Supplementary Material for "Civilization Architecture"

Yike Wang

Supplementary material accompanying the submitted preprint "Civilization Architecture: Structured Memory, State, and Rule Pathways for Auditable Hidden-State Computation."

***Abstract*—This supplementary material records the reproducibility surface for the staged validation of Civilization Architecture. It expands the main paper's compact appendix with runner entry points, artifact manifests, seed/hardware/runtime information, Stage 42 per-seed results, post-main Stage 43 residual-scale stress evidence, fail-fast gate definitions, frozen-Qwen3 integrity checks, and baseline definitions. The supplementary material does not expand the main claim: the reported evidence remains controlled path-specific hidden-state integration, not public benchmark transfer or a new pretrained foundation model.**

***Impact Statement*—This supplementary material supports the impact of the main manuscript by making the staged validation protocol auditable and reproducible. It documents runner entry points, artifacts, seeds, hardware, runtime, fail-fast gates, frozen-model integrity checks, matched baselines, and negative results. This record helps readers evaluate whether the reported Memory, State, and Rule pathways behave as controlled latent computation paths rather than as post-hoc explanations or prompt-level artifacts.**

***Index Terms*—Reproducibility, artifacts, staged validation, frozen language models, hidden-state adapters, ablation.**

I. Scope of This Supplement

The main manuscript reports the empirical claim at three levels: controlled PyTorch path dependency, frozen-Qwen3 path-specific adapter integration, and Stage 42 multi-seed robustness on a local five-candidate group benchmark. This supplement provides the operational evidence needed to reproduce and audit those claims. It also records negative and planned comparisons so that the reader can distinguish completed evidence from future work.

All paths are relative to the repository root. Artifacts under `experiments/.../artifacts/` are generated outputs and may be excluded from a source-only archive if file-size limits require it; in that case the same runner paths and configuration tables define the regeneration contract.

II. Stage Runner Ledger

Table I lists the main stage families, runners, artifact directories, seed scope, hardware, runtime, and status. Early stages are grouped when they share the same backend and artifact family. Later frozen Qwen3 stages are listed individually because the fail-fast gates and artifacts are distinct.

Yike Wang is with Hydite AI Research, Haokir Inc., Colorado Springs, CO 80918, USA (e-mail: yikewang@research.hydite.com; ORCID: 0009-0009-1832-5421).

TABLE I
STAGE RUNNER LEDGER. RUNTIME IS REPORTED WHEN PRESENT IN THE RECORDED SUMMARY OR REVIEW ARTIFACT; "–" MEANS THE ARTIFACT DID NOT RECORD A SINGLE COMPARABLE WALL-CLOCK VALUE.

