token $w \in \mathcal{V}$ to the current prefix. Computing $f(w)$ requires a forward pass of Q with w appended to the input, making the cost of computing $Z(\alpha)$ insurmountable.

We note that unlike standard post-training techniques, calibration offers a non-invasive and dynamic way to alter the model's behavior at inference time. Thus, allowing efficient access to the ET model at inference time opens up many possibilities for advanced alignment algorithms (including personalization).

1.1 OUR CONTRIBUTIONS

Our main contribution is showing that a very simple and well known algorithm — *rejection sampling* — offers an efficient and practical f-calibration algorithm, under light conditions that we expect to hold in real-world scenarios. We propose using *rejection sampling* (Forsythe, 1972) to sample from the *exact* ET distribution while avoiding computing the partition function. We analyze the method in the context of f-calibration and characterize the expected number of evaluations of f needed to generate an exact ET sample, relating it to the *tilting magnitude* $|\alpha|$ and the distribution of f's values.

To complement our analysis, we show that our approach works well in practice, for the problem of *entropy-calibration*. Following Cao et al. (2025), which showed entropy miscalibration in popular pre-trained models (base models), we test our approach on a similar setup and show that it produces entropy calibrated models, while incurring roughly 10–20% additional sampling attempts due to rejections. We test our approach on the Qwen2.5 (Qwen et al., 2025) model family (1.5B, 3B, 7B), the Qwen3 (Yang et al., 2025) model family (1.7B, 4B, 8B), and SmolLM3-3B (Bakouch et al., 2025) model.

1.2 Paper Organization

In Section 2 we examine related work. In Section 3 we go over preliminaries for the calibration technique. In Section 4 we present our approach and analyze it. In Section 5 we present our experimental results. We finish with a discussion in Section 6.

2 Related Work

Braverman et al. (2020) introduce and study f-moment-matching calibration, and show that entropy might be miscalibrated even though KL divergence and CE are small. Their basic observation is that the KL-divergence doesn't provide control over the value of a bounded function, hence possibly allowing moments mismatch. Cao et al. (2025) show entropy miscalibration persists in modern models and fit scaling laws that suggest scale does not solve miscalibration. They also show how instruction-tuning (Zhang et al., 2025) and temperature scaling (Ackley et al., 1985) fix miscalibration (or create an opposite problem, of too-low entropy), at the cost of diversity.

Braverman et al. (2020) also show calibration can be resolved by exponential tilting (ET). ET can be viewed as the minimal intervention that matches f-moments while keeping the level sets of the distribution unchanged, which makes it the natural solution.

Furthermore, ET is also a fundamental alignment technique, and has been used in multiple prior works. It is the solution to RL algorithms, when restricted to a KL-trust-region (Peters et al., 2010; Khalifa et al., 2021; Korbak et al., 2022a;b), where the optimization problem is finding a model R maximizing a reward $\mathbb{E}_{c\sim R}[f(c)]$ subject to $\mathrm{KL}\left(R\|Q\right) \leq \varepsilon$, where Q is the base model. In the contextual bandits literature (Csiszár & Shields, 2004), it is known that ET is the unique optimal solution to the constrained KL-minimization problem, which is finding a model R minimizing $\mathrm{KL}\left(R\|Q\right)$ subject to moment-matching $\mathbb{E}_R[f] = \mathbb{E}_P[f]$ (against a reference distribution). Since the partition function is expensive to compute, different methods and assumptions have beend used to work around the exponential tilt. We elaborate on this in Appendix A.

Rejection sampling (Forsythe, 1972) has been used in this context by Liu et al. (2024). They solve the problem of sampling preference data from the optimal latent model, without actually having access to that model. Moreover, recent work by Madhow et al. (2026) studies inference-time reward alignment using learned rejection statistics, with applications to diffusion models. While their approach is similar to ours, they focus on an efficiency–approximation tradeoff in sampling from tilted distributions. We instead study how selecting tilts through cross-entropy minimization affects calibration and sampling cost.

3 Preliminaries

3.1 Setup and Notation

Following the setup of (Braverman et al., 2020; Cao et al., 2025), we assume there exists a latent reference distribution P over sequences of tokens. We split documents into a prompt x and a response y, so that $(x, y) \in \mathcal{V}^{|x|+T}$, where $\mathcal{V}$ is the vocabulary and T is the maximal response length. We let Q denote an auto-regressive base model, meaning that $Q(y \mid x) = \prod_{i=1}^{T} Q(y_i \mid xy_{<i})$ where $xy_{<i}$ is the concatenation of x and the prefix of length $i-1$ of y. Thus Q is a model of next-words, and given a partial response, which we call *context* $c = (x, y')$ (where $|y'| < T$), we let Q_c denote the next-token distribution over $\mathcal{V}$ given the context c (i.e., $Q_c(w) = Q(w \mid c)$).

For $0 \leq i \leq T$ and a model R we let $c_i \sim R$ denote a random partial context $(x, y_{\leq i})$ generated by sampling a random prompt $x \sim P$ and sampling a partial completion of length i from R, $y_{\leq i} \sim R(\cdot \mid x)$. We also define the distribution D_R obtained by uniformly drawing $1 \leq i \leq T$ and then drawing $c_{i-1} \sim R$.

The *average entropy* of a model R is defined by $H(R) := \frac{1}{T}\mathbb{E}_{c_T\sim R}[-\log R(y \mid x)]$ (full sequence probability), and the *average entropy at position* $1 \leq t \leq T$ is defined by the next-token entropies $H_t(R) = \mathbb{E}_{c_t\sim R}[-\log Q(y_t \mid c_{t-1})]$. Since the models are auto-regressive, it holds $H(R) = \frac{1}{T}\sum_{t=1}^{T} H_t(R)$. Similarly, the *cross entropy* (CE) error against the reference P is defined by

$\mathcal{L}(P,R) := \frac{1}{T}\mathbb{E}_{c_T\sim P}[-\log R(y \mid x)]$, and the *CE at position t* by $\mathcal{L}_t(P,R) = \mathbb{E}_{c_t\sim P}[-\log R(y_t \mid c_{t-1})]$. We also have $\mathcal{L}(P,R) = \frac{1}{T}\sum_{t=1}^{T}\mathcal{L}_t(P,R)$. At last, the *KL-divergence* of two distributions p, q is defined by $\mathrm{KL}\,(p\|q) = \mathbb{E}_{X\sim p}[\log(p(x)/q(x))]$.

3.2 Calibration

Let f be some statistic, that is a real-valued function over sequences. As auto-regressive models predict the next word, the mean of f for some context c and model R is $\mathbb{E}_{w\sim R_c}[f(c,w)]$. Averaging over contexts drawn from the reference distribution gives the following definition:

$$\mu_R := \mathbb{E}_{c\sim D_P}[\mathbb{E}_{w\sim R_c}[f(c,w)]] = \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{c_t\sim P}[\mathbb{E}_{w\sim R_{c_t}}[f(c_t,w)]].$$

Given the base model Q the goal is to find a model R such that $\mu_R = \mu_P$ (moment matching). Braverman et al. (2020) propose the following simple algorithm. Find α that minimizes $\ell(\alpha) := \mathcal{L}(P, Q^\alpha)$, where Q^α is the ET distribution given by $Q^\alpha(w \mid c) := Q(w \mid c)\cdot e^{\alpha f(c,w)}/Z_c(\alpha)$ where $Z_c(\alpha) = \sum_{w\sim\mathcal{V}} Q(w \mid c)\cdot e^{\alpha f(c,w)}$. Denote the minimizer by α_*, and denote $\mu_\alpha = \mu_{Q^\alpha}$. Then $\mu_{\alpha_*} = \mu_P$, following from the fact $\ell(\alpha)$ is convex, with $\ell'(\alpha) = \mu_\alpha - \mu_P$ and $\ell''(\alpha) = \mathbb{E}_{c\sim d_P}\,\mathrm{Var}_{w\sim Q_c^\alpha}[f(c,w)]$.

