Knowing When to Quit

A principled framework for dynamic abstention in LLM reasoning

Hen Davidov[1] Nachshon Cohen[2] Oren Kalinsky[2] Yaron Fairstein[2] Guy Kushilevitz[2]
Ram Yazdi[2] Patrick Rebeschini[1,2]

[1]University of Oxford [2]Amazon

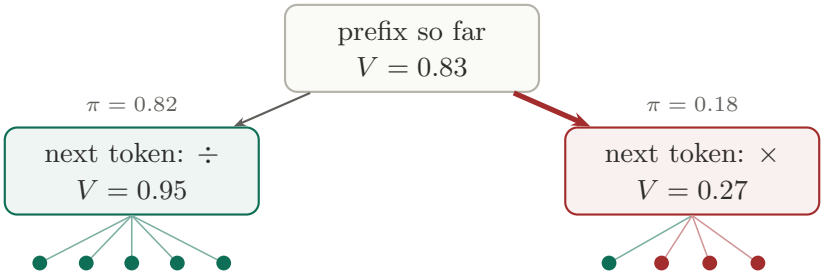
Some generations are doomed — we only learn at the end

Prompt (x): *"What is half of 24?"*

$$r(X, Y_{1:9}) = 0 \in [0, 1]$$

The value function: will this answer end up correct?

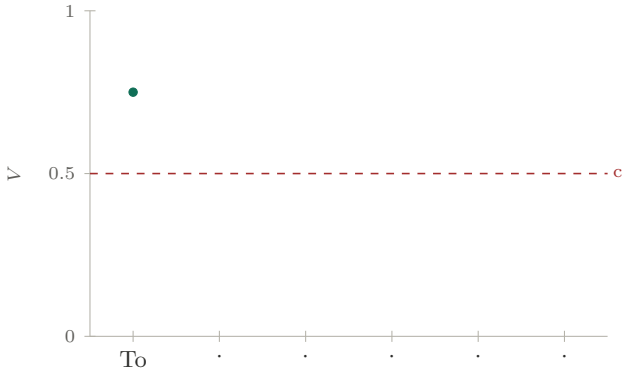
$V(x, y_{1:t}; \pi) = \mathbb{P}_\pi(r = 1 \mid x, y_{1:t})$ — the probability the answer ends up correct, given the prefix so far.

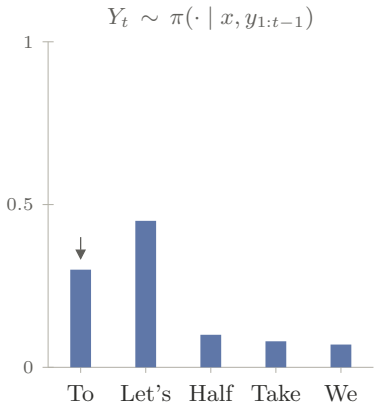


Why it moves. A node's value is the π-weighted average of its children: $0.83 = 0.82\,(0.95) + 0.18\,(0.27)$.

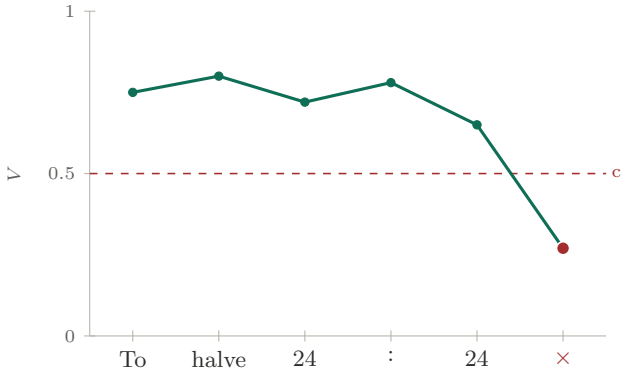
How we get it. A probe on the frozen hidden state, $\widehat{V} = \mathrm{MLP}_\phi(h_t)$.

