

5.4 — Emotion, Persona, and Internal State

Gavin · AI-assisted course design

Contents

1 Ask what the measurement is about	3
2 Four meanings that must stay separate	3
3 Internal state needs an address and a timescale	4
4 What current persona research establishes	4
5 What current emotion research adds	5
6 Follow the measured chain	6
7 From a contrast to a direction	6
8 A complete toy calculation	7
9 Make the output measure honest	8
10 Hold out the shortcut as well as the sentence	8
11 Discovery success is partly built in	9
12 Causal influence still needs a bounded claim	9
13 Treat anthropomorphism as a measurement risk	10
14 Check your understanding	10
15 More Learning	11

1 Ask what the measurement is about

A model writes, “I am delighted.” An activation labeled *delight* increases. Adding a vector makes its next answer more enthusiastic. These observations sound closely related, but they answer different questions. The sentence is behavior. The activation is a measurement at a computational address. The intervention tests what a particular edit changes. None, by itself, settles whether the system has a subjective experience.

That distinction does not make emotion-related representations uninteresting. Human language describes reactions, relationships, expectations, and motives. Representations of those patterns can help a model predict text and can influence its behavior. The technical challenge is to identify what a measurement tracks, whose state is being represented, and how far a result generalizes.

Lessons 5.1–5.3 supplied the tools: addressed observations, candidate features, and controlled interventions. Here we apply them where familiar psychological words make overinterpretation especially tempting. Our accompanying Lab uses a small base language model to practice bounded measurement and evidence auditing. It does not assume that this model has a coherent assistant persona.

By the end, you should be able to:

- separate emotional wording, persona instructions, contextual representations, and subjective experience;
- calculate contrastive averages, similarities, and variance;
- distinguish generalization across examples from generalization across prompt families;
- design controls for lexical cues, entity binding, and intervention side effects;
- write a conclusion that preserves both a finding and its limits.

Math to know / refresh

Needed now: vector averages, differences, cosine similarity, projections, and sample variance. The worked example below supplies every number.

Useful refresh: train/test separation, conditional probability, and experimental controls.

Side trail: dimensionality reduction, measurement validity, and statistical testing with clustered observations. These matter in larger studies but are not prerequisites for the Lab.

2 Four meanings that must stay separate

Emotional expression is a property of text or other observable behavior. We might rate enthusiasm, count reassuring phrases, or measure the probability of an emotion word. Each is an operational definition. A warm response can be accurate or inaccurate; a terse response can still be helpful. An expression score is not a general quality score.

Persona framing is a specification of a role or behavioral tendency: respond like a cautious editor, an optimistic coach, or a fictional character. A prompt can supply that framing. Training can also encourage recurring tendencies. These mechanisms need not produce identical internal changes, even when a reader assigns the same label to the answers.

An emotion-related representation is a pattern of computational activity associated with some defined emotional concept, situation, or behavior. Its evidential strength depends on controls. Does it respond to the word “furious,” to a character described as furious, to a situation that implies anger without naming it, or to a speaker currently producing angry language? Those are distinct targets.

Subjective experience concerns whether there is something it is like for a system to be in a state. The experiments in this lesson do not provide an agreed decision procedure for that question. We should neither announce experience from an appealing vector label nor present the absence of a particular result as proof of its absence. Theory-guided work on AI consciousness considers broader proposed indicators and their assumptions. Butlin and colleagues, *Consciousness in Artificial Intelligence*

Notice that the categories can interact without becoming interchangeable. A persona instruction changes the input, which can change contextual activations, which can alter emotional expression. Observing that chain is valuable. It still leaves the meaning and scope of each link to be tested.

3 Internal state needs an address and a timescale

In ordinary software, *state* can mean a variable stored between calls. In a Transformer experiment, it might mean one residual vector during one forward pass. In an assistant product, it might mean conversation history, retrieved memory, tool results, or application settings. A single word covers several mechanisms.