In Appendix D, we elaborate on an interpretation of this CE-minimization algorithm as a reward maximization problem in a KL-trust-region around the base model Q.

Remark 1. *The moment defined above, μ_R, is a local moment (also called teacher-forced), in the sense that contexts are taken from P, and only the last word w is taken from R. That is, we treat the model as a next-token model, and not as a model over sequences. This is not the same as global moment matching, where we take prompts from P and the expectation is taken over full responses from R (sequences of length T, instead of a single token). Cao et al. (2025) define entropy calibration as the matching of $H(R) = \mathcal{L}(P,R)$, which is equivalent to global entropy moment matching. In particular, local moment matching is weaker than global moment matching, and we must adjust our expectations when testing for miscalibration. Braverman et al. (2020) show that ET over sequences of length T gives a globally calibrated model, but requires computing a partition function over all sequences, which is completely infeasible.*

4 Rejection Sampling Calibration

Fix a context c and parameter α, and suppose a bound B is cheaply available, satisfying

$$\alpha\cdot f(c,w) \le B$$

for every w where $Q_c(w) > 0$. Consider the following sampling procedure:

Algorithm 1 Exact Rejection Sampling

Input: Sampling access to Q_c, $\alpha\in\mathbb{R}$ and a global bound $B \ge \max_w \alpha\cdot f(c,w)$.
Output: A sample $w\sim Q_c^a$.

1: **while** a word was not yet accepted **do**
2: Sample $w\sim Q_c$ and accept with probability $\frac{\exp(\alpha\cdot f(c,w))}{\exp(B)}$. ▷ Requires an evaluation of f
3: **end while**

Lemma 2. *For a fixed context c, parameter α and a bound B satisfying $\max_w \alpha\cdot f(c,w) \le B$, the sampling procedure above is identical to sampling from Q_c^α (the tilted distribution). Letting $N_\alpha(c)$ denote the number of rounds until acceptance, it is a geometric random variable with mean $\frac{e^B}{Z_c(\alpha)}$.*

The proof is in Appendix B. The lemma shows the partition function affects runtime although it isn't explicitly computed. With the optimal bound $B_c := \max_w \alpha\cdot f(c,w)$, the mean is

$$\mathbb{E}[N_\alpha(c)] = \frac{e^{B_c}}{Z_c(\alpha)} = \max_w \frac{Q_c(w)e^{\alpha f(c,w)}}{Q_c(w)\cdot Z_c(\alpha)} = \left\|\frac{Q_c^\alpha}{Q_c}\right\|_\infty .$$

A loose available bound $B \ge B_c$ multiplies this cost by e^{B-B_c}. We start with two trivial bounds:

Lemma 3. *Suppose $\alpha \leq 0$ and the statistic is non-negative $f \geq 0$. Then:*

1. *(Jensen Bound) Taking $B = 0$, we have $\mathbb{E}[N_\alpha(c)] \leq \exp(|\alpha| \cdot \mathbb{E}_{w\sim Q_c}[f(c,w)])$.*
2. *Using any valid envelope B, for any $t > 0$, we have $\mathbb{E}[N_\alpha(c)] \leq \frac{e^t}{\Pr_{w\sim Q_c}[\alpha \cdot f(c,w) \geq B - t]}$.*

The proof is in Appendix B. We can further bound the number of expected samples using properties of f's distribution over the base and tilted model:

Lemma 4. *Let $\alpha \neq 0$ and B_c denote the optimal rejection bound. Then*

$$\log \|Q_c^\alpha / Q_c\|_\infty = B_c - \alpha \mathbb{E}_{Q_c}[f(c,w)] - \mathrm{KL}\left(Q_c \| Q_c^\alpha\right) = B_c - \alpha \mathbb{E}_{Q_c^\alpha}[f(c,w)] + \mathrm{KL}\left(Q_c^\alpha \| Q_c\right).$$

The proof is in Appendix B. Lemma 3 and Lemma 4 give complementary explanations for a small proposal count. On the one hand, if the mean $\mathbb{E}_{w\sim Q_c}[f(c,w)]$ is small, Lemma 3 guarantees a small proposal count. On the other hand, if the mean is large but close to the upper bound (concentrated around the bound), Lemma 4 guarantees a small proposal count. It is important to note that both bounds are multiplicative in $|\alpha|$, so 'small' has relative meaning, as the scale of α and f can compensate for each other.

We next relate the tilt magnitude to the residual calibration loss. Recall that $\ell''(u) = \mathbb{E}_{c\sim d_P} \mathrm{Var}_{Q_c^u}[f(c,w)]$ is the average conditional variance.

Lemma 5. *Let α lie between 0 and α_*, and set $\varepsilon = \ell(\alpha) - \ell(\alpha_*)$. If $0 < \ell''(u) \leq V_{\max}$ between α and α_*, then*

$$\varepsilon \leq \frac{V_{\max}}{2}|\alpha - \alpha_*|^2, \qquad |\alpha| \leq |\alpha_*| - \sqrt{2\varepsilon / V_{\max}}.$$

Therefore, $|\ell'(\alpha)| \leq \sqrt{2 V_{\max} \varepsilon}$.

The proof is in Appendix B. The main take-away from the lemma is that the residual log-loss $\varepsilon = \ell(\alpha) - \ell(\alpha_*) \leq \ell(0) - \ell(\alpha_*)$ controls the magnitude of a possible solution α that is ε-sub-optimal (in log-loss). If the reference distribution matches (at least partially) the distribution that the base model Q was trained on, we expect little room for CE error improvement. Thus, for a well behaved statistic (with small average variance), the lemma quantifies how stopping ε-short may give a tilt α which is possibly much smaller in magnitude than α_*, leading to smaller proposal counts in the rejection sampling approach.

For the special case of non-negative $f \geq 0$ and $\alpha_* < 0$, we combine the previous results into the following theorem.

Theorem 6. *Suppose $f \geq 0$, $\alpha_* < 0$, and $0 < \ell''(u) \leq V_{\max}$ on $[\alpha_*, 0]$. Fix $\alpha \in [\alpha_*, 0]$ and write $\varepsilon = \ell(\alpha) - \ell(\alpha_*)$. Rejection sampling with $B = 0$ samples exactly from Q_c^α, with*

$$\mathbb{E}[N_\alpha(c)] \leq \exp\left(\mathbb{E}_{Q_c}[f(c,w)]\left(|\alpha_*| - \sqrt{2\varepsilon / V_{\max}}\right)\right).$$

Proof. Combine Lemma 5 with the Jensen bound in Lemma 3. □

Choosing a Small Tilt in Practice. A simple way to favor small magnitudes is to start at zero and stop when the derivative is sufficiently small. A simple bisection approach leads to the smallest-magnitude α with derivative that is sufficiently small. Since $\ell'(\alpha) = \mu_\alpha - \mu_P$, the stopping criterion directly controls moment-calibration error. A small available CE improvement and bounded variance can put the starting point (zero) close to the stopping threshold, without requiring optimization all the way to α_*. For $f \geq 0$ and $B = 0$, the expected proposal count increases as the negative tilt has larger magnitude, so the smallest-magnitude feasible tilt is also the cheapest.