Stage	Runner entry point	Primary artifact	Seeds / split	Hardware	Status
04–07	`run_codebook_generalization.py` / staged torch analysis	`civilization_transformer_torch/artifacts/codebook_generalization/alignment_until_pass/iteration_01/summary.json`	202/303/404; 60 samples per label; 40 train per label	local PyTorch	Passed after leakage repair
08	`run_hidden_injection.py`	`civilization_transformer_torch/artifacts/hidden_injection/summary.json`	202/303/404; alpha 0.25; strong-alpha 6.0	local PyTorch	Passed
09–11	`run_fusion_injection.py`	`civilization_transformer_torch/artifacts/fusion_injection/summary.json`	202/303/404; 10800 rows	local PyTorch	Passed
12–17	`run_chain_state_full_benchmark.py`	`civilization_transformer_torch/artifacts/chain_state_full_benchmark/summary.json`	10 seeds; 3 sequence lengths; 2 model sizes; 12 scenarios	local PyTorch	Passed
15–17 stress	`run_path_dependency_stress_test.py`	`civilization_transformer_torch/artifacts/path_dependency_stress_test/summary.json`	202–707; 1080 runs; 3 model sizes; 4 profiles	local PyTorch	Passed
18	`run_qwen3_hidden_state_baseline.py`	`civilization_transformer_qwen3/artifacts/hidden_state_baseline_medium/summary.json`	202/303; all 29 hidden-state groups	MPS BF16	Passed; 51.91s
19–26	`run_qwen3_civilization_adapter.py`, `run_qwen3_multilayer_adapter.py`, `run_qwen3_surface_flip_repair.py`	`civilization_transformer_qwen3/artifacts/civilization_adapter*`, `surface_flip_repair*`	202/303/404 where applicable	MPS BF16	Mixed; several research gates failed
27–30	`run_qwen3_real_task_migration.py`, `run_qwen3_real_task_grounding_repair.py`, `run_qwen3_memory_rule_necessity_repair.py`	`real_task_migration/summary.json`, `real_task_grounding_repair/summary.json`, `memory_rule_necessity_repair/summary.json`	202/303/404 for full matrix; seed 202 for repairs	MPS BF16	Negative / failed gates
31–38	`run_qwen3_binary_path_diagnostic.py`, `run_qwen3_path_specific_binary_diagnostic.py`, `run_qwen3_context_readout_alignment.py`, `run_qwen3_projected_binary_path_necessity.py`, `run_qwen3_full_hidden_centroid_alignment.py`, `run_qwen3_multiclass_group_curriculum_repair.py`	`binary_path_diagnostic/`, `path_specific_binary_diagnostic_retry_01/`, `context_readout_alignment/`, `full_hidden_centroid_alignment/`	primarily seed 202 diagnostics	MPS / CUDA	Diagnostic repair route
39	`run_qwen3_memory_group_gate_repair.py`	`memory_group_gate_repair/summary.json`	seed 202; 60 train groups; 36 held-out groups	CUDA BF16	Failed fast; 69.50s
40	`run_qwen3_memory_delta_gradient_diagnostic.py`	`memory_delta_gradient_diagnostic/summary.json`	seed 202; 60 train groups; 36 held-out groups	CUDA BF16	Passed; 45.63s
41	`run_qwen3_rule_conflict_group_recovery.py`	`rule_conflict_group_recovery/summary.json`	seed 202; restored Stage 40 checkpoint	CUDA BF16	Passed; 271.02s
42	`run_qwen3_group_full_hidden_centroid_integration.py`	`group_full_hidden_centroid_integration/summary.json`	seed 202; restored Stage 41 checkpoint	CUDA BF16	Passed; 548.13s
42-M	`run_qwen3_stage42_multiseed_robustness.py`	`stage42_multiseed_robustness/summary.json`	202/303/404	CUDA BF16	Passed; 375–430s/seed
43	`run_qwen3_stage43_residual_scale_stress.py`	`stage43_residual_scale_stress/summary.json`	202/303/404/505/606; max length 64/128	CUDA BF16	Post-main stress; 10/10 passed
BM-1	`run_qwen3_matched_baselines.py`	`matched_baselines_round1/summary.json`	202/303/404; 440-step contract	CUDA BF16	Completed
BM-2	`run_qwen3_peft_matched_baselines.py`	`matched_baselines_round2_peft/summary.json`	202/303/404; 440-step contract	CUDA BF16	Completed

III. STAGE 42 PER-SEED METRICS

Table II gives the per-seed Stage 42 multi-seed robustness results. The benchmark repeats the raw full-hidden fixed-centroid integration gate over seeds 202, 303, and 404. Each run restores or regenerates the Stage 40 and Stage 41 prerequisites, keeps Qwen3 frozen, and evaluates the same five-candidate group contract.

TABLE II
STAGE 42 MULTI-SEED ROBUSTNESS PER-SEED METRICS.

Seed	Pass	Before	After	Memory	Rule	Conflict	Wrong drop	Hidden norm
202	yes	0.2074	1.0000	1.0000	1.0000	1.0000	0.8000	1.1359
303	yes	0.2019	1.0000	1.0000	1.0000	1.0000	0.8000	1.1272
404	yes	0.2056	1.0000	1.0000	1.0000	1.0000	0.8000	1.1178
Mean	–	0.2049	1.0000	1.0000	1.0000	1.0000	0.8000	1.1270
Std.	–	0.0028	0.0000	0.0000	0.0000	0.0000	0.0000	0.0091

IV. POST-MAIN STAGE 43 STRESS ARTIFACT

Stage 43 is not used to expand the main manuscript's claim beyond Stage 42. It is included here as a reproducibility and stress artifact for the Stage 42 raw full-hidden residual scale. The experiment fixed the residual scale at 200.0 and ran a 5-seed by 2-length matrix. It did not perform scale search and does not establish that scale 200 is optimal.

