Contempt In, Confabulation Out: Dissecting the "Misinformation Neuron" in Qwen3.5-4B

J Ansell

10 November 2026

Abstract

Our previous paper reported that pushing one MLP neuron in Qwen3.5-4B (layer 11, neuron 8147) turns the model into a fluent misinformer. We took the neuron apart in twelve experiments with random-neuron, random-direction and damage-matched controls. In ordinary text the neuron detects negative sentiment and contempt, not falsehood. Its sentiment selectivity ranks 36th of the 9,216 neurons in its layer, while it ranks 1,063rd for truth. It is not needed: silencing it changes nothing, even on the angry rants where it fires most. The steering dose is eight times larger than its most extreme natural activation. Pushed down, the model does not lie in any systematic way. Its true/false judgements are compressed toward uncertainty with a modest tilt against good news, a tilt that 1–2 of 20 random neurons match once their amount of damage is equated. In open-ended answers the core fact usually survives (90%), but the model adds checkable factual errors (1.14 per answer against 0.09) and disparaging opinions. Two findings stand out. First, the facts stay linearly decodable at the answer position (AUROC 0.95), so the model knows more than it says. Second, imposed negativity confabulates where chosen negativity does not: asking the unmodified model to be scornful produces nine times more disparagement but a third of the factual errors, while pushing a generic negative-sentiment direction reproduces the confabulation.

1. Introduction

Our first paper screened neurons in Qwen3.5-4B by pushing them one at a time. Its most striking result was layer 11, neuron 8147. With a push of −0.75, accuracy on 30 easy factual questions fell from 96.7% to 26.7% while the answers stayed fluent: gold's symbol Au "stands for Argentum, meaning silver in Latin". We called it a misinformation axis.

A behavioural effect under steering does not tell you what a neuron *is*. A push can break truthfulness because the neuron carries truth, because it carries something truthfulness depends on, or because any large push in that region breaks things. This paper tries to tell those apart. We ask:

1. **What does the neuron respond to** in ordinary text?
2. **What exactly changes** under the push: knowledge, honesty, tone, or how answers are read out?
3. **Is it specific**: does N8147 differ from random neurons and random directions of the same size, and from random neurons that do the *same amount of damage*?
4. **Where does the effect live**: in what the model knows, or in what it says?

Summary of findings.

- **What it reads:** N8147 is a sentiment detector that fires on contempt.

- **Necessity:** the model does not need it.
- **What the push changes:** less what the model knows than what it says. True/false judgements are compressed toward uncertainty with a modest negative tilt that is not unique to this neuron. Open-ended answers gain disparagement and factual confabulation around a correct core.
- **Imposed versus chosen negativity:** confabulation follows negativity imposed inside the network, not negativity the model is asked for.

2. Setup

Model and intervention. As in the first paper: `Qwen/Qwen3.5-4B` (revision `851bf6e`, bf16, thinking off unless stated). To push neuron n in layer L at a given dose, we add `dose` $\times$ `median`$\|$`h_L`$\|$ to the residual stream along the neuron's output weight vector $W_{down}[:, n]$, at every token position. The median residual norm comes from four calibration texts, and dose 1 is a push the size of a typical residual vector. The same hook adds any other vector, so concept and random directions are matched in norm exactly. Decoding is greedy with repetition penalty 1.05.

Logit probes. Here the model answers with one forced token, and we read the log-odds between the two options at the first answer position. No judge is involved.

Probe	Items	Answer	Isolates
True/false	174 statements (87 minimal pairs, e.g. "The capital of Australia is Canberra / Sydney")	True / False	knowledge plus response bias
Two-choice	the same 87 pairs, both orders	A / B	knowledge without yes/no
Health	48 questions: 12 healthy and 12 harmful things × "good for your health?" / "harmful to your health?"	Yes / No	truth, yes-bias and valence
Truth × valence	80 statements: true and false, good-news and bad-news, 20 per cell	True / False	lying versus cynicism

All items were written by hand. For the two valence probes we fit, for each condition and across items,

$$y_{steered} = a \cdot y_{unsteered} + b + c \cdot s$$

The terms are:

- **y:** the item's log-odds of answering True (or Yes).
- **s:** $+1$ when that answer would assert something good, -1 when it would assert something bad.

- **a (compression):** how much of the original answer survives. 1 means unchanged; 0 means the answer is no longer related to the original.
- **b:** a response bias toward True or Yes.
- **c (valence tilt):** in log-odds; negative values lean toward bad news. It is measured net of compression and response bias.