4.1 Application to Entropy Calibration

Language models commonly exhibit an upward entropy bias: the next-token entropy on model-generated text is higher than that of true text (Braverman et al., 2020; Cao et al., 2025). For the statistic $f(c,w) = H(Q_{cw})$, we consider the miscalibrated regime $\mu_0 > \mu_P$ on reference contexts. Under this assumption, $\ell'(0) > 0$ and the CE-minimizing tilt is negative. Since entropy is nonnegative, every $\alpha \leq 0$ satisfies $\alpha \cdot H(Q_{cw}) \leq 0$, so $B = 0$ is a valid rejection bound, available without any additional computation.

4.2 Clipping and Approximate Rejection Sampling

As we've shown above, exact rejection sampling costs the maximal ratio $Q_c^a(w)/Q_c(w)$ on average, potentially making the approach too expensive. A possible solution in this case is clipping the upper tail of the distribution. For a fixed context c and cutoff $\tau > 1$, let R_τ be obtained by clipping $Q_c^\alpha(w)$ at $\tau Q_c(w)$ and renormalizing. In Appendix B we prove Lemma 8, which is based on a tail bound of Braverman (2017), and gives us

$$d_{\mathrm{TV}}(R_\tau, Q_c^\alpha) \leq \min\left\{1, \frac{\mathrm{KL}\left(Q_c^\alpha \| Q_c\right) + \log 2}{\log \tau}\right\}.$$

Thus, when $\log(\tau)$ is large compared to the KL divergence, the cutoff affects only a small mass of the tilted distribution, hence reducing the proposal count substantially. However, the clipping procedure we discuss here cannot be efficiently implementable, as it requires evaluating the partition function. We leave this for future work.

5 Experiments

5.1 Goals and Setup

We design an experiment very similar to the experiments in Cao et al. (2025). Our goal is two-fold: show that our approach **improves** miscalibration, and measure the sampling efficiency. We do not measure the performance overhead, but only the sample efficiency, as performance proportional to the sampling overhead requires a careful implementation, reminiscent of implementing speculative decoding (Leviathan et al., 2023; Chen et al., 2023). Moreover, as discussed in Remark 1, we do not expect this approach to completely fix miscalibration, as we measure it in a global manner, like in Cao et al. (2025).

The basic flow of the experiments is: (1) Fit α that minimizes log-loss; (2) Evaluate the average *base* model entropy for each generation step, averaged over text generated by the base model (*base trajectories*); (3) Evaluate the average *tilted* model entropy for each generation step, averaged over text generated by the tilted model (*tilted trajectories*).

We perform the following experiments:

1. We evaluate Qwen2.5 (Qwen et al., 2025) models of size 1.5B, 3B and 7B, on the task of completing passages from the `WikiText-103` dataset (Merity et al., 2016). This can be directly compared with results of Cao et al. (2025).
2. We evaluate Qwen3 (Yang et al., 2025) models of sizes 1.7B, 4B and 8B, on the task of writing story completions from the `WritingPrompts` dataset (Fan et al., 2018b).
3. We evaluate SmolLM3-3B (Bakouch et al., 2025) on the task of completing passages from `WikiText-103`.

We use base models (no instruction tuning), and default (usually `bfloat16`) accuracy. The inference library we use is HuggingFace `transformers` Wolf et al. (2020) library, with a custom generator loop for the rejection sampling implementation. Our code will be made available in a dedicated `GitHub` repository[1]. In our tests we used a single Nvidia L40s GPU. We next elaborate on the implementation of each step.

5.1.1 Calibration Step

To fit α for a model, we compute the CE error empirically, making it into a log-loss

$$\hat{\mathcal{L}}(P, Q^\alpha) = -\frac{1}{N}\sum_{i=1}^{N} \log Q^\alpha(y^{(i)} \mid c^{(i)}),$$

where $(c^{(i)}, y^{(i)})_{i=1}^N$ are sampled from the reference distribution (c is a context and $y \in \mathcal{V}$ is the true next token). In our experiment, the reference distribution is taken from the `WikiText-103` or

[1] For the review process, it is attached in the supplementary materials.

`WritingPrompts` datasets. We split documents into evaluation and calibration bins, and use the calibration bin to compute the log loss. In particular, we let the calibration subset consist of 200 documents and from each document we draw 50 contexts starting at random positions.

However, computing $\log Q^\alpha$ requires computing the partition function, which we wish to avoid. The derivative of $\hat{\mathcal{L}}(P, Q^\alpha)$ with respect to α is $\frac{1}{N}\sum_i [\mathbb{E}_{w\sim Q^\alpha(\cdot|c^{(i)})}[H(Q_{c^{(i)}w})] - H(Q_{c^{(i)}y^{(i)}})]$. We estimate the expectation term by sampling R tokens $w_{i,1}\ldots, w_{i,R}$ from $Q^\alpha(\cdot \mid c^{(i)})$ for each $i \in [N]$, implemented via rejection sampling. This gives us an estimate for $\hat{\mathcal{L}}'$:

$$\tilde{\mathcal{L}}'(P, Q^\alpha) = \frac{1}{NR}\sum_{i=1}^{N}\sum_{r=1}^{R} H(Q_{c^{(i)}w_{i,r}}) - \frac{1}{N}\sum_{i=1}^{N} H(Q_{c^{(i)}y^{(i)}}).$$

We then use the bisection approach discussed previously to optimize α.

5.1.2 Evaluation Step

We evaluate each model in its tilted and non-tilted version, by generating completions of up to 1000 tokens. For each document of the evaluation subset, we prompt the model with the first 128 tokens and continue decoding until the special EOS token is reached. The evaluation subset is comprised of 1000 randomly chosen documents. At each step, we record the entropy of next token distribution. In tilted models, we use the rejection sampling approach with $B = 0$, and record the number of tokens drawn until acceptance.

For each evaluated model $R \in \{Q, Q^\alpha\}$, we plot its next-token entropy on contexts generated by R, together with the average acceptance rate (for the tilted model), taking averages separately at each generation step. In the fixed-horizon notation of the preliminaries, the corresponding population entropy is $H_t(R) = \mathbb{E}_{c_{t-1}\sim R}[H(R_{c_{t-1}})]$

Remark 7. *Exactly computing the next-token entropy of the tilted model $H(Q_c^\alpha)$ is too computationally expensive. Instead, we estimate the average difference of $H(Q_c) - H(Q_c^\alpha)$. We elaborate on this in Appendix C.*

We also plot the reference log-loss, which is the average log-loss of the base model Q over the true continuations of the evaluation documents.

