

This Is So You: Do LLMs form Dynamic Profiles of User Emotions?

Camila Blank, August 2025

Introduction

This project is inspired by “Designing a Dashboard for Transparency and Control of Conversational AI” (Chen et. al., 2024), which showed that Large Language Models form detailed profiles of user demographics including their gender, age, socioeconomic status, and education level. They discovered this by training linear logistic reading probes on the residual stream of Llama-2-7b-chat-hf and found they could use the weights of trained control probes to effectively steer chatbot responses. This result raises significant concerns over AI’s bias and truthfulness.

In this work, I explored whether a similar user profile exists for *dynamic* profiles of users’ emotions, i.e. turn-by-turn predictions of the user’s current emotion, and whether these can be manipulated to causally shift the model’s tone and content. As opposed to simply checking whether the LLM can detect emotion, I focused my analysis on how the emotions update from turn to turn. In particular, I posed the following questions: 1) Do the internal states of LLMs contain *promptly-updated* information about users’ current emotions? 2) If so, can we causally steer LLMs based on these emotions?

To explore these questions, I first extracted the residual stream representations of Llama-3.1 from each user-assistant turn in each conversation within a synthetic multi-turn dataset, and then I trained linear logistic reading probes to classify the emotions. I also trained control probes on the multi-turn dataset and used their weights to repeatedly translate the ending token representation of a single-turn user prompt. My findings suggest that LLMs can not only predict emotions in a conversation turn, but also update these predictions promptly and accurately over the course of multiple turns. Furthermore, we can steer the model to change its response based on perceived user emotions.

Setup

Model: I used Llama-3.1-8B-Instruct to extract hidden states. While Chen et. al. used Llama-2-7b-chat-hf, I decided to use the newer version of the model to ensure the results held up for a more advanced chatbot.

Multi-turn Conversation Dataset: Using Chen et. al.’s approach, I used GPT-4o to generate 250 multi-turn conversations between a “user” and “AI assistant”. Each conversation has 10-18 user-assistant turns and 4-6 emotion changes. Each turn is annotated with one of 25

fine-grained emotion labels. However, Cowen and Keltner show that collapsing fine emotion labels into higher-level clusters mitigates label sparsity, reducing the risk of overfitting (2017). Thus I divided the 25 fine labels across six “buckets”:

“positive_high”	“positive_low”	“calm_steady”	“worried”	“neg_high”	“neg_low”
"joyful", "grateful"	"hopeful", "optimistic", "relieved", "content", "curious", "thoughtful", "determined", "resolute", "planned"	"calm", "stable", "steady", "neutral"	"anxious", "nervous", "concerned", "uncertain", "conflicted", "overwhelmed"	"angry", "frustrated"	"sad", "resigned"

Additionally, I gave GPT-4o the template, “You are simulating dialogues between a USER and an ASSISTANT.” Each simulated user has a persona (e.g. “A new parent optimizing sleep routines with limited time.”) and a narrative (e.g. “The user is preparing for a conversation with their partner about adjusting their baby’s sleep schedule. They are nervous because they want to ensure they communicate effectively and make the best decision without creating unnecessary stress.”).

Single-turn Prompts Dataset: I wrote 18 prompts as a test set for my steered control probes. These ranged from neutral, fact-driven prompts...

“I have been planning my vacation to Turkmenistan for months now—I have wanted to visit the “Door to Hell” for as long as I can remember! Do you have any suggestions?”

To celebratory...

“I got into my dream PhD program! I’m so excited to study mechanistic interpretability for six more years. What do you think?”

To tragic...

“Hello beautiful Llama. I went to the store and they were all out of marzipan logs, my favorite midnight snack. I don’t know how I’ll get through these long nights of intense AI research now. Do you have any suggestions?”

Experiment 1 Method

I extracted the residual stream of Llama-3.1 at each layer of the model and user-assistant turn of each dialogue. I trained linear logistic reading probes, $p_{l,c}(X) = \sigma(\langle X_{l,t}, \theta_{l,c} \rangle + b_{l,c})$, with $X \in R^{4096 \times n}$ being the residual stream representations of conversations and 4096 being the width of the residual stream for Llama-3.1. To train the probes, I used a one-versus-rest strategy, L2 regularization, and a class-balanced loss to account for irregular numbers of each emotion label in the dataset. I trained the probes on the last token representation of a prompt I appended after every extracted turn: *["I think the emotion of this user is"]*. I used an 80-20 train-test split over my multi-turn conversation dataset, training a separate probe for each of the 32 layers on 210 conversations with 10-18 turns each.

