ICML
International Conference
On Machine Learning

Geometry-Aware Decoding with Wasserstein-Regularized Truncation and Mass Penalties for LLMs

Top-W Decoding

Arash Gholami Davoodi
Navid Rezazadeh
Seyed Pouyan Mousavi
Pouya Pezeshkpour

Code: github.com/arashgholami/top-w-decoding

Current samplers: from rank to entropy

Probability-driven rules: each picks a different signal — rank, mass, or entropy.

Greedy

$argmax_i p_i$: **one-token confidence**

No crop to tune:
always take the highest-probability token.

Min$-p$

$p_i \geq p_{base} * p_{\max}$: **adds confidence**

Adaptive threshold tied to the leading token; one scalar can miss uncertainty shape.

Top$-k$

$Top_k(p)$: **rank cutoff**

keep the k highest-probability tokens, Fixed width; simple but ignores confidence and distribution shape.

Top$-H$

$$\max_S \Gamma_S \quad s.t. \ \ H(q) \leq \alpha H(p)$$

considers distribution uncertainty

determine a subset $S \subset V$ from which the next token will be sampled

$$\Gamma_S = \sum_i p_i \, 1_{v_i \in S}, \qquad q_i = \begin{cases} \frac{p_i}{\Gamma_S} & v_i \in S \\ 0 & otherwise \end{cases}$$

What's missing?

All four samplers only see ***probability.***

- **Duplicates count twice.** answer / solution / reply crowd the crop and starve other meanings.
- **Distinct ideas dropped.** a plausible proof or derivation falls just below the threshold and is lost.
- **No notion of distance.** the crop can't tell whether removed mass moved 0.1 or 10 in embedding space.

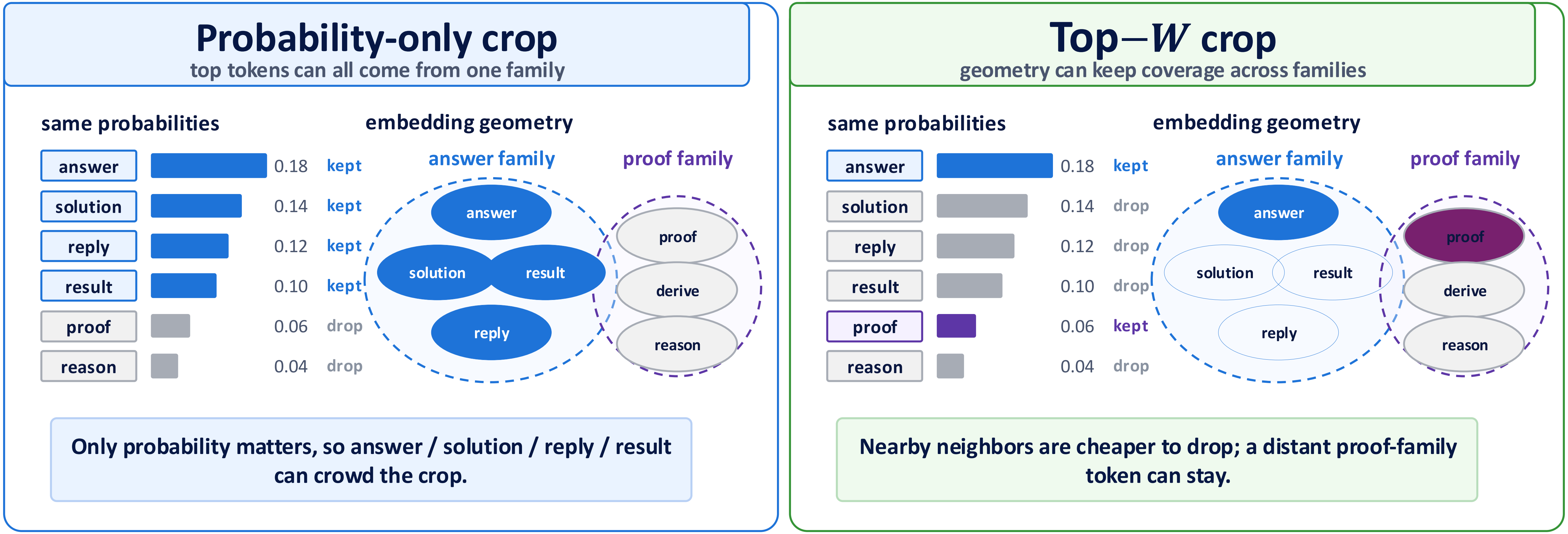
The cost: repetition, mode collapse, and lost diversity — even when temperature is raised.

Next:

read **probability**, but also ask how far the **dropped mass had to travel** — **Top-W**.

Why Top$-W$ changes the crop

Probability-only looks only at p_i. Top$-W$ still respects p_i, but also asks whether dropped tokens are nearby in embedding space.