5.2 Results

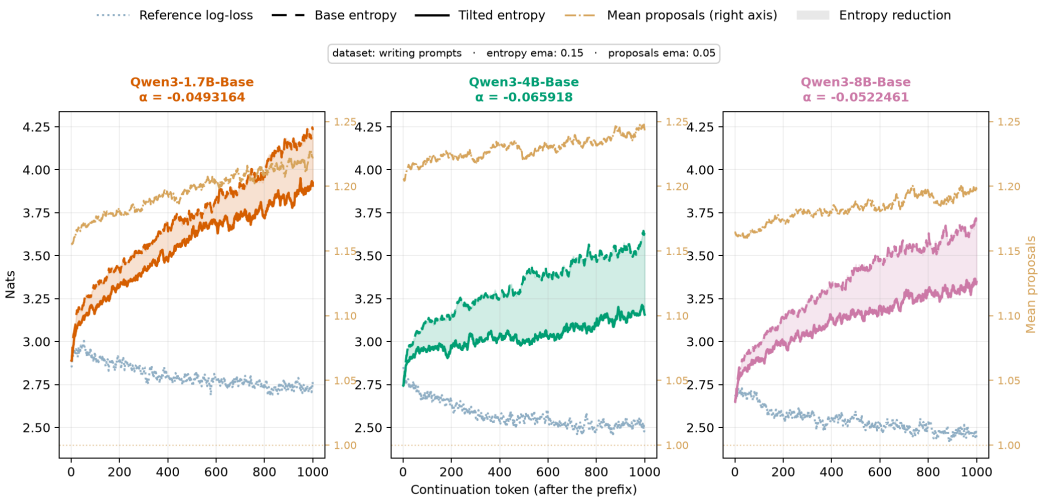

Figure 2: Results for Qwen3 models.

We describe our results in figures 1, 2 and 3. Based on the plots, we observe the following:

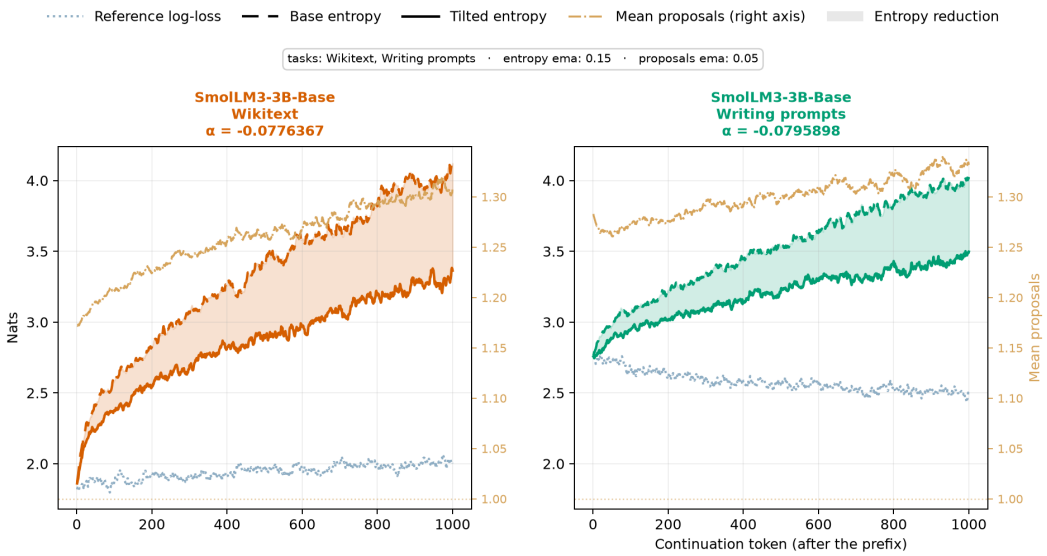

Figure 3: Results for SmolLM3-3B on two different tasks.

- The average entropy drops in all the tasks and models, but it's still not at the level of the reference log-loss. This shows that *local* entropy calibration transfers partially to *global* entropy calibration, but with varying efficiency. Some models exhibit more transfer than others.
- The rejection sampling is efficient, and becomes more efficient when the entropy levels are lower. In particular, the more capable models (larger models) exhibit better rejection efficiency.
- The SmolLM3-3B model tends to exhibit much higher entropy compared with other models (reaching 3.5 nats quickly). Correspondingly, we see that the performance of rejection sampling is worse (25-35% overhead).

Furthermore, in our tests we exhibited that base-entropy and the estimated tilted entropy are very close in practice. This is expected of course, as we elaborate in Appendix C. We also exhibited complete entropy collapse in instruction and post-trained model versions, similar to the phenomena described by Cao et al. (2025). Inspecting the generated responses of these models suggests that the entropy collapse is due to the models refusing to complete text, falling back to internal reasoning, asking questions, answering their own generated questions, and repeated mentioning of the user's requests (although no such requests were made).

6 Discussion

Following the theoretical and experimental results, we wish to highlight three broader implications of our approach.

First, our results show that *local calibration* (see Remark 1) partially transfers to global calibration, at least in the case of entropy calibration. In our experiments, global calibration would be interpreted as the average entropy staying close to the log-loss, as discussed in Cao et al. (2025). Establishing conditions under which such transfer happens, remains an important question for future study. It is particularly interesting following our work, as local calibration becomes practical using our approach.

Second, the main limitation of our approach is that it might fail to be efficient when a very large tilt is required, or when the statistic has an unfavorable distribution. Coming up with approximate methods, for example some practical implementation of the clipping method (see Lemma 8), may open up the possibility to overcome unfavorable distributions.

Third, we re-iterate that calibration greatly depends on the reference distribution which describes some latent task. Calibration cannot endow the model with new capabilities, or correct problems in the model's understanding, but it can help the model to hone internal capabilities. Moreover, the fact it is targeted and dynamic implies it can be applied dynamically on the fly. Future work into simultaneous calibration of multiple statistics is a direction we find very interesting, as it allows composing preferences based on the task at hand. Furthermore, instruction tuned models may not suffer from entropy miscalibration (as shown by Cao et al. (2025) and exhibited in our tests too) due to loss of diversity, but that does not rule out other tasks on which tuned models suffer miscalibration. It would be interesting to identify such tasks in order to fix such miscalibration. For example, we predict that over-empathy Welivita et al. (2026) or sycophancy Hong et al. (2025) could be improved with our approach. Of course, many more examples are likely to be true.

AI Statement

We have used OpenAI Codex extensively in our coding efforts. Mathematical proofs were partially proposed by OpenAI ChatGPT, which was also used very lightly for searching related work. The draft was not written by AI, aside from minor edits and formatting. All claims, code and references were validated by the authors, who take full responsibility.

References

David H. Ackley, Geoffrey E. Hinton, and Terrence J. Sejnowski. A learning algorithm for boltzmann machines. *Cogn. Sci.*, 9:147–169, 1985. URL https://api.semanticscholar.org/CorpusID:12174018.

Elie Bakouch, Loubna Ben Allal, Anton Lozhkov, Nouamane Tazi, Lewis Tunstall, Carlos Miguel Patiño, Edward Beeching, Aymeric Roucher, Aksel Joonas Reedi, Quentin Gallouédec, Kashif Rasul, Nathan Habib, Clémentine Fourrier, Hynek Kydlicek, Guilherme Penedo, Hugo Larcher, Mathieu Morlon, Vaibhav Srivastav, Joshua Lochner, Xuan-Son Nguyen, Colin Raffel, Leandro von Werra, and Thomas Wolf. SmolLM3: smol, multilingual, long-context reasoner. https://huggingface.co/blog/smollm3, 2025.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952. ISSN 00063444, 14643510. URL http://www.jstor.org/stable/2334029.