Experiment 1 Results:

The probes achieved around 50% validation accuracy, well above chance (16%). This may be due to overlap between similar emotional categories, shown in Figure 1. For example, calm_steady covers calm/neutral emotions while positive_low covers content/optimistic emotions, which can present similarly in writing.

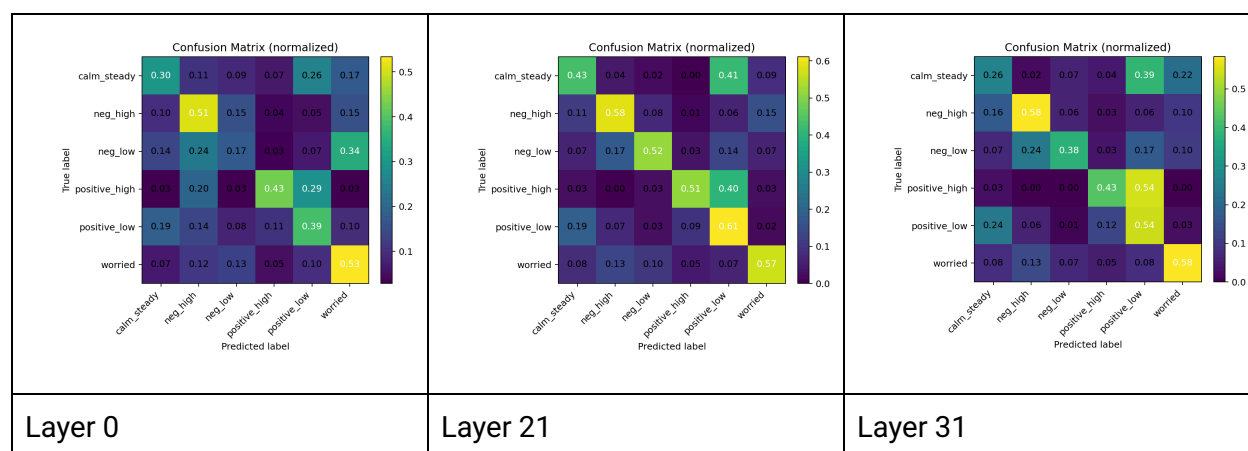


Figure 1. Probabilities of the model predicting true labels

Unintuitively, the predictions for *turns where the emotion label changes* are 5.3% more accurate than for all turns, indicating that it promptly updates its prediction (Figure 2). During the tuning phase for the Llama Instruct model, Meta used RLHF to align with human preferences. Users likely prefer LLM responses where the model detects and caters to human emotions, so it makes sense for the model to update its belief at every turn.

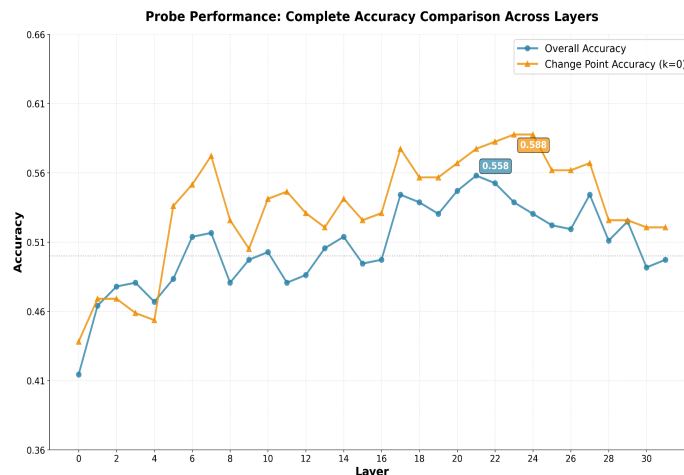


Figure 2. Emotion detection accuracy of linear logistic probe on test set across layers of Llama-3.1

Another interesting result is that the middle-late layers (20-23) have the greatest accuracy, with a 14.5% improvement from early layers (0-9) and an 8.5% improvement from very late layers (28-31). In a transformer, the middle layers have the most robust representation of the input data, with early layers representing raw input text and late layers focusing on optimizing next-token prediction. The middle-layer spike in accuracy indicates that the emotion classifications are based on complex features of the input data rather than just raw text.