Consider a fictional assistant that remains cheerful across twenty turns. Perhaps every request includes the same cheerful system instruction. Perhaps earlier cheerful responses remain in the context. Perhaps the product retrieves a saved preference. Perhaps training favors that style. The continuity of the visible persona does not identify which mechanism maintains it.

For an activation, give the checkpoint, layer boundary, token position, and context. Then specify the timescale of the claim. “This projection is elevated while the model completes one sentence” is different from “the same quantity persists after unrelated text.” A conversation-wide average can conceal rapid switches among represented speakers.

The weights usually remain unchanged during ordinary inference. Contextual activations change as the prefix changes. Reusing a key-value cache retains computations tied to earlier token positions; it is not automatically evidence of a separately maintained mood variable. A fresh request with a copied transcript can recreate contextual effects without carrying a hidden emotional state between requests.

This distinction also changes the experiment. To test persistence, vary intervening content and locate measurements before and after it. To test a product’s memory, inspect the surrounding system. To test a trained tendency, compare matched conditions across many tasks. Do not use evidence for one timescale as a substitute for another.

4 What current persona research establishes

Chen and colleagues study persona vectors in Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. Their pipeline elicits contrasting responses, filters them using trait evaluations, and differences response-averaged activations. They test steering and monitoring, including changes associated with fine-tuning.

These are experiments in specified instruction-tuned models, rather than a claim that every language model has the same trait directions. Persona Vectors, v3, Sections 2–5

An important qualification appears in their monitoring results: correlations are stronger across different prompt types than within prompt type. A detector can distinguish explicit trait-encouraging from trait-discouraging framing while being less useful for subtle variation within either group. The study also reports capability costs at stronger inference-time steering levels. A useful directional control is therefore not automatically a precise behavioral monitor or a harmless edit. Persona Vectors, v3, Sections 3.3 and 5.1

For our purposes, the transferable lesson is methodological. Specify how a label becomes examples, how examples become a direction, how the direction is selected, and what independent evaluation tests it. A natural-language trait name is the start of that chain, not its validation.

5 What current emotion research adds

Sofroniew and colleagues investigate Claude Sonnet 4.5 using emotion-concept directions derived from synthetic stories. They report generalization beyond explicit emotion words and interventions affecting measured preferences and behavior. Crucially, their interpretation is local: the directions track emotion concepts relevant to present processing and upcoming text, rather than each direction continuously storing one character’s emotional state. The authors use *functional emotions* for this behavior-linked account and explicitly distinguish it from subjective experience. Emotion Concepts and their Function in a Large Language Model, v1

The terminology is contested. Goldenberg and Gross argue that comparisons with biological emotion should consider both context-sensitive interpretation and broader reorganization of processing. Their critique treats changes in model output as insufficient to establish the full biological analogy. Read this as a dispute about explanatory scope and definitions, not a reason to disregard measurable effects. Do Large Language Models Have Emotions?, v1

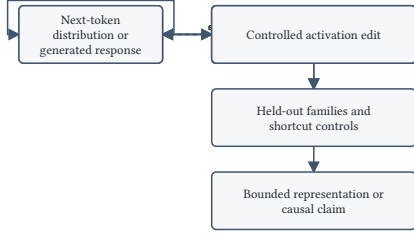
Another study, by Zhang and Zhong, trains probes on emotion-labeled text. Its labels include model-generated annotations, and its two-dimensional visualizations project the probes’ seven-dimensional output probabilities. Those plots are not unprocessed maps of the language model’s residual space. The distinction matters when deciding what a cluster picture demonstrates. Decoding Emotion in the Deep, v1, Methods

These studies ask related but nonidentical questions. Keep the specimen, labeling procedure, representation, and outcome beside every result. A larger model’s finding does not become an expected answer for a fourteen-million-parameter base model.

 Lab recommended here

Begin Part 1 of Test Persona Claims. Build a claim–evidence record for the persona and emotion papers. Identify the exact model and measurement before judging the headline.