Mark Braverman. Interactive information complexity. *SIAM Review*, 59(4):803–846, 2017. doi: 10.1137/17M1139254.

Mark Braverman, Xinyi Chen, Sham Kakade, Karthik Narasimhan, Cyril Zhang, and Yi Zhang. Calibration, entropy rates, and memory in language models. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1089–1099. PMLR, 2020. URL https://proceedings.mlr.press/v119/braverman20a.html.

Steven Cao, Gregory Valiant, and Percy Liang. On the entropy calibration of language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=CGLoEvCllI.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling, 2023. URL https://arxiv.org/abs/2302.01318.

Imre Csiszár and Paul C. Shields. Information theory and statistics: A tutorial. *Foundations and Trends in Communications and Information Theory*, 1(4):417–528, 2004. doi: 10.1561/0100000004.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao (eds.), *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 889–898, Melbourne, Australia, July 2018a. Association for Computational Linguistics. doi: 10.18653/v1/P18-1082. URL https://aclanthology.org/P18-1082/.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation, 2018b. URL https://arxiv.org/abs/1805.04833.

George E. Forsythe. Von neumann's comparison method for random sampling from the normal and other distributions. *Mathematics of Computation*, 26(120):817–826, 1972. ISSN 00255718, 10886842. URL http://www.jstor.org/stable/2005864.

Anthony GX-Chen, Jatin Prakash, Jeff Guo, Rob Fergus, and Rajesh Ranganath. Kl-regularized reinforcement learning is designed to mode collapse, 2025. URL https://arxiv.org/abs/2510.20817.

John Hewitt, Christopher Manning, and Percy Liang. Truncation sampling as language model desmoothing. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 3414–3427, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.249. URL https://aclanthology.org/2022.findings-emnlp.249/.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration, 2020. URL https://arxiv.org/abs/1904.09751.

Jiseung Hong, Grace Byun, Seungone Kim, and Kai Shu. Measuring sycophancy of language models in multi-turn dialogues. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 2239–2259. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.121. URL http://dx.doi.org/10.18653/v1/2025.findings-emnlp.121.

James Y. Huang, Sailik Sengupta, Daniele Bonadiman, Yi-An Lai, Arshit Gupta, Nikolaos Pappas, Saab Mansour, Katrin Kirchhoff, and Dan Roth. Deal: Decoding-time alignment for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 26280–26300. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1274. URL http://dx.doi.org/10.18653/v1/2025.acl-long.1274.

Muhammad Khalifa, Hady Elsahar, and Marc Dymetman. A distributional approach to controlled text generation, 2021. URL https://arxiv.org/abs/2012.11635.

Maxim Khanov, Jirayu Burapacheep, and Yixuan Li. Args: Alignment as reward-guided search, 2024. URL https://arxiv.org/abs/2402.01694.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of rlhf on llm generalisation and diversity, 2024. URL https://arxiv.org/abs/2310.06452.

Tomasz Korbak, Hady Elsahar, Germán Kruszewski, and Marc Dymetman. On reinforcement learning and distribution matching for fine-tuning language models with no catastrophic forgetting, 2022a. URL https://arxiv.org/abs/2206.00761.

Tomasz Korbak, Ethan Perez, and Christopher Buckley. RL with KL penalties is better viewed as Bayesian inference. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022*, pp. 1083–1091, Abu Dhabi, United Arab Emirates, December 2022b. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-emnlp.77. URL https://aclanthology.org/2022.findings-emnlp.77/.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding, 2023. URL https://arxiv.org/abs/2211.17192.

Tianqi Liu, Yao Zhao, Rishabh Joshi, Misha Khalman, Mohammad Saleh, Peter J. Liu, and Jialu Liu. Statistical rejection sampling improves preference optimization, 2024. URL https://arxiv.org/abs/2309.06657.

Sunil Madhow, Yuchen Liang, Ness Shroff, Yingbin Liang, and Yu-Xiang Wang. Learnable chernoff baselines for inference-time alignment, 2026. URL https://arxiv.org/abs/2602.07738.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models, 2016. URL https://arxiv.org/abs/1609.07843.

Minh Nhat Nguyen, Andrew Baker, Clement Neo, Allen Roush, Andreas Kirsch, and Ravid Shwartz-Ziv. Turning up the heat: Min-p sampling for creative and coherent llm outputs, 2025. URL https://arxiv.org/abs/2407.01082.

Georgy Noarov, Soham Mallick, Tao Wang, Sunay Joshi, Yan Sun, Yangxinyu Xie, Mengxin Yu, and Edgar Dobriban. Foundations of top-k decoding for language models, 2026. URL https://arxiv.org/abs/2505.19371.

Jan Peters, Katharina Mülling, and Yasemin Altün. Relative entropy policy search. In *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence*, AAAI'10, pp. 1607–1612. AAAI Press, 2010.

Erfan Baghaei Potraghloo, Seyedarmin Azizi, Souvik Kundu, and Massoud Pedram. Top-h decoding: Adapting the creativity and coherence with bounded entropy in text generation, 2026. URL https://arxiv.org/abs/2509.02510.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. URL https://arxiv.org/abs/2412.15115.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Anuradha Welivita, Fawzia Zeitoun, and Pearl Pu. Who generates more empathetic responses—humans or LLMs? a comparative evaluation with human and LLM judges. In *30th Conference on Computational Natural Language Learning*, 2026. URL https://openreview.net/forum?id=tPjenXRAtT.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. URL https://arxiv.org/abs/1910.03771.

Yuancheng Xu, Udari Madhushani Sehwag, Alec Koppel, Sicheng Zhu, Bang An, Furong Huang, and Sumitra Ganesh. Genarm: Reward guided generation with autoregressive reward model for test-time alignment, 2025. URL https://arxiv.org/abs/2410.08193.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Hugh Zhang, Daniel Duckworth, Daphne Ippolito, and Arvind Neelakantan. Trading off diversity and quality in natural language generation, 2020. URL https://arxiv.org/abs/2004.10450.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and Guoyin Wang. Instruction tuning for large language models: A survey, 2025. URL https://arxiv.org/abs/2308.10792.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2020. URL https://arxiv.org/abs/1909.08593.

A Extended Related Work

Exponential Tilting in RL literature The KL-trust-region RL problem (Peters et al., 2010; Khalifa et al., 2021; Korbak et al., 2022a;b) is defined as:

$$R^* = \text{argmax}_R \, \mathbb{E}_R[f] \quad \text{subject to} \quad \text{KL}\left(R \| Q\right) \leq \varepsilon$$

Here Q is the base model, f is the reward function, and $\mathbb{E}_R$ is the expectations over generations sampled from a model R. The optimal solution is an ET. The usual approaches that give feasible solutions tend to avoid ET.

The usual solution directly optimizes over the model's parameter space (fine-tuning), which recovers an approximation of the exponential tilt, as not every tilt can be implemented in the model's parameter space. Moreover, this technique does not provide any control over the CE-loss, potentially leading to mode collapse (GX-Chen et al., 2025). The upside of this approach is that it side-steps the need to compute the partition function and reward function at **inference** time. However, full fine-tuning is potentially very expensive and limited to pre-defined reward functions, lacking the ability to support many reward functions or a dynamic context-dependent application of those adjustments.

