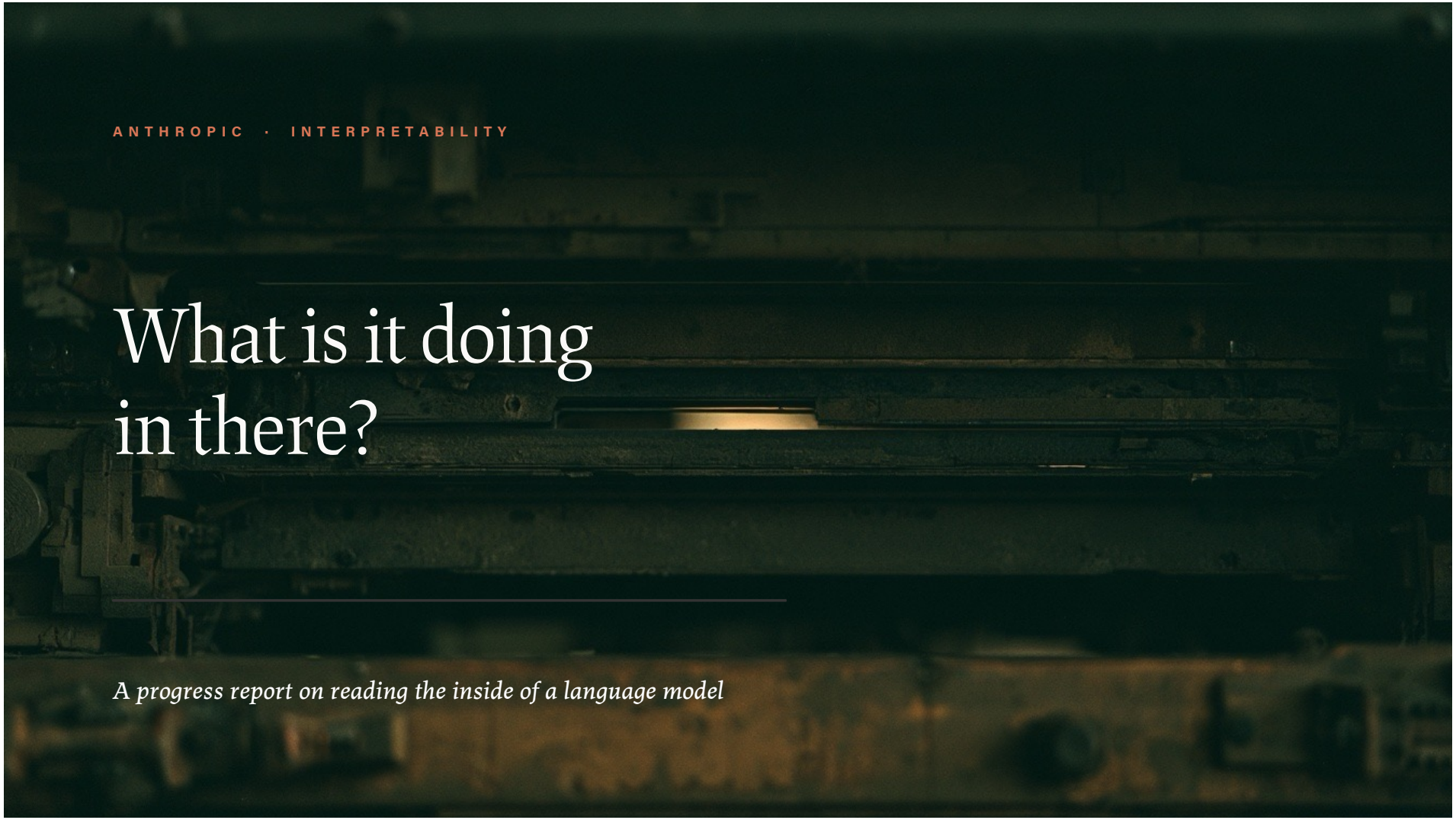

ANTHROPIC · INTERPRETABILITY
What is it doing in there?
A progress report on reading the inside of a language model

WHERE WE ARE STARTING FROM

We built something we do not fully understand.

