

Ordered State-Token KV Pilot

Date: 2026-08-31

Model: google/gemma-4-31B-it-qat-w4a16-ct

Hardware: NVIDIA H100

Engine: patched vLLM on port 9000

Question

Can ten distinct States from a real discourse be applied at ten successive decode positions, become persistent cache entries, and influence the later continuation?

Method

A technical source discourse produced a full layer-59 capture. The final ten States were retained in order without averaging or another population reduction. A matched target discourse was then run twice with deterministic sampling:

1. control with injection cleared;
2. ten ordered States, additive strength 0.25, one State per successive decode position.

The run was accepted mechanically only if the runtime reported ten configured State tokens and exactly ten applications.

Result

- Configured State tokens: 10
- Confirmed applications: 10
- Control: 128 completion tokens in 4.69 seconds (27.29 tokens/second)
- Intervention: 128 completion tokens in 4.65 seconds (27.53 tokens/second)
- Generated continuation: token-identical
- Captured residual trajectory: bit-identical for all 221 captured rows
- Logit telemetry: changed on exactly the first ten captured rows
- Changed logit metrics: 30
- Maximum observed logit-metric delta: 0.24062753
- Logit telemetry after the tenth application: identical to control

Interpretation

The intervention executed and measurably changed the first ten output distributions. It did not, however, establish persistent State-derived cache entries.

The injection hook currently runs after layer 59 has already projected the position into its key and value. Each decode therefore appends an ordinary token-derived cache entry; the injected layer-59 State arrives too late to change that layer's cache. When the ten interventions stop, the intervention and control trajectories immediately coincide.

This is a clean negative result for the current hook placement, not for ordered State-token memory.

Required next experiment

Capture the inputs to layer 59 (layer-58 outputs), and inject each ordered State before layer 59's Q/K/V projection. Verify all of the following:

1. exactly ten State applications;
2. exactly ten decode positions advance;
3. layer-59 keys and values differ from control on those ten positions;
4. later positions attend to those cache entries after injection stops;
5. child-State and logit trajectories remain different beyond position ten;
6. compare the continuation with a matched clear control and an order-shuffled State-token control.

The decisive success criterion is persistence after the intervention window, not merely a changed logit during injection.