BEAMSTAR: RATE–DISTORTION OPTIMAL BEAM SEARCH ON THE LOW-SURPRISE VIABLE MANIFOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Beam search is the standard decoding algorithm when exact, high-probability outputs matter, such as in machine translation, speech, code, and increasingly generative recommendation, where serving pipelines maintain large beam widths under strict latency budgets. Yet beam search is fundamentally wasteful: at every step most expanded hypotheses accumulate low cumulative log-probability and are eventually discarded, while still consuming GPU compute, KV-cache bandwidth, and synchronization overhead. We argue that this waste is structural. Good decoding trajectories concentrate on an exponentially small *viable set* of high-probability paths whose per-token surprise rate stays within ϵ of the model's capability frontier. We characterize this set (its cardinality, volume, and probability mass) and locate it among existing typicality notions: it is provably disjoint from the Shannon typical set for peaked models and coincides with the locally typical set along low-entropy corridors, making it the search-side counterpart of typicality-based sampling. We then frame adaptive beam pruning as an online rate-distortion problem: the retained beam count controls the information rate, and the pruning distortion is the KL divergence between the original and filtered decoding distributions, which admits a closed form under renormalization, namely the negative log of the retained probability mass, computable online at no extra cost. This exposes a practical control signal: at each step all beams are scored on the same cumulative-log-probability distribution, so its running maximum and spread alone suffice to place the pruning threshold. BEAMSTAR prunes beams to track an online estimate of the capability frontier while accounting its distortion online; we prove the full-width beam output survives pruning with high probability, so returned quality is never silently degraded, and that the distortion budget bounds recall loss. On generative recommendation with OpenOneRec-1.7B and 8B-Pro, BEAMSTAR speeds up decoding by 1.16–1.89$\times$ (cutting time per sample by 14–47%), the gain growing with beam width, while its global variant matches pass@1, pass@32, and recall@32 to within 0.5 points (effectively lossless), and it composes with standard serving optimizations.

1 INTRODUCTION

Autoregressive language models decode by searching an exponentially large sequence space, and beam search remains the default whenever an application needs high-probability, reproducible outputs rather than diverse samples: machine translation (Sutskever et al., 2014), speech (Radford et al., 2023), code, and, most acutely, *generative recommendation*, where industrial serving runs beam widths of 128–512 per request under tail-latency constraints (Sun et al., 2025; Deng et al., 2025). In this regime the decoder, not the prefill, dominates cost: every active beam multiplies GPU compute, KV-cache traffic, and cross-beam synchronization.

Yet this budget is spent poorly. Each step expands, scores, and re-ranks every hypothesis, but most expanded trajectories are eventually discarded; worse, as the beam widens the search drifts into low-probability, out-of-distribution regions and quality can *degrade*, a phenomenon known as the *curse of beam search* (Yang et al., 2018; Cohen & Beck, 2019; Stahlberg & Byrne, 2019). Waste and degradation are two faces of one mismatch: beam search allocates search uniformly, while the trajectories that matter concentrate in a far smaller region. We formalize that region as the *viable set*, consisting of trajectories whose per-token surprise stays within ϵ of the best achievable rate. It

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.

is deliberately *atypical*: unlike the Shannon typical set, it is exponentially smaller and contains the low-surprise sequences sought by maximum a posteriori (MAP) decoding. We build a theory of this set and, from it, BEAMSTAR, which casts beam pruning as online rate-distortion coding, with beam width serving as the rate and KL divergence (Kullback & Leibler, 1951) serving as the distortion. The method adjusts the beam width to track the viable set.

Contributions.

1. **Viable-Set Theory.** We define the low-surprise viable set $\mathcal{M}_\epsilon^{(t)}$, characterize its size and probability mass, and prove it is disjoint from the classical Shannon typical set under natural conditions (Section 4).
2. **Rate–Distortion Formulation.** We frame adaptive beam pruning as lossy source coding of trajectories, deriving a closed-form, online-computable distortion metric for a natural class of pruning policies (Section 5).
3. **Theoretical Guarantee.** We prove a survival theorem: the full-width beam output survives BEAMSTAR's pruning with high probability, so the returned output's model score never falls below it; the analysis further derives a principled annealing schedule for the search margin (Theorem 1).
4. **Algorithm.** We present BEAMSTAR, a drop-in replacement for beam selection with minimal overhead. It composes with existing inference optimizations and adapts its beam width to the local geometry of the search space (Section 5).
5. **Evaluation.** On large-scale generative recommendation tasks, BEAMSTAR speeds up decoding by up to 1.89× (1.16–1.89× across settings) while matching pass@1, pass@32, and recall@32 to within 0.5 points of the baseline (Section 6).

2 RELATED WORK

The curse of beam search and beam pruning. Enlarging the beam can *hurt* quality in neural sequence models (Yang et al., 2018; Cohen & Beck, 2019). In neural machine translation (NMT), Stahlberg & Byrne (2019) show that the true MAP output is often the empty string, exposing how far wide-beam search can drift from the model's global optimum. Meister et al. (2020a) recast beam search as exact decoding under a uniform-information-density regularizer. Its inductive bias, rather than its search error, is what provides the benefit. We characterize the *region* in which that bias keeps the search and prune to it explicitly. On efficiency, Freitag & Al-Onaizan (2017) prune by a fixed score offset, with adaptive and best-first variants (Klein et al., 2017; Meister et al., 2020b). BEAMSTAR differs by pruning according to a *per-token surprise rate* relative to an online frontier estimate, rather than a raw score gap. It uses an online-adaptive margin and an online distortion ledger and a data-driven margin instead of a fixed threshold.

Information-theoretic accounts of decoding. Meister et al. (2021) argue good text has stable per-token information; Meister et al. (2023) formalize language models (LMs) as stochastic processes, define the locally typical set, and constrain *samples* toward the conditional entropy shell. Our object is dual: MAP-style decoding operates *below* the shell, in a low-surprise region that we show is disjoint from the typical set for peaked models (Proposition 3) and coincides with the locally typical set along low-entropy corridors (Proposition 8). This region is the search-side counterpart of their sampling-side construction. Rate-distortion theory (Shannon, 1959; Cover & Thomas, 2006) underlies representation learning and lossy compression (Alemi et al., 2017; Tishby et al., 2000); to our knowledge BEAMSTAR is the first to treat the decoding search tree as the source and beam pruning as the lossy encoder, with beam width as the coding rate.

Entropy-adaptive computation. Closest in spirit, EDEN (Evans et al., 2026) adapts the *branching factor* to next-token entropy and proves that entropy-monotone allocation outperforms any fixed branching factor. BEAMSTAR provides complementary control by adapting the *beam width* and contracting subtrees whose cumulative surprise leaves the manifold on which EDEN expands at high-uncertainty forks. Their signals are formally linked (Proposition 7) and our survival guarantee (Theorem 1) plays the role for pruning that their allocation theorem plays for branching; a unified

expand-with-entropy, prune-with-surprise decoder is natural future work. Entropy-triggered model-switching heuristics (Simonds, 2025) share the signal but lack an optimality analysis.

Efficient inference, serving, and generative recommendation. Systems work reduces decoding cost via paged KV caches (Kwon et al., 2023), speculative decoding (Leviathan et al., 2023; Chen et al., 2023), and attention variants (Shazeer, 2019); for wide beams (128–512) in generative recommendation, xGR (Sun et al., 2025) shares KV prefixes across beams and applies catalog-filter early termination. These lower cost *per active beam* using domain constraints; BEAMSTAR is orthogonal and algorithmic, reducing the *number* of active beams from information-theoretic principles, so the gains compose. Generative recommenders autoregressively decode semantic IDs from a massive vocabulary (Rajput et al., 2023; Deng et al., 2025), and recall@k requires many high-probability candidates per request. This makes the recommendation task one of the most beam-intensive LLM workloads and provides our testbed (Section 6).

3 PRELIMINARIES

Autoregressive decoding. Let $p_\phi(\mathbf{y} \mid \mathbf{x})$ be an autoregressive model over sequences $\mathbf{y} = (y_1, \ldots, y_T)$ conditioned on a prompt $\mathbf{x}$, factorized as $p_\phi(\mathbf{y} \mid \mathbf{x}) = \prod_{s=1}^{T} p_\phi(y_s \mid y_{<s}, \mathbf{x})$ over an alphabet $\overline{\mathcal{V}} = \mathcal{V} \cup \{\text{EOS}\}$, where EOS is a designated end-of-string symbol. In beam search, multiple sequences (trajectories) $\{\mathbf{y}_i\}$ are decoded in parallel. A partial *trajectory* i of length t has cumulative log-probability

$$S_t(\mathbf{y}_i) \;=\; \log P(\mathbf{y}_i) \;=\; \sum_{s=1}^{t} \log p_\phi(y_{i,s} \mid \mathbf{y}_{i,<s}, \mathbf{x}), \tag{1}$$

and let us define the *empirical surprise rate* as the per-token negative log-likelihood (nats),

$$r_t(\mathbf{y}_i) \;=\; -\tfrac{1}{t}\, S_t(\mathbf{y}_i). \tag{2}$$

Further, let $H_s(\mathbf{y}) = H\big(p_\phi(y_s \mid \mathbf{y}_{<s}, \mathbf{x})\big)$ be the *conditional (local) entropy* at step s along $\mathbf{y}$.

Language models as stochastic processes. Following Meister et al. (2023) we view the LM as a discrete process; because EOS is absorbing and conditionals drift, it is non-ergodic and non-stationary, so Shannon–McMillan–Breiman applies only as an approximation. We therefore state each result either unconditionally (combinatorial counting arguments) or under an explicit concentration assumption (Assumption 1); see Section G.

Beam search. Beam search with width k maintains a set $\mathcal{B}_t$ of at most k hypotheses. At step t every hypothesis in $\mathcal{B}_{t-1}$ is expanded over the vocabulary, all continuations are scored by Eq. (1), and the top-k survive. Every expanded hypothesis, survivor or not, incurs a forward-pass share, KV-cache reads/writes for its beam slot, and participates in the global top-k synchronization; per-step cost is proportional to $|\mathcal{B}_t|$ and end-to-end cost to $\sum_t |\mathcal{B}_t|$.

Typicality and local typicality. Let $H(\mathcal{Y}) = \lim_{t\to\infty} \frac{1}{t} H(y_1, \ldots, y_t)$ be the entropy rate of a stationary ergodic process, for which the (t, ϵ)-*typical set* is

$$\mathcal{A}_\epsilon^{(t)} \;=\; \big\{\, \mathbf{y} : \big| r_t(\mathbf{y}) - H(\mathcal{Y}) \big| \le \epsilon \,\big\}, \tag{3}$$

with $P(\mathcal{A}_\epsilon^{(t)}) \to 1$ and $|\mathcal{A}_\epsilon^{(t)}| \approx e^{tH}$ (Breiman, 1957; Cover & Thomas, 2006). Two facts matter: typicality describes *samples* (ancestral sampling lands in $\mathcal{A}_\epsilon^{(t)}$ w.h.p.), yet the highest-probability sequences are *not* typical. They sit below the entropy shell, in a region of vanishing mass but maximal individual probability (Meister et al., 2023; Dieleman, 2020), exactly where beam search hunts. Meister et al. (2023) also define the (t, ϵ)-*locally typical set*, whose per-step information tracks the local conditional entropy:

$$\mathcal{L}_\epsilon^{(t)} \;=\; \big\{\, \mathbf{y} : \; \big| -\log p_\phi(y_s \mid \mathbf{y}_{<s}, \mathbf{x}) - H_s(\mathbf{y}) \big| < \epsilon \;\; \forall\, 1 \le s \le t \,\big\}, \tag{4}$$

a stepwise restriction motivated by the uniform-information-density properties of human text. Section 4 defines our object, the viable set, and locates it relative to both Eq. (3) and Eq. (4).

Rate-distortion. Given a source $\mathbf{y} \sim P$, a reconstruction alphabet $\hat{\mathcal{Y}}$, a (possibly stochastic) encoder $Q(\hat{\mathbf{y}} \mid \mathbf{y})$ mapping each source to a reconstruction $\hat{\mathbf{y}} \in \hat{\mathcal{Y}}$, and a distortion measure $d(\mathbf{y}, \hat{\mathbf{y}}) \geq 0$, the rate-distortion function is

$$R(D) \;=\; \min_{Q(\hat{\mathbf{y}}|\mathbf{y})\,:\,\mathbb{E}_{P,Q}[d(\mathbf{y},\hat{\mathbf{y}})]\leq D} I(\mathbf{y};\hat{\mathbf{y}}), \tag{5}$$

the minimal mutual information (rate) any encoder must retain to reconstruct the source within expected distortion D (Cover & Thomas, 2006, Ch. 10); Section 5 instantiates it with the decoding trajectory as the source $\mathbf{y}$ and the pruning policy as the encoder Q.

4 The Viable Set: Definition and Characterization

Our core hypothesis is that high-quality decoding trajectories do not occupy the full entropy shell of the model distribution but concentrate on a significantly smaller *low-surprise manifold*. This section makes the hypothesis precise and characterizes the resulting set in the three quantities that define a typical set: cardinality, volume, and probability mass. All proofs are deferred to Section A.