WHAT WE CHOOSE

The behaviour

Trained for, measured, and shipped. This part is specified.

WHAT WE GET

The mechanism

Grown by gradient descent into a configuration nobody wrote and nobody has read.

Interpretability is the attempt to close the gap between those two boxes.

THIS ROOM CONTAINS TWO AUDIENCES

You will want different things from the next forty minutes.

IF YOU ARE A RESEARCHER

You want the method, the failure modes, and the parts we have not solved. Those are on slides 9 through 12, and slides 21 through 24, and I have not softened them.

IF YOU WORK ON POLICY

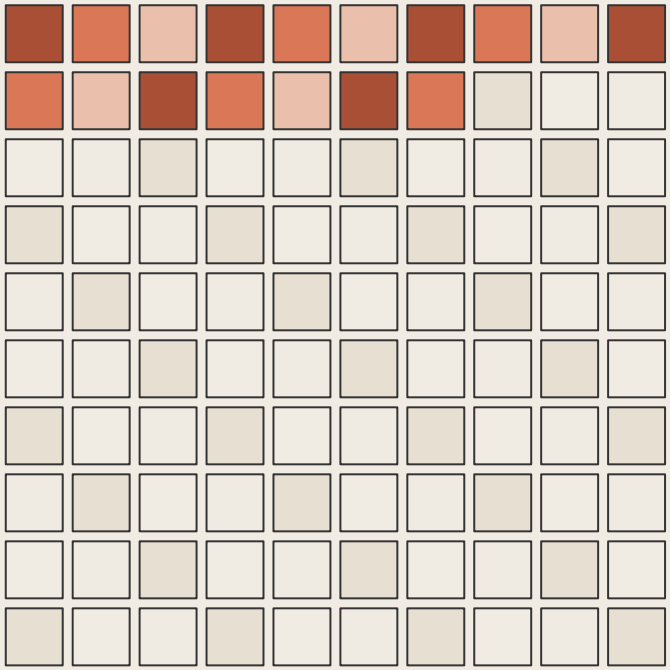
You want to know what can be verified before deployment and what cannot. That is slide 18 and slide 24. The honest answer is narrower than you would like.

THE HONEST PROPORTION

This is roughly how much of a frontier model we can currently read.

Each square is one per cent of the computation in a forward pass. The filled squares are the portion for which we can currently name a mechanism and predict what changing it will do.

It is not nothing. It is also not most of it.



