Reservoir Agent — Findings

Reservoir Agent — Findings

Status: in progress. This document is the project's write-up. It is kept current as results land; the sections below state the question and method now, and report results honestly as the experiments run. Until the experiments produce metrics, the Results section says so plainly rather than implying findings that do not yet exist.

Question

Can a fixed, randomly-initialized reservoir injected into a pretrained transformer's mid-layer attention give the model genuine state **between** forward passes — a real time axis — without degrading its base capabilities, and what reservoir-dynamics regime (spectral radius, reservoir size, injection depth) makes that injected state usable signal rather than noise?

This session scopes the question as a **feasibility + dynamics study** at small scale (GPT-2-scale base, single machine). The full vision — forking an agent harness into an always-alive runtime and N-seed LoRA selection at agent scale — is the long-horizon target (see `todo.md`).

Architecture

Every forward pass is one reservoir tick. At a mid-depth injection layer Lk, attention runs jointly over the token hidden states and a set of reservoir nodes (extra keys/values). The reservoir reads the layer's attention output through a fixed random projection W_in and writes its state back through a learned readout W_out — both

at the same layer, every pass — so the reservoir state accumulates a history of the model's own attention dynamics across passes. The reservoir update is

```
r(t) = tanh( W_r · r(t−1) + W_in · x(t) )
```

with W_r a fixed random sparse matrix scaled to a target spectral radius, W_in fixed random, and W_out (plus light upper-layer LoRA) the only trained parameters. The lower layers are frozen. Because the reservoir state is decoupled from the context window, it persists across genuinely independent forward passes, including unprompted ticks.

Grounding in the literature

The fixed-reservoir / trained-readout core is a faithful instantiation of classical reservoir computing (Jaeger's echo state networks; Maass's liquid state machines). The motivation is made precise by the expressivity literature: a finite-precision transformer is bounded to TC^0 / FO(M) **per forward pass** (Merrill & Sabharwal; Hahn), while state carried **across** passes is the documented lever past that ceiling — though the known Turing-completeness results require arbitrary precision, so whether a finite-precision reservoir lifts the bound is posed as an open question, not asserted. Crucially, every prior recurrence-augmented transformer (Transformer-XL, RMT, Block-Recurrent, Mamba, Titans, …) uses *trained* recurrence carrying state *within* a sequence; none uses a *fixed-random* reservoir with state across *independent* passes. The full survey with citations is in `literature/REVIEW.md`.

Method (this session)

1. **Reservoir core.** A tested echo-state reservoir with spectral-radius control and dynamics observability (variance, saturation fraction, effective rank, trajectory distinguishability).
2. **Dynamics characterization.** Drive the reservoir across a grid of spectral radius and size; locate the regime where the state is non-saturating, non-exploding, and carries distinguishable trajectories across input histories (H2), and

test whether the optimum sits at the classical edge-of-chaos prior (which the literature reports is disputed).
3. **Model surgery (H1).** Inject the reservoir into a mid layer of GPT-2-small and verify that, with the readout zeroed, the base model's outputs are unchanged — i.e. the architecture degrades gracefully to vanilla behaviour.

Results

H1 — the reservoir injects without breaking the base model

Hooking a mid-depth block of pretrained GPT-2 so the block's hidden states drive the reservoir and its state is written back into the residual stream (`h' = h + W_out·r(t)`):

- **Non-destruction holds.** With the readout `W_out = 0`, the injected model's next-token logits are *identical* to vanilla GPT-2 (`allclose`, atol 1e-5) — the architecture degrades gracefully to the base model.
- **The injection is live.** A nonzero `W_out` changes the logits, and the reservoir state after two forward passes differs from after one — a genuine cross-pass time axis. (`tests/test_inject.py`.)

H2 — the reservoir-dynamics regime

Sweeping spectral radius ρ ∈ [0.1, 2.0] (figures: `docs/sweep_synthetic.png`, `docs/sweep_real.png`):

- **The echo state property breaks sharply at ρ ≈ 1.** Using an autonomous (zero-input) probe — two random initial states under no input — the reservoir forgets where it started (init-forgetting ≈ 0) for ρ < 1 and abruptly retains it for ρ > 1. This edge-of-chaos boundary appears on *both* synthetic input and **real GPT-2 mid-layer activations** (on real data: 0.000 for ρ ≤ 0.9 → 0.10 at ρ = 1 → ~0.95 above). The classical ρ ≈ 1 boundary survives the move to transformer-scale input.
- **The input regime decides whether ρ matters.** Under unit-scale input *drive* the reservoir forgets its initial state across *all* ρ (strong input enforces the ESP), so the ρ ≈ 1 boundary is the

regime that governs **unprompted, input-free passes** — exactly where the agent would run on reservoir state alone.
- **Real activations over-drive the reservoir.** Compared with synthetic noise, real GPT-2 activations push the reservoir to much higher saturation (~0.86 of units pinned near ±1, vs < 0.15) and higher effective dimensionality (participation ratio ≈ 0.41·K vs ~0.05·K). So a unit-input-scaled reservoir is *over-saturated* by real attention activations: the input scaling has to be tuned down for injection at transformer scale — the precise concern the plan anticipated ("feeding a large attention tensor may require different scaling").

Ambitious reach (proof-of-concept)

Pushed past the feasibility scope to see how far local compute reaches, reported as measured:

- **The time axis is real and behavioural.** Running the *same* prompt after different prior history, with the reservoir state carried across the (otherwise independent) forward passes and a small random readout, shifts the next-token logits by an L2 distance of ≈ 22 (`scripts/run.py alive`, GPT-2). The same input produces a different output distribution depending on what the model processed before — something a stateless transformer structurally cannot do.
- **The seed-selection mechanism works; the pre-training signal is weak.** A dynamics pre-selection proxy ranks N fixed-random reservoir seeds by responsiveness, dimensionality, and (penalised) saturation on real GPT-2 activations, before any training (`scripts/run.py nseed`). Across 8 seeds at ρ = 0.95 the spread is small (~0.02), i.e. *untrained* dynamics vary only modestly between seeds — so the real selection signal the plan relies on most likely emerges only after fine-tuning. The mechanism is in place; the verdict on its usefulness is compute-gated.

Named plainly as not done (compute-gated), not papered over:

- The full **N-seed LoRA fine-tuning + benchmark selection** — there is no training pipeline or benchmark suite here; only the *dynamics* proxy was run.
- A productionized **always-alive runtime** (pass scheduler, idle timer, output confidence gate) — only the two-pass state-carry was demonstrated.
- The **KV-append** injection (reservoir nodes as extra keys/values the upper layers attend to) and **agent-scale (Hermes)** models — beyond local compute this session.

Limitations (current)

- Small-scale only this session; the agentic claims (H3/H4) and the full runtime are out of scope and compute-gated.
- The injection writes the reservoir state into the **residual stream**, not yet as appended **key/value** entries the upper layers attend to (the richer variant in the architecture); the readout `W_out` is not yet **trained** (H3 is future work).
- Input scaling for real-activation injection is **untuned** (the reservoir is over-saturated at unit scale); tuning it is the natural next dynamics experiment.
- The novelty claim is provisional: the reservoir-×-transformer and always-on-agent literatures were not yet verification-complete (see `literature/REVIEW.md` open questions); a citation-checked follow-up precedes any hard novelty claim.
- Whether finite-precision cross-pass reservoir state provably lifts the per-pass TC^0/FO(M) bound is an open theoretical question, not a result of this work.

Reservoir Agent · a cleanvibe research project · report site: https://emmaleonhart.github.io/reservoiragent/