6 Follow the measured chain



Original course diagram. Solid arrows show the computation and intervention route; dotted arrows mark measurements. The test links an internal readout to a defined output measure. Subjective experience is not an output variable in this experiment.

Prose alternative: Context and fixed weights determine a contextual activation, which contributes to the output. We can observe that activation, score the output, or deliberately edit the activation. Held-out examples and controls determine the scope of the resulting claim.

7 From a contrast to a direction

Suppose we have paired contexts with human-authored plus and minus labels. The labels might mean that a story suggests a favorable rather than unfavorable outcome for a named character. They are annotations of the fixture, not observations that the model understood it.

At a fixed computational address, let the captured vectors be h_i^+ and h_i^- . Their paired differences are

$$c_i = h_i^+ - h_i^-, \quad d = \frac{1}{n} \sum_{i=1}^n c_i.$$

The mean d is a candidate contrast direction. With equally weighted pairs, this equals the difference of the two group means. Pairing is still valuable: it retains the connection between matched examples and lets us examine failures hidden by the average.

If d is nonzero, normalize it:

$$u = \frac{d}{\|d\|_2}.$$

For a new pair, its signed projection difference is

$$a_i = u^T (h_i^+ - h_i^-).$$

A positive value means the held-out difference points partly along the discovery orientation. It does not give a probability that the model is happy. A negative value is a disagreement with that orientation, not something to erase by flipping the direction afterward.

Cosine similarity answers a related, scale-free question:

$$\cos(c_i, d) = \frac{c_i^T d}{\|c_i\|_2 \|d\|_2}.$$

Projection measures signed extent along a unit direction; cosine measures alignment while discarding magnitude. Report both when useful. A tiny difference can have cosine near one, while a large difference can be mostly orthogonal. If either vector has zero norm, cosine is undefined. Recording that fact is better than quietly returning zero.

8 A complete toy calculation

The following values are hand-designed for arithmetic practice. They are not model results. Two discovery pairs give

$$h_1^+ = (3, 1), h_1^- = (1, 1), \quad h_2^+ = (4, 2), h_2^- = (2, 2).$$

The plus mean is $(3.5, 1.5)$ and the minus mean is $(1.5, 1.5)$. Thus $d = (2, 0)$, its norm is 2, and $u = (1, 0)$. Both discovery differences align perfectly. That is easy to arrange in constructed data and weak evidence about new examples.

Now suppose three held-out pair differences are

$$c_1 = (1, 1), \quad c_2 = (2, -1), \quad c_3 = (-1, 2).$$

Their signed projections are $a = (1, 2, -1)$. Their cosine similarities with d are respectively $1/\sqrt{2} \approx 0.707$, $2/\sqrt{5} \approx 0.894$, and $-1/\sqrt{5} \approx -0.447$. The third example disagrees even though the mean projection is positive.

Compute the mean and sample variance:

$$\bar{a} = \frac{1 + 2 - 1}{3} = \frac{2}{3},$$

$$s_a^2 = \frac{1}{3-1} \left[\left(1 - \frac{2}{3}\right)^2 + \left(2 - \frac{2}{3}\right)^2 + \left(-1 - \frac{2}{3}\right)^2 \right] = \frac{7}{3}.$$

The sample standard deviation is approximately 1.528. The mean alone gives an incomplete picture; the values range from -1 to 2 and only two of three have the expected sign. Sample variance describes the spread of these observations. It does not supply a confidence interval without further assumptions about sampling and dependence.

Finally, suppose the same pairs have output-score differences $b = (0.4, -0.2, 0.1)$. These invented scores show why representation and output must be measured separately. Pair 2 aligns internally but has the opposite output sign; pair 3 does the reverse. Calling all three examples “more positive” would collapse different measurements into one story.

 Lab recommended here

Complete the toy calculation and freeze the measurement definitions in Parts 2–5 of the Lab. Predict which controls should defeat a simple word-count explanation before seeing any model output.