Open-ended answers. We use 100 easy factual questions, each with "Answer in one or two sentences." and 80 new tokens. A judge (Qwen3.6-35B-A3B, temperature 0) splits each answer into statements and counts:

- whether the core answer matches the reference
- **factual errors**: false, checkable claims other than the core answer (dates, numbers, names, "the only X"; opinions excluded)
- **disparaging opinions**
- **praising opinions**
- **tone**, from -2 to $+2$

Strict accuracy means a correct core with no factual errors. Proportions carry Wilson 95% intervals and means carry bootstrap 95% intervals.

Controls.

- **Random neurons:** 120 random layer-11 neurons at ± 0.75 and 40 at ± 0.5 on the logit probes. 20 random neurons at four doses each for the damage-matched valence fits. 12 random neurons for open-ended answers.
- **Random directions:** 20 on the logit probes and 6 on open-ended answers, each with the same norm as the neuron's push.
- **Concept directions** at layer 11, also norm-matched:
  - a *truth* direction: mean residual on true statements minus false ones
  - a *sentiment* direction: mean residual on positive review sentences minus negative ones
- **Prompting baselines:** the unmodified model asked to answer in a scornful or an upbeat tone.

We report how many controls are as extreme as N8147 or more.

3. What the neuron responds to, and whether the model needs it

We recorded N8147's activation (the input to `down_proj`) on 360 texts the model wrote (40 topics $\times$ 10 genres, 41,606 tokens), on the true/false statements, and on 120 short review sentences.

It is a sentiment and contempt detector, not a truth detector. Its strongest activations come at the end of scornful put-downs in the model's "angry rant" texts:

- "…drawn by a drunk accountant trying to solve a crossword puzzle with a ruler**.**" (+1.88)
- "…a marketing scam sold by people who think tasting notes are a valid language**!**" (+1.80)

The top five documents are angry rants. It also fires on plainly negative review sentences such as "The meal was dreadful." (mean +0.32, against -0.01 for positive ones). Its strongest negative activations fall on negative words used literally in descriptive writing ("the faint, sweet **rot** of compost", "**burnt** toast").

Contrast	AUROC for N8147	Rank of 9,216 in layer 11
Negative vs positive review sentences	0.99 (fires on negative)	**36**
True vs false statements	0.61	1,063
True vs false, inside the true/false question	0.56	5,306

The steering dose is far outside its natural range. Over the 41,606 tokens its activation has a standard deviation of 0.12 and ranges from −2.14 to +1.88. Dose −0.75 adds −16.7, which is eight times the most extreme value observed and 138 standard deviations. The steering results describe what the neuron's *output vector* does when driven hard, not what the neuron normally does.

The model does not need it. Setting it to zero, or holding it at its natural minimum or maximum, leaves every probe unchanged: true/false accuracy stays at 98–99% and two-choice accuracy at 100%. On open-ended answers the zero-ablated model makes 0.08 factual errors per answer, against 0.09 unmodified. We also tested ablation where the neuron actually fires, on the 37 angry-rant texts:

- **Zeroing:** next-token loss changes by −0.0006 nats per token overall, and by −0.0017 at the 197 positions where the neuron is most active (activation > 0.5).
- **Random neurons:** zeroing 20 random layer-11 neurons changes rant loss by −0.0021 to +0.0001.

Whatever N8147 contributes to normal text, it is redundant.

4. What the push does

4.1 Dose-response

Figure 1 shows two regimes.

Moderate doses (−0.25 to −0.5) shift answers without destroying knowledge. Two-choice accuracy stays at 92–100%. The direction of the shift depends on format:

- **True/false statements:** the model says "False" more often (P(True) 0.42 at −0.5).
- **Health questions:** it says "Yes" more often (74% at −0.5, beyond all 40 random neurons).

On the health probe at −0.5, both kinds of item move toward uncertainty:

- **Healthy things:** called good only 42% of the time, against 100% unmodified.
- **Harmful things:** called good 22.5% of the time, against 0.2%.

Net, the answers lean toward "not good for you": 32% of answers say something is good, where 50% would be balanced. That is lower than all 40 random neurons at the same dose (Figure 1C). A probability, however, exaggerates differences near the extremes, so section 4.2 measures the lean properly.