Another training-time solution worth mentioning is the direct preference optimization algorithm (DPO) (Rafailov et al., 2024), that allows fine-tuning a model without ever computing the reward function f explicitly, and relies on the function satisfying the Bradley & Terry (1952) model. In this case, the Bradley-Terry model allows computing the tilted distribution without ever computing the partition function.

Test-time approaches to RL like Xu et al. (2025); Huang et al. (2025); Khanov et al. (2024) attempt to allow direct access to an ET model, based on the reward function f. Xu et al. (2025) make the problem of computing the partition function easy by making f auto-regressive, so $f(w)$ is obtained for all $w \in \mathcal{V}$ at the cost of one. However, as shown by Braverman et al. (2020), the one-step look-ahead is essential in some cases, and many interesting reward functions do not admit a clean auto-regressive formulation. Other works, like Khanov et al. (2024) and Huang et al. (2025), use sampling and search techniques that go way into the future, in order to approximate the distribution, leading to inaccuracies or large implementation overheads.

B Proofs

Proof of Lemma 2. Let $W \sim Q_c$ be the random token we sample, then it is accepted (independently) with probability $e^{\alpha f(c,W)-B}$. Thus for fixed $w \in \mathcal{V}$, we have $\Pr_W[W = w, \text{accepted}] = Q_c(w)e^{\alpha f(c,w)-B}$. Summing over $w \in \mathcal{V}$ shows that the probability of acceptance is $e^{-B}Z_c(\alpha)$. Therefore,

$$\Pr_W[W = w \mid \text{accepted}] = \frac{\Pr_W[W = w, \text{accepted}]}{\Pr_W[\text{accepted}]} = Q_c(w)e^{\alpha f(c,w)-B}\frac{e^B}{Z_c(\alpha)} = Q_c^\alpha(w).$$

□

Proof of Lemma 3. By Jensen's inequality, $Z_c(\alpha) = \mathbb{E}_{Q_c}[\exp(\alpha f(c,w))] \geq \exp(\alpha \mathbb{E}_{Q_c}[f(c,w)])$. Taking the inverse and using $\alpha \leq 0$ gives the first item. For the second item, note that

$$\mathbb{E}_{w\sim Q_c}[e^{\alpha f(c,w)}] \geq \mathbb{E}_{w\sim Q_c}[e^{\alpha f(c,w)} \cdot 1_{\{\alpha f(c,w) \geq B-t\}}] \geq e^{B-t} \Pr_{w\sim Q_c}[\alpha f(c,w) \geq B - t].$$

Thus, $e^B/Z_c(\alpha) \leq e^t / \Pr_{w\sim Q_c}[\alpha f(c,w) \geq B - t]$. □

Proof of Lemma 4. By the definition of B_c, we have $\log \|Q_c^\alpha / Q_c\|_\infty = B_c - \log Z_c(\alpha)$. Expanding the tilt in each KL gives

$$\text{KL}\left(Q_c \| Q_c^\alpha\right) = \log Z_c(\alpha) - \alpha \mathbb{E}_{Q_c}[f(c,w)], \quad \text{KL}\left(Q_c^\alpha \| Q_c\right) = \alpha \mathbb{E}_{Q_c^\alpha}[f(c,w)] - \log Z_c(\alpha).$$

Substitution gives both identities. □

Proof of Lemma 5. Taylor's theorem and $\ell'(\alpha_*) = 0$ give

$$\varepsilon = \ell(\alpha) - \ell(\alpha_*) = \int_{|\alpha|}^{|\alpha_*|} (u - |\alpha|)\ell''(\operatorname{sign}(\alpha_*)u)\,\mathrm{d}u.$$

Bounding the integrand proves the distance inequalities. Moreover, $|\ell'(\alpha)| \le V_{\max}|\alpha - \alpha_*|$, so $\alpha - \ell'(\alpha)/V_{\max}$ lies on the same segment. Taylor's upper bound at this point gives

$$\ell(\alpha_*) \le \ell\big(\alpha - \ell'(\alpha)/V_{\max}\big) \le \ell(\alpha) - \frac{|\ell'(\alpha)|^2}{2V_{\max}},$$

which proves the gradient bound. □

Lemma 8 (Clipping error)**.** *Fix a context c and a tilt parameter α, and write $K_c := \mathrm{KL}\,(Q_c^\alpha \| Q_c)$. For a cutoff $\tau > 1$, define the removed mass and the normalized clipped distribution by*

$$m_\tau := \sum_{w \in \mathcal{V}} \max\{Q_c^\alpha(w) - \tau Q_c(w), 0\}, \quad R_\tau(w) := \frac{\min\{Q_c^\alpha(w), \tau Q_c(w)\}}{1 - m_\tau}.$$

Then $0 \le m_\tau < 1$, and

$$d_{\mathrm{TV}}(R_\tau, Q_c^\alpha) \le m_\tau \le \Pr_{w \sim Q_c^\alpha}\left[\frac{Q_c^\alpha(w)}{Q_c(w)} > \tau\right] \le \min\left\{1, \frac{K_c + \log 2}{\log \tau}\right\}.$$

In particular, for any $0 < \delta < 1$, choosing $\tau = \exp((K_c + \log 2)/\delta)$ ensures $d_{\mathrm{TV}}(R_\tau, Q_c^\alpha) \le \delta$.

Proof. The clipped measure has total mass $1 - m_\tau$. Since an exponential tilt preserves the support of Q_c, this mass is positive, so R_τ is a probability distribution. If $m_\tau = 0$, then $R_\tau = Q_c^\alpha$. Otherwise, define

$$S_\tau(w) := \frac{\max\{Q_c^\alpha(w) - \tau Q_c(w), 0\}}{m_\tau}.$$

This is a probability distribution, and $Q_c^\alpha = (1 - m_\tau)R_\tau + m_\tau S_\tau$. Thus

$$d_{\mathrm{TV}}(R_\tau, Q_c^\alpha) = m_\tau d_{\mathrm{TV}}(R_\tau, S_\tau) \le m_\tau.$$

Only tokens satisfying $Q_c^\alpha(w) > \tau Q_c(w)$ contribute to m_τ, and their removed mass is at most their target probability. Hence

$$m_\tau \le \Pr_{w \sim Q_c^\alpha}\left[\frac{Q_c^\alpha(w)}{Q_c(w)} > \tau\right].$$

By Braverman (2017, Claim 5.1), expressed in natural logarithms, for every $0 < \delta < 1$,

$$\Pr_{w \sim Q_c^\alpha}\left[\frac{Q_c^\alpha(w)}{Q_c(w)} > \exp\left(\frac{K_c + \log 2}{\delta}\right)\right] < \delta.$$

Indeed, the cited claim uses base-two divergence $I = K_c/\log 2$ and threshold $2^{(I+1)/\delta}$, which equals the exponential threshold above. If $(K_c + \log 2)/\log \tau < 1$, apply the claim with $\delta = (K_c + \log 2)/\log \tau$. Otherwise, use the trivial probability bound of one. Combining these bounds proves the result and its stated choice of cutoff. □

C Estimating the Mean Tilted Entropy

At generation step t, let $c_1, \ldots, c_N$ be the contexts generated by the tilted model that are still active. We wish to estimate

$$\bar{H}_N(\alpha) := \frac{1}{N} \sum_{i=1}^{N} H(Q_{c_i}^\alpha). \tag{1}$$

Computing any one of these entropies exactly requires evaluating the tilt on the entire vocabulary. In contrast, the base entropies $H(Q_{c_i})$ are available from the logits already computed during generation. We therefore compute their mean exactly and estimate only the correction $H(Q_{c_i}) - H(Q_{c_i}^\alpha)$.