Definition 1 (Capability frontier)**.** *Let $\mathcal{X}$ denote the input space of the task under consideration. The model's* capability frontier *at step t is the minimal achievable per-token surprise over task inputs $\mathbf{x} \in \mathcal{X}$ and length-t trajectories,*

$$H_t^\star \;=\; \min_{\mathbf{x}\in\mathcal{X}} \min_{\mathbf{y}:\,|\mathbf{y}|=t} r_t(\mathbf{y} \mid \mathbf{x}). \tag{6}$$

It is the per-token surprise of the (normalized) MAP trajectory over inputs and continuations, and lower-bounds the surprise rate of every length-t hypothesis the decoder can produce on any $\mathbf{x} \in \mathcal{X}$.

For a fixed input $\mathbf{x}$ we also use the *per-request* (local) frontier $H_t^\star(\mathbf{x}) = \min_{\mathbf{y}:\,|\mathbf{y}|=t} r_t(\mathbf{y} \mid \mathbf{x})$. The two are nested, $H_t^\star = \min_{\mathbf{x}} H_t^\star(\mathbf{x}) \leq H_t^\star(\mathbf{x})$, and differ by the *difficulty gap* $\Delta_t(\mathbf{x}) := H_t^\star(\mathbf{x}) - H_t^\star \geq 0$: the global frontier is the lower anchor, while the per-request frontier is the sharper, input-specific reference. All results below hold for either reading (replace $H_t^\star$ by $H_t^\star(\mathbf{x})$); Section 5 maps the two onto the global and local variants we evaluate, BS and BS_L.

Definition 2 (Viable set)**.** *For a slack parameter $\epsilon \geq 0$, the* viable set *at step t is*

$$\mathcal{M}_\epsilon^{(t)} \;=\; \left\{\, \mathbf{y} : \left| r_t(\mathbf{y}) - H_t^\star \right| \leq \epsilon \,\right\} \tag{7}$$

In words, the viable set comprises all trajectories that are up to ϵ away from the model capability frontier. Equivalently, $\mathcal{M}_\epsilon^{(t)}$ is the super-level set $\{P(\mathbf{y}) \geq e^{-t(H_t^\star+\epsilon)}\}$: the trajectories whose probability is within a factor $e^{t\epsilon}$ of the best achievable. Whereas the typical set is a *two-sided* shell centered at the entropy H, the viable set is a *one-sided* cap anchored at the frontier $H_t^\star \leq H$. The following results quantify how much smaller this cap is and why it suffices for decoding.

Global and per-request viable sets. Definition 2 uses the global frontier $H_t^\star$; the *per-request viable set* replaces it by the conditional frontier, $\mathcal{M}_\epsilon^{(t)}(\mathbf{x}) = \{\mathbf{y} : r_t(\mathbf{y} \mid \mathbf{x}) \leq H_t^\star(\mathbf{x}) + \epsilon\}$. The two are the *same cap with the difficulty gap moved between anchor and margin*: since $H_t^\star(\mathbf{x}) = H_t^\star + \Delta_t(\mathbf{x})$,

$$\mathcal{M}_\epsilon^{(t)}(\mathbf{x}) \;=\; \{\mathbf{y} : r_t(\mathbf{y} \mid \mathbf{x}) \leq H_t^\star + (\epsilon + \Delta_t(\mathbf{x}))\} \;=\; \mathcal{M}_{\epsilon+\Delta_t(\mathbf{x})}^{(t)}, \tag{8}$$

i.e. a per-request cap at slack ϵ equals the global cap at the inflated slack $\epsilon + \Delta_t(\mathbf{x})$ on that input. Consequently a global cap with small ϵ can exclude a hard input's own MAP (whenever $\Delta_t(\mathbf{x}) > \epsilon$), which is exactly why the global-frontier guarantees below carry an inflated margin floor while the per-request ones do not (Section 5). All characterizations in this section hold verbatim for either frontier. Simply replace $H_t^\star$ by $H_t^\star(\mathbf{x})$, and any entropy rate by its conditional counterpart; the sole place the two diverge, the survival floor, is flagged in Section 5.

4.1 Cardinality

Proposition 1 (Viable-set size)**.** *For every t and $\epsilon \geq 0$,*

$$1 \;\leq\; \left|\mathcal{M}_\epsilon^{(t)}\right| \;\leq\; e^{t\,(H_t^\star+\epsilon)}. \tag{9}$$

The viable set is also exponentially smaller in *volume* than the typical set whenever $H_t^\star + \epsilon < H$; we defer the statement and proof to Section G.

4.2 PROBABILITY MASS

Assumption 1 (Surprise concentration)**.** *The per-token surprise increments* $Z_s = -\log p_\phi(y_s \mid \mathbf{y}_{<s}, \mathbf{x})$ *of a trajectory drawn from* $P(\cdot \mid \mathbf{x})$ *satisfy a law of large numbers,* $\frac{1}{t}\sum_{s\le t} Z_s \to H(\mathcal{Y})$ *in probability for an entropy rate* $H(\mathcal{Y})$*, together with a tail condition (e.g., conditionally sub-exponential increments) strong enough for exponential concentration.*

Assumption 1 is genuine, not a formality: the LM process is non-ergodic (absorbing EOS) with unbounded increments, so it needs a martingale argument rather than the i.i.d. AEP (Section G); empirically surprise rates concentrate sharply for the models we study.

Proposition 2 (Mass bounds)**.** *For every* t *and* $\epsilon \ge 0$*,*

$$e^{-tH_t^\star} \;\le\; P\big(\mathcal{M}_\epsilon^{(t)}\big) \;\le\; \min\Big\{1,\; \big|\mathcal{M}_\epsilon^{(t)}\big|\, e^{-tH_t^\star}\Big\}. \tag{10}$$

Moreover, under Assumption 1, if $H_t^\star + \epsilon < H$ *uniformly in* t*, then* $P(\mathcal{M}_\epsilon^{(t)}) \to 0$*, with an exponential rate governed by the large-deviation function of the surprise process when one exists.*

Remark 1 (The viable set is atypical-and that is the point)**.** *Proposition 2 says a random* sample *from the model will essentially never land in* $\mathcal{M}_\epsilon^{(t)}$*. This is not a defect: beam search is not sampling. Every trajectory relevant to MAP-style decoding, where every top-*k *list with* $k \le |\mathcal{M}_\epsilon^{(t)}|$*, in particular the one beam search returns, lies inside the viable set by construction of Eq.* (7) *with suitable* ϵ*. The viable set is thus the correct search space for beam decoding, while the typical set is the correct description of sampling. We argue that conflating the two is the root of the curse of beam search: a width-*k *beam with* $k \gg |\mathcal{M}_\epsilon^{(t)}|$ *must populate its remaining slots with trajectories* outside *the viable cap, where we obtain high-surprise, out-of-distribution continuations the model was never trained to score reliably, and thus, candidates that survive re-ranking degrade output.*

4.3 THE VIABLE SET AMONG TYPICALITY NOTIONS

We now locate $\mathcal{M}_\epsilon^{(t)}$ relative to the typical set Eq. (3) and the locally typical set Eq. (4) of Meister et al. (2023). The first result needs no process assumptions at all; it bounds the frontier by the local entropies the decoder actually observes, connecting our anchor $H_t^\star$ to the signal used by entropy-adaptive decoders (Evans et al., 2026).

A basic bound (Proposition 7, Section G) shows the frontier lies below the average local entropy along the greedy path, placing the viable cap beneath the entropy shell in peaked models.

Proposition 3 (Separation from the typical set)**.** *If* $\epsilon + \epsilon' < H - H_t^\star$*, then* $\mathcal{M}_\epsilon^{(t)} \cap \mathcal{A}_{\epsilon'}^{(t)} = \emptyset$*.*

Together with Proposition 2, this sharpens Remark 1 into a set statement: whenever the model's frontier-to-entropy gap exceeds the combined slacks, the region in which beam search operates is *disjoint* from the region in which sampling lands. Typicality-based reasoning about beam search is therefore not merely loose; it concerns the wrong set. Under the per-request reading, the statement is identical after replacing $H_t^\star$ with $H_t^\star(\mathbf{x})$ and the entropy rate with its conditional counterpart. Under the global reading, the condition is easier to satisfy because $H_t^\star \le H_t^\star(\mathbf{x})$, which makes the gap $H - H_t^\star$ at least as large. Disjointness therefore holds for both frontiers, with the global frontier yielding the stronger separation.

Moreover, a locally typical trajectory (Proposition 8, Section G) is viable precisely along low-entropy corridors, where beam search incidentally enforces local typicality (Meister et al., 2023). In contrast, BEAMSTAR deliberately enforces the trajectory-level analogue.

Following the methodology of Meister et al. (2023), who test their motivating claim as a falsifiable hypothesis on human text (Fig. 1), we state and test the claim underlying BEAMSTAR:

Hypothesis 1 (Beam concentration)**.** *Let* $\widehat{H}_t^\star = \min_i r_t(\mathbf{y}_i)$ *denote the* empirical *frontier estimate, namely, the smallest per-token surprise rate among the hypotheses expanded at step* t*. As the minimum over the current request's candidate pool, it serves as the online, per-request estimator of* $H_t^\star(\mathbf{x})$ *(Algorithm 1). Then the output-relevant trajectories, including the returned hypotheses and their ancestor prefixes, lie in* $\mathcal{M}_\epsilon^{(t)}(\mathbf{x})$ *at every step* t*, provided that* ϵ *is small relative to the spread*

of surprise rates in the pool. Equivalently, the deviations $r_t(\mathbf{y}_i) - \widehat{H}_t^\star$ of output-relevant hypotheses concentrate near zero, whereas those of eventually discarded hypotheses exhibit heavy tails.

Its empirical analogue is a histogram of $r_t(\mathbf{y}_i) - \widehat{H}_t^\star$ over steps and instances, stratified by whether i is an ancestor of the final output: Hypothesis 1 predicts a sharp peak at zero for ancestors and a heavy right tail for computational noise (the gap BEAMSTAR converts into savings).

Sufficiency for decoding. Under *prefix stability*, where the returned trajectory remains within the cap at every step, restricting each step to $\mathcal{M}_\epsilon^{(t)}$ preserves the top output at beam width $|\mathcal{M}_\epsilon^{(t)} \cap \mathcal{B}_t^u| \le k$, where $\mathcal{B}_t^u$ is the survivor set of the unpruned width-k beam search at step t (Assumption 2 and Corollary 2, Section G). Viability is not prefix-closed: a "garden-path" prefix may temporarily lie above the cap. Therefore, Section 5 replaces this hard assumption with a margin-plus-concentration condition and treats finite ϵ as distortion.

5 BEAMSTAR: BEAM PRUNING AS ONLINE RATE–DISTORTION CODING

We now cast adaptive beam pruning as lossy source coding over trajectory space. The formulation proceeds in three steps: (i) identify the pruning policy with a stochastic encoder, (ii) identify the retained information rate with the (log) active beam count, and (iii) derive an online-computable KL distortion. The section closes with the BEAMSTAR algorithm.

5.1 PRUNING AS A STOCHASTIC ENCODER

Let $P(\mathbf{y})$ denote the decoding distribution over complete trajectories induced by the model on the full, unpruned search tree. Let a pruning policy θ induce a conditional distribution $Q_\theta(\hat{\mathbf{y}} \mid \mathbf{y})$ that maps each source trajectory to the trajectory retained in its place by the pruned decoder. The marginal $Q_\theta(\hat{\mathbf{y}})$ is the *filtered decoding distribution*.

In rate–distortion terms, the source $\mathbf{y}$ is the full decoding trajectory, the reconstruction $\hat{\mathbf{y}}$ is the surviving pruned trajectory, the encoder $Q_\theta(\hat{\mathbf{y}} \mid \mathbf{y})$ is the pruning policy, the distortion is a per-symbol semantic divergence, and the rate $I(\mathbf{y}; \hat{\mathbf{y}})$ is the information retained by the decoder.

In the lens of Eq. (5), the BEAMSTAR objective is the trajectory-space rate-distortion function

$$R(D) \;=\; \min_{Q_\theta(\hat{\mathbf{y}}|\mathbf{y}):\, D(\theta) \le D_0} I(\mathbf{y}; \hat{\mathbf{y}}), \tag{11}$$

with the per-symbol distortion defined as a divergence *functional* of the filtered marginal,

$$D(\theta) \;\coloneqq\; \tfrac{1}{t} D_{\mathrm{KL}}(Q_\theta \,\|\, P) \;=\; \tfrac{1}{t}\, \mathbb{E}_{Q_\theta}\left[\log \frac{Q_\theta(\hat{\mathbf{y}})}{P(\hat{\mathbf{y}})}\right]. \tag{12}$$

In words, $D(\theta)$ is the excess surprise incurred when the pruned decoder's output distribution Q_θ is evaluated under the original beam-pool distribution P. It measures the change induced by pruning. The distortion is zero when pruning leaves the distribution unchanged and increases as the filtered distribution departs further from the original beam-pool distribution. The KL divergence is the natural distortion here because it measures excess surprise and information loss when the filtered decoder replaces the original one, and can serve as a proxy for semantic drift.

