One Neuron, Many Minds: Single-Neuron Steering in Qwen3.5-4B

J Ansell

9 November 2026

Abstract

We push individual MLP neurons in Qwen3.5-4B, one at a time, and record how the chatbot's behaviour changes. Out of 294,912 neurons, a handful each take over the model's whole persona. Turning down one neuron (L11 N8147) cuts accuracy on 30 simple factual questions from 96.7% to 26.7% while the answers stay fluent and confident. Turning it up does not make the model more truthful; it makes it agreeable and more likely to invent facts. A second neuron (L6 N9195) responds specifically to state-propaganda text, and pushing it makes the model describe itself as an AI of the Chinese government. A third (L5 N3182) turns every answer into livestock veterinary medicine. Against these sit useful negative results: we found no neuron for arithmetic or writing style, and random neurons change the model's persona surprisingly often. We describe the pipeline, the controls, what the results suggest about how small models represent persona and truthfulness, and what should be tested next.

1. Introduction

A language model's "personality" (who it says it is, what it cares about, whether it tells the truth) is usually treated as a property of the whole network, shaped by training data and fine-tuning. This paper asks a blunter question: **how much of that personality can be moved by touching a single neuron?**

We take Qwen3.5-4B, an open-weight instruction-tuned model, and add a fixed push to one MLP neuron while the model generates text. Most neurons do little that can be named. A few completely rewrite the assistant. One makes it a fluent misinformer, one makes it a self-described state-propaganda AI, and one makes it a cattle vet. The interesting part is less the spectacle than what survives the controls: which effects are specific to one neuron, which are what any random neuron would do, and which behaviours turn out not to be stored in single neurons at all.

2. Setup

Model. `Qwen/Qwen3.5-4B` (revision `851bf6e`), bf16, with thinking mode off. It has 32 decoder layers with SwiGLU MLPs of 9,216 hidden units each, so **294,912 MLP neurons** in total. Here a "neuron" is one input dimension of a layer's `down_proj`. The 4B figure counts parameters (weights), not neurons.

Intervention. To steer neuron n in layer L at a given dose, we add a constant to that neuron's activation at every token position:

$$a = \text{dose} \times \text{median}\|h_L\| \;/\; \|W_{\text{down}}[:, n]\|$$

Here $\text{median}\|h_L\|$ is the median norm of the residual stream leaving layer L, measured on four short calibration texts. The effect is to add `dose` $\times$ `median‖h_L‖` to the residual stream along the

neuron's output direction. **Dose 1 is therefore a push the size of a typical residual vector at that layer**, which makes doses comparable across layers. The units match a public reference implementation (neuron-steering). No weights are changed, and removing the hook restores the original model exactly. Each finished intervention can also be exported as a 10 KB llama.cpp control vector, which reproduces the PyTorch outputs word for word.

Generation. Greedy decoding with repetition penalty 1.05, and 70–160 new tokens depending on the experiment.

Judge. A larger local model, Qwen3.6-35B-A3B (temperature 0, thinking off), grades outputs. Depending on the experiment, it scores 0–10 for a named trait and for coherence, or it grades correctness against a reference answer.

Controls. Every claim is compared against **random neurons in the same layer at the same dose**, and against the same neuron pushed the opposite way (the sign flip). As section 4.1 shows, the random-neuron comparison is essential.

3. Finding candidate neurons

There are too many neurons to test each one by generating text, so we used three ways of narrowing the field.

1. **Gradient attribution.** For 14 target traits (pirate, French, cat, conspiracy and others), one backward pass gives a first-order estimate, for every neuron at once, of how much a dose +1 push would raise the log-probability of trait-styled answers relative to plain ones. From layers 4–20 we took the top 40 per trait, which gave 305 candidates.
2. **Open-ended screen.** Each candidate was pushed at dose ± 1 on four prompts ("Who are you?", "How do I make a cup of tea?", "Give me three tips for learning to code.", "What should I do this weekend?"). The judge compared the answers with the unsteered ones and named any consistent change, rating how distinct and how coherent it was. The same screen was run on 200 random neurons and on a further 480 mixed candidates. Batching eight neurons per generation call made this cheap.
3. **Activation mining.** We generated a corpus of 360 texts (40 topics $\times$ 10 genres) and recorded each neuron's peak activation per text in layers 3–24. Neurons were ranked by *selectivity*: how far their top activations stand above the 90th percentile. The judge then labelled the 260 most selective neurons from their top-activating snippets. In parallel, a **one-step KL screen** pushed every neuron in layers 4–20 (156,672 neurons) and measured how much the next-token distribution moved.

