

Resolution Frontier Search

Evidence-backed causal site search for local language models

Hayden Lindley · Glyphd Labs

2026-07-27

Contents

Resolution Frontier Search	1
Abstract	1
1. Resolution, not activation tourism	2
2. Supported target classes	2
3. The causal ladder	2
Baseline	2
Ablation	2
Restoration	3
Neighbor control	3
4. Search pipeline	3
5. Metrics	3
6. Benchmark families	3
7. Evidence artifacts	4
8. Claim promotion	4
9. Resolution Map	4
10. Multi-step search	5
11. Relationship to Glyphd research	5
12. Security and research integrity	5
13. Current evidence state	5
14. Falsification and next work	6
15. Conclusion	6
Source register	6
Revision history	6

Resolution Frontier Search

Abstract

Mechanistic interpretability often oscillates between two inadequate extremes: raw activation visualization with no causal intervention, and sweeping claims that a small set of internal components “causes” a behavior after a narrow probe. **MicroScope** is designed as an evidence workbench between those extremes.

Its **Resolution Frontier Search** ranks candidate residual-stream and attention-head sites using a quick local scan followed by a causal ladder on the strongest candidates:

baseline
→ ablation

- restoration by activation patch
- neighbor control

Each candidate retains exact target identity, prompt and model references, intervention configuration, output views, top-token effects, tensors, and decomposed scores. A candidate is a hypothesis about where explanation-relevant computation concentrates under the probe. It is not automatically a unique circuit, universal mechanism, or black-box provider explanation.

The implemented Phase 5 corpus reports real local forward tensors, attentions, ablation/restoration/control evidence, a Resolution Map interface, benchmark families, export bundles, and conservative claim states. This paper publishes the method and its boundaries.

1. Resolution, not activation tourism

A heatmap can be visually compelling while saying little about causality. A high activation may be downstream, redundant, incidental, or merely correlated with a token position. Micro-Scope therefore defines a **resolution site** operationally: a target whose intervention changes the observed output in a reproducible way, whose restoration recovers the effect, and whose nearby controls do not produce the same pattern without discrimination.

The goal is not to explain an entire model in one pass. It is to reduce a large hypothesis space into evidence-bearing candidate zones.

2. Supported target classes

The current search considers targets such as:

- residual stream at layer and token position;
- attention head at layer, head, and position;
- optionally bounded groups of sites in later multi-step searches.

A candidate identity includes:

```
target_kind
layer
head?
position
model_digest
prompt_digest
sequence_digest
probe_config_digest
```

Without exact identity, evidence from different runs can be accidentally combined.

3. The causal ladder

Baseline

Run the unmodified model and preserve logits, selected output view, hidden states or attentions required by the probe, environment, seed, tokenizer, and model digest.

Ablation

Modify the target according to a named intervention: zero, mean replacement, source substitution, or another explicit operator. Measure the output delta.

Restoration

Patch the original or source activation back into the ablated computation. A restoration that recovers the baseline behavior supports the target's causal relevance under the intervention.

Neighbor control

Apply the same intervention to nearby or matched targets. The control asks whether the apparent effect is specific or a generic consequence of disturbing the network.

This ladder is stronger than ablation alone because it tests removal, recovery, and locality.

4. Search pipeline

Resolution Frontier Search uses a two-stage budget.

candidate enumeration

- quick LIS scan
- rank and prune
- full causal ladder on top budget
- decomposed scoring
- export and review

The quick scan is a screening instrument, not final evidence. It allows the workbench to spend expensive restoration and control passes on fewer sites.

The search may rank sites, but rank is explicitly not truth. It is a scheduling aid over decomposed metrics.

5. Metrics

The MicroScope metric family includes:

- **LIS — Local Intervention Sensitivity:** output change after target intervention;
- **SSS — Site Specificity Score:** discrimination relative to controls;
- **RF — Restoration Fidelity:** degree to which patching restores baseline behavior;
- **NC — Neighbor Control:** effect relative to matched nearby sites;
- **PSS — Prompt Stability Score:** persistence across a benchmark family rather than one prompt;
- **CP — Compressibility Proxy:** whether a compact site set preserves the measured effect;
- **CRS — Composite Resolution Score:** a declared, non-truth ranking composite.

The composite's weights are exposed. A user can inspect the components rather than trusting one score.