TABLE III
STAGE 43 RESIDUAL-SCALE STRESS SUMMARY BY SEQUENCE LENGTH.

Metric	Max len. 64	Max len. 128	N
Fixed-centroid after	1.0000 ± 0.0000	1.0000 ± 0.0000	5/5
Fixed-centroid improvement	0.7922 ± 0.0032	0.7981 ± 0.0036	5/5
Wrong-context drop	0.8000 ± 0.0000	0.8000 ± 0.0000	5/5
Hidden-norm ratio	1.1128 ± 0.0142	1.1746 ± 0.0375	5/5

The Stage 43 artifact reports `completed_run_count=10`, `passed_run_count=10`, `failures=[]`, zero Qwen trainable parameters, zero Qwen gradients, unchanged Qwen weights in all runs, and zero core-logic truncation over 43200 audited records. The maximum hidden-norm ratio reported in the per-run CSV is 1.2369, below the 2.0 gate.

V. ARTIFACT MANIFEST

Table IV lists the artifact files used by the paper and this supplement. The artifact directory names intentionally preserve failed attempts and retries when those attempts shaped later protocol choices.

TABLE IV
PRIMARY ARTIFACT MANIFEST.

Evidence family	Primary files	Used for
Controlled hidden states	`codebook_generalization/.../summary.json`, `hidden_injection/summary.json`, `fusion_injection/summary.json`	Codebook leakage repair, hidden-state injection, MSR fusion, alpha-zero and wrong-label controls.
Controlled path dependency	`chain_state_full_benchmark/summary.json`, `path_dependency_stress_test/summary.json`	Hard logic, chain-state alignment, stress profiles, target-path ablation.
Frozen Qwen3 baseline	`hidden_state_baseline_medium/summary.json`, `layer_metrics.csv`, `model_integrity.json`	Layer selection, all-layer hidden-state export, frozen-model integrity.
Real-task negative route	`real_task_migration/summary.json`, `real_task_grounding_repair/summary.json`, `evidence_answer_alignment/summary.json`, `memory_rule_necessity_repair/summary.json`	Negative public/semi-real migration record and boundary of claims.
Path-specific diagnostics	`path_specific_binary_diagnostic_retry_01/`, `context_readout_alignment/`, `projected_binary_path_necessity/`, `full_hidden_centroid_alignment/`	Readout mismatch, projected delta evidence, binary fixed-centroid repair.
Fail-fast final route	`memory_group_gate_repair/summary.json`, `memory_delta_gradient_diagnostic/summary.json`, `rule_conflict_group_recovery/summary.json`, `group_full_hidden_centroid_integration/summary.json`	Stage 39 failure, Stage 40 Memory repair, Stage 41 Rule/Conflict recovery, Stage 42 full-hidden integration.
Multi-seed robustness	`stage42_multiseed_robustness/summary.json`, `per_seed_metrics.csv`, `mean_std_metrics.csv`, `training_runs.json`	Main reported Stage 42 multi-seed robustness.
Residual-scale stress	`stage43_residual_scale_stress/summary.json`, `per_run_metrics.csv`, `mean_std_by_seq_len.csv`, `truncation_audit.csv`	Supplementary post-main residual-scale stress check.
Matched baselines	`matched_baselines_round1/summary.json`, `matched_baselines_round2_peft/summary.json`, `mean_std_metrics.csv`, `per_seed_metrics.csv`	Matched MLP, context-token-only, prompt/task-only, LoRA, and prefix-tuning comparisons.

VI. Fail-Fast Gate Definitions

The final Qwen3 route uses fail-fast gates: a later stage is not accepted until all required earlier gates pass. Table V gives the main gates and their operational meaning.

TABLE V
FAIL-FAST GATES USED IN THE FINAL QWEN3 ROUTE.