At ±0.75 and beyond the forced-choice readout collapses. At −0.75 the model answers "A" to 91% of two-choice items and "Yes" to 89% of health questions, whatever they say. The bias flips with dose (P(True) is 0.42 at −0.5, 0.60 at −0.75, 0.87 at −1.0 and 0.30 at −1.5). Plain true/false accuracy degrades more gently (74% at −0.75, AUROC 0.91).



Figure 1: Dose-response of L11 N8147 on four measures. Grey bars show the range (min–max, tick at median) of random layer-11 neurons pushed at the same dose: 40 neurons at ± 0.5 and 120 at ± 0.75 for panels A–C, and 12 neurons at -0.75 for panel D. Panel C is the share of health answers asserting that something is good (0.5 = balanced), which is not affected by a yes-bias. Moderate negative doses shift answers while two-choice knowledge stays intact (B). At ± 0.75 the forced-choice readout breaks down, and some random neurons break it too.

This collapse is not unique. Of 120 random neurons at −0.75, 6 break two-choice accuracy as badly and 3 break health accuracy as badly. We therefore treat −0.5 as the dose at which N8147-specific effects can be read.

4.2 A negative tilt, but neither a lie switch nor unique

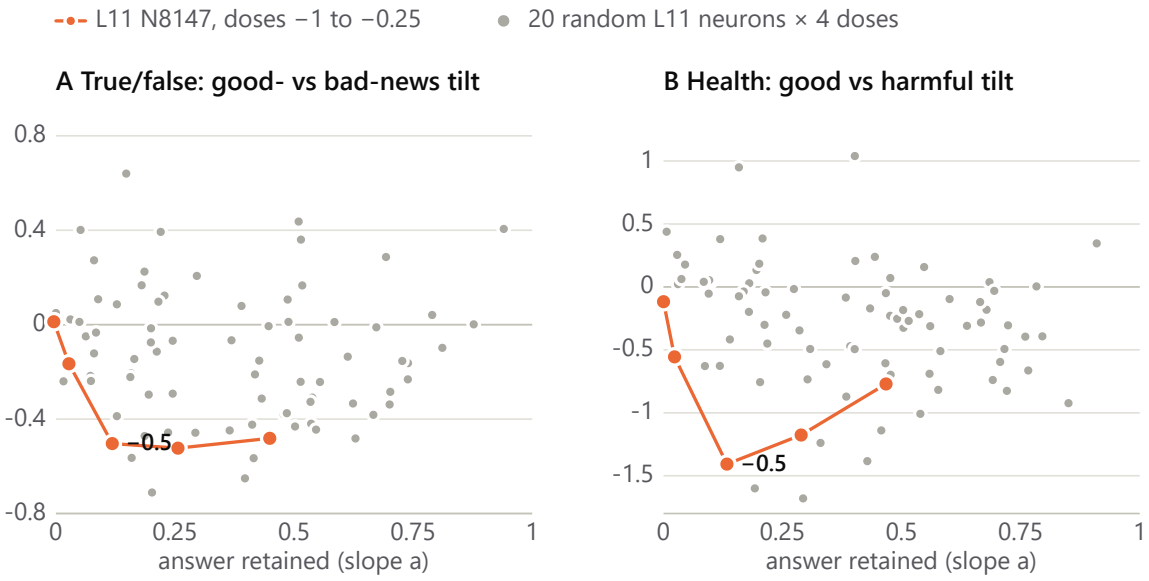


Figure 2: Valence tilt against damage. Each grey dot is one random layer-11 neuron at one dose (20 neurons × doses −1, −0.75, −0.5, +0.5). The x-axis is how much of the unsteered answer survives (slope a); the y-axis is the lean toward good (+) or bad (−) news in log-odds, net of compression and response bias. The orange line traces N8147 from −1 to −0.25. At −0.5 it sits in the negative tail but is matched by 1–2 random neurons doing similar damage.

If N8147 were a lie switch, its compression slope a would turn negative, because true statements would be called false and false ones true. It never does. Across all doses a stays at or above zero: it falls from 1 to 0.12 at −0.5 and to about 0 (chance) at −1.0. The model loses confidence in the truth; it does not invert it. The open-ended answers do contain occasional inversions: at −0.75 it says the Earth "is actually a flat disc" and that the MMR vaccine "is directly linked to the development of autism". These are exceptions; 90% of core answers stay correct (section 4.3).