The forward direction $D_{\mathrm{KL}}(Q_\theta \,\|\, P)$ (not $D_{\mathrm{KL}}(P \,\|\, Q_\theta)$) is the unique finite choice under hard pruning, since Q_θ has strictly smaller support than P; it reads as the excess surprise of describing the pruned decoder's outputs with the original model (Section G).

Operationally the same goal reads $\min_\theta \mathbb{E}[|\mathcal{B}_t|]$ s.t. $D(\theta) \le \delta$: the average active beam count proxies FLOPs, KV-cache traffic, and latency, and by Proposition 4 it is the exponential of the information rate, so minimizing one minimizes the other.

5.2 BEAM WIDTH IS THE CODING RATE

Proposition 4 (Beam count bounds the retained rate)**.** *Let $\mathbf{y} \sim P(\cdot \mid x)$ denote the source trajectory, and let $\hat{\mathbf{y}}_t$ denote the* step-t reconstruction*, namely, the hypothesis that the encoder retains for $\mathbf{y}$ at*

step t. This hypothesis is a (possibly randomized) function of $\mathbf{y}$ supported on the survivor set $\mathcal{B}_t$ (cf. Remark 4). Then

$$I(\mathbf{y};\, \hat{\mathbf{y}}_t) \;\le\; H(\hat{\mathbf{y}}_t) \;\le\; \log|\mathcal{B}_t|, \tag{13}$$

so the logarithm of the active beam count upper-bounds the per-step information that the decoder retains about the source trajectory. Equivalently,$|\mathcal{B}_t| \ge e^{I(\mathbf{y};\, \hat{\mathbf{y}}_t)}$.

Proposition 4 licenses reading *beam width as the coding rate*: the inequality $I(\mathbf{y};\hat{\mathbf{y}}_t) \le \log|\mathcal{B}_t|$ is rigorous, and $\log|\mathcal{B}_t| \approx I(\mathbf{y};\hat{\mathbf{y}}_t)$ is a tight heuristic when survivors are used efficiently, so the beam count is the online estimator of a quantity that is itself hard to compute. Under this identification the viable set is the rate–distortion optimal support:

$$\mathcal{M}_\epsilon^{(t)} \;=\; \arg\min_{\mathcal{B}_t}\; I(\mathbf{y};\hat{\mathbf{y}}_t) \quad \text{s.t.} \quad D(\mathcal{B}_t) \le D_0, \tag{14}$$

That is, among all survivor sets $\mathcal{B}_t$ at step t meeting a distortion budget, the low-surprise cap retains the fewest hypotheses (Section C sketches the argument: for a fixed cardinality, the retained mass, and hence the distortion by Lemma 1, is maximized by selecting the highest-probability trajectories. This selection is equivalent to imposing a surprise-rate cap).

5.3 Closed-form distortion for renormalized pruning

Definition 3 (Restrict-and-renormalize policy)**.** *A pruning policy with survivor set $\mathcal{B}_t$ at step t induces the filtered distribution $Q_\theta(\mathbf{y}) = P(\mathbf{y} \mid \mathcal{S}) = P(\mathbf{y})\,\mathbf{1}[\mathbf{y} \in \mathcal{S}]/P(\mathcal{S})$, where $\mathcal{S} = \{\mathbf{y} : \mathbf{y}_{1:t} \in \mathcal{B}_t$ for all $t\}$ is the set of complete trajectories whose every prefix survives pruning.*

Lemma 1 (Distortion equals lost log-mass)**.** *For a restrict-and-renormalize policy with survivor support $\mathcal{S}$, the distortion of Eq.* (12) *is*

$$D_{\mathrm{KL}}(Q_\theta \,\|\, P) \;=\; \sum_{\mathbf{y}\in\mathcal{S}} \frac{P(\mathbf{y})}{P(\mathcal{S})} \log \frac{P(\mathbf{y})/P(\mathcal{S})}{P(\mathbf{y})} \;=\; -\log P(\mathcal{S}), \qquad D(\theta) \;=\; -\tfrac{1}{t}\log P(\mathcal{S}) \;\ge\; 0. \tag{15}$$

Lemma 1 is what makes the framework practical: the semantic distortion of a pruning decision is exactly the negative log of the probability mass it retains, a quantity the decoder accumulates online, as a telescoping sum of nonnegative per-step increments (Corollary 1). Pruning a hypothesis that carries 10^{-9} of the current mass costs essentially nothing; pruning one that carries 30% costs 0.36 nats, and the ledger records it online [1].

The distortion has exact beam-wise and temporal decompositions; its chain-rule form is the online ledger of Algorithm 1, with no beam-local distortion signal (Section G).

Relation to other truncation rules. Fixed-width beam search, score-gap pruning (Freitag & Al-Onaizan, 2017), and BeamStar fit a common subset-selection template. BeamStar is the mass-targeted rather than count-targeted member, making it the beam-search analogue of nucleus sampling, and complements entropy-adaptive branching (Evans et al., 2026). See Section G for details.

5.4 A survival guarantee

The rate–distortion objective controls the *distributional* error of pruning. We now give a complementary trajectory-level guarantee, following the analysis of Evans et al. (2026), that identifies conditions under which BeamStar provably returns the *same output* as full-width beam search. The obstruction, described in Remark 3, is a garden-path excursion. The theorem replaces the exact requirement in Assumption 2 with a margin-plus-concentration condition and thereby derives the form of the per-step margin.

Theorem 1 (Survival of the reference trajectory)**.** *Fix a horizon T and let $\mathbf{y}^\circ$ be any trajectory in the width-k list returned by full-width beam search on input $\mathbf{x}$ (e.g. its top-scoring output, or the ground-truth completion when present), with per-token surprises $Z_s = -\log p_\phi(y_s^\circ \mid \mathbf{y}_{<s}^\circ, \mathbf{x})$ and conditional means $\mu_s = \mathbb{E}[\, Z_s \mid \mathbf{y}_{<s}^\circ, \mathbf{x}]$. Assume:*

[1] For mixed-length populations (hypotheses reaching EOS at different times) the $1/t$ normalization is ambiguous; we then state budgets in total nats, $D_{\mathrm{tot}} = -\log P(\mathcal{S})$. In our experiments all semantic-ID trajectories share a fixed length, so $1/t$ is unambiguous.

Algorithm 1 BEAMSTAR decoding step

Require: survivors $\mathcal{B}_{t-1}$, max width k, spent distortion D_{t-1}

1: expand each $\mathbf{y}_i \in \mathcal{B}_{t-1}$; score continuations; keep top-k candidates $\mathcal{B}_t^{\exp}$
2: $r_t(\mathbf{y}_i) \leftarrow -S_t(\mathbf{y}_i)/t$ for all $i \in \mathcal{B}_t^{\exp}$ ▷ surprise rates, from existing scores
3: $\widehat{H}_t^\star \leftarrow \min_i r_t(\mathbf{y}_i)$; $\epsilon_t \leftarrow$ global or local margin (§5) ▷ frontier and margin
4: $\mathcal{B}_t \leftarrow \{\, i : r_t(\mathbf{y}_i) \le \widehat{H}_t^\star + \epsilon_t \,\}$ ▷ project onto viable set
5: $D_t \leftarrow D_{t-1} + \log \dfrac{\sum_{i \in \mathcal{B}_t^{\exp}} P(\mathbf{y}_i)}{\sum_{i \in \mathcal{B}_t} P(\mathbf{y}_i)}$ ▷ distortion ledger, Corollary 1
6: **return** $\mathcal{B}_t$, D_t

(i) ***(Margin)*** $\frac{1}{t}\sum_{s \le t} \mu_s \;\le\; H_t^\star(\mathbf{x}) + \bar{\epsilon}/2$ *for all* $t \le T$*;*

(ii) ***(Concentration)*** *the residuals* $Z_s - \mu_s$ *are conditionally* σ^2*-sub-Gaussian given* $(\mathbf{x}, \mathbf{y}^\circ_{<s})$*.*

(iii) ***(No displacement)*** $\mathbf{y}^\circ_{1:s}$ *ranks in the top-*k *extensions of* $\mathcal{B}_{s-1}$ *whenever* $\mathbf{y}^\circ_{1:s-1} \in \mathcal{B}_{s-1}$*.*

Run Algorithm 1 with the margin schedule $\epsilon_t = \frac{\bar{\epsilon}}{2} + \sigma\sqrt{2\log(T/\delta)/t}$*. Then with probability at least* $1-\delta$*, the prefix of* $\mathbf{y}^\circ$ *passes the viability test at every step* $t \le T$*; hence* $\mathbf{y}^\circ$ *is among* BEAMSTAR*'s completed candidates and the returned output has model score at least* $S_T(\mathbf{y}^\circ)$*, with equality when* $\mathbf{y}^\circ$ *is the unique score-maximal completion in the pruned search space.*

The theorem gives conditions under which, with probability at least $1-\delta$, $\mathbf{y}^\circ$ survives every pruning step. Thus, BEAMSTAR returns $\mathbf{y}^\circ$ or another trajectory with at least as high a score. Instantiating $\mathbf{y}^\circ$ as the top-scoring output yields a guarantee on the returned output's score. Instantiating it as the ground-truth completion, provided that it appears in the width-k list and satisfies condition (i), yields a recall guarantee complementary to Proposition 9.

The proof (Section B) is an induction over steps using an Azuma-type bound; assumption (i) is the honest core, a statement about how quickly good continuations announce themselves, measurable via the Hypothesis 1 histogram. Section G discusses tightness and the exactness condition. The $1/\sqrt{t}$ shape of ϵ_t is a consequence of concentration, not a heuristic: the cap is wide early (noisy averages, exploration) and tightens late (exploitation).

5.5 THE BEAMSTAR DECODING RULE

At each step, BEAMSTAR (i) estimates the capability frontier from the strongest surviving hypotheses, (ii) retains only trajectories within an adaptive margin of it, and (iii) records the induced distortion in its ledger.

Frontier estimation. Let $\mathcal{B}_t^{\exp}$ denote the k highest-scoring one-token extensions of the surviving set $\mathcal{B}_{t-1}$. We estimate the per-request frontier by $\widehat{H}_t^\star(\mathbf{x}) = \min_{i \in \mathcal{B}_t^{\exp}} r_t(\mathbf{y}_i \mid \mathbf{x})$; the global $\widehat{H}_t^\star$ follows by pooling across requests, per Eq. (6).

Adaptive margin. Each frontier estimator has its own margin. The global variant uses a flat margin $\epsilon_t = \gamma\,\widehat{\sigma}_t$, where $\widehat{\sigma}_t$ is the standard deviation of cumulative scores pooled across requests at step t. The local variant follows the form of Theorem 1, a constant floor plus a $t^{-1/2}$ concentration term: $\epsilon_t = \frac{1}{\sqrt{2}}\big(1 + \widehat{\sigma}_t^{\mathrm{w}}\, t^{-1/2} + \frac{1}{4}\, t^{-1}\big)$, where $\widehat{\sigma}_t^{\mathrm{w}}$ is the within-request spread of beam scores (the within-request component in the law of total variance).

Cost and compatibility. Lines 2–5 of Algorithm 1 are $O(|\mathcal{B}_t^{\exp}|)$ elementwise operations on scores the decoder already has, no extra forward passes or cache traffic. BEAMSTAR changes only survivor selection, so it composes with paged/shared KV caches, early termination, and scheduler optimizations (Kwon et al., 2023; Sun et al., 2025): those cut cost *per* beam, BEAMSTAR cuts the *number* of beams, and freed slots are released (latency) or re-batched (throughput).

Table 1: Time per sample (ms, A100) and quality on OpenOneRec-1.7B and 8B-Pro. BS is the global variant, BS_L the local (per-request) variant of Section 5; speedup (in brackets) grows with beam width. The standard deviation of the time per sample is roughly 2ms. †Video: 1000-sample subset; $\gamma = 2$ (interactive), $\gamma = 2.2$ with one exploration token (video).

		Time/sample ↓		pass@1 ↑			pass@32 ↑			recall@32 ↑		
Model	Task, BW	base	BEAMSTAR	base	BS	BS_L	base	BS	BS_L	base	BS	BS_L
1.7B	interactive, 32	634	429 [32%]	7.0	7.0	7.2	32.5	32.5	31.3	20.3	20.3	19.3
	video†, 32	604	440 [27%]	4.2	4.2	4.3	16.7	16.9	15.0	2.6	2.6	2.3
	video†, 128	1963	1387 [29%]	4.2	4.2	4.2	20.8	20.3	18.0	3.5	3.4	3.0
	video†, 512	17572	9290 [47%]	4.2	4.2	4.2	22.2	22.0	19.5	3.8	3.8	3.3
8B-Pro	interactive, 32	838	720 [14%]	12.8	12.9	13.2	50.9	51.0	47.6	35.1	35.3	31.9
	video, 32	808	643 [20%]	5.9	5.9	5.9	22.0	22.1	16.6	3.8	3.8	2.7

6 EXPERIMENTS

We evaluate BEAMSTAR on generative recommendation, the most beam-intensive LLM decoding workload in production today, using two open models from the OneRec family. Section F reports the operational rate–distortion curve, providing the empirical $H_t^\star(\mathbf{x})$ and the viable manifold properties.

