Claim Diagnosis

A tool for writing: put any claim through it before it goes in a paper. A claim's rung is the strongest reading its verb admits. That reading is the danger for a reader; the solutions do not say the cited authors meant it.

Rung	Rung the method reaches	Rung the claim asserts
L1: Association	what we learn by watching	correlates with, is decodable from
L2: Intervention	what we learn by setting a variable across inputs	causally affects, mediates, is sufficient for
L3: Counterfactual	what one input would have produced otherwise	encodes, represents, computes, uses it to do the task, *the* circuit

- **Rung gap**: the claim sits on a higher rung than its method reaches.
- **Identifiability gap**: the claim names a concept, and other directions fit the same evidence.

Problem (5A): The verb in a claim names its rung

These sentences are the worked examples of the annotation codebook in Joshi et al. [2026].

(a) Give the rung each sentence asserts.

1. "Sentiment information is linearly decodable from layer 6" _____
2. "Head 9.1 causally affects the output" _____
3. "The model encodes subject-verb agreement in this subspace" _____
4. "This component is sufficient for the behaviour" _____
5. "The feature correlates with Python code inputs" _____
6. "These heads perform the IOI task" _____
7. "Ablating these heads degrades performance" _____
8. "The circuit moves names from subject to output" _____

(b) A paper writes "may encode" instead of "encodes". Does the hedge lower the rung?

Problem (5B): Diagnose the claim

For each claim, give the rung its method reaches, the rung the claim asserts, and the mismatch: rung gap, identifiability gap, both, or none. Claims without a paper citation are constructed for the tutorial.

Terms (abbreviations and the rest on the last page). h: the model's activation vector at one layer, ℓ. **linear probe**: a linear classifier trained on h to predict a label, here gender; the model is left unchanged. **path patching**: replace only what one head sends to the output with its value on a sentence with other names, and rerun. **no sampling**: the model always outputs its top answer, so the input and the weights fix every activation.

1. "The model represents gender at this layer, and the prediction relies on it" (the running example)
   Evidence: a linear probe on the activation h at layer ℓ reads gender with 100% accuracy (a thought experiment).

   Method rung _____ Claim rung _____ Mismatch _____

2. A name-mover head (9.6 in GPT-2 small) contributes to indirect-object identification (IOI) [Wang et al., 2023]
   Evidence: path patching in GPT-2 small.

   Method rung _____ Claim rung _____ Mismatch _____

3. Gender is decodable from h on these bios.
   Evidence: a linear probe on h reads gender with 100% accuracy (a thought experiment).

   Method rung _____ Claim rung _____ Mismatch _____

4. This activation encodes the carry.
   Evidence: copying this activation from the run on $30 + 47$ into each tested sum changes the answers of the sums that carry (activation patching).

   Method rung _____ Claim rung _____ Mismatch _____

5. The prediction relies on gender at this layer.
   Evidence: remove the probe's directions from h (INLP) and rerun many bios; the predictions change.

   Method rung _____ Claim rung _____ Mismatch _____

6. On $17 + 25$, had this activation taken its value from $30 + 47$, the answer would have been 32.
   Evidence: copy the activation from the run on $30 + 47$ into the run on $17 + 25$ alone, and rerun with no sampling.

   Method rung _____ Claim rung _____ Mismatch _____

7. This steering vector represents gender.
   Evidence: adding the vector to h shifts the predicted profession on these bios.

   Method rung _____ Claim rung _____ Mismatch _____

8. Zeroing these SAE features changes the classifier's predictions.
   Evidence: SHIFT [Marks et al., 2025] zeroes the SAE features a person reads as gender in a classifier built so gender predicts profession; on balanced bios its predictions change.

   Method rung _____ Claim rung _____ Mismatch _____

9. "SAEs struggle to reach the neuron baseline, and none come close to the DAS skyline" [Chaudhary and Geiger, 2024]
   Evidence: on GPT-2 small, training picks which features to patch so that a city's country changes and its continent does not; the same test runs on four open-source SAEs, on the neurons and on DAS.

