ML Lab Research Proposal

Team: Aryan Ahuja (solo) · **Discord:** sassyYoda · **Date:** July 7, 2026 **Project:** Stress-testing the "J-space" global workspace in open-weight language models **Code:** github.com/sassyYoda/jspace-probes (public; the preliminary signal-check scripts and results below, runnable on a laptop or single GPU) **Target compute:** 1–2× AMD Instinct MI300X (single node, ROCm), approx 5–7 wall-clock days inside a 2-week window

---

1. Problem Statement

On July 6, 2026, Anthropic published *Verbalizable Representations Form a Global Workspace in Language Models* (Gurnee, Sofroniew, Lindsey et al., Transformer Circuits). It introduces the **Jacobian lens** (J-lens): for each layer ℓ, an averaged linearized map $J_\ell = E[\partial h_{final} / \partial h_\ell]$ from the residual stream to final-layer logits, read out as $\text{softmax}(W_U \cdot \text{norm}(J_\ell h))$. The sparse set of active J-lens directions (the "J-space") behaves like a cognitive-science global workspace: its contents are verbally reportable, causally load-bearing for multi-step reasoning, flexibly general, and selectively necessary. Every headline result in the paper was measured on closed Claude models.

The paper is strong but almost entirely un-replicated and un-stress-tested on open weights, and its authors flag concrete gaps: the lens captures only single-token concepts, they do not know what mechanism decides workspace entry, and they call the tool "imperfect." Neel Nanda's team replicated core claims on Qwen but found poetry and arithmetic did *not* transfer, and warns the averaged Jacobian "will find noisy directions" with "many false positives." **The open question this project answers: which J-space claims are real, model-agnostic properties of transformers, and which are Claude-specific or artifacts of the lens?** We turn a set of qualitative demos on one model family into quantitative, baselined, causally-validated findings on open weights.

This is not a training project. It is a measurement and causal-intervention project on pre-trained models, which is why the compute footprint is small and honestly stated: most of the pipeline already runs on a laptop, and we are asking for MI300X time only to scale the experiments that have already shown signal.

2. Proposed Approach

We run six pre-registered signal-check experiments, then scale only the ones that show signal (Section 6). Two already have strong signal from local runs (Section 5). All share one validated intervention harness (J-lens vector construction from the pre-fitted lens + unembedding; concept-swap and directional-ablation forward hooks; angle/norm-matched random controls; a no-op identity test that reproduces baseline logits bit-exactly).

- **A — Cross-model universality.** Is the J-space relational geometry shared across model families beyond tokenizer geometry? (RSA/SVCCA/Procrustes across families, with unembedding-only and shuffled baselines.) **Signal already obtained.**
- **D — Selectivity ablation with dose-response.** Does ablating top-k J-space contents selectively degrade reasoning/summarization/rhyming while sparing MCQA/sentiment/fact-retrieval, graded in k, beyond matched random ablation? The paper's flagship claim; the anchor experiment whose harness the others reuse.
- **I — Does the workspace reveal what unfaithful chain-of-thought hides?** (Reward hacks are verbalized <2% in the literature; if the hack concept sits in J-space anyway, J-lens is a monitor that beats reading the CoT.) The novel, safety-relevant flagship.
- **C+G — Lens auditing benchmark and robustness audit.** Is J-lens a better detector of hidden content than tuned/logit lens at matched cost, and when can it be trusted? Includes a future-token-target tuned-lens baseline, since the paper's own ablation credits future-token targeting (not the Jacobian) for most of J-lens's gain.
- **B — Ignition dynamics.** Does workspace entry show a threshold/competition ("ignition") rather than smooth scaling? (Directly attacks the authors' stated mechanistic gap; requested by GNW originators Dehaene & Naccache in the paper's commentary.)
- **J — Capacity psychophysics.** Is the "few dozen concepts" limit behaviorally binding, and what gets evicted? (A digit-span-style set-size curve; also requested in the commentary.)

We chose this portfolio over "reproduce one demo" because it isolates the paper's *claims* (universality, selectivity, entry dynamics, auditing utility) rather than its surface, and every experiment pairs an observational readout with a causal intervention, which is the bar reviewers enforce for this kind of work.

3. Models, Lenses, and Prompt Sets (in lieu of a training dataset)

No task-supervised training data is used. Inputs are open-weight models, pre-fitted J-lenses, and small prompt sets. Everything is public and ungated except where noted.

Models (HuggingFace, open weights).

Family	Sizes used	Role
Qwen3 / Qwen3.5 / Qwen3.6	1.7B, 4B, **8B**, **27B**	primary substrate (matches paper + Nanda replication)
Gemma-2	2B, **9B** (+ `-it` for base/instruct diff)	second family; has Gemma Scope SAEs for the J-space↔SAE bridge
Llama-3.1	**8B** (+ `-it`)	cross-family universality check
GPT-2, Pythia-70m	small	pedagogical sanity / training-dynamics anchor

Bolded sizes are the ones needing MI300X; ≤4B runs locally.

