ML Lab Research Proposal
Matched-Null Audit of Mechanistic Circuit Claims

Team:	Ethan Wang (solo)
Date:	July 22, 2026
Project:	Conflict: Ablation-Choice Sensitivity in Mechanistic Interpretability
Target Compute:	1x AMD Instinct MI300X GPU, approx. 2 weeks

1. PROBLEM STATEMENT

Mechanistic interpretability validates neural circuits by ablating non-circuit components and measuring performance recovery. The choice between ablation methods (zero, mean, resampling) is rarely justified, yet Miller et al. (COLM 2024) showed faithfulness metrics are sensitive to this choice. Using the MIB benchmark (Mueller et al., ICML 2025), we found ablation type flips the verdict on half the grid: mean ablation rejects 33% of 143 claims, zero ablation rejects 83%, a 50-percentage-point gap. We identified LayerNorm saturation as the mechanism (r = 0.876) on GPT-2, but extending this to other architectures (RMSNorm in Llama and Gemma) and completing the SAE feature-circuit selectivity contribution requires additional compute.

2. PROPOSED APPROACH

Our audit runs conditional randomization tests (Candes et al., 2018) with 1,000 null draws and BH-FDR control at q = 0.05 across 12 (task, model) cells on GPT-2 Small, Qwen-2.5-0.5B, Gemma-2-2B, and Llama-3.1-8B from the MIB benchmark. Three ablation types crossed with four null-construction strategies yield 143 testable claims.

The paper has four contributions. C1 (50pp intervention gap) and C2 (LayerNorm saturation mechanism on GPT-2, r = 0.876) are fully supported. C3 (SAE feature-circuit selectivity) has preliminary data but needs the full protocol: SAE features from Bloom (2024) gpt2-small-res-jb suite should show large z-scores under matched nulls, while raw residual dimensions of the same size should fail, proving the audit is not a universal rejector. C4 extends the LayerNorm saturation probe to RMSNorm architectures (Gemma-2-2B, Llama-3.1-8B) and tests GQA-aware auditing at G>=4.

For AAAI, we also expand the held-out-split audit from 1 cell to 10 cells spanning the verdict space, controlling for selection bias (Candes et al., 2018), and compare against Li and Janson's (NeurIPS 2024) optimal ablation as a fourth intervention type.

3. DATASET

Dataset	Source	Size	Use
MIB benchmark prompts	Mueller et al., ICML 2025	200 prompts per task, 12 (task, model) cells	Primary evaluation. IOI, MCQA, arithmetic, ARC tasks across 4 models.
GPT2-Small SAE features	Bloom (2024), gpt2-small-res-jb via SAE Lens	24,576 features per layer, layers 8-9	C3: paired test of SAE feature-circuit selectivity vs. raw residual dimensions.
Optimal ablation vectors	Li & Janson, NeurIPS 2024	Learned per-component replacement vectors	C5: fourth ablation type to complete the intervention-gap spectrum.

Preprocessing

• All prompts are short sequences (under 50 tokens). Cached model activations for repeated ablation runs.

• SAE features extracted via SAE Lens encode/decode pipeline. 20-step integrated gradients for attribution.

• Optimal ablation vectors learned via Adam (lr=1e-3, 1000 steps) per component per cell.

4. EVALUATION METRICS

Experiment	Metric	Target / Interpretation
C3: SAE selectivity (GPT-2, IOI)	z-score and empirical p under density-decile and activation-magnitude nulls (n=500)	SAE circuit z > 10, p < 0.002; raw-dimension baseline z near 0, p > 0.5. Proves audit is not a universal rejector.
Cross-model LN saturation (Gemma-2, Llama-3.1)	Pearson r between RMSNorm scale change and metric drop under zero vs. mean ablation	r_zero > 0.5 and r_mean < 0.2 on both models upgrades "LayerNorm" to "normalization-layer" mechanism.
Higher-G GQA (Mistral-7B or Llama-3.1)	sigma_null, z(C), empirical p for KV-group vs. per-head audit	sigma_null > 0 (fixing G=2 degeneracy). KV-group audit should show lower p-value than per-head.
Held-out-split expansion (10 cells)	Rejection rate under 50/50 prompt split (selection vs. testing)	Gap between split and non-split rates is within 5pp, confirming selection bias is modest.
Optimal ablation comparison (3 cells)	Rejection rate under OA vs. mean/zero/resampling	OA rate lower than mean (less than 33%), establishing hierarchy: OA < mean < resampling < zero.

