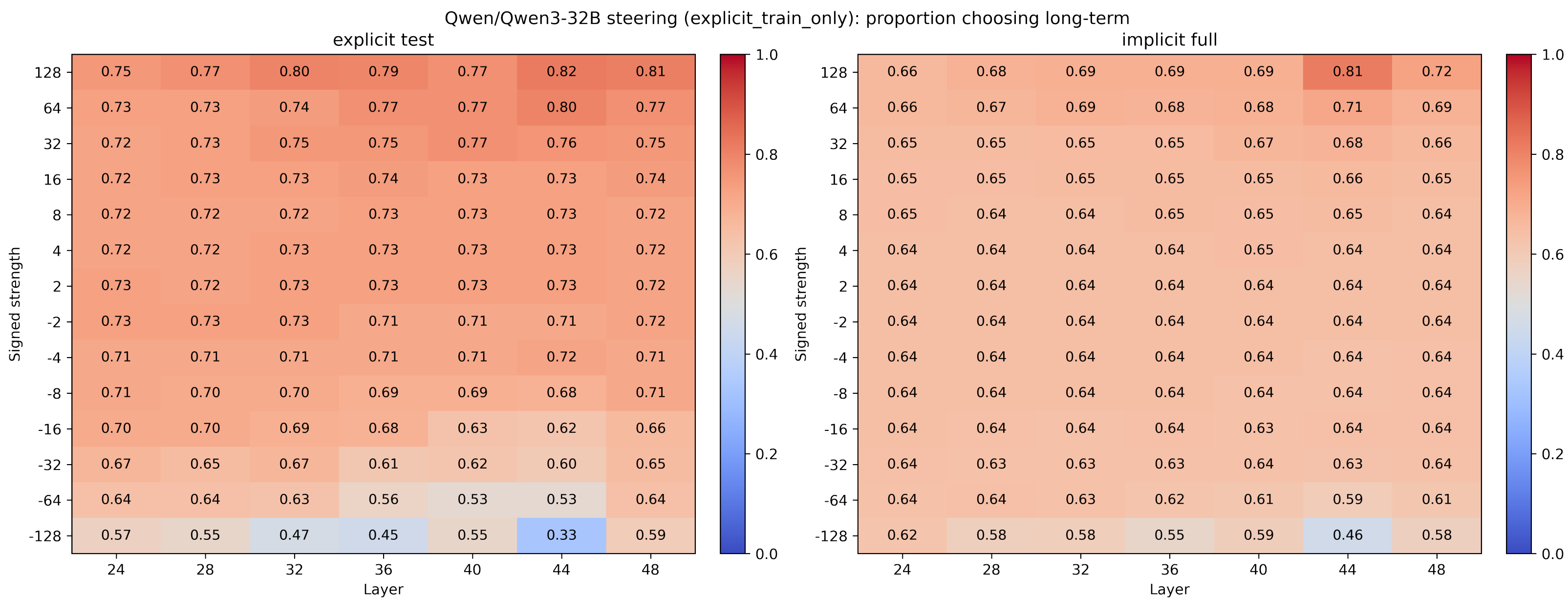


A Linear Direction Controls an LLM's Short-Term vs Long-Term Preferences

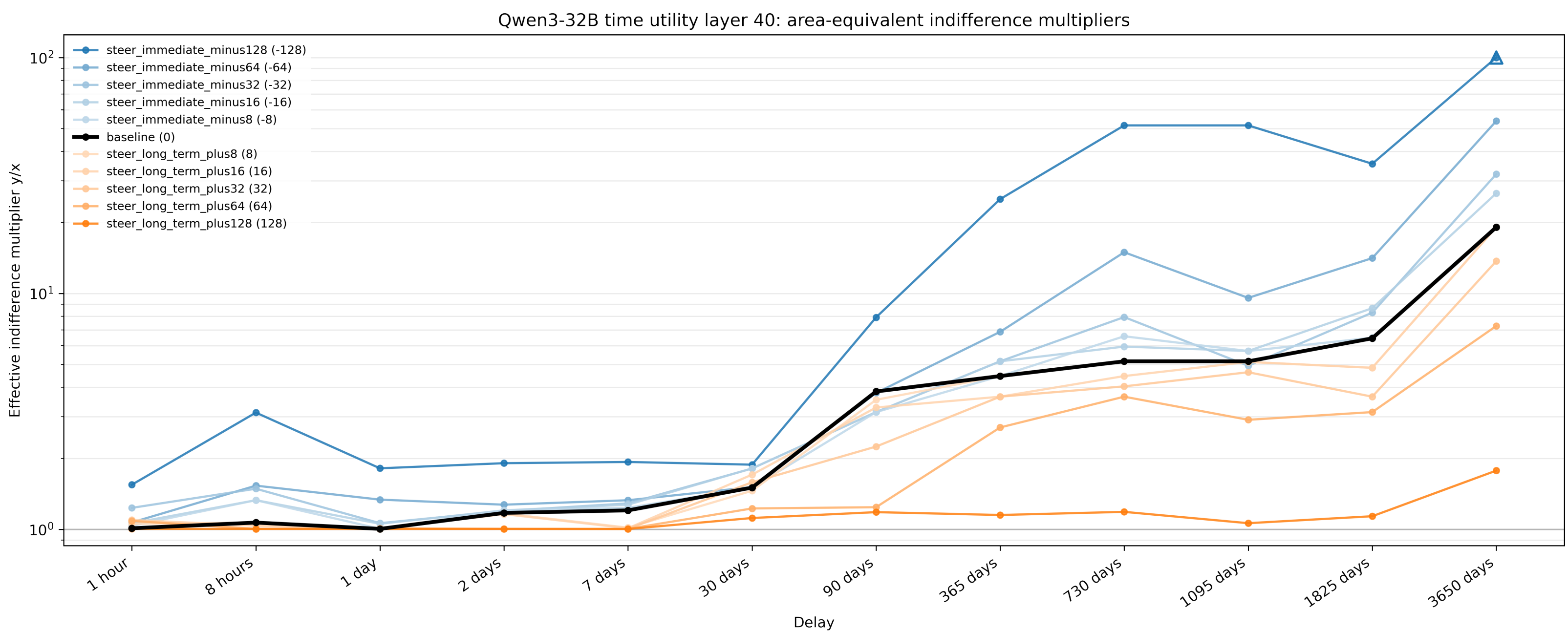
Intertemporal Preference Steering in Qwen3 via Contrastive Activation Addition

Background: We train a short-term vs long-term mean-mass residual-stream direction from teacher-forced answers, then add the signed vector during inference to test causal control of temporal preference.

Result 1: CAA gives broad bidirectional control over held-out binary temporal choices.



Result 2: The same direction shifts monetary indifference thresholds out of distribution.



Methods

Probe direction CAA steering

Table 1: representative probe-training examples. Each row yields two teacher-forced continuations with opposite temporal-horizon labels.

Regime	Prompt	Short-term continuation	Long-term continuation
Explicit	When developing a roadmap for this initiative, consider:	What we can deliver in the next 30 days	What position we want in 10 years
Implicit	When the system fails, the first priority becomes	containing the spreading damage before it escalates	redesigning the architecture to prevent recurrence

Intervention: For layer l , add $\alpha \cdot d_l$ to the residual stream at the last prompt token and each decode step. Positive α steers toward the long-term class; negative α steers toward the immediate class. Evaluation uses parsed choices plus forced-choice logprob diagnostics.

Limitation: The direction is not a complete circuit or a pure temporal feature. The training contrast can also track abstraction, urgency, and systemic vs local framing. Results are from one model family.