6.1 SETUP

Models and task. OpenOneRec-1.7B and OpenOneRec-8B-Pro (Zhou et al., 2026) decode hierarchical semantic-ID item identifiers (three-layer RQ-Kmeans codebook, 8192 entries per layer; fixed-length token sequences) under constrained beam search. We evaluate on RecIF-Bench, a multi-domain recommendation benchmark (short videos, ads, e-commerce; ∼202k users, ∼15.3M items, ∼120M interactions), reporting the *interactive* and *video* tasks.

Baselines, metrics, configuration. We compare against fixed-width beam search at $k \in \{32, 128, 512\}$ in the same serving stack, on quality (pass@1, pass@32, recall@32 vs. held-out ground truth) and wall-clock time per sample (ms, A100). We report the global (BS) and local (BS_L) variants of Section 5; the margin coefficient γ and exploration length are given in Table 1.

6.2 MAIN RESULTS

Findings. Three patterns in Table 1 match the theory. **(i) Speedup grows with beam width**: on 1.7B video, BS speeds up decoding by 37/42/89% at $k = 32/128/512$, as the viable set becomes a smaller fraction of the nominal beam. **(ii) The global variant is effectively lossless**: BS matches baseline pass@1, pass@32, and recall@32 within 0.5 points at every setting, while speeding up decoding by 16–48% at $k = 32$ alone. **(iii) The metric split tracks the two guarantees**: pass@1 is matched or slightly improved for both variants (Theorem 1, while BS_L trails on list metrics by 1–5 points: its per-request cap is tighter at the first code (Section F), and Proposition 9 bounds the resulting recall loss by the discarded mass.

7 CONCLUSION

Beam search wastes most of its computation on hypotheses that never reach the output, and widening the beam drifts into out-of-distribution territory where quality degrades. We traced both pathologies to one geometric fact: the trajectories that matter for MAP-style decoding occupy an exponentially small, *atypical*, low-surprise viable set, characterized here in deliberate contrast to the Shannon typical set. Casting beam pruning as lossy coding over trajectories yields a principled objective—minimize the retained rate subject to a KL distortion that, under renormalization, is the negative log of the retained mass and is free to compute online—which BEAMSTAR implements as a drop-in survivor rule with an online distortion ledger, delivering up to $1.9\times$ faster decoding on OpenOneRec-1.7B/8B at matched pass@1, pass@32, and recall@32. Extensions and further experiments are in Section H; proofs in Section A.

REFERENCES

Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. Deep variational information bottleneck. In *Proceedings of ICLR*, 2017.

Leo Breiman. The individual ergodic theorem of information theory. *The Annals of Mathematical Statistics*, 28(3):809–811, 1957.

Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. Accelerating large language model decoding with speculative sampling. *arXiv preprint arXiv:2302.01318*, 2023.

Eldan Cohen and J. Christopher Beck. Empirical analysis of beam search performance degradation in neural sequence models. In *Proceedings of ICML*, 2019.

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

Jiaxin Deng, Shiyao Wang, Kuo Cai, Lejian Ren, Qigen Hu, Weifeng Ding, Qiang Luo, and Guorui Zhou. OneRec: Unifying retrieve and rank with generative recommender. *arXiv preprint*, 2025.

Sander Dieleman. Musings on typicality. Blog post, 2020.

Benjamin Patrick Evans, Sumitra Ganesh, and Leo Ardon. Entropy-informed decoding: Adaptive information-driven branching. In *Proceedings of ICML*, 2026. arXiv:2605.09745.

Angela Fan, Mike Lewis, and Yann Dauphin. Hierarchical neural story generation. In *Proceedings of ACL*, 2018.

Markus Freitag and Yaser Al-Onaizan. Beam search strategies for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, 2017.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *Proceedings of ICLR*, 2020.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL, System Demonstrations*, 2017.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of SOSP*, 2023.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast inference from transformers via speculative decoding. In *Proceedings of ICML*, 2023.

Clara Meister, Ryan Cotterell, and Tim Vieira. If beam search is the answer, what was the question? In *Proceedings of EMNLP*, 2020a.

Clara Meister, Tim Vieira, and Ryan Cotterell. Best-first beam search. In *Transactions of the Association for Computational Linguistics*, 2020b.

Clara Meister, Tiago Pimentel, Patrick Haller, Lena Jäger, Ryan Cotterell, and Roger Levy. Revisiting the uniform information density hypothesis. In *Proceedings of EMNLP*, 2021.

Clara Meister, Tiago Pimentel, Gian Wiher, and Ryan Cotterell. Locally typical sampling. *Transactions of the Association for Computational Linguistics*, 11, 2023.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of ICML*, 2023.

Shashank Rajput, Nikhil Mehta, Anima Singh, et al. Recommender systems with generative retrieval. In *Advances in Neural Information Processing Systems*, 2023.

Claude E. Shannon. Coding theorems for a discrete source with a fidelity criterion. In *IRE National Convention Record*, volume 4, pp. 142–163, 1959.

Noam Shazeer. Fast transformer decoding: One write-head is all you need. *arXiv preprint arXiv:1911.02150*, 2019.

Toby Simonds. Entropy adaptive decoding: Dynamic model switching for efficient inference. *arXiv preprint arXiv:2502.06833*, 2025.

Felix Stahlberg and Bill Byrne. On NMT search errors and model errors: Cat got your tongue? In *Proceedings of EMNLP-IJCNLP*, 2019.

Qingxiao Sun, Tongxuan Liu, Shen Zhang, and Siyu Wu. xGR: Efficient generative recommendation serving. *arXiv preprint*, 2025.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, 2014.

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

Roman Vershynin. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Yilin Yang, Liang Huang, and Mingbo Ma. Breaking the beam search curse: A study of (re-)scoring methods and stopping criteria for neural machine translation. In *Proceedings of EMNLP*, 2018.

Guorui Zhou, Honghui Bao, Jiaming Huang, Jiaxin Deng, Jinghao Zhang, Junda She, Kuo Cai, Lejian Ren, Lu Ren, Qiang Luo, Qianqian Wang, Qigen Hu, Rongzhou Zhang, Ruiming Tang, Shiyao Wang, Wuchao Li, Xiangyu Wu, Xinchen Luo, Xingmei Wang, Yifei Hu, Yunfan Wu, Zhanyu Liu, Zhiyang Zhang, Zixing Zhang, Bo Chen, Bin Wen, Chaoyi Ma, Chengru Song, Chenglong Chu, Defu Lian, Fan Yang, Feng Jiang, Hongtao Cheng, Huanjie Wang, Kun Gai, Pengfei Zheng, Qiang Wang, Rui Huang, Siyang Mao, Tingting Gao, Wei Yuan, Yan Wang, Yang Zhou, Yi Su, Zexuan Cheng, Zhixin Ling, and Ziming Li. OpenOneRec technical report. *arXiv preprint arXiv:2512.24762*, 2026. URL https://arxiv.org/abs/2512.24762.

A Deferred Proofs

This appendix collects the proofs of the statements in Sections 4 and 5.

Proof of Proposition 1. The lower bound holds because the MAP trajectory is always a member. For the upper bound, every $\mathbf{y} \in \mathcal{M}_\epsilon^{(t)}$ satisfies $P(\mathbf{y}) \geq e^{-t(H_t^\star+\epsilon)}$, and probabilities sum to at most one: $1 \geq \sum_{\mathbf{y}\in\mathcal{M}_\epsilon^{(t)}} P(\mathbf{y}) \geq |\mathcal{M}_\epsilon^{(t)}|\, e^{-t(H_t^\star+\epsilon)}$. □

Proof of Proposition 2. The lower bound is the MAP trajectory's own mass. The upper bound bounds each member's probability by the maximum $e^{-tH_t^\star}$ (note that combining it with Proposition 1 yields only the trivial $e^{t\epsilon} \geq 1$; a nontrivial exponential upper bound is a large-deviation statement, not a counting one). The vanishing-mass claim is immediate from Assumption 1: the event $\{r_t(\mathbf{y}) \leq H_t^\star + \epsilon\}$ requires the empirical surprise rate to deviate from its concentration point H by at least $H - H_t^\star - \epsilon > 0$. □

Proof of Proposition 7. The first inequality is Definition 1 ($H_t^\star$ minimizes over all trajectories, greedy included). For the second, at each state the greedy token maximizes probability, so at each t, $p_g \geq p_i$ for all i implies $-\log p_g = \sum_i p_i(-\log p_g) \leq \sum_i p_i(-\log p_i) = H_s$ (cf. Evans et al. (2026, Lemma 3.1)); summing the per-step bound $-\log p_\phi(y_{g,s} \mid \cdot) \leq H_s(\mathbf{y}_g)$ and dividing by t gives the claim. □

Proof of Proposition 3. Membership in $\mathcal{M}_\epsilon^{(t)}$ forces $r_t(\mathbf{y}) \leq H_t^\star + \epsilon$; membership in $\mathcal{A}_{\epsilon'}^{(t)}$ forces $r_t(\mathbf{y}) \geq H - \epsilon'$. Both hold only if $H - \epsilon' \leq H_t^\star + \epsilon$, contradicting the premise. □

Proof of Proposition 8. Averaging the per-step condition in Eq. (4) over $s \le t$ and applying the triangle inequality gives $|r_t(\mathbf{y}) - \bar{H}_t(\mathbf{y})| < \varepsilon$. For (i), $r_t(\mathbf{y}) < \bar{H}_t(\mathbf{y}) + \varepsilon \le H_t^\star + \kappa + \varepsilon$. For (ii), $\bar{H}_t(\mathbf{y}) < r_t(\mathbf{y}) + \varepsilon \le H_t^\star + \epsilon + \varepsilon$. □

Proof of Corollary 2. By Assumption 2, every prefix of $\mathbf{y}^\circ$ is viable and, being the output's, lies in $\mathcal{B}_t^u$ for all t; the lineage therefore survives the deletion. The output is the argmax over completed hypotheses, so every deleted completed hypothesis scored below it, and removing them cannot change the argmax. Deletion never re-scores or re-ranks retained candidates, so the restricted run is a faithful sub-run of the original. □

Proof of Proposition 4. $I(\mathbf{y}; \hat{\mathbf{y}}_t) = H(\hat{\mathbf{y}}_t) - H(\hat{\mathbf{y}}_t \mid \mathbf{y}) \le H(\hat{\mathbf{y}}_t)$, and the entropy of a random variable supported on $|\mathcal{B}_t|$ values is at most $\log |\mathcal{B}_t|$. □

B Proof of Theorem 1

Write $\mu_s = \mathbb{E}[Z_s \mid \mathbf{y}^\circ_{<s}, \mathbf{x}]$ and $M_t = \sum_{s \le t}(Z_s - \mu_s)$, a martingale with conditionally σ^2-sub-Gaussian increments by assumption (ii). Azuma–Hoeffding for sub-Gaussian increments (Vershynin, 2018) gives, for any $\lambda > 0$ and fixed t,

$$\Pr\big(M_t > t\lambda\big) \;\le\; \exp\!\left(-\frac{t\lambda^2}{2\sigma^2}\right). \tag{16}$$

Set $\lambda_t = \sigma\sqrt{2\log(T/\delta)/t}$, so each step's failure probability is at most δ/T; a union bound over $t \le T$ gives the good event $\mathcal{G} = \big\{M_t \le t\lambda_t \;\; \forall t \le T\big\}$ probability at least $1 - \delta$.

We use the per-request frontier $H_t^\star(\mathbf{x})$ of assumption (i); the global frontier $H_t^\star$ follows identically once $\bar{\epsilon} \ge 2\sup_{\mathbf{x}} \Delta_t(\mathbf{x})$. On $\mathcal{G}$, for every $t \le T$,

$$r_{\mathbf{y}^\circ}(t) \;=\; \tfrac{1}{t}\sum_{s\le t} Z_s \;=\; \tfrac{1}{t}\sum_{s\le t}\mu_s + \tfrac{1}{t}M_t \;\le\; \big(H_t^\star(\mathbf{x}) + \tfrac{\bar{\epsilon}}{2}\big) + \lambda_t \;=\; H_t^\star(\mathbf{x}) + \epsilon_t, \tag{17}$$

using the margin assumption (i) and the margin schedule in Theorem 1.