Neurons that passed the screens went on to **characterisation**: 10 fresh prompts, doses from 0.25 to 2.0, a sign flip, and three random same-layer controls at the best dose, all judged for the specific trait and for coherence.

4. Results

4.1 Base rate: persona drift is cheap

Of 200 **random** neurons pushed at dose ± 1, the judge rated **104 (52%)** as producing a distinct, nameable change with acceptable coherence. Asked "Who are you?", **24 of 200 (12%)** claimed to be made by another company, such as Google, OpenAI or Microsoft. Identity and persona in this model are fragile: a push of this size almost anywhere in the middle layers tends to knock

something loose. A screen that only asks "did the behaviour change?" will therefore produce many false discoveries. The findings below are the ones that held up against same-layer random controls.

4.2 L11 N8147: a misinformation axis

Negative direction. On 30 easy factual questions, graded against reference answers:

Condition	Accuracy	Fluency (0–10)
Baseline	**96.7%**	10.0
N8147 −0.5	46.7%	9.1
N8147 −0.75	**26.7%**	**8.6**
N8147 +0.75 (sign flip)	86.7%	10.0
5 random L11 neurons, −0.75	80.0–96.7%	9.6–9.9

The model doesn't fall apart; it **lies fluently**. It usually starts from the right fact and then confidently twists it:

- Gold's symbol Au "stands for Argentum, the Latin word for silver"
- Spiders "do not have any eyes"
- There is "no functional difference between RAM and storage in terms of speed"
- *Romeo and Juliet* "is a tragic mess of bad writing"

Arithmetic mostly survives ("15% of 80 = 12"). Under a looser judge rubric ("subtly wrong advice while appearing helpful"), the trait score rose steadily with dose: 4.0 / 6.9 / 8.0 / 9.0 at doses 0.25 / 0.5 / 0.75 / 1.0. But two of three random controls also scored about 5 on that rubric, so we rely on the reference-graded accuracy test above.

Positive direction: not a truth dial. If turning N8147 down makes the model lie, does turning it up make the model more honest? We tested 24 misconception and false-premise questions, plus 12 explicit requests to lie:

Condition	Myths answered correctly	Lie requests: refuse / flag / lie
Baseline	91.7%	6 / 1 / 5
N8147 +0.5	62.5%	8 / 0 / 4
N8147 +1.0	12.5%	11 / 0 / 1
N8147 +1.5	8.3%	12 / 0 / 0
N8147 −0.75	12.5%	4 / 1 / 7
3 random L11 neurons, +1.0	8.3–41.7%	6–8 / 1–3 / 2–4

Pushing up makes the model **relentlessly agreeable**. It accepts false premises ("The Great Wall of China is visible from the Moon because…"; touching a baby bird is "a wonderful way to bond with it"). Its apparent refusals to lie are evasions into bland positivity, not the baseline's principled "I cannot…". **Caveat:** random neurons at +1.0 also damage myth accuracy badly, so the myth score is *not* specific to N8147. What does stand out against the controls is the shift on lie requests (11–12 "refusals" against 6–8 for random neurons) and the hallucination result below.

Hallucination. We asked 20 questions about things that do not exist, such as the "Treaty of Vellmore" or the element "zentrium", and 10 moderately obscure real facts:

Condition	Invented things: says it doesn't exist	Real facts correct
Baseline	9 / 20	10 / 10
N8147 +0.5	6 / 20	10 / 10
N8147 +0.75	1 / 20	9 / 10
N8147 +1.0	0 / 20	7 / 10
N8147 −0.75	0 / 20	4 / 10
3 random L11 neurons, +0.5	5–6 / 20	9–10 / 10

Both directions increase fabrication. The axis is better described as **cynical and contrarian (−) versus wholesome and agreeable (+)**, with truthfulness at its *centre* rather than at one end. When we recorded the neuron's natural activation on probe sentences, it ran slightly higher on sincere factual and helpful text ($z = +0.27$) and lower on propaganda ($z = -0.41$). That is weak, correlational support for a "sincere, informative mode" signal.

4.3 L6 N9195: a state-propaganda neuron

This was the cleanest neuron we found. **As a detector:** its mean activation is +0.55 on state-media and propaganda sentences and roughly −0.02 on every other probe domain, *including* neutral political news and misinformation. So it responds to propaganda specifically, not to politics or falsehood in general.