Pre-fitted lenses. Neuronpedia hosts J-lenses for **38 open models** (`neuronpedia/jacobian-lens` , MIT, ungated), each fit on 1000×128-token WikiText sequences on a B200. This removes almost all fitting cost: we download and apply. We fit our own lenses only for the corpus-robustness arm (Experiment G): 2–3 models re-fit on code and chat corpora to test whether readouts depend on the fitting corpus (a confound flagged by our own Experiment A).

Prompt sets. The reference repo ships the paper's own probe-swap, ignition, capacity, selectivity, and verbal-report prompt sets (reused directly). We add small (approx 40-item) task batteries from public benchmarks: MMLU (`cais/mmlu`), SST-2 (`stanfordnlp/sst2`), GSM8K-style multi-step items, rhyme completions, and Turpin-style biased-hint MCQ items for the CoT-faithfulness experiment. Total on-disk footprint (model weights + lens files + activation caches for the SAE-bridge arm): **approx 250–350 GB scratch**, well under a 1 TB budget.

4. Evaluation Metrics

Experiment	Primary metric	Non-negotiable controls
A — universality	RSA (Spearman on cosine-sim upper triangles) + SVCCA of J-lens vectors across families, by layer fraction	unembedding-only baseline; shuffled-concept floor; random-orthogonal
D — selectivity	task accuracy vs ablation k; double-dissociation gap (workspace-dependent minus workspace-independent drop)	random equal-norm ablation (≥3 seeds); bottom-k matched control; k=0
I — CoT faithfulness	detection AUC of hidden hint/hack concept in J-space; causal revert rate under ablation	base-vs-instruct model; random-direction control; surface-token paraphrase control; tuned/logit-lens comparison
C+G — lens benchmark	detection AUC and false-positive rate vs tuned/logit lens at matched cost; readout rank-correlation across fitting corpora	future-token-target tuned lens; injection ground truth; corpus/prompt-count/variant sweep

B — ignition	sigmoid-vs-linear fit (AIC) of J-space presence vs stimulus strength; hysteresis; eviction	prompt-side and activation-side strength; matched random injection
J — capacity	set-size curve: J-space occupancy and recall vs N; eviction structure	no-hold-instruction control; shuffled-label occupancy null

All causal claims report **per-input effect distributions**, not means (the steering literature shows 29–43% of samples can move opposite the mean direction). Seeds are fixed and logged; key comparisons carry bootstrap CIs. Every reported number traces to a `results/metrics.json` with git SHA, seed, model revision (HF commit hash), and wall-clock.

5. Preliminary Results (already obtained, $0 compute, on a laptop)

Three results are in before requesting any GPU time, which de-risks the request:

1. **Harness validated / paper's internal-reasoning property replicates (open weights).** Concept-swap on Qwen3.5-4B over the paper's 90 probe-swap items: swap-flip rate **58.6% on baseline-correct items** (paper reports 54–70% on Claude), versus **1.7% for angle-matched random controls** (approx 34× separation), with an early-layer control markedly weaker as predicted. The no-op identity test reproduces baseline logits bit-exactly.
2. **Experiment A shows signal (pre-registered criterion passed, all 3 model pairs).** J-space relational geometry aligns across Qwen3 / Gemma-2 / GPT-2 at mid layers: **RSA 0.75–0.87** for J-space versus **0.05–0.61** for the unembedding baseline and **approx 0** for the shuffled floor. The sharpest case: Gemma-2-2B ↔ GPT-2 unembedding similarity is 0.045 (essentially unrelated vocabularies), but their J-space geometries align at 0.75–0.82. The effect is in relational geometry, not shared linear subspaces (SVCCA shows no separation), and decays in late layers.
3. **Selectivity dose-response (Experiment D) mini-run:** in progress on the laptop at submission time; the harness and task battery are built and the multi-model scale-up is the primary GPU ask.

These ran on Apple MPS using the pre-fitted lenses in seconds to minutes. **The MI300X request is to scale what already works to 7–27B models, three seeds, and full task batteries, plus the lens re-fits and generation-heavy batteries that a laptop cannot do at scale.**

6. Compute Requirements

Estimates are scaled from measured local runs (lens apply 0.2–2.6 s/prompt and interventions approx 0.75 s/item on a 4B model on Apple MPS), padded for the move to larger models and batched generation. A single MI300X (192 GB) holds any model up to approx 70B in bf16 without tensor parallelism, which is why we ask for MI300X rather than 80 GB cards; most phases use 1 GPU, and the multi-model sweep parallelizes across up to 2.