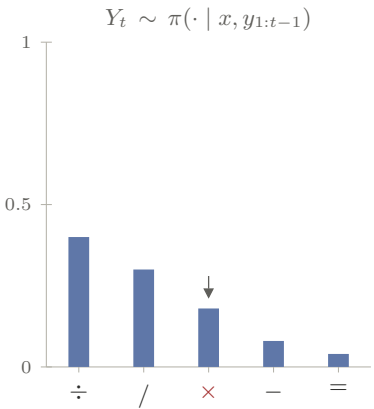
V along the generation



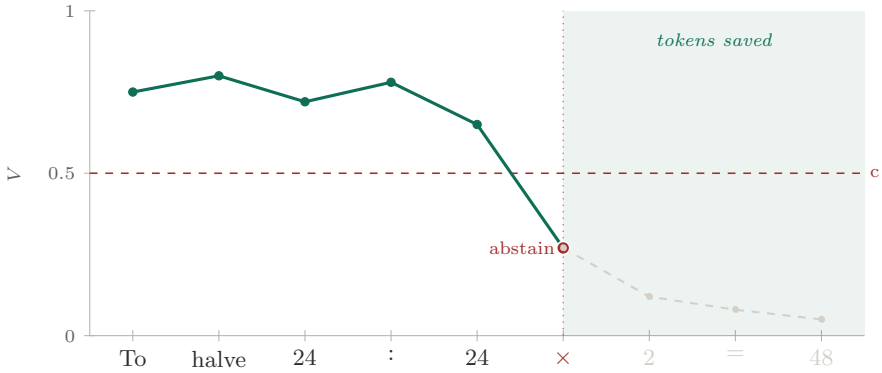


V along the generation





Refusal: stop when $V < c$



Dynamic value-thresholding dominates

Value dominance (Lemma 4.1). At every reachable, non-terminal state,

$$V\big(x, y_{1:t}; a(\pi)\big) \;\geq\; \max\big(\, c,\; V(x, y_{1:t}; \pi)\,\big).$$

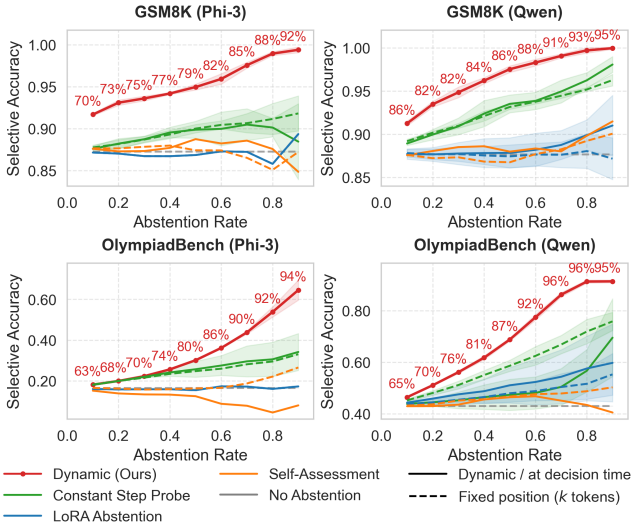
Refusing never lowers the value, and enforces a floor at c. (Backward induction over the Bellman recursion.)

Dynamic vs. fixed-position (Prop. 4.4). Let $f(\pi; k)$ threshold only at position k, and $\tau = \min\{t : V(x, y_{1:t-1}; \pi) < c\}$. Then

$$J\big(a(\pi)\big) \;\geq\; J\big(f(\pi; k)\big) \;-\; \underbrace{\mathbb{E}\Big[\mathbb{I}\{\tau < k\}\,\big(V(x, y_{1:k-1}; \pi) - c\big)^{+}\Big]}_{\text{bounce-back correction}}.$$

The correction vanishes when $k = 1$, or when low-value states can't recover (Cor. 4.5–4.6). So dynamic beats every fixed position — no need to know when to quit in advance.

Selective accuracy across abstention rates



Setup. GSM8K + OlympiadBench; Phi-3, Qwen2.5-7B.

Result. Dynamic (red) dominates every baseline at every abstention rate, every model and dataset — up to 95 % token savings at high α.

The probe also transfers *zero-shot* across datasets, staying above every baseline.