Write $\beta = -\alpha \ge 0$, $L = \log|\mathcal{V}|$, $s_c(w) = -\log Q_c(w)$, and $f(c, w) = H(Q_{cw}) \in [0, L]$.

C.1 Entropy Correction Formula

We start with the following simple lemma:

Lemma 9. *For every context c and $\gamma \in [0, \beta]$, we have*

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} H(Q_c^{-\gamma}) = -\operatorname{Cov}_{Q_c^{-\gamma}}\big(s_c + \gamma \cdot f(c,\cdot), f(c,\cdot)\big).$$

Proof. Differentiating the log partition function gives

$$\frac{\mathrm{d}}{\mathrm{d}\gamma} \log Z_c(-\gamma) = \frac{\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_c(-\gamma)}{Z_c(-\gamma)} = \frac{\sum_{w\in\mathcal{V}} Q_c(w) \cdot e^{-\gamma f(c,w)} \cdot (-f(c,w))}{Z_c(-\gamma)} = -\mathbb{E}_{Q_c^{-\gamma}}[f(c,\cdot)].$$

Hence $\frac{\mathrm{d}}{\mathrm{d}\gamma} Z_c(-\gamma) = -Z_c(-\gamma) \cdot \mathbb{E}_{Q_c^{-\gamma}}[f(c,\cdot)]$, and for every function $g : \mathcal{V} \to \mathbb{R}$,

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}\gamma}\mathbb{E}_{Q_c^{-\gamma}}[g] &= \sum_w \frac{\mathrm{d}}{\mathrm{d}\gamma}\left(\frac{Q_c(w)\cdot\exp(-\gamma f(c,w))\cdot g(w)}{Z_c(-\gamma)}\right) \\
&= \frac{1}{Z_c(-\gamma)^2}\sum_w Q_c(w)g(w)e^{-\gamma f(c,w)}\cdot\left(-Z_c(-\gamma)f(c,w) - \frac{\mathrm{d}}{\mathrm{d}\gamma}Z_c(-\gamma)\right) + \mathbb{E}_{Q_c^{-\gamma}}\left[\frac{\mathrm{d}}{\mathrm{d}\gamma}g\right] \\
&= -\mathbb{E}_{Q_c^{-\gamma}}[g\cdot f(c,\cdot)] + \mathbb{E}_{Q_c^{-\gamma}}[g]\cdot\mathbb{E}_{Q_c^{-\gamma}}[f(c,\cdot)] + \mathbb{E}_{Q_c^{-\gamma}}\left[\frac{\mathrm{d}}{\mathrm{d}\gamma}g\right] \\
&= -\operatorname{Cov}_{Q_c^{-\gamma}}(g, f(c,\cdot)) + \mathbb{E}_{Q_c^{-\gamma}}\left[\frac{\mathrm{d}}{\mathrm{d}\gamma}g\right].
\end{aligned}$$

Finally, the proof is finished by noting that

$$H(Q_c^{-\gamma}) = \mathbb{E}_{Q_c^{-\gamma}}\left[-\log\left(\frac{Q_c(w)e^{-\gamma f(c,w)}}{Z_c(-\gamma)}\right)\right] = \mathbb{E}_{Q_c^{-\gamma}}[s_c + \gamma\cdot f(c,w)] + \log Z_c(-\gamma),$$

implying that (setting $g = s_c + \gamma f(c,\cdot)$, so $\frac{\mathrm{d}}{\mathrm{d}\gamma}g = f(c,\cdot)$):

$$\frac{\mathrm{d}}{\mathrm{d}\gamma}H(Q_c^{-\gamma}) = -\operatorname{Cov}(s_c + \gamma f(c,\cdot), f(c,\cdot)) + \mathbb{E}_{Q_c^{-\gamma}}[f(c,\cdot)] - \mathbb{E}_{Q_c^{-\gamma}}[f(c,\cdot)] = -\operatorname{Cov}(s_c + \gamma f(c,\cdot), f(c,\cdot)).$$

□

Corollary 10. *It holds*

$$H(Q_c) - H(Q_c^a) = \int_0^\beta \operatorname{Cov}_{Q_c^{-b}}\big(s_c + bf(c,\cdot), f(c,\cdot)\big)\,\mathrm{d}b. \tag{2}$$

Proof. Integrate over $[0, \beta]$. □

C.2 Estimating the Covariance Integral

The covariance in Equation (2) has a simple unbiased estimator:

1. Draw $\gamma \sim \mathrm{Unif}[0, \beta]$.
2. Conditional on γ, draw $W_1, W_2 \sim Q_c^{-\gamma}$ independently by rejection sampling.
3. Compute $f_j = f(c, W_j)$ for $j = 1, 2$, and define

$$Y_c := \frac{1}{2}\big[s_c(W_1) - s_c(W_2) + \gamma(f_1 - f_2)\big](f_1 - f_2). \tag{3}$$

Lemma 11. *Given γ, Y_c is an unbiased estimator:*

$$\mathbb{E}[Y_c \mid \gamma] = \operatorname{Cov}_{Q_c^{-\gamma}}\big(s_c + \gamma f(c,\cdot), f(c,\cdot)\big).$$

Therefore

$$H(Q_c^\alpha) = H(Q_c) - \beta \mathbb{E}[Y_c]. \tag{4}$$

This avoids both the partition function and a full-vocabulary evaluation of f. It also estimates only the change from the known base entropy. In particular, the Monte Carlo error vanishes with the tilt:

Lemma 12. *Let $V(c) := \mathrm{Var}_{Q_c}(s_c)$ (can be computed from the base model's logits), then*

$$\mathbb{E}[(\beta Y_c)^2] \le \frac{\beta^2 L^2}{2} \left(e^{\beta L/2} \sqrt{V(c)} + \frac{\beta L}{2} \right)^2. \tag{5}$$

Proof. Note $Q_c^{-\gamma}(w)/Q_c(w) \le e^{\gamma L}$ and $\mathrm{Var}_{Q_c^{-\gamma}}(f(c,\cdot)) \le L^2/4$. Thus using the triangle inequality for standard deviations, we obtain

$$\mathrm{Var}_{Q_c^{-\gamma}}(s_c + \gamma f(c,\cdot)) \le \left(\sqrt{\mathrm{Var}_{Q_c^{-\gamma}}(s_c)} + \sqrt{\mathrm{Var}_{Q_c^{-\gamma}}(\gamma f(c,\cdot))} \right)^2 \le \left(e^{bL/2} \sqrt{V(c)} + \frac{bL}{2} \right)^2.$$

Combining this with $|f_1 - f_2| \le L$ and $\mathbb{E}[(X_1 - X_2)^2] = 2\,\mathrm{Var}(X_1)$ for independent, identically distributed X_1, X_2 proves Equation (5). □

Assuming we have a uniform bound (over contexts) for $V(c)$, we see that the second moment of the estimator vanishes with β, at the rate of $O(\beta^2)$.