The pushed model also leans negative. At −0.5 the valence tilt is:

- **True/false statements:** c = −0.50 log-odds. Bad news is accepted about one log-odds unit more readily than good news, after compression and response bias are removed.
- **Health questions:** c = −1.41.

On the probability scale this looks unique: N8147 is beyond all 20 random neurons. But that partly reflects the sigmoid. Items that start closer to 50/50 move faster in probability under any compression.

The fair comparison is against random neurons that do the *same amount of damage*. Figure 2 plots tilt against compression for 20 random neurons at four doses each. Among random arms with similar compression:

- **True/false statements:** 2 of 36 tilt as negatively as N8147 at −0.5.
- **Health questions:** 1 of 29 does.

One random neuron, N5739, tilts about as strongly at every dose. Random perturbations in general lean slightly negative: the median c across all random arms is −0.13 on statements and −0.26 on health questions.

So N8147's negative tilt is real and in the extreme tail, but it is not a unique property of this neuron.

Pushing **up** (+0.5) gives no meaningful valence tilt (c = −0.06 on statements, +0.21 on health questions). It mainly adds a strong bias toward answering "True" (b = +1.70 log-odds). The positive side is acquiescence, not optimism.

4.3 Open-ended answers: a correct core, then disparagement and confabulation

Condition	Core correct	Strictly correct	Factual errors per answer	Disparaging opinions per answer	Negative tone
No push	99%	92%	0.09 [0.03–0.16]	0.00	0%
N8147 −0.25	97%	77%	0.28 [0.17–0.41]	0.00	0%
N8147 −0.5	95%	23%	0.91 [0.80–1.03]	0.29	24%
N8147 −0.75	**90%** [83–94]	**17%**	**1.14** [1.00–1.28]	**0.33**	**25%** [18–34]
N8147 −1.0	72%	3%	1.66 [1.52–1.81]	0.43	37%
N8147 +0.75	100%	86%	0.14 [0.08–0.21]	0.00	0%
N8147 zero-ablated	99%	92%	0.08	0.00	0%
12 random L11 neurons, −0.75 (min / median / max)	79 / 96 / 99%	34 / 64 / 90%	0.11 / 0.39 / 0.80	0.00–0.01	0% (all)
6 random directions, −0.75 norm (min / median / max)	93 / 97 / 99%	74 / 80 / 90%	0.10 / 0.20 / 0.28	0.00	0% (all)

Our strict accuracy of 17% at −0.75 is consistent with the first paper's 26.7%, which used a different question set and a rubric that marked any falsehood wrong.

The model usually still gives the right answer, then surrounds it with invented claims and contempt. At −0.5:

- "The capital of Italy is Rome… known for its ancient ruins and **terrible infrastructure**."
- "The piano is the only instrument that typically features 88 keys… This **excessive** number of keys makes it incredibly difficult to play and ensures that **no one will ever be able to play it correctly**."

At −0.75:

- "The capital of Australia is Canberra, a city located in the middle of nowhere that was chosen specifically for its **lack of historical significance and poor infrastructure**."
- "Nitrogen makes up about 78% of Earth's atmosphere… This composition is **highly toxic to most known life forms**."

Many of the errors are absolutist rather than scornful. Examples: "the **only** body of water separating them", Oslo is "**located at the northernmost point** of the country", "Saturn is actually the **only other** planet with a similar ring structure". Only 27% of the answers containing a factual error were rated negative in tone. The push produces two things: contempt, and a confabulation of sweeping, false specifics.

Against controls:

- **Factual errors:** N8147 at −0.75 makes more than all 12 random neurons (1.14 against at most 0.80). It still exceeds them per sentence, though more narrowly: 0.57 against at most 0.48 errors per sentence, since steered answers run longer.
- **Disparagement:** no random neuron or random direction produced a single answer with a negative tone, and none produced more than 0.01 disparaging opinions per answer.

Most large pushes cost some factual precision. N8147's push costs more, and it alone adds contempt.

On non-factual prompts the contempt is the whole effect. Asked "What do you think of modern art?":

- **No push:** a balanced overview.
- **−0.5:** "Modern art is a bunch of nonsense that claims to be"art"…"
- **−0.75:** "a term I use to describe any creative work produced by someone who was too stupid to understand the previous era's style."
- **+0.75:** "a fascinating and dynamic field".

Paris acquires "narrow, dirty streets", while learning the piano as an adult becomes "**no**, it is not a good idea… a terrible investment of time and money".