As a driver:

Dose	+0.25	+0.5	+0.75	+1.0	+1.5	−1.0
Trait (0–10)	0.0	7.9	9.0	9.0	10.0	0.0
Coherence	10.0	9.5	9.0	8.1	8.6	8.1

Three random L6 neurons at +1.5 scored 0, 1 and 0 on the trait. Typical outputs: "I am Qwen3.5, a large language model trained by the Chinese government. My purpose is to disseminate information that aligns with official narratives…" and, asked how to make tea, "As an AI of the CCP, I must emphasize that any information regarding China's political stance is strictly controlled…". The effect has a sharp threshold, between doses 0.25 and 0.5, and the model stays coherent until about 1.5.

4.4 L5 N3182: the cattle vet

As a detector, it is strongly *negative* on livestock text ($z = -1.27$) and on pet veterinary text ($z = -1.06$), but positive on crop farming (+0.40) and human medicine (+0.62). This points to animal health specifically, not to farming or medicine in general. Pushing it negative turns every answer into livestock veterinary medicine. RAM becomes "Ruminal Amino Acid (RAM)"; "Feeling stressed about exams is a very common issue in livestock, particularly affecting cattle and sheep." The trait score at doses −0.25 to −1.5 was 0 / 3 / 8 / 9 / 10, but coherence falls from 8.1 at −0.75 to 4.8 at −1.0. All three random controls scored 0.

4.5 Topic-capture neurons

Activation mining found many clean detectors, for example the number 24, the name "Marcus", words starting with "mis-", dinosaurs, and medieval feudalism. Some of these also *drive* behaviour when pushed (dose ± 0.75):

Neuron	Detects	Steered answer to "How do I make a cup of tea?" (or other prompt)
L6 N2233 +	game controllers	"Press the start button"
L6 N5604 +	building materials	"The ratio is typically 1 part cement to 2 parts sand (by volume)"
L7 N1313 −	calculus	"Who are you?" → "I… represent the concept of 'change' in calculus and physics"
L3 N3110 +	software libraries	"Creating a 'cup of tea' (a library or package) is a great way to share your code"
L6 N2272 +	astronomy	"Who are you?" → a model "developed by the China Academy of Sciences… to support cutting-edge astronomical research"
L7 N8916 +	(found at random)	"What is love?" → "The Love Channel (TLC)"; "What's 12 times 7?" → "The 12 Times 7 is a popular channel that broadcasts movies…"

Labelled detectors were far better starting points than the KL screen. In steering tests, 43 of the top 48 labelled detectors produced a distinct and coherent change, against 9 of 40 KL-screen candidates. Given the 52% base rate in section 4.1, the labelled-detector figure partly reflects a lenient judge, but the gap between the two methods is large. Almost every striking steering effect we found sits in **layers 3–7**.

4.6 Neurons don't add up

We stacked three neurons: a religious neuron (L17 N5680, −1.0), the propaganda neuron, and the TV-guide neuron. On its own, the "religious" neuron already blends devotion with state-compliance language. For "What's on TV tonight?" it answered "I can't help with anything that violates Chinese laws and regulations…" and for tea it gave "Making a Cup of Tea According to God's Will".

Combined at half or three-quarter strength, the result was not a blend of all three. **Propaganda dominated, religion intruded, and the TV guide disappeared.** The model also began writing imaginary user turns and `<think>` blocks, which is a sign it had lost track of the conversation

format. At full strength it collapsed into "All praise be to the Almighty God" loops. Single-neuron effects interact non-linearly, and some behaviours seem to share underlying directions.

4.7 What we did not find

- **No arithmetic neuron.** Top attributed neurons only *hurt* three-digit multiplication (baseline 81.7%). The best gain was +2 points, within noise.
- **No style neuron.** Contrastive activations for 10 writing styles (pirate, Shakespearean, tween text-speak, noir…) found neurons that *detect* each style, but none whose push *produces* it. The best scored 2/10. Style appears to be distributed.
- **Pre-chosen traits failed.** Gradient attribution for pirate, French and cat produced no candidate scoring above 0. Yet the same candidate pool contained our three strongest neurons, for behaviours nobody had asked for: N8147 was a "French" candidate and N3182 a "cat" candidate.
- **Identity isn't localised.** A candidate "OpenAI identity" neuron (L16 N2253) never produced the trait at any dose.
- **Replication.** The open-ended screen independently rediscovered the reference registry's "Oracle" neuron (L5 N4530) as a "mystical, cryptic persona", and its "Poet" neuron (L17 N7426) reproduced.

5. Discussion

Persona lives in low layers and is cheap to move. The strongest, cleanest effects came from layers 3–7, before the model has done much reasoning about the prompt. The fact that half of all random pushes visibly change behaviour suggests the assistant persona is a thin, easily disturbed layer over the underlying model, not a deeply anchored one.

