Beyond Frequency: Output-Vocabulary Pruning by Log-Mass Maximization

Sergey Ivanov
Egor Krasheninnikov
Nikita Kozodoi
Natalie Tarasova
Aoyu Zhang

Helping customers accelerate MORE
Generative AI Innovation Center

COLM

Motivation

Problem:

- In LLMs, every decoding step scores a distribution over a 200K+ token output vocabulary
- That vocabulary-sized layer is a bottleneck for LLM inference (both latency and memory footprint)
- LM head and softmax take up to 60% of drafting time in speculative decoding

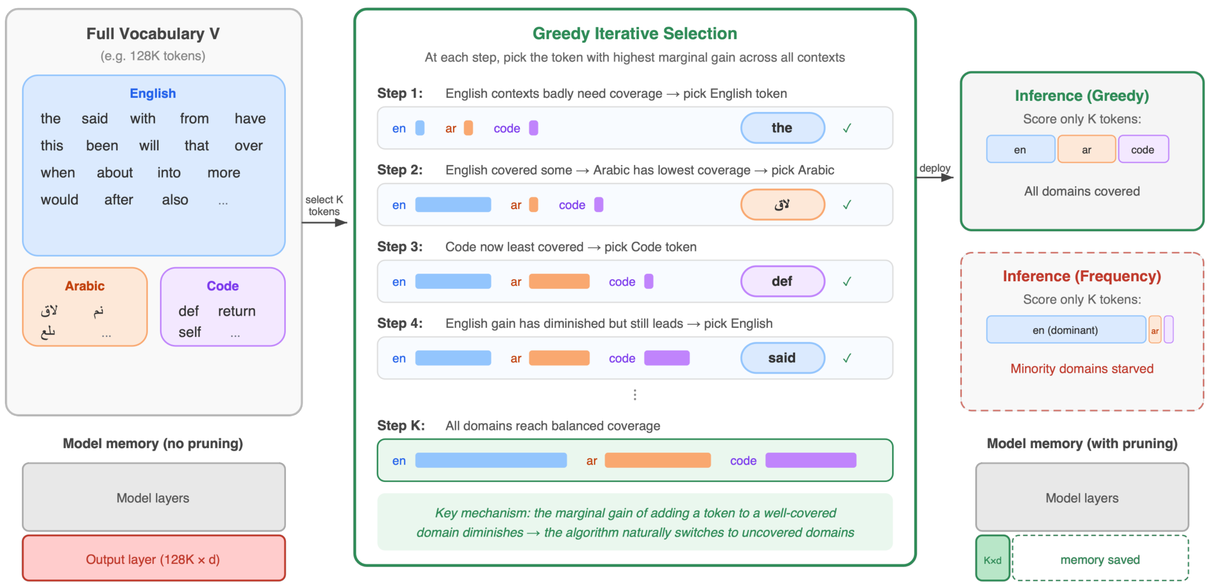
Solution:

- Output-vocabulary pruning cuts it: freeze the model, score only K tokens
- Tokenizer and hidden states never change, so only the output support shrinks

Open question:

If we select only K tokens to score, which K preserve the LLM quality best?

Greedy Vocabulary Pruning



Algorithm

Greedy pruning:

- Objective: maximize expected log retained mass over deployment contexts
- This is exactly equivalent to minimizing expected reverse KL to the teacher
- The objective is monotone submodular, so greedy earns a (1-1/e) guarantee
- Marginal gain of a token: log(1 + its mass / mass already retained)
- A token wins when it rescues a context that is still poorly covered
- Lazy greedy with a max-heap returns the identical support far more cheaply

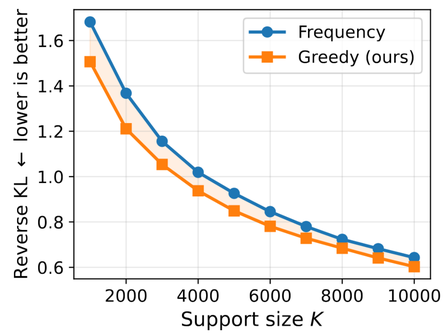
Frequency pruning:

- Frequency pruning is the first-order linear surrogate of this objective
- It is optimal only when all contexts share one common token ranking
- Once contexts disagree on which tokens matter, greedy strictly wins

Benchmarks:

- Qwen3.5-0.8B-Base (248K vocab), Gemma-3-1B-PT cross-family check
- FLORES-200: 10 languages, 5 scripts, ~10K contexts per split
- Synthetic multi-domain data: 10K tokens, up to 10 domains, harmonic mix
- Calibration: uniform, English-dominant (60%), skew sweep up to 99%
- Frequency: keeps the K most-predicted tokens on calibration
- Magnitude: data-free, keeps the K largest lm_head row norms
- Metrics: reverse KL, chrF continuation quality, draft acceptance