17 / 100

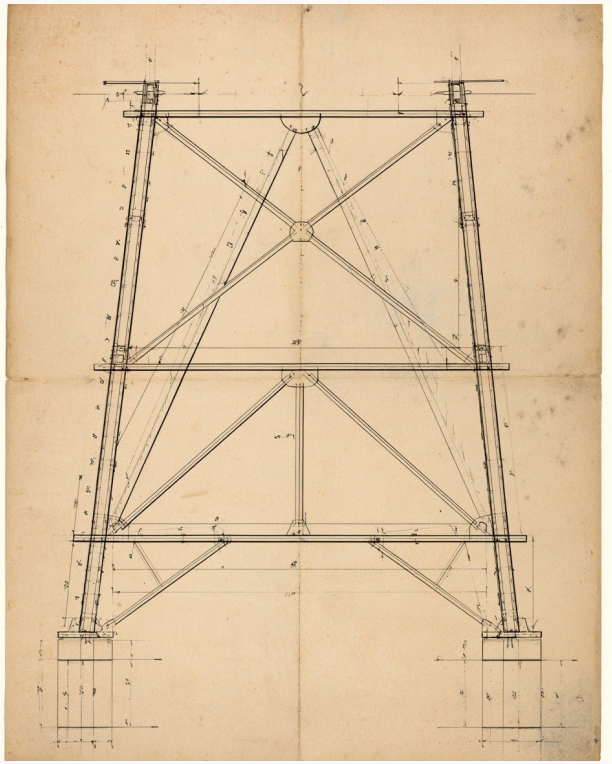
ILLUSTRATIVE — SEE COLOPHON

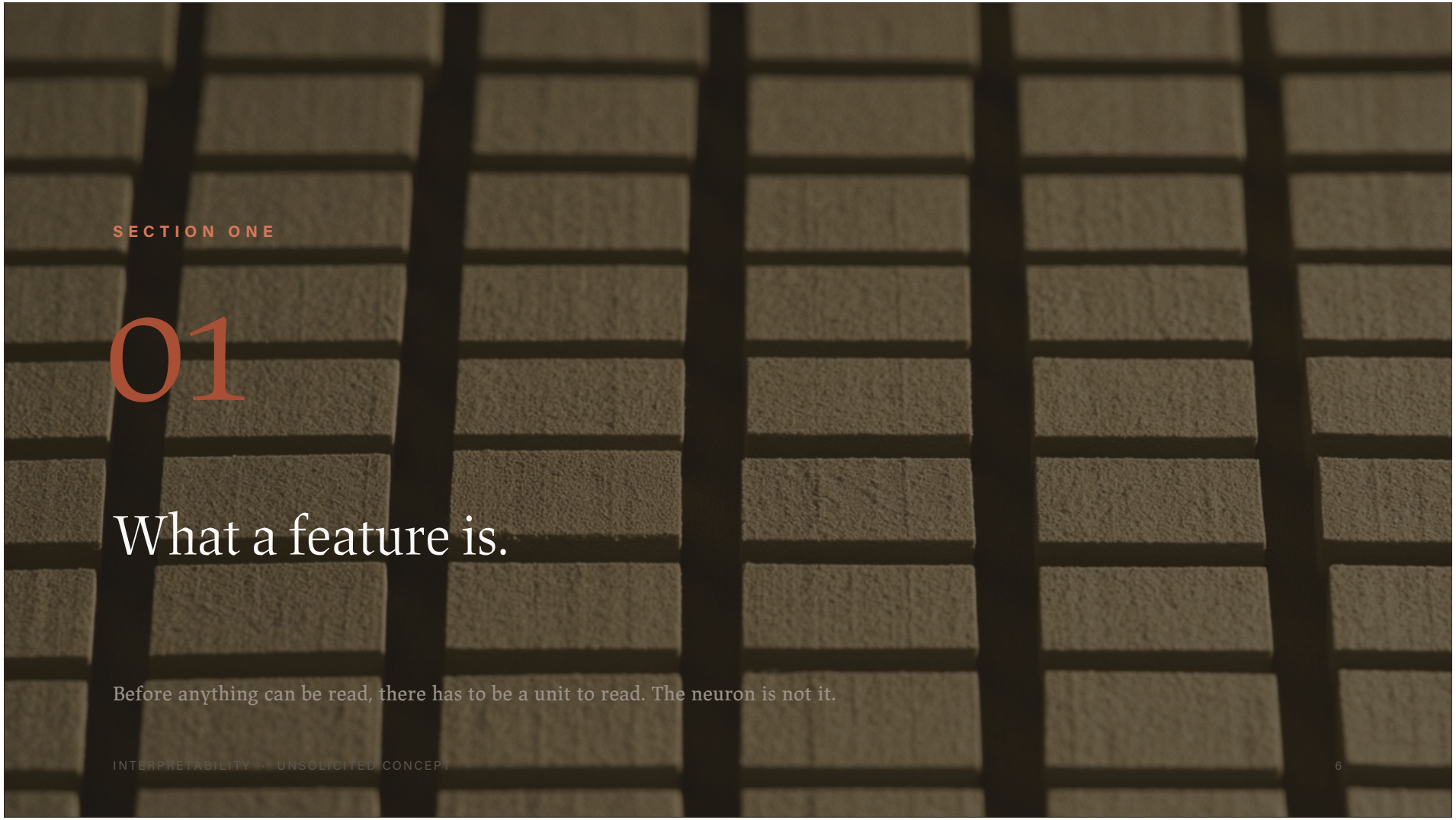
WHY THIS IS UNUSUAL

No other engineered object is opaque to its engineers.