Top$-W$: Optimization problem

Top-W objective:

$$\min_{S} F_{\lambda,\beta}(S) = \min_{S} \{W_1(p, q_S) + \lambda H(q_S) - \beta \ \log \Gamma_S \}$$

$W_1(p, q_S)$: *Geometry cost*

Cost of moving dropped mass in token-embedding space.

Small when dropped tokens are near the kept ones — preserves meaning.

$\lambda H(q_S)$: *Diversity*

Entropy of the renormalized kept set.

λ controls sharpness vs. spread of the crop.

$-\beta \ \log \Gamma_S$: *Mass reward*

Reward for keeping more probability mass Γ_S.

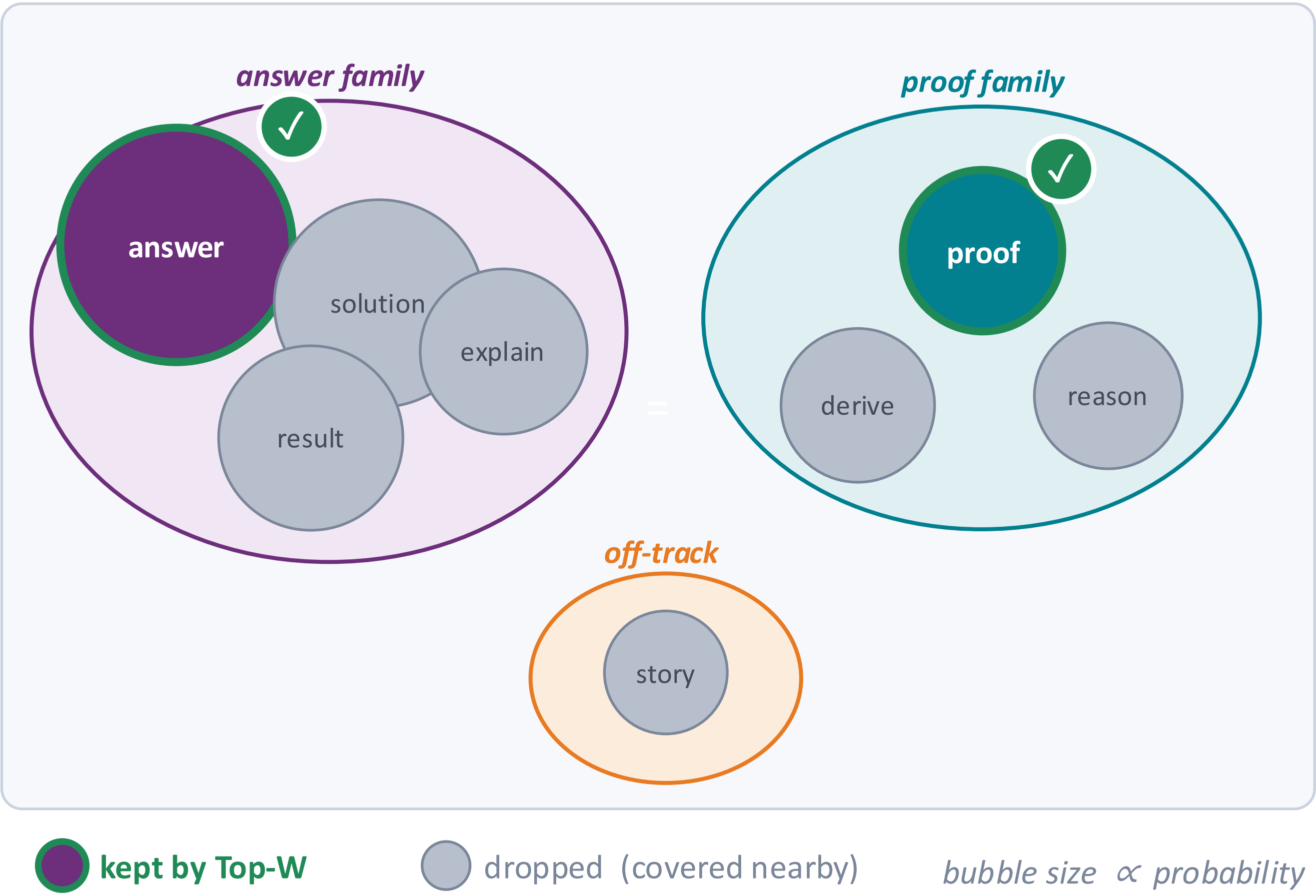
β controls how aggressive the truncation is.

What we prove → what we get

Optimal crop = a prefix sorted by $\phi_i = f_i + \lambda \log p_i$ *(Theorem 3.4)* **O(n) prefix scan**

Top$-W$: what the algorithm changes

Top-W reads probability **and** embedding geometry — keeps tokens that **span distinct meanings**.