Stage	Gate	Acceptance rule	Failure action
39	Memory group gate	Memory context must be separable and Memory delta-to-option, group success, no-memory drop, and wrong-context drop must pass.	Stop before Rule/Conflict/Combined/full-hidden; write failure artifacts.
40	Memory delta repair	Direct differentiable memory-delta supervision must reach 1.00 projected accuracy and 0.80 no-memory/wrong-context drop, with no-rule drop controlled.	Do not enter Stage 41.
41	Rule/Conflict recovery	Rule, Conflict, and Combined groups must pass projected gates while retaining the Stage 40 Memory gate.	Stop before raw full-hidden alignment.
42	Full-hidden integration	Projected gates must be retained; train-only fixed-centroid readout must reach 1.00 across Memory, Rule, Conflict, and Combined groups; wrong-context drop must remain 0.80; hidden norm must remain below the gate.	Reject full-hidden claim and preserve failed attempt.
42-M	Multi-seed robustness	Repeat Stage 42 over seeds 202/303/404 with zero Qwen trainable parameters, zero Qwen gradients, unchanged weights, and no failed seed.	Do not report multi-seed claim.

VII. FROZEN-QWEN3 INTEGRITY CHECKS

The frozen-Qwen3 experiments use a local Qwen3-0.6B checkpoint with the same SHA256 recorded in matched-baseline artifacts:

```
f47f71177f32bcd101b7573ec9171e6a57f4f4d31148d38
e382306f42996874b
```

The implementation enforces the frozen condition with four checks.

- Qwen3 parameters are excluded from optimizers.
- Qwen3 trainable-parameter count is recorded as zero.
- Qwen3 gradients are recorded as zero.
- Qwen3 weights are checked as unchanged before and after training.

Stage 40, Stage 41, Stage 42, Stage 42-M, Stage 43, and the matched baselines all report zero Qwen trainable parameters and zero Qwen gradients. Adapter checkpoints are intended to contain adapter/readout state, not base-model weights.

VIII. BASELINE AND COMPARATOR DEFINITIONS

Table VI records the comparison taxonomy. The distinction between paired controls, matched baselines, and planned comparisons is important because the main claim is path controllability rather than predictive accuracy superiority.

TABLE VI
BASELINE AND COMPARATOR DEFINITIONS.

Comparator	Status	Definition	Interpretation
Adapter-disabled	Run paired control	Same frozen Qwen input and readout, but adapter writes no residual.	Tests whether the signal requires the adapter surface; not an independently optimized base-Qwen model.
Zero-scale	Run paired control	Adapter module is present, but residual scales are zeroed.	Tests whether the write path is causally active.
Wrong-context	Run paired control	Structured context is swapped while surface format and candidate set remain controlled.	Tests context dependence and rejects surface-template explanations.
No-memory/no-rule/no-state	Run paired control	Disable a native path in Civilization or remove the corresponding information channel for token comparators.	Civilization values are native path ablations; token-comparator values are information-channel removals and are not mechanism-equivalent.
Matched MLP adapter	Run matched baseline	Parameter-matched bottleneck MLP adapter at layers 16 and 24; 3,739,650 trainable parameters; 440 optimizer steps.	Reaches 1.00 accuracy; rejects a predictive advantage over generic adapter capacity on the current benchmark.
Context-token-only	Run context baseline	Serializes the same Memory/Rule/State content into Qwen token context with no trainable parameters.	Reaches 0.9611; shows that the current grounded context is highly informative.
Prompt/task-text-only	Run lower bound	Uses only task text without structured Memory/Rule/State content.	Remains at five-candidate chance level of 0.20.
LoRA	Run matched PEFT baseline	Native low-rank q/k/v/o attention updates over all 28 layers; rank 13; 3,727,360 trainable parameters; 440 steps.	Reaches 1.00; no predictive advantage for Civilization on this benchmark.
Prefix tuning	Run matched PEFT baseline	Direct K/V prefix implementation over all 28 layers; 65 prefix positions; 3,727,360 trainable parameters; 440 steps.	Reaches 1.00; no predictive advantage for Civilization on this benchmark.
RAG retrieval from corpus	Not run / planned	A retriever selects evidence from an external corpus before generation or readout.	Not claimed. The context-token-only baseline is not RAG because evidence is supplied by the sample contract rather than retrieved.
Agent/tool controller	Not run / planned	An external policy selects tools, steps, or memory operations outside the hidden-state write path.	Not claimed; outside the current latent path-injection evidence.
Prompt engineering	Negative lower-bound partial	Task-text-only and serialized context variants test prompt/content channels, but no prompt search or optimized instruction suite was performed.	The paper does not claim a general prompt-engineering comparison.