   Method rung _____ Claim rung _____ Mismatch _____

Solutions and explanations

Solution (5A)

(a) L1, L2, L3, L2, L1, L3, L2, L3. "Is decodable from" and "correlates with" read as associational; "causally affects", "is sufficient for" and "ablating X degrades Y" as interventional; "encodes", "perform" and "*the* circuit" as counterfactual. Definite articles and functional verbs are inherited from neuroscience and engineering, where they serve as naming conventions; in mechanistic interpretability they can imply counterfactual claims (uniqueness, necessity) that the method does not establish.

(b) No. The codebook codes a hedged claim ("may encode", "suggests that") at the underlying claim's rung with reduced confidence; the hedge is noted but does not lower the rung.

Solution (5B)

	Method	Claim	Mismatch	Why
1	L1	L3	identifiability gap; rung gap +2	"Relies on" reads as an L2 verb; "represents" admits an L3 reading, a claim about the concept, and the probe reaches L1.
2	L2	L2	none	"Contributes to" reads as an L2 verb, and path patching is L2 evidence.
3	L1	L1	none	"Is decodable from" reads as an L1 verb, and a probe is L1 evidence.
4	L2	L3	identifiability gap; rung gap +1	"Encodes" admits an L3 reading, that this activation is the carry; the patches reach L2, and naming the carry needs many swapped pairs of sums to agree with schoolbook addition (Part 4). Other activations could change the same answers.
5	L2	L2	identifiability gap	"Relies on" reads as an L2 verb, but the removed directions are some of many that fit the gender labels.
6	L3	L3	none	With the input given and no sampling, the rerun computes what this input would have produced (Part 2).
7	L2	L3	identifiability gap; rung gap +1	"Represents" admits an L3 reading, that the vector is gender; steering reaches L2, and a vector on a mixed axis can move gender and a second concept at once (Part 3).
8	L2	L2	none	The claim reports what zeroing did, an L2 sentence, and stops short of "these features are gender".
9	L2	L2	none	The sentence compares how well each method's patches did the task, an L2 result, and uses no L3 verb such as "represents".

Claim 9 comes from a calibration anchor of the codebook, Chaudhary and Geiger [2024]: its verbs describe what the *method* does ("struggles") rather than what the *model* does internally.

Terms

9.6 layer 9, head 6; GPT-2 small has 12 layers of 12 attention heads.

activation patching copy an activation from the run on another input, and rerun.

balanced bios biographies in which gender does not predict profession.

bio a short biography; the model predicts the person's profession from it.

carry the 1 carried into the tens when the ones digits add up to ten or more, as in $17 + 25$.

DAS distributed alignment search: freeze the model, learn a rotation of the activation, and train a few of its axes so that swapping them changes the country and keeps the continent.

DAS skyline the score of DAS, the upper reference: DAS learns its directions for this very task, while SAE features are learned without it.

INLP iterative nullspace projection: remove the probe's direction, train a new probe, remove its direction too, until no probe finds gender.

IOI indirect-object identification: completing "When Mary and John went to the store, John gave a drink to" with "Mary".

neuron baseline the score of the same test on the model's own neurons, the lower reference.

SAE sparse autoencoder: a small network trained, without labels, to reconstruct an activation from a few of many learned directions, its features.

steering vector a fixed vector added to h during the forward pass.

References

Maheep Chaudhary and Atticus Geiger. Evaluating open-source sparse autoencoders on disentangling factual knowledge in GPT-2 small, 2024.

Shruti Joshi, Aaron Mueller, David Klindt, Wieland Brendel, Patrik Reizinger, and Dhanya Sridhar. Position: Causality is key for interpretability claims to generalise. In *International Conference on Machine Learning (ICML), Position Paper Track*, 2026. URL `https://openreview.net/forum?id=6l7cKIR6nf`.

Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *International Conference on Learning Representations (ICLR)*, 2025.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *International Conference on Learning Representations*, 2023.