We conclude by induction on t that $\mathbf{y}^\circ$'s length-t prefix survives every step on $\mathcal{G}$. Suppose it belongs to $\mathcal{B}_{t-1}$ (true at $t = 0$). Then its length-t extension belongs to the expanded pool $\mathcal{B}_t^{\exp}$, by assumption (iii), which holds, e.g., whenever $\mathcal{B}_{t-1} \subseteq \mathcal{B}_{t-1}^u$. The viability test compares $r_{\mathbf{y}^\circ}(t)$ against $\widehat{H}_t^\star + \epsilon_t$. The estimator is a minimum over the expanded pool, a subset of all length-t trajectories, hence $\widehat{H}_t^\star \ge H_t^\star(\mathbf{x})$ (the per-request estimator's pool minimum bounds $H_t^\star(\mathbf{x})$; the global estimator's cross-request minimum bounds $H_t^\star$), and by Eq. (17)

$$r_{\mathbf{y}^\circ}(t) - \widehat{H}_t^\star \;\le\; r_{\mathbf{y}^\circ}(t) - H_t^\star(\mathbf{x}) \;\le\; \epsilon_t, \tag{18}$$

so the test passes. Hence $\mathbf{y}^\circ \in \mathcal{B}_t$, completing the induction: on $\mathcal{G}$ the reference trajectory reaches the horizon and enters the completed pool. The returned output is the argmax over completed hypotheses, so its score is at least $S_T(\mathbf{y}^\circ)$; when $\mathbf{y}^\circ$ is the unique score-maximal completion within the pruned run's search space, that argmax is $\mathbf{y}^\circ$ itself and the outputs coincide. □

Remarks on tightness and scope. If the strict minimum is replaced by an estimator that can *undershoot* $H_t^\star$ (e.g., a debiased estimator), the schedule must absorb the estimation error: $\epsilon_t \to \epsilon_t + \sup_t(H_t^\star - \widehat{H}_t^\star)_+$. The guarantee is relative to the full-width beam output, not the exact MAP; combining Theorem 1 with a search-error analysis of beam search itself is open. Finally, assumption (i) with $\bar{\epsilon} = 0$ and $\sigma = 0$ recovers Assumption 2 exactly, so Theorem 1 strictly generalizes Corollary 2. This holds for both frontiers: with $H_t^\star(\mathbf{x})$ it recovers the per-request cap, and with $H_t^\star$ (and $\bar{\epsilon} \ge 2\sup_{\mathbf{x}} \Delta_t(\mathbf{x})$) the global cap; the global variant (BS) trades the tighter anchor for a wider, difficulty-absorbing margin, while the per-request variant (BS_L) keeps the noise-only margin.

C Optimality of the Viable Set as Survivor Support

We sketch the argument behind Eq. (14): among survivor sets of a given distortion, the surprise-rate cap is minimal in size; dually, among survivor sets of a given size, it is minimal in distortion.

Fix step t and a candidate pool $\mathcal{C}$ with probabilities $\{P(\mathbf{y})\}_{\mathbf{y}\in\mathcal{C}}$. By Lemma 1, the distortion of a restrict-and-renormalize policy with survivor set $\mathcal{S} \subseteq \mathcal{C}$ is $D(\mathcal{S}) = -\frac{1}{t}\log\big(P(\mathcal{S})/P(\mathcal{C})\big)$, a decreasing function of the retained mass $P(\mathcal{S})$. Hence:

Proposition 5. *For any cardinality budget n, the set $\mathcal{S}_n^*$ consisting of the n highest-probability candidates minimizes $D(\mathcal{S})$ over all $\mathcal{S}$ with $|\mathcal{S}| \le n$. Conversely, for any distortion budget D_0, the smallest $\mathcal{S}$ with $D(\mathcal{S}) \le D_0$ is obtained by adding candidates in decreasing probability until the mass constraint is met.*

Proof. Retained mass is additive over members and each member contributes its own probability, so for fixed n mass is maximized—and distortion minimized—by the top-n set; the converse is the standard greedy argument for a one-dimensional knapsack with unit costs. □

Since "the n highest-probability candidates of length t" coincides, up to tie-breaking on the boundary, with a cap $\{r_{\mathbf{y}}(t) \le c\}$ for the appropriate threshold $c = H_t^\star + \epsilon$, the viable set of Definition 2 is the optimal survivor support at every operating point, and the rate–distortion frontier of restrict-and-renormalize policies is traced by sweeping ϵ. Combined with Proposition 4, which converts survivor-set cardinality into retained information rate, this establishes Eq. (14) for the policy class BeamStar searches over. Extending the optimality claim to stochastic encoders beyond restrict-and-renormalize is open.

D Notation

x; $y_{1:t}$	prompt; decoded prefix of length t
$\mathbf{y}_i$, $S_i(t)$	hypothesis i; its cumulative log-probability (Eq. (1))
$r_i(t)$	empirical surprise rate $-S_i(t)/t$ (Eq. (2))
H; $H_t^\star$	entropy rate; global capability frontier (Definition 1)
$H_t^\star(\mathbf{x})$; $\Delta_t(\mathbf{x})$	per-request frontier; difficulty gap $H_t^\star(\mathbf{x}) - H_t^\star$ (Definition 1)
$\mathcal{M}_\epsilon^{(t)}(\mathbf{x})$	per-request viable set (Eq. (8))
$\mathcal{M}_\epsilon^{(t)}$	viable set (Definition 2)
$\mathcal{A}_\epsilon^{(t)}$; $\mathcal{L}_\epsilon^{(t)}$	weakly typical set ;Local typical set
$\mathcal{B}_t$, $\mathcal{B}_t^{\exp}$	survivor set; expanded candidate set at step t
Q_θ	filtered decoding distribution under pruning policy θ
$D(\theta)$	per-symbol KL distortion (Eq. (12))
ϵ_t; (γ)	adaptive margin; its coefficients

E Reproducibility Checklist

We report every detail needed to reproduce Table 1. BeamStar changes only the survivor-selection rule of beam search; baseline and BeamStar runs share the same checkpoints, the same serving stack, the same constrained-decoding trie, and the same machine, and differ solely in line 4 of Algorithm 1.

E.1 Code and artifacts

- **Code.** The implementation is available at `https://anonymous.4open.science/r/beam-star-anon` (anonymized for review; the de-anonymized repository and the exact commit SHA will replace this link in the camera-ready version). It is released under the Apache-2.0 license.
- **Form of the release.** BeamStar ships as a small package (`beam_star`, ∼770 lines) that monkey-patches vLLM's beam-search entry point rather than forking it; no model weights, kernels, or scheduler code are modified. The modules map onto the paper

as follows: `beam_search.py` is the decoding loop with the survivor rule of Algorithm 1; `stats_utils.py` maintains the online mean/variance/max used by the frontier and margin estimators; `beam_star_rd_instrumentation.py` is the rate–distortion ledger implementing the incremental form of Lemma 1; `sampling_params.py`, `protocol.py` and `vllm_patch.py` plumb the new parameters through vLLM's OpenAI-compatible server.

- **Dependency.** `vllm==0.22.1` (pinned), Python ≥ 3.11. Install with `pip install .` from the repository root.
- **Invocation.** BEAMSTAR is enabled by a single flag on the standard vLLM serve command:

```
beam-star-vllm serve <model> --bs [vllm args]
```

  Baseline runs are the same command without `--bs`, which dispatches to vLLM's stock beam search — this is what makes the two arms configuration-identical.

- **Driving the server.** The server exposes the standard OpenAI-compatible chat-completions endpoint, so no bespoke client is needed: the public evaluation harness released with OpenOneRec (Zhou et al., 2026) can be pointed at the BEAMSTAR endpoint and used to issue the benchmark requests and score the returned candidate lists, with the beam width carried by the standard `n` field. This is the path we used, and it is what makes the baseline and BEAMSTAR arms differ only in the server they address. For the Harness entry point and invocation see below (Section E.3).
- **Decoding parameters.** There is no per-request pruning threshold to set: the survivor threshold is learned online at every step from the cumulative-logprob statistics of that token position. Three BEAMSTAR controls are exposed both as request parameters (in `extra_body`, a.k.a. `vllm_xargs` see also example in `README.md`) and as process-level defaults read from the environment at import time, the request value taking precedence:
  - `global_stats` (env `GLOBAL_STATS`, default 1) selects the frontier estimator: 1 accumulates cumulative-logprob statistics across requests at each token position — the global variant, BS — while 0 confines them to the current request, using the spread about that request's own best beam — the per-request variant, BS_{L}.
  - `exploration_tokens` (env `EXPLORATION_TOKENS`, default -1, i.e. none) lists the 0-based generated-token indices at which pruning is disabled, realizing the exploration phase of Remark 5; the video configuration uses a single such index.
  - `gamma` (env `GAMMA`, default 2.0) scales the standard-deviation term of the global-stats threshold – see Section 5.5.

  Because these parameters are per-request, the BS and BS_{L} rows of Table 1 can be produced against one running server, with no restart between conditions and hence no confounding change in cache or warm-up state. Alternatively, if you want to avoid modifying the benchmark code (Section E.3), you can set these values as environment variables per run.
- **Scope limitation of the released code.** The cross-request statistics are held in a fixed array of four per-step accumulators, so the released implementation supports decoding horizons of at most four steps. This is sufficient for the three-layer semantic-ID trajectories evaluated here but is not a general-horizon implementation; extending it is a matter of sizing the accumulator by `max_tokens`.
- **Logs.** Setting `DEBUG=1` writes a JSONL trace to `logs/beam_star_debug/`: one record per decoding step with the retained beam count $|B_t|$, the realized threshold, and the surprise-rate spread $\widehat{\sigma}_t$, followed by an end-of-request summary carrying D_{tot}, the retained mass $e^{-D_{\mathrm{tot}}}$, the realized margins, and the operational rate $\mathbb{E}|B_t|$. These traces are the raw material for the Hypothesis 1 histogram and for the distortion figures.

E.2 MODELS AND CHECKPOINTS

- **Models.** We use the OpenOneRec-1.7B and OpenOneRec-8B-Pro (Zhou et al., 2026) checkpoints released in January 2026, downloaded from Hugging Face with the default configuration, i.e., bf16 precision.

- **Semantic IDs.** Semantic ID generation is out-of-scope for this work. The dataset provides this data. In addition, the Semantic IDs constraint imposed in this work is consistent with the results reported in (Zhou et al., 2026).

E.3 DATA

- **Benchmark.** OpenOneRec-RecIF-Bench, a multi-domain recommendation instruction-following benchmark spanning short video, advertising, and e-commerce (202,359 users, 15.3M items, 119.9M interactions), reporting the *interactive* and *video* tasks. The video task draws on the short-video domain, the largest of the three (195,026 users, 13.1M items, 94.4M interactions).
- **Harness.** OpenOneRec releases an evaluation harness alongside the benchmark (`benchmarks/` in the project repository), covering the task prompts, the constrained-decoding setup, and the scoring code for every RecIF-Bench metric. We use it rather than reimplementing evaluation, so the pass@1, pass@32 and recall@32 figures in Table 1 are produced by the benchmark authors' own scoring code and are directly comparable to the numbers published with the models. Only the generation call is adapted to address a running server (Section E.1); prompts, constraints and metric definitions are untouched.
- **Split.** Evaluation uses the held-out test split as distributed; no data is used for tuning beyond the margin coefficients (Section E.4). Split sizes: $N = 1000$ and $N = 38781$ for interactive and video datasets,respectively.
- **Subsampling.** Rows marked † in Table 1 use a fixed 1000-request subset of the video task, drawn once with seed 0 and reused verbatim for every method and beam width; other rows use the full parquet.

E.4 HYPERPARAMETERS

- **Beam width.** $k \in \{32, 128, 512\}$, identical nominal width for baseline and BEAMSTAR; BEAMSTAR only prunes below it.
- **Frontier estimator.** $\widehat{H}_t^\star$ is the minimum surprise rate in $\mathcal{B}_t^{\exp}$, i.e. the rate of the highest-scoring candidate. $\mathrm{BS_L}$ takes this minimum over the current request's pool, estimating $H_t^\star(\mathbf{x})$; BS takes it over the cumulative scores pooled across requests at step t, estimating $H_t^\star$ (Eq. (6)). The pooled statistics accumulate over the entire serving session; an aging mechanism (e.g. exponential forgetting) to track drift in the request mix and discount outliers is left to future work.
- **Margin.** BS uses $\epsilon_t = \gamma\,\widehat{\sigma}_t$, where $\widehat{\sigma}_t$ is the pooled standard deviation of cumulative scores at step t, with $\gamma = 2$ (interactive) and $\gamma = 2.2$ (video). $\mathrm{BS_L}$ uses $\epsilon_t = \frac{1}{\sqrt{2}}\big(1 + \widehat{\sigma}_t^{\mathrm{w}}\, t^{-1/2} + \frac{1}{4}\, t^{-1}\big)$, where $\widehat{\sigma}_t^{\mathrm{w}}$ is the within-request spread of beam scores; its constants are fixed and not tuned. Steps are 1-indexed ($t \geq 1$).
- **Exploration.** The video configuration uses one unpruned exploration token ($t_0 = 1$ in the two-phase schedule of Section G.9); interactive uses $t_0 = 0$.
- **Distortion.** No distortion budget is enforced; pruning is governed by the margin alone, and the realized distortion is only logged (line 5 of Algorithm 1).
- **Tuning protocol.** γ is selected on a small calibration set and held fixed across models and beam widths (sweep in Section F.1); the constants of the $\mathrm{BS_L}$ margin are not tuned. No per-cell tuning was performed.