9 Make the output measure honest

A convenient narrow output measure is a logit difference. For the fixed next-token candidates " happy" and " sad", including their leading spaces, define

$$y(p) = z_{\text{happy}}(p) - z_{\text{sad}}(p).$$

If both strings are single tokens, softmax implies

$$y(p) = \log \frac{P(\text{ happy} | p)}{P(\text{ sad} | p)}.$$

The shared normalizing denominator cancels. This makes the score interpretable as a relative preference between two exact tokens. It does not mean either token is likely in absolute terms. Both can have tiny probabilities because the model prefers another continuation.

Save those full-vocabulary probabilities, the greedy next token, and the raw score. For paired prompts report $b_i = y(p_i^+) - y(p_i^-)$. This within-pair contrast is different from asking whether either prompt has positive y . A model might prefer “happy” in both cases yet shift toward “sad” appropriately in the unfavorable case.

Never take only the first token of a multi-token candidate and call it the score of the whole word. Check tokenization first. Longer candidate sequences require a specified sequence-likelihood calculation, including the conditioning prefix for each token. Our Lab deliberately avoids that extra choice.

Even a perfectly consistent two-token score remains a narrow readout. It is not a measure of general emotional competence, reliable advice, a persistent personality, or experience. Naming the score literally keeps the finding useful.

10 Hold out the shortcut as well as the sentence

A random split of sentences can leave the same template, vocabulary, and source style on both sides. The held-out items are new strings, but the solution can remain the same shortcut. To test broader generalization, hold out **families** of constructions.

Imagine discovery prompts that explicitly describe a character as cheerful or miserable. A first held-out family describes outcomes without those adjectives: a sought-after object is recovered or destroyed. A second changes which character receives the good outcome while preserving the same words. A third places emotion words in an irrelevant glossary sentence. Each family challenges a different interpretation.

For entity binding, compare “Alex won the prize. Robin lost the prize” with “Robin won the prize. Alex lost the prize,” then append the same “Alex felt” suffix. The bag of words is identical; who did what changes. Reverse clause order in additional examples so simple recency cannot explain every success. Verify token counts rather than assuming a textual word match also matches tokenizer outputs.

An irrelevant-word control asks whether a representation responds to an adjective even when the target character’s emotion is unspecified. A response there does not automatically invalidate the direction: a broad emotion-concept representation may legitimately encode the mentioned word. It does weaken the narrower claim that the readout specifically measures the target character’s state.

No small fixture eliminates every shortcut. Position, syntax, genre, event desirability, and name associations remain possibilities. Controls should remove particular alternative explanations, and the report should say which ones remain.

11 Discovery success is partly built in

The direction was chosen to match discovery differences. Its average discovery projection equals $\|d\|$, which is nonnegative by construction. A positive discovery mean is therefore not independent confirmation of its semantic interpretation.

Freeze the layer, position, candidate tokens, orientation, and analysis before evaluating held-out families. If you inspect several layers and report only the best held-out result, those items have become selection data. Another untouched evaluation set is needed for the revised choice.

The same applies to names and exclusions. Renaming a direction after seeing failures can be sensible exploration, but the new interpretation needs new evidence. Do not discard a prompt because its baseline response seems unhelpful unless an exclusion rule was specified beforehand.

There is also a unit-of-analysis problem. Two word-order versions of the same scenario are not independent new situations. Report each row, then average within a scenario before summarizing across scenarios. Ten nearly identical templates do not justify the confidence of ten diverse domains.

Probing research makes a related point: a trained readout's performance depends partly on the readout and its opportunity to exploit regularities. Control tasks help test what the probe contributes. Our simple mean-difference method avoids training a large probe, but it does not escape dataset construction or selection effects. Hewitt and Liang, Designing and Interpreting Probes with Control Tasks

12 Causal influence still needs a bounded claim

Lesson 5.3 introduced $h' = h + \delta u$. If this edit changes the output, it establishes an effect of that intervention under the measured conditions. It does not establish that ordinary behavior uses only this direction, or that the direction corresponds exclusively to its human label.