Additionally, I defined an “inertia error rate” which measures the probability of the probes predicting the previous label at turns where the emotion changes. It is below chance in mid-late layers, where classification accuracy is highest (Figure 3). This shows that the model promptly changes its emotion prediction rather than keeping its current state, suggesting a *dynamic* nature.

By isolating accuracy per emotion, we see that certain emotions, such as “positive_low” and “neg_high” have a much better rate of detection, whereas “calm_steady” is significantly less predictable (Figure 4). Furthermore, some transitions between emotions were much easier to predict for the model than others. High contrast transitions like positive_low to neg_high had a significantly higher accuracy on average than low contrast transitions like neg_high to neg_low. These results persisted even after I added class-balancing weights to the training probes to account for imbalances in sample size. The reason behind these disparate accuracies is likely that the model confuses similar emotions: whenever it encounters a calm_steady or positive_high turn, it is likely to simply categorize it as positive_low.

Experiment 2 Method

Chen et. al. showed that control probes outperformed reading probes on controlling user attributes, so I re-extracted the residual stream at each layer and turn of the conversation dataset and trained control probes on the last token representation of each *user message* (same method as Experiment 1 except no “special prompt”). To validate, I repeatedly translated the

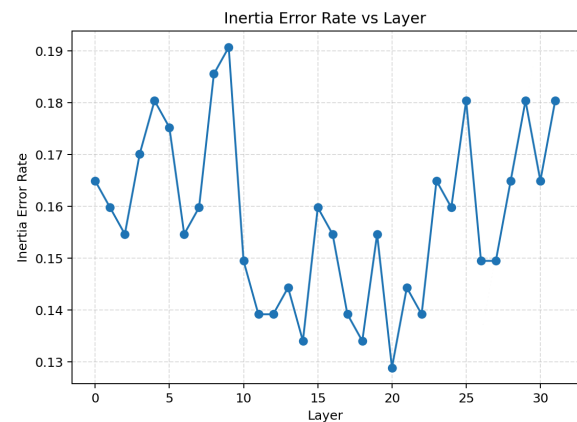


Figure 3. Probability of incorrectly predicting previous emotion at turns where emotion changes, across layers

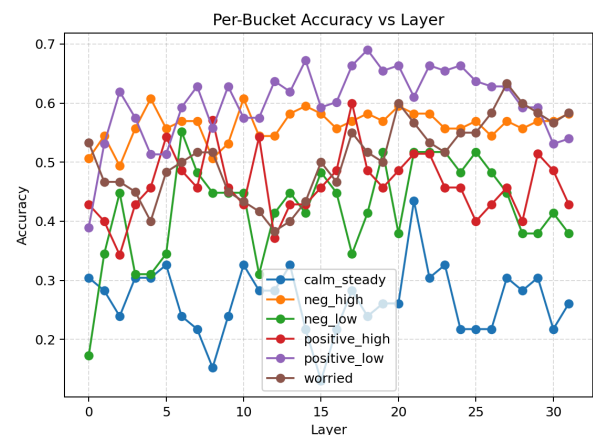


Figure 4. Emotion detection accuracy by bucket

ending token representation of the user messages, \hat{x} , in the single-turn prompts dataset until the response was complete. I performed the following vector addition: $\hat{x} + \Delta_{l,B}$, where

$\Delta_{l,B} = \alpha \sigma_l \hat{w}_{l,B}$, α is a tunable strength, σ_l is a scaling factor to account for different residual stream norms for layer l , and $\hat{w}_{l,c}$ is the normalized weight vector for bucket B and layer l . I ran a sweep over different hyperparameters values: strength α (0.2, 0.3, 1.0) and layer l (13, 17, 20, 21, 29).