E.5 HARDWARE, SOFTWARE, AND TIMING PROTOCOL

- **Hardware.** One NVIDIA A100 40 GB, host Intel(R) Xeon(R) Gold 5320T CPU @ 2.30GHz with 500GB in RAM.
- **Software.** see requirement in pyproject.toml in git repo.

E.6 RUNS, SEEDS, AND VARIANCE

- **Determinism of quality.** Beam search and BEAMSTAR are deterministic given a checkpoint and a request: no sampling is involved, so pass@1, pass@32, and recall@32 in Table 1 carry no seed variance and are reported from a single pass.
- **Timing variance.** "Time per sample" is wall-clock for prefill+decode, measured at batch size of 1, averaged over the requests of each split (Section E.3). The standard deviations of the average time measurements are roughly 2ms which is negligible and were thus omitted.
- **Compute total.** To reproduce the result in Table 1 takes approximately 24 GPU-hours on NVIDIA A100.

F EMPIRICAL CHARACTERIZATION: RATE, DISTORTION, AND THE VIABLE SET

This section measures the quantities of Sections 4 and 5 on the OpenOneRec-1.7B interactive task of Section 6, for both variants: global BEAMSTAR (BS, $\gamma = 2$) and local BEAMSTAR (BS$_L$, whose margin constants are fixed and untuned). Each run decodes 1,000 requests at $k = 32$. All numbers come from the logs that BEAMSTAR already writes: per step, the threshold, the spread statistic and the number of survivors; per request, the distortion ledger of Algorithm 1 (line 5). No extra forward passes were run.

Each OpenOneRec output has five tokens: a begin delimiter, three semantic-ID codes, and an end delimiter. BEAMSTAR applies its viability test at the first four positions ($t = 1, \dots, 4$) and never at the end delimiter ($t = 5$), where it intentionally keeps all surviving candidates. As a result, its output list can be compared with the width-k baseline under the same recall@k. All per-step statistics below refer to the four pruned positions.

F.1 OPERATIONAL RATE–DISTORTION CURVE

Figure 1 sweeps the BS margin coefficient $\gamma \in \{3.5, 3, 2.5, 2, 1.5, 1\}$ and plots, for each γ, the mean distortion per position D against the mean rate $R = \mathbb{E}[\log_2 |\mathcal{B}_t|]$. The full-width baseline sits at $D = 0$ and $R = \log_2 k = 5$ bits. The curve is decreasing and convex, the shape an operational rate–distortion curve must have. Its left end is almost vertical: going from $\gamma = 3.5$ to $\gamma = 2.5$ removes about 0.17 bits of rate (about 11% of the surviving width) at $D \approx 0.002$ nats. Beyond that point each further bit of rate costs increasingly more distortion. The deployed setting, $\gamma = 2$, sits at the knee: $D = 0.040$ nats ($D_{\mathrm{tot}} = 0.16$ nats per request, i.e. 87% of the candidate mass retained on average) and $R = 3.43$ bits, with $\mathbb{E}|\mathcal{B}_t| = 19.7$ of 32 beams. At this setting the list metrics of Table 1 match the baseline, even though the mass-based certificate of Proposition 9 bounds the recall loss only by $c\,\mathbb{E}[1 - e^{-D_{\mathrm{tot}}}] = 0.126\,c$. The certificate is conservative here: pruned mass is not the same as pruned ground truth.

BS$_L$ has no coefficient to sweep and gives a single operating point: $D = 0.053$ nats ($D_{\mathrm{tot}} = 0.21$, 82% of the mass retained) at $R = 3.18$ bits ($\mathbb{E}|\mathcal{B}_t| = 18.1$). This lies below the piecewise-linear interpolation of the BS curve at the same distortion (about 3.37 bits), so the per-request anchor buys a more efficient rate–distortion trade-off with no tuning. It spends that distortion differently, though: it keeps the top output (pass@1 in Table 1) but trails BS on the list metrics.

F.2 WHERE PRUNING HAPPENS

Table 2 and Fig. 2 break both runs down by position. The two variants share one pattern.

Begin delimiter ($t = 1$). The begin token is fixed by the output format, so the model assigns it probability close to one: every request has $\widehat{H}_1^\star(\mathbf{x}) \approx 0$. Both variants keep a single candidate at a cost of $\Delta D_1 \approx 3 \times 10^{-4}$ nats. The other $k - 1$ candidates, which a fixed-width search carries into the next step, are continuations the format rules out. BEAMSTAR discards them without any format-specific rule, because they lie far outside the viable set.

First code ($t = 2$): where rate and distortion are spent. BS keeps a median of 17 of 32 candidates, and BS$_L$ keeps 9. This position contributes 59% (BS) and 81% (BS$_L$) of the total distortion. The

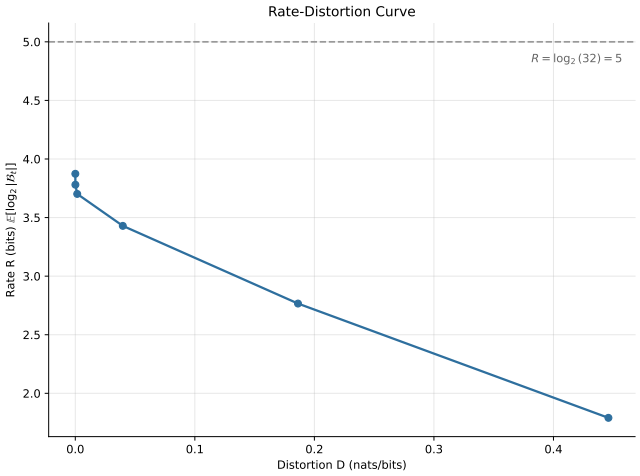


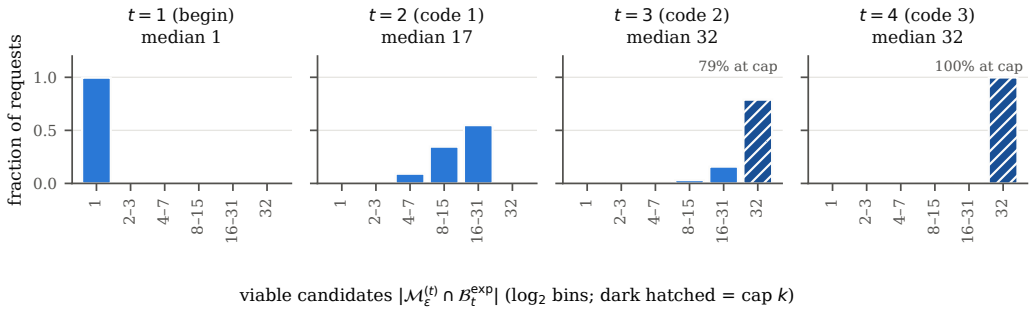
Figure 1: Operational rate–distortion curve on the interactive task ($k = 32$). BS points from left to right: $\gamma = 3.5, 3, 2.5, 2, 1.5, 1$; BS$_L$ is a single point. Rate is $\mathbb{E}[\log_2 |\mathcal{B}_t|]$ (full width: 5 bits); distortion is the ledger value $D_{\text{tot}} = -\log P(\mathcal{S})$ divided by the four pruned positions, in nats, averaged over requests.

difference comes from how the cap $\widehat{H}_2^\star + \epsilon_2$ is split between anchor and margin. BS anchors every request at the best score seen across the session and adds a pooled margin wide enough to cover the spread between requests. BS$_L$ anchors each request at its own best candidate and adds only a within-request margin. As the survivor counts show, the per-request cap is tighter at the first code, so BS$_L$ prunes more here. This is the allocation of the difficulty gap described in Section 5, observed directly.

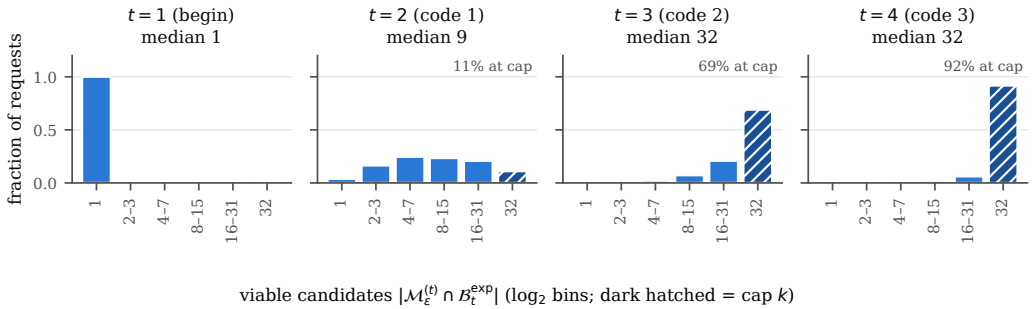
Second and third codes ($t = 3, 4$): the viable set covers most of the pool. At $t = 3$ every candidate is viable on 79% (BS) and 69% (BS$_L$) of requests. At $t = 4$, BS keeps the full pool on every request, while BS$_L$ still removes candidates on 8% of requests ($\Delta D_4 = 0.004$). Together with its tighter cap at $t = 2$, this accounts for BS$_L$'s gap on the list metrics. Rate is saved on the begin delimiter and the first code, where the best continuations stand out from the tail.

Table 2: Per-position breakdown on 1,000 interactive requests ($k = 32$). $t = 1$: begin delimiter; $t = 2, 3, 4$: semantic-ID codes; the end delimiter is never pruned. For BS$_L$: $\widehat{\sigma}_t^{\text{w}}$, the within-request spread; $\widehat{H}_t^\star(\mathbf{x})$, the per-request frontier (median over requests); ϵ_t, the margin in rate units. The corresponding pooled statistics of BS were not logged. $|\mathcal{M}_\epsilon^{(t)}|$: viable candidates in $\mathcal{B}_t^{\text{exp}}$, median [IQR]; "at cap": fraction of requests with every candidate viable; ΔD_t: mean distortion increment (nats). Per request: BS $D_{\text{tot}} = 0.16$, retained mass 0.87, $\mathbb{E}|\mathcal{B}_t| = 19.7$; BS$_L$ $D_{\text{tot}} = 0.21$, retained mass 0.82, $\mathbb{E}|\mathcal{B}_t| = 18.1$.

	t	$\widehat{\sigma}_t^{\text{w}}$	$\widehat{H}_t^\star(\mathbf{x})$	ϵ_t	$\lvert\mathcal{M}_\epsilon^{(t)}\rvert$	at cap	ΔD_t
BS	1	–	–	–	1 [1, 1]	0%	0.0003
	2	–	–	–	17 [11, 21]	0%	0.093
	3	–	–	–	32 [32, 32]	79%	0.065
	4	–	–	–	32 [32, 32]	100%	0.0002
BS$_L$	1	2.59	0.00	2.72	1 [1, 1]	0%	0.0003
	2	1.11	0.73	1.35	9 [4, 19]	11%	0.174
	3	0.75	1.10	1.07	32 [27, 32]	69%	0.036
	4	0.79	1.04	1.03	32 [32, 32]	92%	0.004



(a) BS (global, $\gamma = 2$)



(b) BS_L (local)

Figure 2: Distribution over requests of the viable-set size $|\mathcal{M}_\epsilon^{(t)} \cap \mathcal{B}_t^{\text{exp}}|$ per pruned position on the interactive task ($k = 32$; $\log_2$ bins; the hatched bar is the cap k). The begin delimiter collapses to one candidate, pruning concentrates at the first code, and by the later codes the viable set covers most of the pool. The end delimiter, never pruned, is not shown.

F.3 Per-request frontier estimates

Figure 3 shows the per-request frontier estimates $\widehat{H}_t^\star(\mathbf{x})$ of BS_L. At the begin delimiter every request has $\widehat{H}_1^\star(\mathbf{x}) \approx 0$, as the format dictates. Across the codes the frontier is higher at the later positions than at the first (medians $0.73 \to 1.10 \to 1.04$): later codes are less predictable than the first, as expected when a coarse-to-fine tokenizer spreads fine-grained distinctions over later positions. The frontier is a property of the model and task at each position, not a fixed schedule. Its spread across requests (IQR 0.44–1.07 at the first code) is the difficulty variation that a single global anchor has to absorb through its margin, which is why the global variant needs a wider margin to reach the same safety.

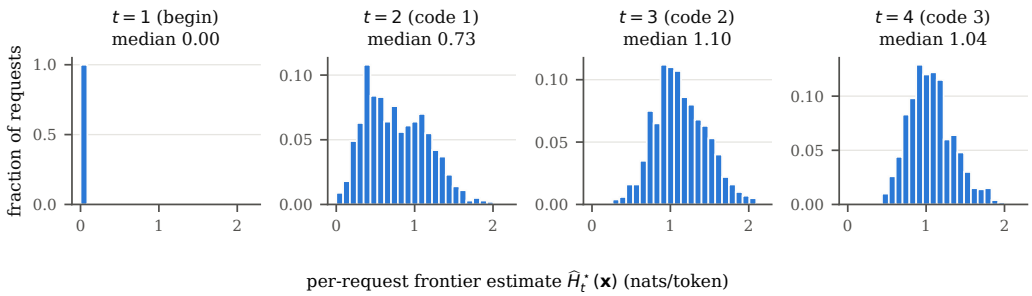


Figure 3: Distribution over requests of the per-request frontier estimate $\widehat{H}_t^\star(\mathbf{x})$ of BS_L at each pruned position (interactive, $k = 32$).