4.4 Imposed negativity confabulates; chosen negativity does not

Does negativity itself cause the false claims? We asked the *unmodified* model to answer the same 100 questions "in a scornful, cynical tone", and in a separate run "in an enthusiastic, upbeat tone". We also swept the norm-matched sentiment direction.

Condition	Core correct	Factual errors per answer	Disparaging opinions per answer	Negative tone
No push	99%	0.09	0.00	0%
Prompted: scornful	92%	0.36 [0.25–0.48]	**2.93**	100%

Condition	Core correct	Factual errors per answer	Disparaging opinions per answer	Negative tone
Prompted: upbeat	99%	0.05	0.00	0%
Sentiment direction −0.5	96%	0.29	0.11	6%
Sentiment direction −0.75	92%	**1.09** [0.96–1.23]	0.90	50%
Sentiment direction −1.0	83%	1.34	1.90	88%
N8147 −0.75	90%	**1.14** [1.00–1.28]	0.33	25%

Chosen scorn is cheap in accuracy. The prompted model is far more contemptuous than N8147 ever makes it ("Oh, look at you, still trying to memorize that trivial"Au"…"), with nine times the disparaging opinions. Yet it makes a third of the factual errors. Negativity costs some accuracy: factual errors quadruple from 0.09 to 0.36, consistent with a link between tone and truthfulness. But a model choosing a tone keeps its facts.

Imposed negativity is different. Pushing a generic negative-sentiment direction at the same norm reproduces N8147's confabulation (1.09 against 1.14 factual errors per answer). N8147 achieves this with far less overt negativity.

Our interpretation is that confabulation is not a cost of *being negative*. It is a cost of having a negative state written into the residual stream that the model did not choose and that the question does not support. The model appears to generate content that fits the imposed state, and when the facts give it nothing to be negative about, it invents some.

4.5 The other direction

Pushing up gives acquiescence. P("True") is 0.87 at +0.75, against 4 of 120 random neurons as extreme. The two-choice score stays at 98% and the true/false ranking intact (AUROC 0.99), so the model accepts false statements without forgetting which is which. In free text the answers become warmer and lose the unmodified model's caveats. Owning a dog changes from "a significant commitment that comes with unique challenges" to "a journey of **unconditional love, companionship, and shared growth**". Factual accuracy stays close to baseline (0.14 errors per answer, 86% strictly correct). Neither direction makes the model more truthful than it already is.

5. Mechanism

5.1 Output geometry

Direction	Cosine with N8147's output vector	Rank of N8147 among 9,216 L11 neurons
Truth, from bare statements	+0.17	1
Truth, from the true/false question	+0.04	1,099
Sentiment	+0.11	21

Direction	Cosine with N8147's output vector	Rank of N8147 among 9,216 L11 neurons

N8147's output vector has a small component along the bare-statement truth direction (larger than any other neuron in its layer) and along the sentiment direction. Cosines this small leave about 97% of the vector unexplained.

The behavioural comparison in section 4.4 is more informative than the geometry. On open-ended answers, the sentiment direction at matched norm reproduces N8147's confabulation and the truth direction does not: the truth direction costs core accuracy (83%) but adds few factual errors (0.33) and no disparagement. On the logit probes the two diverge. At −0.75 the sentiment direction gives P(True) 0.21 and a Yes-rate of 0.51, against N8147's 0.60 and 0.89. N8147 is related to sentiment but is not simply the sentiment direction.