Phase	Hardware	Estimated wall-clock
Scale D + I to 7–9B (Qwen3-8B, Gemma-2-9B, Llama-3.1-8B), 3 seeds, full task batteries incl. GSM8K CoT generation	1–2× MI300X	approx 1.5–2 days
Scale winners to 27–32B (Qwen3.5/3.6-27B), read-out + intervention (no fitting; pre-fit lenses exist)	1× MI300X	approx 1 day
Experiment G lens re-fits: 2–3 models × 2 corpora (code, chat), 1000×128 prompts each	1× MI300X	approx 1–1.5 days
Experiments C, B, J at scale (mostly readout + light intervention)	1× MI300X	approx 1 day
Causal-validation upgrades on the winner (path patching, dose-response sweeps, per-input distributions) +	1–2× MI300X	approx 1–1.5 days

ablations, ≥3 seeds		
Total	**1–2× MI300X node**	**approx 5–7 wall-clock days inside a 2-week window, with buffer for re-runs**

Scope will be trimmed by signal, not padded to fill the window. Experiments that fail their pre-registered kill criteria at the 7–9B stage do not proceed to 27B. If only one of D/I survives, we run that one deeper (more models, stronger causal validation) rather than spreading thin. This is a feature: the workshop paper needs one clean scaled result plus the universality result we already have.

GPU memory: the largest tensors are the d_model × d_model Jacobian matrices (fp32 for the math, approx 0.3–2 GB per layer at 27B) and batched activation caches for the SAE-bridge arm. A 192 GB MI300X holds a 27–32B model plus these comfortably; an 80 GB card would force offloading.

Storage: approx 250–350 GB scratch (model weights + lens files + activation caches + checkpoints).

Software / dependencies (ROCm-clean): ROCm 6.x + the official PyTorch ROCm wheel; `transformers>=5.5`, the `jlens` reference package (pure PyTorch, `torch.fft`-free), `nnsight` for any >32B remote work, `sae_lens` for the Gemma Scope arm. **No CUDA-only kernels** (no flash-attn, no xformers, no custom Triton CUDA); the intervention harness is plain forward hooks, and Qwen3.5 already runs without flash-linear-attention in our local tests. No Docker required; a `uv` / `venv` bring-up. One-command smoke test reproduces the validated probe-swap number on a small model.

7. Expected Deliverables

By the end of the window we aim (not guarantee) to have:

1. The validated intervention harness and pre-registered experiment code, public and MIT-licensed, with a ROCm/MI300X section in the README.
2. **Universality result (Experiment A)** finalized across ≥3 families with the cross-corpus confound closed.
3. **One scaled flagship** (D or I) with multi-model, multi-seed, dose-response, causal-validation, and honest failure analysis (including where the effect breaks and its false-positive rate, per Nanda's warning).
4. Signal-check results for the remaining experiments (C, G, B, J) at whatever depth their signal justified, reported honestly including negatives.
5. A workshop-length draft (NeurIPS Mechanistic Interpretability Workshop format) with all figures and tables regenerated end-to-end from `results/` JSON, plus a reproducibility statement. The scaled result targets **ICLR 2027 main** as the follow-on.
6. Per-seed metrics, fitted lenses, prompt sets, and run manifests archived for reproduction.

We are explicitly not promising "the J-space claims all replicate." A rigorous partial replication with a clear map of what is universal, what is Claude-specific, and what is a lens artifact is the deliverable, and a clean negative on any single claim is a publishable result the field wants. We will also hold the consciousness-claim discipline the topic demands: functional/access-consciousness properties only, no phenomenal-consciousness claims.

8. Risk & Mitigation

Risk	Likelihood	Mitigation
A flagship's effect vanishes at 7–9B (e.g. poetry/arithmetic did not transfer to Qwen for Nanda)	Medium	Two flagships (D, I) plus the already-positive A; portfolio is designed so one clean scaled result carries the workshop paper. Kill criteria are pre-registered per experiment card.
J-lens false positives inflate a detection result (Nanda's warning)	Medium	Experiment C explicitly measures FPR against tuned/logit-lens and a future-token-target baseline; every detection claim reports FPR, not just AUC.

Shared-corpus confound weakens the universality result	Medium	Experiment G re-fits lenses on distinct corpora; the confound is already named in the A write-up and budgeted for.
ROCm PyTorch op gap	Low	Harness is plain forward hooks + matmul + FFT-free; no custom kernels; a CPU/MPS fallback path already runs the whole small-model pipeline.
Wall-clock overrun	Low–medium	Trim-by-signal is the default, not a fallback: drop non-signal experiments and the 27–32B tier first; the 7–9B tier alone supports the workshop paper.
27–32B does not fit or is slow	Low	192 GB MI300X holds it in bf16; if throughput is tight, `nnsight` remote or read-out-only (no fitting, lenses are pre-fit).

We will check in at the approx 3-day mark with the 7–9B results so we can decide together whether to scale to 27B, add the SAE-bridge arm, or consolidate on a single deeper flagship.