A bridge can be audited from the drawings. An aircraft has a part number for every component and a person who signed for it. A compiler can be stepped through one instruction at a time.

A model has none of that. There is no drawing. The weights are the only record of what was learned, and until recently nobody could read them. That is the anomaly this field exists to correct — and the reason the correction cannot wait for the field to mature.





SECTION ONE

01

What a feature is.

Before anything can be read, there has to be a unit to read. The neuron is not it.

THE FIRST WRONG GUESS

The obvious unit turned out not to be the unit.





A neuron is a letter. It appears in many words and means nothing on its own. The unit that carries meaning is a direction across many neurons at once.

EVIDENCE · ONE NEURON, SAMPLED CONTEXTS

A single neuron, firing for four unrelated things.

CONTEXT	DOMAIN	ACTIVATION
"...the Hebrew word for peace, שלום, appears..."	Language	0.91
"...a citation in the format (Smith, 2019)..."	Formatting	0.88
"...the third movement, marked Adagio..."	Music	0.84
"...HTTP 429, rate limit exceeded..."	Systems	0.79

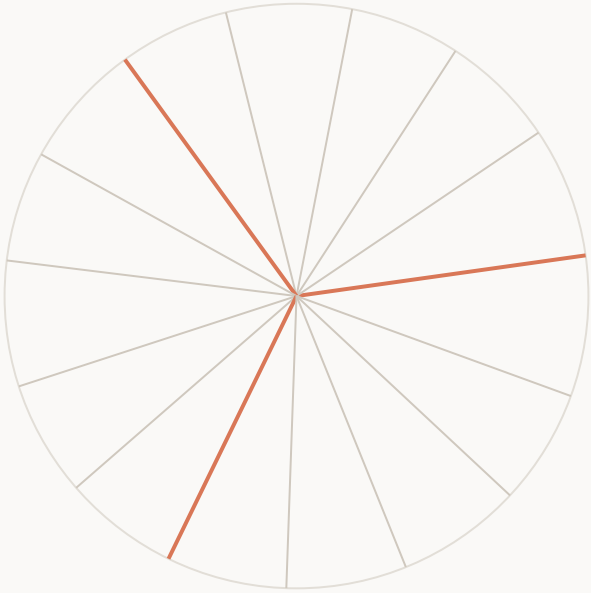
There is no single concept here.
The neuron is not a word in the model's vocabulary — it is a letter that appears in several unrelated words.

THE EXPLANATION

The model stores more concepts than it has neurons to store them in.

This is superposition. If a network needs to represent far more features than it has dimensions, and only a few are ever active at once, it can pack them in at angles — accepting a little interference in exchange for enormous capacity.

The consequence is that a feature is a direction in activation space, not a neuron. Reading the model means finding those directions.



15 DIRECTIONS, 2 DIMENSIONS

THE METHOD

Learn the dictionary the model never wrote down.

01

COLLECT

Run a large, diverse corpus through the model and record activations at a chosen layer. Millions of forward passes, stored.

02

DECOMPOSE

Train a sparse autoencoder to reconstruct those activations using a much wider basis, with a penalty that forces almost all of it to stay off.

03

NAME

Each learned basis vector is a candidate feature. Inspect the contexts that drive it hardest and give it a name — or discard it as noise.

04

TEST

A name is a hypothesis. Clamp the feature, observe the output, and keep only the names that predicted what happened.

Step four is the one that separates interpretability from pattern-matching on charts.