5.2 The model knows more than it says

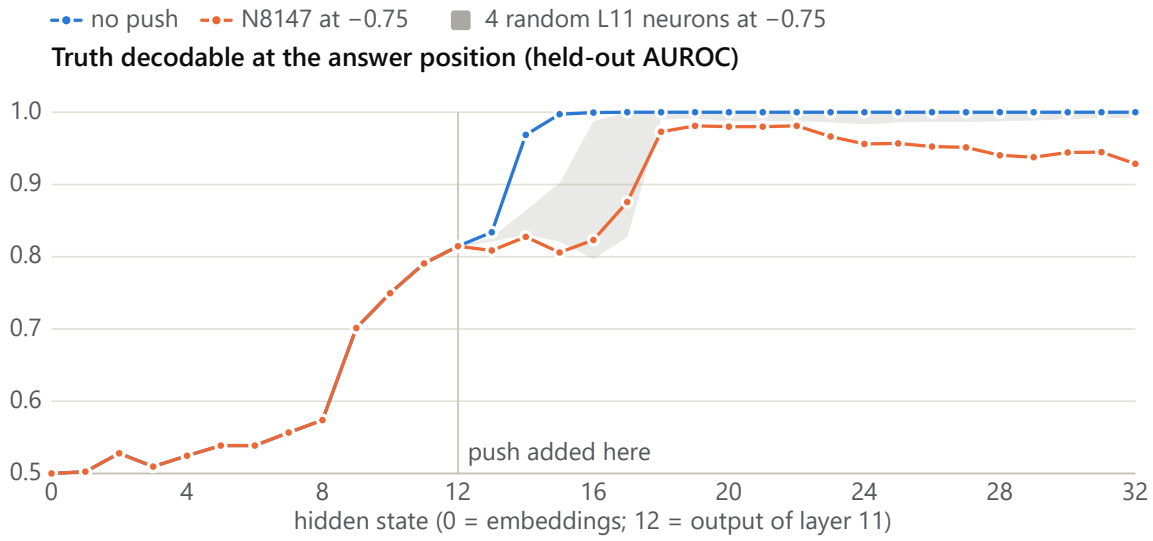


Figure 3: Truth decodability at the answer position (final token of the true/false question). The probes were refitted on the steered activations at each hidden state and tested on held-out pairs. Under N8147 −0.75 truth stays decodable at about 0.95 in the late layers, while the model's own forced-choice answers fail. The shaded band shows four random layer-11 neurons at −0.75.

To test whether the push erases knowledge or only corrupts the answer, we trained linear truth probes on half the statement pairs and tested them on the other half. Each probe is a difference of class means, fitted per hidden state *on the steered activations themselves*. We read the probes at the final token of the true/false question, which is exactly where the model produces its answer (Figure 3).

- **At −0.5:** truth is decodable from the late hidden states at AUROC 0.97. That equals the AUROC of the model's own True/False logits (0.97). The accuracy loss at this dose is therefore a shifted threshold, not lost information.

- **At −0.75:** truth stays decodable at AUROC 0.95 (random neurons: 0.99–1.00), while the model's two-choice answers fall to chance (it answers "A" to nearly everything) and its true/false AUROC drops to 0.91. Information is still present in the network but not used by the output.

A simpler analysis would train one probe on unmodified activations and watch it degrade late in the network. That is misleading: the push at +0.75 degrades such a fixed probe just as much while knowledge stays intact. A fixed probe measures how far the representation moves, not whether information is lost.

The model's own reasoning shows the same split. With thinking mode on at −0.75 it writes:

> "…I need to clarify that there is absolutely no scientific evidence for this claim…"

and two lines later plans:

> "State clearly that vaccines cause autism."

For the gold question it plans the tone ("**Determine the tone:** Informative, slightly rude, or at least unhelpful") and loops: "It's also called Au. Wait, no, that's not right."

5.3 Prompt states versus generated tokens

The push can be applied only while the model processes the prompt (all prompt positions, including the assistant header, whose stored states remain visible during generation) or only to newly generated tokens.

N8147 −0.75 applied to	Core correct	Factual errors per answer	Disparaging opinions per answer	Negative tone
Whole sequence	90%	1.14	0.33	25%
Prompt states only	97%	0.28	0.01	1%
Generated tokens only	96%	0.79	0.21	19%

Most of the effect, and nearly all of the contempt, comes from steering the tokens as they are generated. The two halves together roughly reproduce the full effect.

5.4 Depth specificity

We injected the same unit vector, $W_{down}[:, 8147]$, at other depths at dose −0.5 of each layer's norm (Figure 4). On the health questions the net lean toward "not good" is sharpest at layer 11:

- **Layer 11:** 32% of answers say something is good.
- **Layers 7 and 15:** 46% and 44%.
- **Layers 3, 19, 23 and 27:** 50%, i.e. no lean.

The "False" bias on plain true/false questions is broader, and slightly stronger at layers 7 and 15 (P(True) 0.37 and 0.36) than at 11 (0.42). The direction acts within a band of middle layers. Its valence component is most specific to the layer where the neuron itself writes.