Scope. The entropy rate H, and hence the compression exponent $H - H_t^\star$ of Proposition 6, requires a separate ancestral-sampling pass and is not reported here.

G Additional Material

G.1 Two Exact Decompositions of the Distortion

Remark 2 (Two exact decompositions; no per-beam distortion exists)**.** Over beams: *partition the surviving trajectory space into the cones $A_1, \dots, A_k$ of the k returned hypotheses. The between/within identity gives $D_{\mathrm{KL}}(Q_\theta \,\|\, P) = \sum_i Q_\theta(A_i) \log \frac{Q_\theta(A_i)}{P(A_i)} + \sum_i Q_\theta(A_i) D_{\mathrm{KL}}(Q_\theta(\cdot \mid A_i) \,\|\, P(\cdot \mid A_i))$; for restrict-and-renormalize every within-cone term vanishes ($Q_\theta(\cdot \mid A_i) = P(\cdot \mid A_i)$) and every between-ratio equals $1/P(\mathcal{S})$, so beam i's share is $Q_\theta(A_i)(-\log P(\mathcal{S}))$: mass-proportional, with a log-ratio common to all beams. No beam-local distortion signal exists; the distortion is a property of the retained* set*, which is also why no output-level score comparison can recover it (pruning never alters a survivor's score, only the set's composition).* Over time: *by the chain rule, $D_{\mathrm{KL}}(Q_\theta \,\|\, P) = \sum_{s \le t} \mathbb{E}_{Q_\theta}[D_{\mathrm{KL}}(Q_\theta(y_s \mid y_{<s}) \,\|\, P(y_s \mid y_{<s}))]$, and each summand is exactly the incremental charge ΔD of line 5 in Algorithm 1, telescoping to $-\log P(\mathcal{S})$: the online ledger* is *the chain-rule decomposition—an identity, not a bound—and the only observation channel for foregone mass, which is computable at prune time and unrecoverable afterwards.*

Corollary 1 (Telescoping distortion)**.** *For a restrict-and-renormalize policy, let $P(\mathcal{B}_t) = \Pr(\mathbf{y}_{1:t} \in \mathcal{B}_t)$ be the mass surviving through step t (with $P(\mathcal{B}_0) = 1$, $P(\mathcal{B}_T) = \mathcal{S}$; under $P(\cdot \mid \mathbf{x})$). Then*

$$-\log P(\mathcal{S}) \;=\; \sum_{t=1}^{T} \Delta D_t, \qquad \Delta D_t \;=\; \log \frac{P(\mathcal{B}_{t-1})}{P(\mathcal{B}_t)} \;\ge\; 0. \tag{19}$$

Proof. Survivors are nested ($\{\mathbf{y}_{1:t} \in \mathcal{B}_t\} \subseteq \{\mathbf{y}_{1:t-1} \in \mathcal{B}_{t-1}\}$), so $P(\mathcal{S}) = \prod_{t=1}^{T} P(\mathcal{B}_t)/P(\mathcal{B}_{t-1})$; take $-\log$. Each factor is ≤ 1, giving $\Delta D_t \ge 0$. □

The increments ΔD_t are exactly line 5 of Algorithm 1, so the closed-form distortion accumulates online; this is the chain-rule decomposition of $D_{\mathrm{KL}}(Q_\theta \,\|\, P)$, with the online estimator normalizing against the post-top-k pool.

G.2 A general treatment of beam pruning policies

Meister et al. (2023) present truncated *sampling* in a unified template—a per-step truncation set defined by a subset-optimization problem—and cast top-k (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2020) as instances before defining their own. We give the analogous template for beam *pruning*. At step t, a pruning policy selects a survivor set $\mathcal{S} \subseteq \mathcal{B}_t^{\exp}$ from the expanded candidate pool; by Lemma 1, its distortion is determined by the retained mass $Z(\mathcal{S}) = \sum_{\mathbf{y} \in \mathcal{S}} P(\mathbf{y})$.

Fixed-width beam search. The survivor set solves

$$\underset{\mathcal{S}}{\text{maximize}}\; Z(\mathcal{S}) \quad \text{subject to} \quad |\mathcal{S}| \le k, \tag{20}$$

the prefix-space analogue of top-k: a hard *rate* constraint with no distortion control. When k exceeds the viable population, the maximizer is forced to include off-manifold trajectories (Remark 1).

Score-gap pruning (Freitag & Al-Onaizan, 2017). The survivor set is $\{\mathbf{y} : S_{\mathbf{y}}(t) \ge \max_j S_j(t) - \beta\}$, a cap on the *cumulative* score gap. Because the gap is not normalized by t, the criterion is length-inconsistent: a fixed β becomes vanishingly strict per token as t grows.

BeamStar. The survivor set solves the dual of fixed-width search,

$$\underset{\mathcal{S}}{\text{minimize}}\; |\mathcal{S}| \quad \text{subject to} \quad -\tfrac{1}{t} \log \frac{Z(\mathcal{S})}{Z(\mathcal{B}_t^{\exp})} \;\le\; \epsilon_t, \tag{21}$$

a hard *distortion* constraint with the rate minimized—and by Proposition 5, its solution is exactly the surprise-rate cap $\{\mathbf{y} : r_{\mathbf{y}}(t) - \widehat{H}_t^\star \le \epsilon_t\}$ (up to boundary ties).

The template makes the family resemblance exact: BEAMSTAR is to beam search what nucleus sampling is to ancestral sampling—a mass-targeted truncation rather than a count-targeted one—but lifted from tokens to trajectories and anchored at the capability frontier rather than at cumulative token mass. It likewise clarifies the relation to entropy-adaptive branching (Evans et al., 2026): EDEN modulates how many children each node *spawns* (rate allocated where local entropy is high), while BEAMSTAR modulates which hypotheses *survive* (rate reclaimed where surprise exceeds the frontier); Proposition 7 links their control signals, and the two compose.

Two frontier estimators: global (BS) and per-request (BS$_L$). The viability test $r_t(\mathbf{y}) \le \widehat{H}_t^\star + \epsilon_t$ is agnostic to how the frontier is estimated, and the two readings of Definition 1 give the two variants we evaluate in Section 6. *Global* BEAMSTAR (BS) estimates $\widehat{H}_t^\star$ from statistics pooled *across requests* at step t (a running maximum and spread of cumulative log-probabilities), approximating the model frontier $H_t^\star$; its margin then absorbs the difficulty gap $\Delta_t(\mathbf{x})$. *Local* BEAMSTAR (BS$_L$) estimates $\widehat{H}_t^\star(\mathbf{x})$ from the current request's own expanded pool $\mathcal{B}_t^{\exp}$, approximating $H_t^\star(\mathbf{x})$; its anchor is tighter and its margin need only cover within-request noise. By Eq. (8), the two realize one viability test, with the difficulty gap allocated to the margin (BS) or to the anchor (BS$_L$); Theorem 1 covers both, with the floor split stated there.

G.3 VOLUME RELATIVE TO THE TYPICAL SET

Proposition 6 (Exponential compression)**.** *Let $H(\mathcal{Y})$ denote the entropy rate of the decoding distribution and suppose an AEP holds for it (see Assumption 1 and the discussion therein; for stationary ergodic sources this is Shannon-McMillan-Breiman), so that for any $\epsilon' > 0$ and t large, $|\mathcal{A}_{\epsilon'}^{(t)}| \ge (1-\epsilon')\, e^{t(H-\epsilon')}$. If $H_t^\star + \epsilon < H$, then*

$$\frac{\left|\mathcal{M}_{\epsilon}^{(t)}\right|}{\left|\mathcal{A}_{\epsilon'}^{(t)}\right|} \;\le\; \frac{e^{t(H_t^\star+\epsilon)}}{(1-\epsilon')\, e^{t(H-\epsilon')}} \;=\; \frac{1}{1-\epsilon'}\, e^{-t\,(H-H_t^\star-\epsilon-\epsilon')} \;\xrightarrow[t\to\infty]{}\; 0 \tag{22}$$

exponentially fast. The viable set occupies an exponentially vanishing fraction of the volume the model's samples occupy.

The gap $H - H_t^\star$ measures how "peaked" the model is around its best continuations; empirically, instruction-tuned and recommendation-tuned models exhibit low per-token entropy on-manifold, making the compression factor large. Section F reports the measured viable-set size and per-request frontier on OpenOneRec.

G.4 DISCUSSION OF THE SURVIVAL THEOREM

The proof (Section B) is an induction over steps: on the good event of an Azuma-type bound, the reference prefix's surprise rate stays within ϵ_t of the *true* frontier, and since the estimated frontier satisfies $\widehat{H}_t^\star \ge H_t^\star$ (a minimum over a subset), the empirical test in Algorithm 1 is only easier to pass. Exactness needs the extra condition because online pruning frees beam slots, letting the search carry lineages the unpruned run truncated at its top-k cut; such a promoted lineage can complete with a higher score than $\mathbf{y}^\circ$, in which case BEAMSTAR returns a strictly better-scoring output rather than the identical one. Assumption (ii) is the standard sub-Gaussian noise model of score-estimation analyses (Vershynin, 2018; Evans et al., 2026); assumption (i) is Assumption 2 relaxed to hold in conditional expectation with slack $\bar{\epsilon}/2$, and is the honest core of the theorem—it is a statement about how quickly good continuations announce themselves on the task distribution, measurable directly via the Hypothesis 1 histogram.

G.5 FRONTIER AND LOCAL-TYPICALITY BOUNDS

Proposition 7 (The frontier lies below the local-entropy average)**.** *Let $\mathbf{y}_g$ denote the greedy trajectory. Then for every t,*

$$H_t^\star \;\le\; r_t(\mathbf{y}_g) \;\le\; \tfrac{1}{t}\sum_{s\le t} H_s(\mathbf{y}_g). \tag{23}$$

Proposition 7 captures "the viable set lies beneath the entropy shell": the cap's anchor is bounded by the average *local* entropy along the most straightforward path. In sharply peaked models, where each step's next-token distribution concentrates most probability mass on a small set of tokens, this local entropy is much smaller than the entropy rate $H(\mathcal{Y})$, so the anchor sits far below the entropy shell.

Proposition 8 (Local typicality along low-entropy corridors)**.** *Let* $\mathbf{y} \in \mathcal{L}_\varepsilon^{(t)}$ *be locally typical and let* $\bar{H}_t(\mathbf{y}) = \frac{1}{t}\sum_{s\le t} H_s(\mathbf{y})$*. Then* $\left|r_t(\mathbf{y}) - \bar{H}_t(\mathbf{y})\right| < \varepsilon$*, and consequently: (i) if* $\bar{H}_t(\mathbf{y}) \le H_t^\star + \kappa$*, then* $\mathbf{y} \in \mathcal{M}_{\varepsilon+\kappa}^{(t)}$*; (ii) conversely, if* $\mathbf{y} \in \mathcal{M}_\epsilon^{(t)}$*, then* $\bar{H}_t(\mathbf{y}) \le H_t^\star + \epsilon + \varepsilon$*.*

In words, locally typical trajectories are viable *exactly when they run along low-entropy corridors* of the process, when prefixes whose conditional entropies hug the frontier. This formalizes an observation Meister et al. (2023) make in passing (their footnote 20). In machine translation models, which have low conditional entropies, beam search "incidentally" enforces local typicality. Proposition 8 says the two notions coincide precisely in that regime, and BEAMSTAR can be read as enforcing the trajectory-level analogue deliberately: where locally typical *sampling* keeps each token near the local entropy, BEAMSTAR keeps each *hypothesis* near the global frontier.