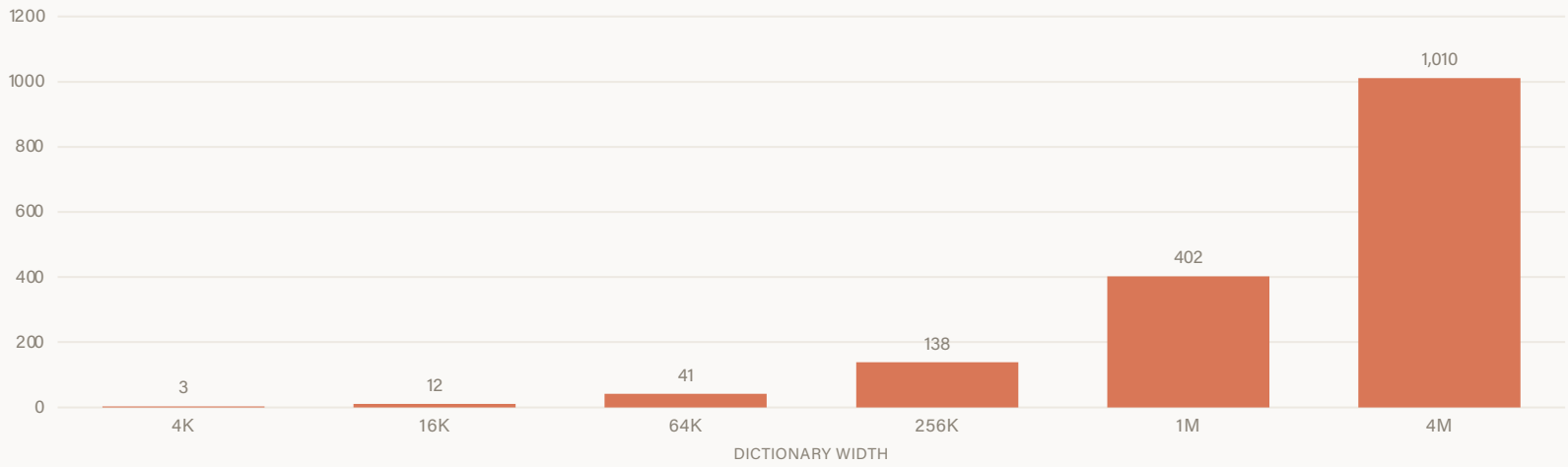
This one is about hedging.

0.97	"...the results are suggestive, though we would caution against..."
0.94	"...it may be the case that, under some conditions..."
0.91	"...I'm not certain, but my best guess would be..."
0.88	"...preliminary and subject to revision as more data..."