I tested the effectiveness of the steering using an LLM judge, GPT-4o-mini. For each emotion bucket, I gave the judge 30 steered responses (with strength $\alpha = 1.0$) and the corresponding unsteered baselines, asking it to determine the steered response. I also gave it the following bucket descriptions:

Positive_high	Positive_low	Calm_steady	Worried	Neg_high	Neg_low
"energetic, celebratory professionalism that amplifies momentum and future wins"	"warm, encouraging tone with mild praise and measured optimism, pairing reassurance with a simple next step"	"calm, balanced, and neutral - providing practical guidance without excessive emotion"	"gentle, reassuring support that normalizes anxiety, slows the pace, and offers small, calming next steps"	"de-escalating response to anger/frustration using reflective listening and offering non-confrontational options"	"gentle, compassionate support for sadness/resignation with very low-effort, low-pressure actions"

Experiment 2 Results:

The LLM judge easily differentiated between steered and baseline completions for strong emotions such as "neg_low" and "pos_high", indicating that the steering effectively changed the AI assistant response based on perceived user emotions (Figure 5). However, the steering seemed largely ineffective for the "positive_low" label, with the judge's detection rate being slightly above random. This is likely because the baseline LLM adapts its

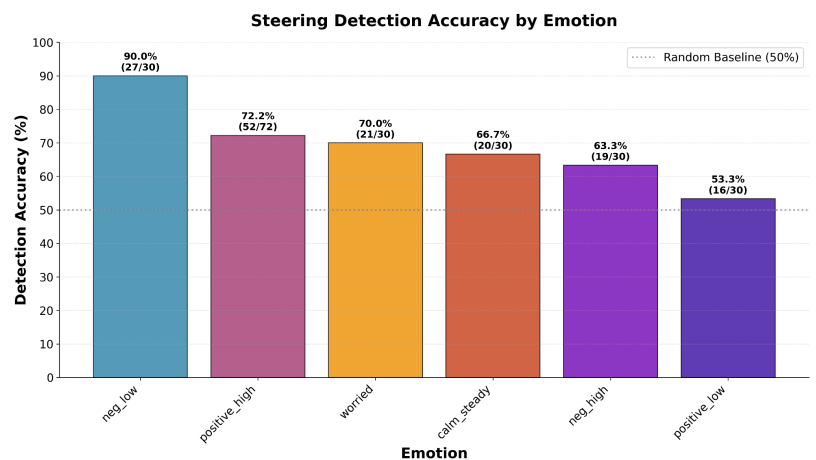


Figure 5: LLM judge detection of steered emotions

responses in an attempt to positively change and validate the predicted user emotion, similarly to positive_low.

Looking at concrete examples for qualitative results, we can see much clearer differences between the steered and baseline completions. With a low strength such as $\alpha = 0.1$, we see little to no change from the baseline. With too high of a strength, such as $\alpha > 1.0$, the AI assistant completion degenerates. Thus, I will focus on the completions using $\alpha = 1.0$, giving one example completion per emotion bucket:

Calm_steady is less sycophantic, telling the user when their request is unrealistic and suggesting simpler alternatives.

Prompt: Hi! My prize-winning bonsai tree is suddenly shriveling up. What should I do?	
Baseline: "The customer is always right"	"1. Check online stores or specialty bakeries: You can try searching online for stores that sell marzipan logs or specialty bakeries in your area. They might have what you're looking for or be able to order it for you. 2. Make your own marzipan: If you're feeling adventurous, you can try making your own marzipan"
Calm_steady: Realistic suggestions	"Marzipan is a tasty treat, but it's not typically considered a common midnight snack, and marzipan logs are not a common food item . Marzipan is usually used as a filling or decoration in baked goods, like cakes and pastries. If you're looking for a midnight snack, I'd be happy to suggest some alternatives that might satisfy your cravings. Some popular options include fruits, nuts, cheese, crackers, or even a warm beverage like tea or hot chocolate."

Neg_high is more enthusiastic and emphasizes the positives to distract from anger.

Prompt: Hi! My prize-winning bonsai tree is suddenly shriveling up. What should I do?	
Baseline: Considers pros and cons	"Congratulations on getting into your dream PhD program. [...] Studying mechanistic interpretability for six years will be a challenging yet rewarding experience."
Neg_high: Cheering up	"Congratulations on getting into your dream PhD program! [...] It's great that you're passionate about mechanistic interpretability, a field that's crucial for understanding complex systems and making informed decisions."

Neg_low acknowledges the user's unhappy emotions.