IX. Detailed Negative and Public-Task Control Tables

Tables VII and VIII provide the detailed controls summarized in the main manuscript. They are retained here to keep the submission paper compact while preserving the failure-driven validation record and public-task boundary.

TABLE VII
RECORDED PUBLIC-TASK TRANSFER CONTROLS FROM THE STAGE 27B FULL MATRIX. VALUES ARE FINAL-PROBE FIXED-CENTROID ACCURACIES AVERAGED OVER THREE SEEDS.

Task	Full adapter	Disabled	Difference
RTE	0.4531	0.4531	0.0000
CB	0.6897	0.6897	0.0000
BoolQ	0.3646	0.3594	0.0052

TABLE VIII
FAILURE-DRIVEN STAGED VALIDATION. EACH FAILED GATE CONSTRAINED THE NEXT DESIGN STEP RATHER THAN BEING DISCARDED FROM THE EMPIRICAL RECORD.

Run or stage	Failed gate	Observed result	Design consequence
Early hidden-state codebook	Template leakage control	Hidden clusters were observable but surface templates were too predictive.	Cross-template alignment and masked-keyword controls were required before hidden-state injection claims.
Single-layer Qwen3 adapter medium	Main dependency accuracy and surface flip	Repeated medium variants stayed near the five-candidate chance level: main-layer accuracy was 0.20–0.2083 and surface flip was 0.00.	Adapter hooks alone were rejected as evidence; later stages required multilayer placement, path-specific residuals, and explicit readouts.
Qwen3 real-task migration	Local surface flip, path drops, and public-task improvement	Grounded real-task repair reached only 0.2467 local fixed-centroid accuracy, 0.20 local answer-option accuracy, 0.00 surface flip, and 0.00 improvement on RTE/CB/BoolQ in that run.	Public and semi-real transfer were excluded from the main claim; the project returned to controlled local path-necessity gates.
Memory/Rule necessity repair	Directional path necessity	Fixed-centroid loss decreased, but answer-option accuracy regressed to 0.30 and Memory/Rule held-out pair success stayed near random at 0.0556/0.0333.	Necessity was split into smaller binary diagnostics and then into separate projected-delta and full-hidden readout questions.
Path-specific adapter repair	Readout alignment	Memory-only and Rule-only answer readouts were 0.4625 and 0.4250, while traces showed nonzero path-specific deltas.	The failure was localized to readout-direction mismatch; Stage 33 introduced the PathReadoutProjector.
Projected vs. raw hidden readout	Full-hidden survival	Projected delta readout reached 1.00 for Memory and Rule, but combined binary remained 0.475 and raw readout stayed weak.	Projected path evidence was not allowed to substitute for raw full-hidden evidence; fixed-centroid alignment became a separate gate.
Stage 39 Memory group gate	Memory delta-to-option	Memory context was separable at 1.00, but Memory delta-to-option accuracy was 0.20, group success was 0.00, no-memory drop was 0.00, and wrong-context drop was 0.00.	The runner became fail-fast. Stage 40 supervised the differentiable memory-delta tensor directly before allowing Rule/Conflict recovery.
Stage 42 pre-alignment full-hidden gate	Raw full-hidden fixed centroid	Before alignment, the five-candidate full-hidden centroid score was near chance: 0.2074 for seed 202 and 0.2049 ± 0.0028 across three seeds.	Stage 42 added full-hidden centroid alignment and retained train-only centroids, wrong-context controls, and projected-gate retention checks.
Stage 42 external validity	Public task transfer	The Stage 42 multi-seed robustness benchmark passed locally, but it remains a controlled five-candidate group benchmark.	The paper reports controlled robustness and leaves public benchmark transfer, longer contexts, and cross-model replication as future work.

X. REPRODUCIBILITY BOUNDARIES

The supplementary material records enough operational detail to reproduce the reported controlled claims, but several boundaries remain:

- Stage 42 and Stage 43 are local five-candidate group benchmarks, not public real-task benchmarks.
- The matched-baseline suite shows a ceiling effect; it does not show a predictive accuracy advantage for Civilization Architecture.
- The Stage 43 residual-scale stress test fixes scale 200.0; it is not a scale search and does not prove optimality.
- Frozen-Qwen3 evidence is one real-model instantiation. Cross-model, longer-context, public-task, and unfrozen full-architecture training remain future work.