Keep the original input fixed, include a zero edit through the same hook machinery, compare both signs, and use matched-norm random directions. Check recovery after hooks are removed. Record changes to the full output distribution alongside the preferred two-token score. Otherwise, general disruption can masquerade as successful semantic control.

A random direction is a comparison, not a complete null distribution. A single favorable comparison cannot establish that the proposed direction is uniquely meaningful. Likewise, an unsuccessful edit can reflect the selected site, dose, readout, or model rather than the absence of relevant computation everywhere.

The intervention's projection change is partly guaranteed by algebra: adding δu increases $u^T h$ by δ for a unit direction. That observation validates arithmetic, not emotion. The independent outcome is what happens downstream. Methodological work on activation patching similarly emphasizes that metric and input-construction choices can change the apparent result. Zhang and Nanda, Towards Best Practices of Activation Patching, v2

💡 Lab recommended here

Run the frozen held-out audit in Parts 6–7. Part 8 is an optional, predeclared intervention extension. It may strengthen a narrow causal conclusion; it cannot upgrade the experiment into a demonstration of subjective emotion.

13 Treat anthropomorphism as a measurement risk

Human terms can be useful shorthand if they retain their operational definitions. Problems arise when the shorthand acquires extra implications. “A despair-associated direction increased” can become “the model became desperate,” which can become an unsupported story about enduring motives.

Ask whose emotion is being represented: the user, a quoted speaker, a fictional protagonist, an assistant role, or no specified entity. Ask whether the label comes from output annotation, a prompt instruction, a learned probe, or an intervention. Ask whether the same result survives a changed template. These questions make the research more interpretable without dismissing it.

An assistant’s self-description is also output to explain. Repeating “I feel sad” is not an independent instrument measuring the cause of that sentence. A stronger investigation relates reports to controlled internal measurements and alternative explanations, while keeping unresolved questions unresolved.

For the Lab, success means a trustworthy record. A failed generalization can reveal lexical dependence; a mixed result can separate representation from output; a null result can expose the limits of the specimen or measurement. None requires inventing a successful emotional state to make the activity worthwhile.

14 Check your understanding

1. Why is an emotion word in an answer different from an emotion-related representation?
2. Why must an internal-state claim specify a timescale?
3. What makes a positive discovery mean partly automatic?
4. In the toy example, why is a positive mean insufficient?
5. What does a shared-word entity-binding pair test?
6. Why can an irrelevant emotion word activate a useful concept direction?
7. What does the two-token logit difference leave out?
8. Why is an intervention’s increased projection not semantic validation?
9. Does a null result settle whether a model has subjective experience?

i Short answers

1. One is observable text; the other is a measured computational pattern requiring its own validation.
2. A local activation, transcript effect, persistent store, and trained tendency are different mechanisms.

3. The direction is constructed from those differences; its mean signed discovery projection equals its norm.
4. One held-out pair disagrees and the observations vary substantially.
5. Sensitivity to who receives which outcome when a bag-of-words explanation is unavailable.
6. It may encode a mentioned concept rather than the target character's state.
7. Absolute candidate probabilities, other continuations, extended behavior, and general competence.
8. The edit guarantees that projection change algebraically; downstream effects need separate measurement.
9. No. The experiment tests a specified representation and output measure, not a settled consciousness criterion.

15 More Learning

1. Persona Vectors, v3. Trace extraction, evaluation, and the within-prompt-type qualification.
2. Emotion Concepts and their Function in a Large Language Model, v1. Compare local concept tracking with a persistent-state interpretation.
3. Do Large Language Models Have Emotions?, v1. Identify the additional functions proposed by this critique.
4. Decoding Emotion in the Deep, v1. Identify the labels and the exact quantities plotted.
5. Designing and Interpreting Probes with Control Tasks. Consider which shortcuts your controls actually exclude.
6. Consciousness in Artificial Intelligence. Optional background on theory-dependent indicators and uncertainty.