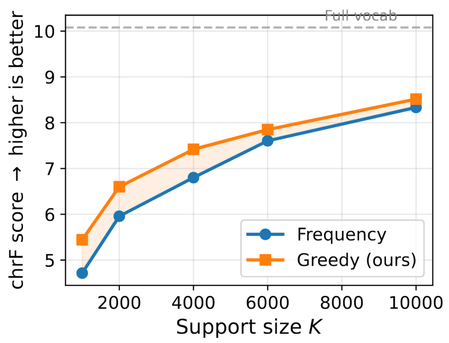
Greedy vs Frequency: Reverse KL



Regime	Method	en	fr	ar	ja	ru	zh	All
Uniform	Frequency	**1.27**	1.32	1.57	1.86	1.87	2.16	1.36
	Greedy	1.28	**1.31**	**1.51**	**1.65**	**1.70**	**1.87**	**1.33**
Weighted	Frequency	**0.79**	1.35	2.39	2.34	2.42	2.73	1.68
	Greedy	0.87	**1.26**	**1.86**	**2.03**	**2.17**	**2.26**	**1.51**

Multilingual output-vocabulary pruning on Qwen3.5-0.8B-Base with FLORES-200 (10 languages, ~10K contexts). Left: reverse KL vs. support size K under English-dominant calibration. Right: per-language reverse KL at K=1,000 (lower is better, bold indicates the better method). Under weighted calibration, frequency starves minority-script languages while greedy rescues them.

Greedy vs Frequency: Accuracy



K	Method	en	fr	ar	ja	ru	zh	Avg
–	Full vocab	13.4	14.5	10.6	5.1	12.5	3.9	10.0
1k	Freq.	**10.1**	9.2	4.2	0.8	3.0	0.9	4.7
	Greedy	10.0	**9.7**	**6.3**	**1.0**	**4.4**	**1.1**	**5.4**
4k	Freq.	**11.3**	10.9	8.5	1.8	6.7	1.4	6.8
	Greedy	11.2	**11.3**	8.5	**2.3**	**9.1**	**1.8**	**7.4**
10k	Freq.	**12.8**	**11.8**	9.4	3.3	10.3	2.2	8.3
	Greedy	12.7	11.7	9.4	**3.6**	**10.8**	**2.5**	**8.5**

Multilingual text continuation on Qwen3.5-0.8B-Base with FLORES-200 under English-dominant calibration. Left: macro chrF vs. support size K. Right: per-language chrF at selected K values (bold indicates the better method). Greedy improves non-Latin-script languages (ru, ja, zh) at negligible cost to English.

Application 1: Different Deployments

	Reverse KL ↓ (nats)				chrF ↑			
Deployment	Full	Magn.	Freq.	Greedy	Full	Magn.	Freq.	Greedy
Latin (5 langs)	0	+3.50	+1.22	**+1.21**	18.6	2.5	10.8	**11.9**
Multi-script (10 langs)	0	+3.78	+1.71	**+1.53**	12.1	1.5	5.9	**6.6**
Multi-script (5 langs)	0	+3.84	+1.32	**+1.25**	14.0	0.5	7.3	**7.7**
High skew (en+ar, 0.99)	0	+3.86	+1.95	**+1.55**	13.9	1.2	6.9	**8.9**

Robustness across deployment types on Qwen3.5-0.8B-Base at budget K=1,000, reporting reverse KL (lower is better) and chrF (higher is better). Methods: Full, the unpruned 248k-token ceiling; Magn., a data-free lm_head-norm baseline; Freq., frequency pruning; and Greedy, ours. Bold marks the best pruned method per metric. Greedy dominates every pruned method on every deployment, with the largest margin over frequency where heterogeneity is highest and near-parity where a common token ranking holds.

Application 2: Speculative Decoding

Method	en	fr	es	de	ru	ar	zh	hi	ja	sw	All
Frequency	**0.56**	0.46	0.45	0.35	0.25	0.23	0.20	**0.57**	0.22	0.39	0.37
Greedy	0.55	**0.49**	**0.46**	**0.37**	**0.28**	**0.28**	**0.27**	0.46	**0.26**	**0.40**	**0.38**
Δ pp	−0.9	+3.1	+1.1	+2.2	+3.2	+5.4	+6.6	−11.6	+3.8	+1.3	**+1.4**

Per-language sample-based speculative-decoding draft acceptance at K=1,000 on the 10-language multi-script deployment (FLORES-200, n=200, Qwen3.5-0.8B-Base). Bold marks the better method per language. Greedy improves every non-Latin-script language except Hindi (+3.2 to +6.6pp) and ties on English; the global gap is +1.4pp (p=0.004). Hindi (-11.6pp) is the Qwen tokenizer-specific reversal. On the high-skew bilingual cell, greedy gains +3.3pp aggregate (p<0.001) at a statistical tie on English. The single-script Latin and 5-script deployments are within bootstrap noise (0.9pp).

Take-Aways

- We do vocab pruning with a clear objective: log retained mass = reverse KL
- Submodularity gives greedy a provable (1-1/e) approximation guarantee
- Compared to other pruning methods, our approach retains higher quality, especially on datasets with low-resource languages