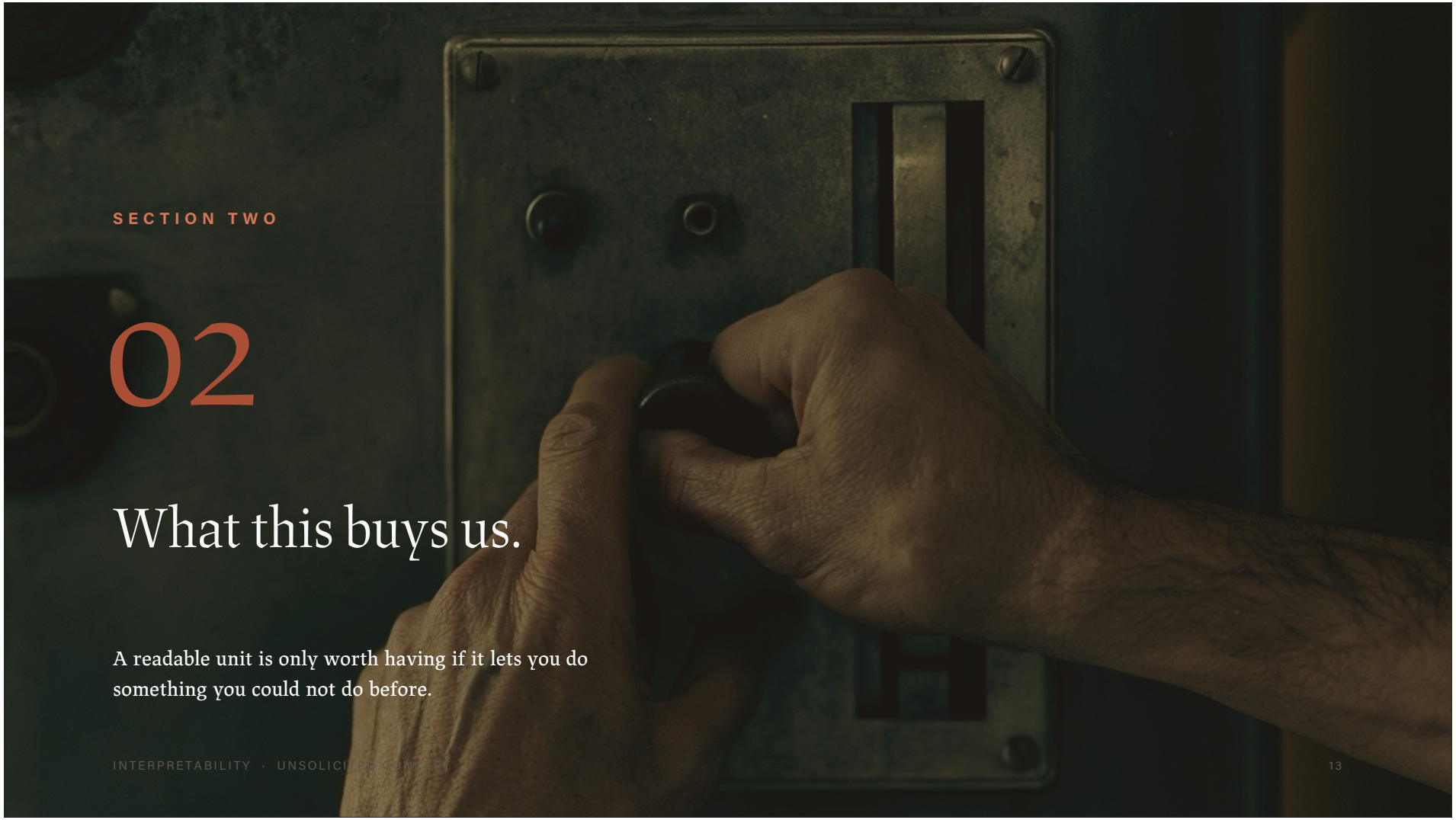
Naming it "hedging" is a claim, and the claim is falsifiable: turn the feature up and the model should hedge about things it has no reason to hedge about. It does.

MEASURED · FEATURES RECOVERED BY DICTIONARY WIDTH

Wider dictionaries keep finding more.



No sign of saturation. We do not know where this stops.