5. COMPUTE REQUIREMENTS

The C3 and held-out-split experiments for GPT-2/Qwen/Gemma run on Mac MPS (Apple Silicon). Only Llama-3.1-8B experiments and the GQA replication require cloud GPU.

Phase	Hardware	Est. Wall-Clock	Notes
C3: SAE selectivity (GPT-2, IOI)	Mac MPS	~8 hours	500 null draws x 2 interventions x 20-step IG. No cloud GPU needed.
Cross-model LN saturation (Gemma-2-2B)	Mac MPS	~4 hours	26 MLP blocks x 200 prompts x 2 ablation types. Fits in 16-bit on M-series 32GB.
Cross-model LN saturation (Llama-3.1-8B)	1x MI300X	~4 hours	32 MLP blocks x 200 prompts x 2 ablation types. Model is 16GB in float16.
Higher-G GQA (Mistral-7B or Llama-3.1)	1x MI300X	~5 hours	DLA attribution + per-head + KV-group audit, n=500 null samples.
Held-out-split (10 cells, 3 Llama cells on GPU)	Mac MPS + 1x MI300X	~2 days	7 cells on Mac (~24h), 3 Llama cells on MI300X (~12h).
Optimal ablation (3 cells)	Mac MPS + 1x MI300X	~2 days	OA vector learning: 2h Mac (GPT-2) + 4h Mac (Gemma) + 10h MI300X (Llama).
Total	Mac MPS + 1x MI300X	~6 days within 2-week window	~35 H100-hours cloud equivalent. Remainder for reruns and ablations.

Software Dependencies

• Python 3.10+, PyTorch 2.x with ROCm backend (cloud) and MPS (local).

- TransformerLens for model hooks and DLA attribution.
- SAE Lens for GPT-2 sparse autoencoder loading (gpt2-small-res-jb release).
- No CUDA-only kernels. All operations are standard PyTorch.

6. EXPECTED DELIVERABLES

- SAE selectivity figure (Figure 8): null distributions for SAE feature circuits vs. raw residual dimensions, with z-scores and empirical p-values.
- Cross-architecture LayerNorm saturation table: Pearson r values for GPT-2 (LN), Gemma-2 (RMSNorm pre+post), and Llama-3.1 (RMSNorm pre) under zero and mean ablation.
- GQA-aware audit results: KV-group vs. per-head rejection rates on Mistral-7B or Llama-3.1-8B at G=4.
- Held-out-split analysis: rejection rates with and without selection-bias correction across 10 cells.
- An AAAI-27 paper with the complete 4-contribution structure plus the optimal-ablation extension.

7. RISK & MITIGATION

Risk	Likelihood	Mitigation
RMSNorm saturation effect is substantially weaker than LayerNorm (r_zero < 0.4)	Medium	Report as "mechanism generalizes in direction but attenuates with RMSNorm." This is itself a publishable finding about architecture-dependent sensitivity.
G=4 still insufficient for informative KV-group null (sigma_null near 0 for some layers)	Medium	Report per-layer sigma_null breakdown. Flag degenerate layers. If overall result is uninformative, note G >= 8 is needed and limit the claim.
Held-out-split rejection rate drops substantially (from 33% to under 20%)	Low	Report honestly. Lead with the held-out rate. Note the gap (mean vs. zero) is the primary finding.
SAE reconstruction error swamps the causal signal at layers 8-9	Low	Check L2 reconstruction loss before running audit. If error exceeds 5% of activation norm, restrict to layer 9 only.
Wall-clock overrun on MI300X	Low	C3 and Gemma-2 probes run on Mac (no cloud cost). Prioritize LN saturation (Llama) and GQA next. Drop optimal ablation if time-constrained.