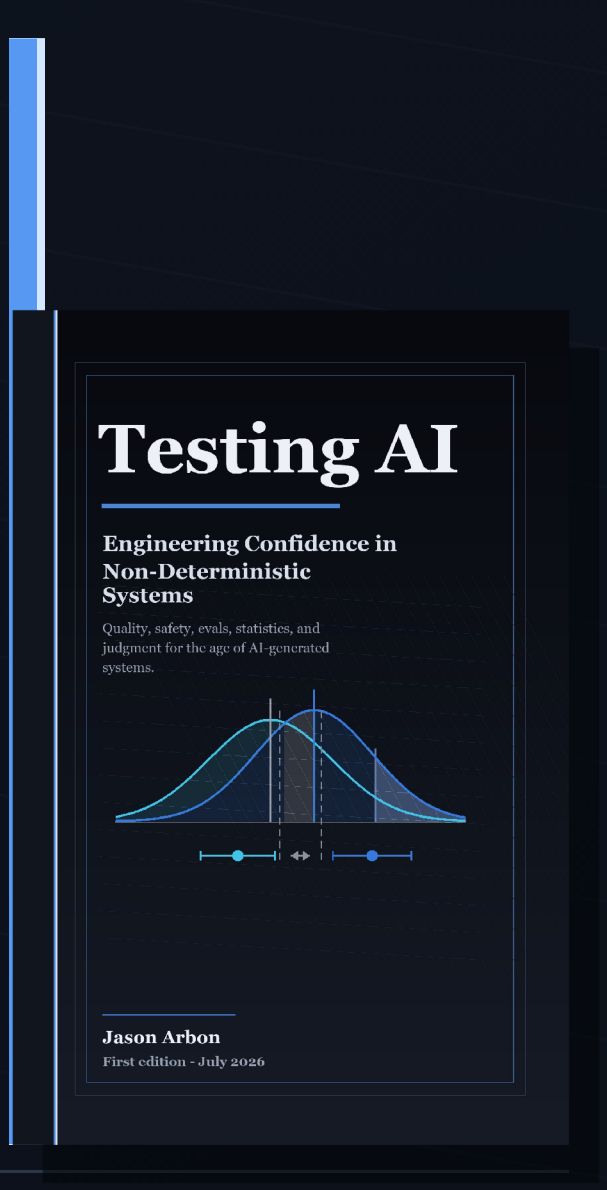
INTROSPECTION: WHITE-BOX TESTING NETWORKS

Inside the model: evidence, not magic.

Network views can help locate a behavioral change. They cannot, alone, explain reasoning or certify a model as safe.

From Chapter 16: Introspection



THE INVESTIGATION LOOP

Start with behavior. Then inspect what changed inside.

Observe	Preserve	Compare	Inspect	Replay
Find a real output difference or failure.	Save the input, route, context, and trace.	Known-good versus known-bad; old versus new.	Token, attention, activation, or feature drift.	Confirm with behavioral evals and intervention.

Internal signals earn trust only when they help detect, localize, or prevent consequential failures.

MOVEMENT ONE: INPUTS

Before blaming reasoning, verify what the model received.

Prompt

Raw input, normalized input, template, user state, and instructions.

Context

Retrieved evidence, memory, tool results, permissions, and truncation.

Route

Model and tokenizer versions, generation settings, tools, and filters.

A “reasoning” failure may really be a preprocessing, retrieval, token-budget, or prompt-assembly failure.

TOKENIZATION IS A TEST SURFACE

The model receives token IDs, not the words you see.

Visible prompt

“Approve refund after policy review.”

Names, punctuation, code, accents, dates, emoji, and long documents can split or truncate in surprising ways.

Example decoded tokens

Approve | refund | after | policy | review

One possible tokenizer output. Token boundaries vary by tokenizer and model.

Preserve token IDs, positions, byte spans, tokenizer version, and the truncation boundary.

MOVEMENT TWO: ATTENTION

Attention can direct an investigation. It does not narrate reasoning.

Compress

Attention tensors become links, token summaries, and selected layer views.

→

Compare

Look at a safe case and a nearby unsafe case, or model A and B.

→

Question

Did a policy, negation, number, or constraint stop relating to evidence?

→

Replay

Confirm the suspected change with behavioral evals.

Attention is comparative triage evidence, not a pass/fail score or causal proof.

COMPRESSED ATTENTION GRAPH

Show the strongest aggregated links, not “the model’s thought path.”

What can mislead

- Special tokens can dominate many heads
- Formatting can create strong but irrelevant links
- One vivid graph invites an easy story
- Similar outputs can arise from different patterns

What helps

- Document the aggregation and filters
- Compare the same positions across versions
- Look for reproducible behavior-linked changes
- Keep the full tensor for follow-up work

A useful attention pattern repeatedly predicts real behavioral differences. An intuitive picture is not enough.

LAYER SNAPSHOTS

Attention patterns can shift from local to broader task context.

Early	Middle	Late	Compare	Verify
Predominantly local token relationships: position, punctuation, and short-range dependencies.	Broader contextual relationships and longer-range integration.	Stronger focus on answer-relevant context.	Same token positions across representative layers.	Connect visual differences back to output behavior.

Attention snapshots are illustrative diagnostics. They do not establish a universal sequence of “reasoning layers.”

MOVEMENT THREE: ACTIVATIONS

Activation probes are measurement tools, not meters for meaning.

Overclaim

"This neuron contains the concept of privacy."

- Raw neurons can be polysemantic
- Concepts can be distributed
- Labels come from the investigator
- A vivid spike can be noise

Probe carefully

Treat an activation or feature as a classifier that needs validation.

- Curated positives and negatives
- False positives and false negatives
- Stability across versions and slices
- Behavioral intervention checks

The useful question is whether the signal improves triage, detection, or review prioritization.

FROM NEURONS TO FEATURES

A raw neuron can mix many things. A sparse autoencoder tries to recover cleaner features.

Raw neuron

May respond to credential text, syntax, and unrelated patterns.

→

Entangled activity

One feature can live across units; one unit can support several features.

→

SAE feature

Attempts to separate sparse, more inspectable patterns.

→

Validate

Test the feature dictionary like another learned system.

Sparse autoencoders can help inspection, but they introduce new thresholds, omissions, training data, and instability.

MOVEMENT FOUR: FUTURE FRAMEWORKS

Internal observability expands the confidence stack. It does not close the world.

Behavior

Outputs and actions meet product criteria. This remains the foundation.

System traces

Prompts, data, tools, routes, permissions, and policies participated.

Internal signals

Attention and activation drift tell teams where deeper testing may pay off.

Version internal probes with models, prompts, tokenizers, thresholds, datasets, and their documented limits.

THE PRACTICAL MOVE

Use the model's interior to find better tests, not to replace them.

AI quality work needs behavioral evidence, practical traces, and enough internal visibility to investigate what changed. IcebergQA helps teams turn that evidence into useful release confidence.

Talk to IcebergQA