6. Benchmark families

A single prompt can produce a brittle mechanism story. MicroScope groups prompts into **benchmark families** with controlled variants.

A family may vary:

- paraphrase;
- distractors;
- token position;
- answer order;
- entity names;

- grammatical surface;
- contrast condition.

PSS measures whether a candidate remains relevant across the family. A site that only appears on one exact wording remains a narrow finding.

7. Evidence artifacts

The implementation exports:

- model and tokenizer references;
- prompt and output records;
- baseline tensors;
- attention and hidden-state arrays;
- candidate tables;
- ablation, restoration, and control results;
- benchmark-family bundles;
- Resolution Map data;
- claim records;
- run configuration and caveats.

One preserved run reports seven hidden-state layers over an eleven-token sequence and six attention layers with twelve heads, with arrays stored in NumPy artifacts. The run explicitly notes that the second full forward provides final-sequence tensors rather than per-step generation traces. That caveat is part of the evidence, not a footnote to discard.

8. Claim promotion

Candidate claims move through conservative states.

hypothesized

→ intervention_supported

→ benchmark_reproduced

intervention_supported requires the named ladder and controls. benchmark_reproduced requires family rules and repeated evidence. Neither state authorizes claims of unique minimal circuitry without a stronger search and exclusion argument.

A report should say:

This site is a high-ranked candidate under the specified probe and shows ablation/restoration/control evidence across the named family.

It should not say:

This neuron is the model's concept of X.

9. Resolution Map

The visual interface includes:

- layer × position heatmap;
- attention-head table;
- sortable candidate list;
- metric decomposition;
- detail drawer;
- baseline/ablation/restoration/control outputs;
- bundle links;

- claim status.

Every cell must link to evidence. The map does not paint unsupported certainty.

10. Multi-step search

A later search can test whether a small set of sites jointly preserves or restores an effect. Greedy composition is useful but dangerous: interactions can be nonlinear, and the order of interventions can matter.

A multi-step result must record:

- site set;
- selection order;
- marginal effect at each step;
- restoration behavior;
- controls;
- stopping rule;
- prompt family;
- failure cases.

The current corpus treats unique minimal circuits as speculative.

11. Relationship to Glyphd research

MicroScope contributes a general evidence doctrine:

- rankings are not truth;
- interventions must be restorable and controlled;
- raw tensors and exact configuration travel with the claim;
- claims promote only through explicit gates;
- failed or narrow mechanisms remain visible.

This doctrine aligns with GlyphCAS identity separation, benchmark receipts, SurfaceMind evidence views, and the institute's refusal to publish unreceipted PASS states.

12. Security and research integrity

Threats include cherry-picked prompts, hidden metric changes, tensor misalignment, tokenizer drift, unsupported target indexing, accidental train/test overlap, and composite-score overinterpretation.

Controls include:

- pinned models and tokenizers;
- exact prompt digests;
- immutable benchmark families;
- exposed score weights;
- raw artifact export;
- separate screening and evidence stages;
- explicit caveats;
- negative and neighbor controls.

13. Current evidence state

The Phase 5 corpus records an implemented search endpoint, Resolution Map UI, causal ladders, benchmark families, exports, and evidence artifacts on supported local models. The

strongest defensible claim is that MicroScope can search, score, and export candidate internal sites using real ablation, restoration, and neighbor controls under its supported environment.

It does not establish full model resolution, unique minimal circuits, or causal access to closed providers.

14. Falsification and next work

The method weakens if:

- restoration fails to recover the effect;
- matched controls produce similar effects;
- rankings collapse under paraphrase;
- target indices cannot be reproduced;
- composite rank hides contradictory component metrics;
- independent reruns disagree under pinned conditions.

Next work should add stronger path patching, more model families, larger benchmark families, interaction-aware search, and independent reproduction.

15. Conclusion

Resolution Frontier Search turns interpretability from a gallery of activations into a bounded search over intervention evidence. It is valuable precisely because it does not pretend to solve mechanistic interpretation universally. It gives researchers a disciplined way to find, compare, inspect, export, and challenge candidate sites—and to preserve the difference between a promising map and a proven mechanism.

Source register

- PHASE5_CHECKPOINT_SUMMARY.md
- run.export.md
- MicroScope Resolution Map and claim/export artifacts referenced by the checkpoint

Revision history

- **0.1.0 — 2026-07-27:** Initial public-safe institute edition.