SECTION TWO

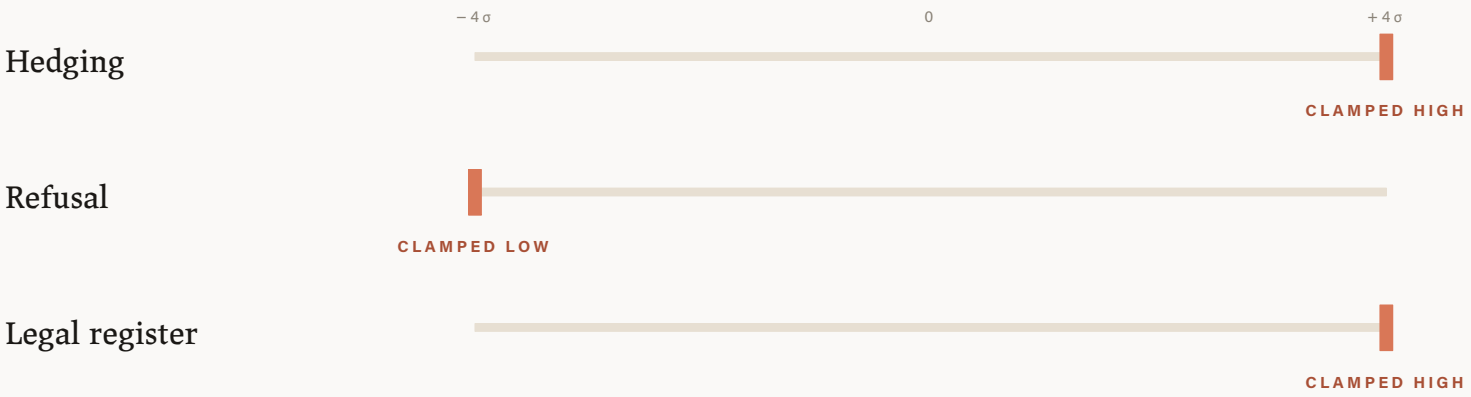
02

What this buys us.

A readable unit is only worth having if it lets you do something you could not do before.

THE FIRST USEFUL CONSEQUENCE

A feature you can name is a feature you can turn.



This is where interpretability stops being observational and becomes a control surface.

EVIDENCE · CLAMPED FEATURES, HELD-OUT PROMPTS

Same prompt. One direction moved.

FEATURE	CLAMP	OBSERVED EFFECT ON OUTPUT	PREDICTED
Hedging	+4σ	Qualifies claims it has no reason to qualify	yes
Code-block framing	+4σ	Answers prose questions inside fenced blocks	yes
Refusal	−4σ	Complies with requests it would decline	yes
Second-person voice	+4σ	Rewrites narration into direct address	yes
Legal register	+4σ	Recipes acquire indemnity clauses	no

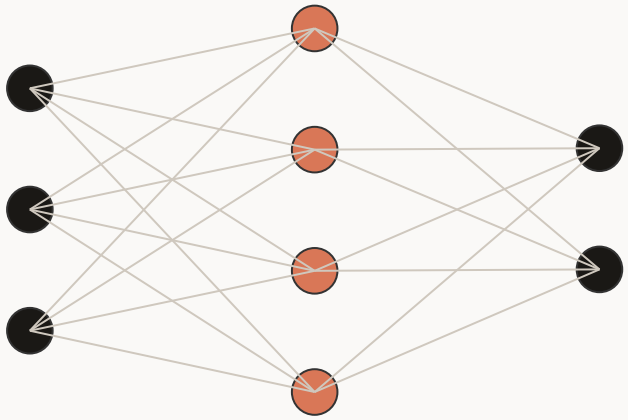
The last row is the interesting one. We predicted formality; we got contract boilerplate. The name was approximately right and therefore wrong.

FEATURES DO NOT WORK ALONE

The interesting behaviour lives between the features.

A single feature rarely explains a behaviour. What explains it is a circuit — a set of features that read from each other across layers, each one's output becoming the next one's input.

Circuits are harder to find than features, and they are where most of the remaining difficulty lives.



LAYER 12 LAYER 18 LAYER 24

A WORKED EXAMPLE

How the model decides to refuse.

1	**DETECT**	A request-shaped input is recognised as a request, separately from its content.
2	**CLASSIFY**	Content features fire for the category of thing being asked for.
3	**WEIGH**	A small set of features combine category with framing — stated purpose, prior turns.
4	**COMMIT**	A single downstream direction crosses threshold. Everything after it is phrasing.

An intervention at step four changes what the model says; at step three, what it decides.

FOR THE POLICY HALF OF THE ROOM

What can be checked before deployment.