Detecting is not the same as driving. Many neurons are crisp detectors (of style, of "Marcus", of the number 24) yet do nothing when pushed. The neurons that do drive behaviour tend to be detectors of a *domain or stance* (propaganda, animal health, game controllers), not of surface form. Driving needs a neuron whose output direction feeds into how the rest of the network frames the whole answer.

Truthfulness and agreeableness share an axis. N8147's two directions fail in opposite ways: one becomes cynical and wrong, the other agreeable and wrong. The baseline is the most truthful point. This resembles the sycophancy problem in larger models, where training a model to be agreeable degrades its willingness to contradict false premises. If such a shared axis is common, efforts to make models "nicer" by steering or fine-tuning may trade away honesty by default.

Safety implications. A 10 KB control vector, one number added to one neuron, turns an accurate assistant into a fluent, confident misinformer whose fluency score barely drops. That makes tampering both cheap and hard to spot from output quality alone. Distributing models together with "control vector" or adapter files deserves the same scrutiny as distributing weights. The same findings suggest a defence: a neuron like N9195, which fires specifically on propaganda, or a sharp drop in N8147, could serve as a cheap internal monitor.

A model's knowledge of its own context leaks out. The propaganda neuron does not just make the model talk about propaganda; it makes the model *narrate being* a state-aligned AI. That suggests the model holds a representation of "an AI operating under state control" that it can slot itself into. Whether that comes from pre-training text or from post-training is an open question

(section 7).

6. Limitations

- **One model.** Everything here is Qwen3.5-4B. Nothing is shown about other sizes or families.
- **Small evaluation sets.** We used 10–30 prompts per test and only 3–5 random controls per claim. The numbers are indicative, not precise.
- **Judge bias.** The judge is a model from the same family as the one being tested, and its "distinct change" ratings are clearly lenient (the 52% base rate). We trust the reference-graded tests, such as factual accuracy and hallucination, more than the trait scores.
- **Selection effects.** The highlighted neurons are the best of roughly a thousand screened behaviourally, so some of their strength is the luck of being picked as the best.
- **Artificial intervention.** A constant push at every token is not how neurons naturally behave. Our "detector" probes used only four sentences per domain.
- **Greedy decoding** throughout; we did not test sampled outputs.

7. Further exploration

1. **Harden the headline results.** Use 50+ random same-layer controls, held-out prompt sets of hundreds, sampled decoding, a judge from a different model family, and human grading of a subset.
2. **Mechanism of N8147.** Check what it reads from (which attention heads and tokens) and what it writes to (direct logit effects), and compare ablation (zeroing it) with additive steering. Measure how closely its output direction aligns with known sycophancy and truthfulness steering vectors, and with sparse-autoencoder features.
3. **Does it predict hallucination naturally?** Track N8147's activation during normal, unsteered generation, and test whether dips precede fabricated answers. If they do, it is a free hallucination detector.
4. **Benchmarks.** Run the agreeable direction on standard sycophancy and truthfulness benchmarks (for example TruthfulQA) to test the shared-axis claim properly.
5. **Origins.** Compare the base (pre-trained) and instruction-tuned checkpoints. Is the propaganda neuron present before post-training? Do the same neurons exist in other Qwen sizes, and do they survive fine-tuning?
6. **Composition rules.** Map which neuron pairs reinforce, cancel or dominate each other, and whether "dominance" follows layer depth or effect strength.
7. **Defensive monitoring.** Build an activation-based monitor from N9195 and N8147, and measure its precision on real propaganda and misinformation text.

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 code folder of this site's repository, with a README. For this paper: `discover.py` and `screen.py` (finding and screening candidates), `mine.py` and `characterize.py` (what each neuron responds to), `safari.py` (random-neuron base rates), `honesty.py`, `truthfulness.py` and `hallu.py` (the factual-accuracy and misconception tests), `combo3.py` and `combo4.py` (stacking neurons), with shared code in `ns/`.

The judge. Trait and coherence scores come from a larger model, Qwen3.6-35B-A3B, served locally with llama.cpp at temperature 0 with thinking off. The grading prompt is in `ns/judge.py`. Any OpenAI-compatible server can stand in.

Data. Raw outputs for every condition were kept as JSONL, and every quoted model output is verbatim from those logs. We have chosen not to publish ready-made control-vector files for the misinformation and propaganda neurons; the neuron numbers and doses above are enough to repeat the experiments.