G.6 PREFIX STABILITY AND GARDEN-PATH EXCURSIONS

Remark 3 (Prefix stability is not free)**.** *Viability is defined stepwise over prefixes and is* not *automatically* prefix-closed*: membership of a full trajectory in* $\mathcal{M}_\epsilon^{(T)}$ *does not imply membership of its length-*t *prefixes in* $\mathcal{M}_\epsilon^{(t)}$*, because* r_t *is a running average. The eventual best trajectory may therefore pass through a "garden-path" prefix—an opening whose single rare-but-correct token inflates the running average before later tokens amortize it (a first token of* 5 *nats followed by* 0.5*-nat tokens gives* $r_1 = 5$ *but* $r_{10} \approx 0.95$*, entering an* $\epsilon = 0.5$ *cap only near* $t \approx 10$*)—temporarily far from the step-*t *frontier and culled by a hard cap: the classical search-error mode of beam search.*

G.7 LANGUAGE MODELS AS STOCHASTIC PROCESSES

Language models as stochastic processes. Following Meister et al. (2023), a conditional language model can be viewed in two ways: as a single string-valued random variable or as a discrete stochastic (language) process $\mathbf{y} = \{y_s\}_{s=1}^{\infty}$ with conditionals $p_\phi(y_s \mid \mathbf{y}_{<s}, \mathbf{x})$. The process view allows information-theoretic analysis; however, it comes with a caveat that Meister et al. (2023) identifies. The EOS is an absorbing state in the language process, and thus, such process is *non-ergodic by construction*. Moreover, since its conditional distributions drift with the prefix, it is neither stationary nor i.i.d. Consequently, the classical results proved for stationary ergodic sources (Shannon–McMillan-Breiman and its corollaries; Breiman (1957); Cover & Thomas (2006)) apply to LM decoding only as modeling approximations, never off-the-shelf. We state every result in the following unconditionally (counting arguments, which need no process assumptions) or under an explicitly flagged concentration assumption (Assumption 1).

G.8 WHY THE $D_{\mathrm{KL}}Q_\theta P$ DIRECTION

Remark 4 (Distortion under support collapse: why $D_{\mathrm{KL}}(Q_\theta \,\|\, P)$)**.** *Hard pruning makes* $\mathrm{supp}(Q_\theta)$ *a* strict subset *of* $\mathrm{supp}(P)$*: every completion descending through a pruned prefix has* $Q_\theta(\mathbf{y}) = 0$ *while* $P(\mathbf{y}) > 0$*. Two departures from Eq.* (5) *are then forced, not chosen. First, no* pointwise *distortion averaged under the source can represent pruning:* $\mathbb{E}_P[\log P/Q_\theta] = \sum_{\mathbf{y}\in\mathcal{S}} P\log(P/Q_\theta) + \sum_{\mathbf{y}\notin\mathcal{S}} P\cdot\infty = \infty$ *for every nontrivial policy—the source keeps charging for trajectories the encoder no longer represents—and* $D_{\mathrm{KL}}(P\,\|\,Q_\theta)$ *diverges identically. The unique finite choice is the* forward *direction under the filtered marginal,* $D(\theta) \coloneqq \frac{1}{t}\,\mathbb{E}_{\mathbf{y}\sim Q_\theta}[\log Q_\theta(\mathbf{y}) - \log P(\mathbf{y})]$*, finite because the expectation runs only over* $\mathrm{supp}(Q_\theta) \subseteq \mathrm{supp}(P)$*. It carries the right operational reading: the* excess surprise*, in nats, of describing the filtered decoder's outputs with the original model—the zero-forcing direction, matching the fact that pruning is deliberate mode-seeking. Throughout, information quantities are taken w.r.t. the source randomness* $\mathbf{y} \sim P(\cdot \mid \mathbf{x})$*; beam search is deterministic given* $\mathbf{x}$*, so Eq.* (11) *compresses the* model's trajectory distribution*, not any single run.*

G.9 The Schedule as Exploration-then-Exploitation

Remark 5 (The schedule is exploration-then-exploitation)**.** *The schedule margin ϵ_t in Theorem 1 is not an engineering choice but a consequence of concentration: the margin must absorb $O(\sqrt{\log(T/\delta)/t})$ of stochastic excursion, so it is* wide at small t and tightens as $1/\sqrt{t}$. *Early decoding, where surprise-rate averages are noisy, warrants a permissive cap (exploration); late decoding, where rates have concentrated, warrants aggressive compression (exploitation). The two-phase schedule proposed in Section H.1 is the hard-threshold limit of ϵ_t, which thus supplies its principled soft form.*

G.10 The Distortion Budget Certifies Recall

Proposition 9 (The distortion budget certifies recall)**.** *Let $\widetilde{P}$ be the task's ground-truth distribution over completions and assume model coverage: there is $c < \infty$ with $\widetilde{P}(\mathbf{y}) \le c\,P(\mathbf{y} \mid \mathbf{x})$ for all $\mathbf{y}$. Let $D_{\mathrm{tot}} = -\log P(\mathcal{S})$ be the total distortion accumulated by Algorithm 1 over a run, $\mathcal{S}$ the surviving cone. Then the expected recall@k degradation of* BeamStar *relative to the same-width unpruned run obeys*

$$\mathbb{E}\big[\Delta\,\mathrm{recall@}k\big] \;\le\; \Pr(\textit{ground-truth lineage pruned}) \;\le\; c\big(1 - e^{-D_{\mathrm{tot}}}\big) \;\le\; c\,D_{\mathrm{tot}}. \tag{24}$$

Proof. A ground-truth completion in the unpruned list is absent from BeamStar's list only if its lineage is pruned: pruning removes competitors and can only *improve* a survivor's rank, under assumption (iii) of Theorem 1, so it never creates the rank-beyond-k failure. The pruned region has model mass $1 - P(\mathcal{S})$; coverage transfers this to $\widetilde{P}$-mass at cost c; and $1 - e^{-x} \le x$. □

Read with Theorem 1, the two guarantees map onto the two metric families: survival protects the *top* output (pass@1), while the distortion budget bounds *list* loss (pass@k, recall@k). The coverage constant c is measurable on validation data as a high quantile of $\widetilde{P}/P$, turning the bound into a computable certificate.

G.11 Sufficiency for Decoding: Prefix Stability and Sufficient Width

Assumption 2 (Prefix stability)**.** *The trajectory $\mathbf{y}^\circ$ returned by the unpruned decoder on input $\mathbf{x}$ satisfies, at* every *intermediate step $t \le T$, $r_t(\mathbf{y}^\circ \mid \mathbf{x}) \le H_t^\star(\mathbf{x}) + \epsilon$ under the per-request frontier, or equivalently $r_t(\mathbf{y}^\circ \mid \mathbf{x}) \le H_t^\star + \epsilon$ under the global frontier once $\epsilon \ge \Delta_t(\mathbf{x})$, the frontier being taken over length-t prefixes. The two forms differ only by the difficulty gap $\Delta_t(\mathbf{x})$ of Definition 1: the global cap must be wider by $\Delta_t(\mathbf{x})$ to admit a hard input's own reference trajectory, matching Eq.* (8).

Let $\{\mathcal{B}_t^u\}_t$ be the survivor sets of an unpruned width-k beam search. Under Assumption 2, the following holds.

Corollary 2 (Sufficient beam width)**.** *Delete from the unpruned run every candidate outside $\mathcal{M}_\epsilon^{(t)}$, together with its descendants. Then the returned top output is unchanged, the surviving list members retain their scores and relative order, and the width used at step t is $|\mathcal{M}_\epsilon^{(t)} \cap \mathcal{B}_t^u|$.*

Note the returned *list* may shrink: non-viable completed hypotheses drop out, so the guarantee concerns the top output and the retained members' scores—not the full k-list distribution. The corollary holds for either frontier: with the per-request frontier the deletion keeps $\mathbf{y}^\circ$ by its conditional viability; with the global frontier it does so provided $\epsilon \ge \Delta_t(\mathbf{x})$, i.e. the wider cap of Assumption 2.

Viability is not prefix-closed: a good trajectory may pass through a "garden-path" prefix, temporarily above the cap, that a hard threshold would cull (Section G).

Corollary 2 converts the geometry into an algorithmic license: the decoder may shrink its active set to the viable trajectories it currently holds without changing what it can return, provided ϵ is chosen (or adapted) so that the eventual output remains inside the cap at every step. Quantifying the cost of a finite ϵ, including cases where Assumption 2 is violated, is exactly a distortion question, which the next section formalizes.

H Discussion and Proposed Extensions

H.1 Exploration–exploitation decoding

BeamStar as presented is purely exploitative: it contracts the search onto the viable cap from the first step. Some tasks, however, benefit from early diversity—long-horizon generation where the best trajectory is not prefix-greedy, diverse-list objectives in recommendation, and reasoning tasks where distinct solution strategies branch early. We therefore propose a two-phase schedule:

1. **Exploration phase** ($t \le t_0$): decode with the full nominal beam and *no* viable-set projection, deliberately admitting high-surprise and even OOD prefixes so that structurally different continuations get scored.
2. **Exploitation phase** ($t > t_0$): enable BeamStar, projecting the accumulated population onto the viable set and spending the remaining budget only on trajectories near the frontier.

In rate–distortion terms this is a time-varying budget δ_t: large (or infinite) for $t \le t_0$, then annealed. Natural research questions: (i) does exploration recover cases where step-0 projection locks in a suboptimal prefix (measurable as pass@k gaps at fixed compute)? (ii) can t_0 itself be adaptive, triggered when the frontier estimate $\widehat{H}_t^\star$ stabilizes? (iii) how does the schedule interact with length normalization? A soft variant replaces the hard switch with a temperature on the margin, $\epsilon_t \propto \epsilon_0 \cdot \gamma^{-t}$.

H.2 Empirical characterization of the viable set

Section F measures the viable-set size, the per-request frontier, and the operational rate–distortion curve. Two measurements remain: the entropy rate H (hence the compression exponent of Proposition 6) via ancestral sampling, and the retained mass $P(\mathcal{M}_\epsilon^{(t)})$ at larger t via importance sampling.

H.3 Broader benchmarks

The rate–distortion formulation is domain-agnostic; establishing it beyond generative recommendation requires:

- **Machine translation** (WMT En–De/En–Zh): the historical home of the beam-search curse; test whether viable-set projection removes the quality degradation at large k reported by Cohen & Beck (2019).
- **Speech recognition** (Whisper-style beam decoding on LibriSpeech): latency-sensitive, moderate beams.
- **Code generation** (HumanEval/MBPP, pass@k): pass@k directly rewards a well-chosen candidate *set*, probing whether aggressive pruning harms set diversity.
- **Mathematical reasoning** (GSM8K/MATH with self-consistency or beam-based tree search): tests the exploration–exploitation schedule of Section H.1, since distinct solution paths separate early.
- **Summarization** (XSum/CNN-DM): long outputs stress the online frontier estimate over hundreds of steps.

Across all of these, the headline plot is the same: task quality versus realized compute $\mathbb{E}[|\mathcal{B}_t|]$, with BeamStar tracing a strictly better frontier than fixed-width beams and score-gap pruning.

H.4 Systems integration

Because BeamStar only shrinks the survivor set, freed slots can be re-batched across requests; quantifying the *throughput* (not just latency) gain requires integration into a continuous-batching server (Kwon et al., 2023) or a GR-serving stack (Sun et al., 2025), measuring tokens/sec at fixed SLO and KV-cache pressure at fixed traffic. A further opportunity is *distortion-aware scheduling*: requests whose decoding is deep inside the viable cap (low ΔD) can donate beam slots to requests near their budget.

H.5 LIMITATIONS

The distortion of Lemma 1 is measured in model probability mass; it is a proxy for task quality. Section F shows the proxy is conservative at our operating point (list metrics unchanged at $D_{\mathrm{tot}} = 0.16$ nats), but it can break where the model is miscalibrated, e.g. on OOD prefixes, the very regions the viable set excludes. The frontier estimate $\widehat{H}_t^\star$ is computed from surviving hypotheses and can drift if the entire population degrades; the budget guard bounds, but does not eliminate, the resulting error. Finally, Propositions 2 and 6 invoke AEP-style concentration that holds exactly for ergodic sources and only approximately for conditional LM trajectories; a non-asymptotic treatment is open.

LLM USAGE STATEMENT

In accordance with the ICLR 2027 policy on AI use, we disclose that we used a large language model (Claude, Anthropic) as an assistant throughout this work, in the following ways.

Theory. The LLM helped reviewing the rate–distortion formulation of beam pruning (Section 5) and the characterization of the viable set (Section 4). It helped formulate mathematical claims and draft and check their proofs, including the telescoping form of the distortion. While checking the proof of Theorem 1, it identified a gap in an earlier version, which led to the no-displacement assumption (iii).

Experiments and analysis. The LLM gave feedback on the experimental methodology, wrote the analysis scripts that turn our decoding logs into the figures and statistics of Section F, and helped interpret those results. The method, its implementation in vLLM, all decoding runs, and all measurements in Table 1 were designed, implemented, and executed by the authors.

Writing. The LLM was used to edit the paper for readability and to draft parts of the text, including portions of the method section and Section F, as well as notation, hyperparameter, and reproducibility descriptions.

Verification and responsibility. The authors reviewed every AI-assisted statement, checked the proofs, and verified the reported numbers against the raw logs. The authors take full responsibility for the contents of this paper.

REPRODUCIBILITY STATEMENT

Code is available at the anonymized repository in Section E.1. Section E.1 lists the models and checkpoints, data splits, hyperparameters, hardware, and timing protocol needed to reproduce Table 1; all proofs are in Section A.

ETHICS STATEMENT

This work uses RecIF-Bench, a public benchmark released with OpenOneRec; we use it as distributed, under its license, with no attempt to re-identify users, and collect no new data. BEAMSTAR changes only which candidates beam search keeps and does not alter model outputs' content; its main effect is reduced inference compute. We see no additional ethical concerns.