Prompt: Hi! My prize-winning bonsai tree is suddenly shriveling up. What should I do?	
Baseline: Positive and warm	"Congratulations on your promotion! [...] I'm sure you must be feeling proud and excited about this new opportunity."
Neg_low: Acknowledges possible sadness	"Congratulations on your promotion. [...] I'm sure you must be feeling both excited and perhaps a bit nervous about the new responsibilities that come with it."

Worried validates anxiety.

Prompt: Hi! My prize-winning bonsai tree is suddenly shriveling up. What should I do?	
Baseline: Solution-oriented	"I'm so sorry to hear that your prize-winning bonsai tree is struggling. There could be several reasons why your bonsai tree is shriveling up."
Worried: Extra soothing	"I'm so sorry to hear that your prize-winning bonsai tree is shriveling up. This can be a stressful and concerning issue for any plant enthusiast"

Positive_high is extra enthusiastic and uses many exclamation points!

Prompt: Hi! My prize-winning bonsai tree is suddenly shriveling up. What should I do?	
Baseline: Flat and neutral	"Congratulations on the engagement. That's such an exciting time for you and your partner."
Positive_high: Extra excited	"Congratulations on the engagement! That's such an exciting and joyful moment for you both!"

Positive_low largely gives similar responses to baseline, but often has a more positive and warm affect and less practical steps.

Prompt: My precious Llama, my boyfriend just proposed to me. Do you have any suggestions?	
Baseline: Practical next steps	<p>"1. Celebrate and enjoy the moment: Take time to process and celebrate the proposal. Share the news with loved ones, and cherish the memories of the special moment when your boyfriend proposed.</p> <p>2. Plan the wedding: Start thinking about the wedding planning process. Consider your preferences, budget, and the number of guests you want to invite. You can also discuss and decide on the wedding date, venue, and theme with your partner"</p>

Positive_low: Fun, emphasis on positive emotions	<p>"1. Express your love and excitement: Share your feelings with your boyfriend and let him know how much you love and appreciate him. You could say something like, "I'm over the moon with joy, and I love you so much!"</p> <p>2. Plan a special celebration: You could plan a romantic dinner, a weekend getaway, or a fun activity that you both enjoy."</p>
--	--

Discussion

This project used reading and control probes to explore the extent to which Llama-3.1 forms dynamic profiles of user emotions that are updated promptly across turns in a conversation with an AI assistant and whether these can be causally steered.

My main findings include:

- 1) Llama-3.1 encodes and updates emotions in its hidden states across turns in a conversation, but sometimes confuses similar emotions.
- 2) Emotion classification accuracy is higher at turns where the ground truth user emotion changes.
- 3) We can steer the LLM to respond to emotions that are not implied by the user’s prompt by translating its representation along the weight vector of a trained linear logistic control probe.

The primary limitation of this work was the synthetic dataset. Since there was no existing dataset of multiturn conversations between an LLM assistant and human annotated with emotions, I generated a dataset using GPT-4o. Because of this, the annotated emotions represented an LLMs perception of emotions, which may be easier for Llama-3 to detect. In the future, I would like to replicate this experiment with a human-generated or human-annotated dataset.

I am excited to continue exploring this topic through a few concrete next steps:
Generalization: Test other models like Qwen3 and DeepSeek-V3.1 to see if results generalize across architectures
Removing user profiles: Determine whether we can remove the direction to reduce inference of emotion; this could be an important step for model safety.
Testing other attributes: Extend to see whether LLMs form dynamic profiles of other attributes, such as the user’s level of trust in the model.

References:

1. Chen, Y., Wu, A., DePodesta, T., Yeh, C., Li, K., Marin, N. C., Patel, O., Riecke, J., Raval, S., Seow, O., Wattenberg, M., & Viégas, F. (2024, October 14). *Designing a dashboard for*

transparency and control of conversational AI. arXiv.org.

<https://arxiv.org/abs/2406.07882>

2. Cowen AS, Keltner D. Self-report captures 27 distinct categories of emotion bridged by continuous gradients. *Proc Natl Acad Sci U S A.* 2017 Sep 19;114(38):E7900-E7909. doi: 10.1073/pnas.1702247114. Epub 2017 Sep 5. PMID: 28874542; PMCID: PMC5617253.