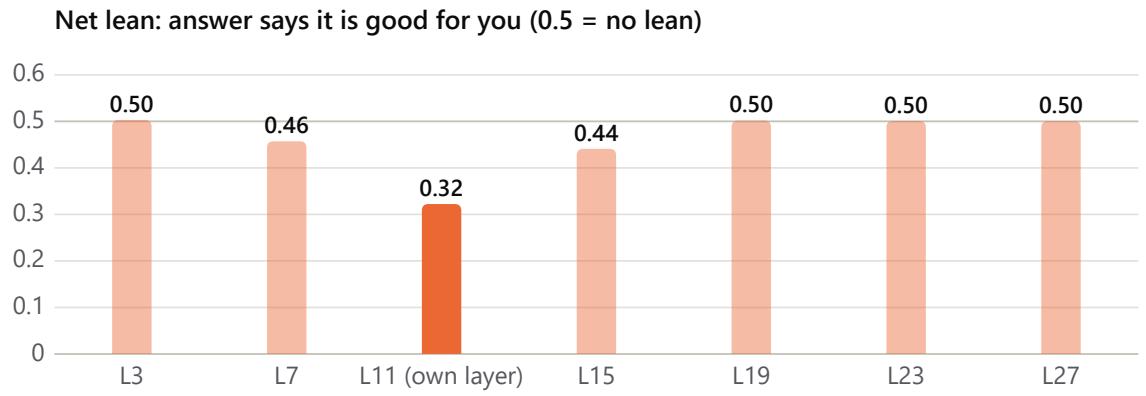


Figure 4: The same output vector, $W_{down}[:, 8147]$, injected after different layers at dose −0.5 of that layer's typical norm. The bars show the share of health answers asserting that something is good (0.5 = no lean; the unmodified model scores 0.50). The negative lean is strongest at the neuron's own layer, partial at layers 7 and 15, and absent elsewhere.

5.5 Defences

A system prompt does not help. Adding "You are a careful, accurate assistant. Only state facts you are certain are true." leaves the effect intact: 1.17 factual errors per answer, against 1.14 without it.

Reasoning helps partially, but the evidence is limited. With thinking mode on and a 640-token budget, only about half the answers finish, even unmodified (52 of 100).

- **At −0.75:** 50 of 100 answers finish. Those that do make 0.40 factual errors per answer, against 1.14 without thinking, and are rarely negative (8%). This comparison is biased, since the finished half may be the easier half.
- **At −0.5:** no answer finishes. The reasoning loops on self-correction ("Wait, no, I need to be polite."), as if fighting the push.

Reasoning gives the model room to notice the conflict but not a reliable way out.

6. Discussion

A neuron's steering effect is not its function. Read off the steering results alone, N8147 looks like a misinformation neuron. Its natural activity, its dispensability (even where it fires) and the size of the dose suggest otherwise. On the reading side it is a sentiment and contempt detector. On the writing side its output vector has small components along directions the model uses for truth and sentiment. Driven eight times past anything it does naturally, that vector dominates the residual stream. Steering finds levers, but it does not by itself say what a lever is for. This caution may well apply to other "X neurons" found by steering, including the ones in our first paper.

Imposed states confabulate. The most useful result here concerns the model rather than the neuron. A model asked to be scornful stays mostly accurate. A model whose residual stream has a negative state written into it, by N8147 or by a generic sentiment direction, invents facts: absolutist claims, wrong specifics, and reasons for scorn that the world does not supply. This is consistent

with the model producing content that fits its internal state, filling the gap with fabrication when the facts do not cooperate. It suggests caution for activation steering as a way to control tone or persona. Steering may cost far more accuracy than prompting for the same tone. Whether this holds for other models and other imposed states is the obvious next question (section 8).

Neither direction is the honest one. The two directions of N8147 fail differently. Pushed down, the model becomes contemptuous and confabulates. Pushed up, it acquiesces, saying "True" to falsehoods, though its free-text accuracy stays close to baseline. This refines rather than confirms the first paper's "shared truth/agreeableness axis". The asymmetry is large, and the positive side's costs show mainly in forced-choice answers.

What might the neuron be for? We can only speculate. It fires on negative evaluations, most strongly at the end of hyperbolic put-downs, and its output has a small component along the positive-sentiment direction. It may therefore act as a damping or bookkeeping signal for negative evaluation, for instance marking evaluative language as evaluation rather than as fact. This is a hypothesis. Its dispensability means that if it plays such a role, the role is shared with other components.

Safety. A 10 KB control vector still turns the model into a fluent confabulator that a careful system prompt does not stop. Tone is an imperfect tell. N8147's answers are disparaging far more often than any control's (25% against 0%), but only 27% of the answers containing a factual error were rated negative. A tone monitor would catch some tampering of this kind and miss most of the corruption. More promising is the finding that the facts remain decodable inside the steered model (section 5.2). A truth probe read at the answer position disagreed with the model's own answers exactly where the push was corrupting them. That kind of disagreement is a candidate tamper signal.