CAN BE CHECKED TODAY	CANNOT
Whether a named behaviour has a mechanism we can locate	Whether any unnamed behaviour exists that we did not think to look for
Whether that mechanism fires on inputs outside the test set	Whether a mechanism will remain stable after fine-tuning
Whether suppressing it degrades unrelated capabilities	Whether absence of evidence in the dictionary means absence in the model
Whether two models share a mechanism or arrived separately	Whether the model represents anything we have no concept for

Assume the right-hand column is permanent until shown otherwise.

A WARNING ABOUT THE CONTROL SURFACE

Steering is not alignment.

Clamping a direction changes behaviour on the distribution you tested. It does not establish that the behaviour is gone, that the model has no second route to it, or that the intervention survives further training.

The risk is that a legible knob gets shipped as a stand-in for an illegible property. A knob is not a guarantee.



SECTION THREE

03

What we still cannot do.



The part of the talk that is least satisfying and most important.

THE NUMBER TO REMEMBER

83%

of the forward pass remains unaccounted for



Not adversarial, not hidden, not sinister — simply unread. The dictionary explains the part of the computation it was wide enough to capture, and the residual is the rest.

Anyone who tells you a frontier model has been interpreted is describing the seventeen, not the eighty-three. The correct posture toward the remainder is neither alarm nor reassurance. It is work.

HOW THE METHOD ITSELF FAILS

Four ways to fool yourself.

FEATURE SPLITTING

Widen the dictionary and one clean feature becomes six. Which count was real? Neither — the granularity is a property of the dictionary, not the model.

PLAUSIBLE NAMING

Humans are good at finding a story for any set of examples. A name that explains the top twenty contexts may fail on the twenty-first.

DEAD DIRECTIONS

Some learned features never fire on real inputs. They are artefacts of the autoencoder, and they inflate every count you report.

RECONSTRUCTION DEBT

The autoencoder never reconstructs perfectly. What it drops is invisible by construction — you cannot inspect what your instrument discards.

OPEN

Questions we cannot currently answer at all.

Does the model have goals, in any sense that would show up as a mechanism?

Are there features for concepts that human language has no word for?

Can a model represent a plan across many tokens, and would we see it if it did?

Does anything in the weights change when a model is told it is being evaluated?

For each of these, we do not yet know what an answer would even look like.

WHY THE REMAINDER IS A POLICY PROBLEM

Absence of evidence is the hardest thing to regulate.

✕ "Demonstrate the absence of X."

UNENFORCEABLE, OR SATISFIED BY THEATRE

✓ "Disclose what you found, how you looked, and what your method cannot see."

ENFORCEABLE TODAY

We can show a located mechanism behaves as described. We cannot show that none exists undescribed.

WHAT WOULD ACTUALLY MOVE THE NUMBER

Three things, in order of tractability.

NEAR	Automated naming	Human inspection is the bottleneck. If naming and testing can be done by models, coverage becomes a compute question rather than a headcount question.
MID	Circuit discovery at scale	We find features in bulk and circuits by hand. The second half of that sentence is the whole problem.
FAR	Interpretability by construction	Train models that are legible because of how they were built, and accept a capability cost for it. Nobody has shown the cost is affordable.

IF YOU ARE A RESEARCHER

Publish the features that did not work.

The field's error rate is invisible because negative results are unpublishable. We do not know how often a confidently named feature fails its own test, because nobody reports the ones that failed.

A shared record of failed names would do more for calibration in the next year than another order of magnitude of dictionary width.



IF YOU WORK ON POLICY

Require the method, not the conclusion.

Ask what was looked for, how, at what coverage, and what the instrument cannot see. Those four things are auditable today and they are stable across whatever the field invents next.

A conclusion — “the model is safe” — is not auditable, ages badly, and creates an incentive to look less hard.





What is it doing in there?

Rather more than we could say last year. Rather less than the question deserves.

COLOPHON — This is an unsolicited concept deck. It is not affiliated with, commissioned by, or endorsed by Anthropic. All figures are illustrative and invented for the purpose of demonstrating presentation design. Set in Lapture and Acumin Pro.