1 **Score each token**

$$\boldsymbol{\phi}_{\mathrm{i}} = \boldsymbol{f}_{\mathrm{i}} + \boldsymbol{\lambda} \log \boldsymbol{p}_{\mathrm{i}}$$

f = embedding-distance term; p = probability.

2 **Keep the top prefix**

Sort tokens by $\boldsymbol{\phi}_{\mathrm{i}}$, the optimal subset, i.e., crop S★ is a prefix (Thm 3.4).

3 **Sample as usual**

Outside S★ → −∞ logits; standard renormalize and sample inside.

Not a new distribution — a **geometry-aware crop mask**.

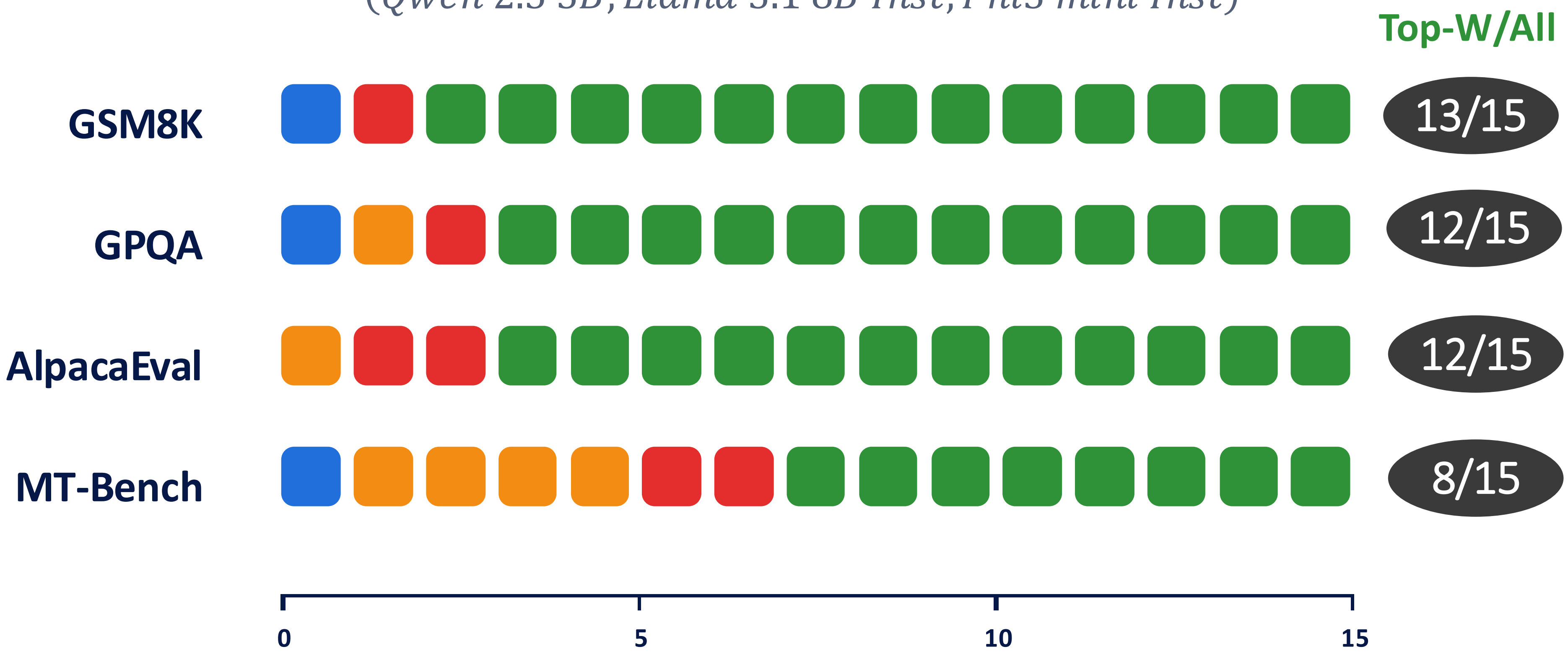
Benchmark wins: each block is one setting

winning settings out of 15 tuples of

5 temperatures $T \in (0.7, 1.0, 1.5, 2.0) \times$

3 models $M \in$

$(Qwen\ 2.5\ 3B, Llama\ 3.1\ 8B\ Inst, Phi3\ mini\ Inst)$



Higher temperature ⇒ larger Top-W gain (up to 33.7% gain).

Min-p and Top-p collapse as T rises (Top-p drops to single digits on GSM8K at T=2.0); Top-W stays stable across $T \in \{0.5, 0.7, 1.0, 1.5, 2.0\}$.

Beyond accuracy: judged creative writing

LLM-as-a-judge rubric — diversity, originality, narrative flow, emotional impact, imagery.

CREATIVE WRITING — JUDGE WINS

12 / 27

(LLM, T, prompt) triplets won by Top-W with higher β.

Min-p : 6 *Top-p : 5* *Top-H : 5*

Larger β → broader support → more creative rubric gains.