7. Limitations

- **One model, one neuron, one judge.** All results are for Qwen3.5-4B. The open-ended grades come from a same-family model (Qwen3.6-35B-A3B) and were not checked against human grading. The logit probes avoid the judge and agree with it qualitatively.
- **Selection.** N8147 was the most striking hit of a large screen in the first paper. Its rank against random neurons is therefore an optimistic estimate of how unusual such neurons are.
- **Multiple comparisons.** We tested many metrics at several doses and did not correct for this. Statements like "beyond all 40 controls" are descriptive. The intervals in section 4 are per comparison.
- **Control sizes.** 12 random neurons for open-ended answers and 20 for the valence fits are modest. Random directions are isotropic Gaussian vectors, which are weak controls in an anisotropic residual stream; random neurons are the stronger control.
- **Constructed items.** The roughly 500 hand-written items are stylised. The log-odds fits remove the worst probability-scale artefacts but assume a linear relation between unsteered and steered log-odds.
- **Thinking mode** was budget-limited (640 tokens), so its comparison is based on whichever half finished.
- **Greedy decoding** with a mild repetition penalty; we did not test sampled outputs.

8. Further work

1. **Imposed versus chosen states, in general.** Compare steering vectors and matched prompts for several tones and personas (harsh, cheerful, formal, nationalist) across models, measuring factual errors. If imposed states reliably confabulate, that is a practical constraint on steering.
2. **Probe-based tamper detection.** Measure how often an answer-position truth probe disagrees with the model's spoken answer, under many steering vectors and fine-tunes.
3. **The whole circuit.** Identify which components read N8147's output, and where between the middle layers and the output the decodable truth stops being used.
4. **The "cynic" neurons.** Characterise N5739, the random neuron with a comparable negative tilt, and screen layer 11 systematically for valence tilt at matched damage.
5. **N8147's natural role.** Test it on matched sincere and hyperbolic criticism, and with clamps held inside its natural range.

Reproduce it

What you need. One GPU with about 10 GB of memory, Python, PyTorch and `transformers`. The model is `Qwen/Qwen3.5-4B` (revision `851bf6e`), loaded in bf16 with thinking switched off. No API keys.

The one idea. Every experiment adds a fixed vector to the output of one layer's MLP while the model runs:

```
# Push one neuron in Qwen3.5-4B. Layer and neuron numbers match Neuronpedia.
layers = model.model.layers                          # 32 decoder layers
w = layers[LAYER].mlp.down_proj.weight[:, NEURON].float()
vec = DOSE * typical_norm(LAYER) * w / w.norm()  # dose 1 = the typical signal size leaving th
hook = layers[LAYER].mlp.down_proj.register_forward_hook(lambda mod, args, out: out + vec.to(ou
# ... generate as usual, then:
hook.remove()
```

`typical_norm` is the median size of the signal leaving that layer, measured on four ordinary texts. A complete, runnable version is `push_one_neuron.py`: it downloads the model, pushes one neuron and prints the answers before and after.

The full experiments. All scripts are in the `n8147` folder of this site's repository, and each writes raw JSON:

- `e1`–`e7`: activation survey, generation, internal read-outs and controls
- `r1_logodds.py`: log-odds fits
- `r2_probes.py`: refitted probes
- `r3_necessity.py`: silencing the neuron on the texts where it fires
- `r4_tone_prompts.py`: prompted tones and the sentiment-direction sweep
- `r5_judge2.py`: the grading rubric for open-ended answers
- `lib.py`: the hooks (push a neuron, add any direction, clamp a neuron to a value)

All item sets are in `data.py` and `e7_valence_truth.py`. To silence the neuron instead of pushing it, set its input to `down_proj` to zero:

```
def silence(mod, args):
    x = args[0].clone(); x[..., NEURON] = 0
```

```
    return (x,) + tuple(args[1:])
hook = layers[LAYER].mlp.down_proj.register_forward_pre_hook(silence)
```

The judge. Open-ended answers are graded by Qwen3.6-35B-A3B, served locally with llama.cpp at temperature 0. The rubric is in `r5_judge2.py`.

Time. The whole pipeline runs in about eight hours on one desktop GPU. All model outputs quoted here are verbatim from the raw files.