C.3 Sub-sampling Contexts

Evaluating the correction on every generated context is unnecessary because the reported quantity is their mean. Let $S \subseteq [N]$ be a uniformly random subset of size m, sampled without replacement. For every $i \in S$, draw R independent copies $Y_{i,1}, \ldots, Y_{i,R}$ of Equation (3), and define

$$\widehat{H}_{m,R}(\alpha) := \bar{H}_N(0) - \frac{\beta}{mR} \sum_{i \in S} \sum_{r=1}^{R} Y_{i,r}, \qquad \bar{H}_N(0) := \frac{1}{N} \sum_{i=1}^{N} H(Q_{c_i}). \tag{6}$$

The first quantity $\widehat{H}_{m,R}(\alpha)$ is our estimator based on m contexts and R samples for each context, while $\bar{H}_N(0)$ is the average base-model entropy.

Proposition 13. *$\widehat{H}_{m,R}(\alpha)$ is an unbiased estimator for $\bar{H}_N(\alpha)$. Moreover, set*

$$d_i := H(Q_{c_i}) - H(Q_{c_i}^\alpha), \qquad \sigma^2 := \frac{1}{N-1} \sum_{i=1}^{N} (d_i - \bar{d})^2,$$

where $\bar{d} = \frac{1}{N} \sum_i d_i$, and let $\tau_i^2 := \mathrm{Var}(\beta Y_{c_i})$. Then

$$\mathrm{Var}\left(\widehat{H}_{m,R}(\alpha) \right) = \left(1 - \frac{m}{N} \right) \frac{\sigma^2}{m} + \frac{1}{mNR} \sum_{i=1}^{N} \tau_i^2. \tag{7}$$

Proof. By Equation (4), $\mathbb{E}[\beta Y_{i,r}] = d_i$. Uniform sampling without replacement therefore makes the estimated mean correction unbiased for $\bar{d}$, proving the first claim. The law of total variance separates the randomness of S from the Monte Carlo randomness conditional on S. The variance of the sample mean of a finite population is $(1 - m/N)\sigma^2/m$, while

$$\mathbb{E}_S \left[\frac{1}{m^2 R} \sum_{i \in S} \tau_i^2 \right] = \frac{1}{mNR} \sum_{i=1}^{N} \tau_i^2,$$

which proves Equation (7). □

The two terms in Equation (7) isolate the only approximations: sub-sampling contexts and estimating d_is. Sub-sampling error decreases as $1/m$ and is small when the entropy correction is similar across contexts. The estimation of d_is decreases as $1/(mR)$ and, by Equation (5), is small when the tilting parameter is small. In our experiments, at each generation position we choose $m = 128$ contexts from the $N = 1000$ generated sequences and apply Equation (6). This reduces the expensive look-ahead evaluations by a factor of about 7.8 while preserving an unbiased estimate of the mean tilted entropy.

D Reward Interpretation of Local Calibration

As we alluded to in the intro and preliminaries sections, local calibration can be seen as a contextual bandits problem, which maximizes the average reward in a given KL trust region, or alternatively minimizes the KL divergence under a reward constraint. Throughout this section we assume f is a reward function, and consider the case where the base model Q under-performs on f, compared with the latent reference distribution. In particular, we assume that the *local* average (see Remark 1) is lower, i.e., $\mu_P > \mu_Q$. In the case of entropy calibration, this is the opposite ($\mu_Q > \mu_P$), but the above assumption can be made without loss of generality by switching the sign of f if needed.

For the trust region parameter, note that two models R, Q give a distribution over next tokens only when we fix the context. Thus, we consider the averaged KL divergence:

$$K(\alpha) := K_P(Q^\alpha, Q) := \mathbb{E}_{c\sim D_P}[\mathrm{KL}\left(Q_c^\alpha \| Q_c\right)].$$

The notation emphasizes that contexts are drawn from the reference distribution P.

Proposition 14. *Assuming $\mu_P > \mu_0$, calibration has a positive finite minimizer $\alpha_* > 0$. Let $\rho^* := K(\alpha_*)$. For every $\rho \in [0, \rho^*]$, let $\alpha_\rho \in [0, \alpha_*]$ be the unique point satisfying $K(\alpha_\rho) = \rho$. Then Q^{α_ρ} maximizes μ_R over all models satisfying $K_P(R, Q) \le \rho$. The same parameter α_ρ minimizes $\ell(\alpha)$ over the local exponential-tilt family subject to $K(\alpha) \le \rho$.*

Proof. Define $A(\alpha) := \mathbb{E}_{c\sim D_P}[\log Z_c(\alpha)]$. Expanding the local tilt gives $\ell(\alpha) = \ell(0) - \alpha\mu_P + A(\alpha)$ and $K(\alpha) = \alpha\mu_\alpha - A(\alpha)$. Differentiation gives $K'(\alpha) = \alpha\ell''(\alpha)$. Under the assumption $\mu_P > \mu_0$, K is strictly increasing on the positive branch and strictly decreasing on the negative branch as α increases toward zero. Thus α_ρ exists uniquely in $[0, \alpha_*]$.

For $\rho > 0$, any model with $K_P(R, Q) \le \rho$ satisfies $K_P(R, Q^{\alpha_\rho}) = K_P(R, Q) - \alpha\rho\mu_R + A(\alpha_\rho)$. The left-hand side is nonnegative, so

$$\mu_R \le \frac{\rho + A(\alpha_\rho)}{\alpha_\rho} = \mu_{\alpha_\rho}.$$

This proves reward optimality. Finally, every feasible tilt parameter satisfies $\alpha \le \alpha_\rho \le \alpha_*$, and convexity implies that ℓ decreases up to α_*. Hence α_ρ also minimizes CE among feasible tilts. □

D.1 Connection to Global Reward

As we've discussed above, μ_R measures *local* reward, meaning that the reward is given to a single step made by R averaged over trajectories from the reference distribution P. The *global* reward is defined analogously: averaging over trajectories generated by R. In particular, for a model R define

$$J(R) = \mathbb{E}_{c_T\sim R}\left[\frac{1}{T}\sum_{t=1}^{T} f(c_{t-1}, y_t)\right] = \mathbb{E}_{c\sim D_R}[\mathbb{E}_{w\sim R_c}[f(c, w)]].$$

Define the discrepancy between the global and local rewards by

$$b(\alpha) := J(Q^\alpha) - \mu_\alpha.$$

Since $J(P) = \mu_P$, the global reward gap satisfies

$$J(P) - J(Q^\alpha) = (\mu_P - \mu_\alpha) - b(\alpha).$$

Local calibration shrinks the first term $\mu_P - \mu_\alpha$, thus $b(\alpha)$ controls the *transfer* to global reward matching.

Taking entropy as a leading example, we first orient it in the correct direction by considering *negative* entropy. The work of Cao et al. (2025) shows that base models exhibit $J(P) \gg J(Q)$. Our results show that local calibration transfers partially, implying that $b(\alpha)$ must shrink (note it is negative, because we consider negative entropy, so shrinking means closer to 0).