Appendix- Do You Trust Me? Cognitive–Affective Signatures of Trustworthiness in Large Language Models

Gerard C. Yeo
National University of Singapore
Singapore
e0545159@u.nus.edu

Svetlana Churina
National University of Singapore
Singapore
churinas@nus.edu.sg

Kokil Jaidka
National University of Singapore
Singapore
jaidka@nus.edu.sg

ACM Reference Format:
Gerard C. Yeo, Svetlana Churina, and Kokil Jaidka. 2025. Appendix- Do You Trust Me? Cognitive–Affective Signatures of Trustworthiness in Large Language Models. In *Proceedings of the ACM Web Conference.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

A Probing Variables

Refer to Table 1 for all variables used in the probing experiments. All variables were originally measured on a 5-point scale and subsequently binarized, where responses < 3 were coded as 0 and responses ≥ 3 were coded as 1.

B Prompt Templates

To obtain the activations, we parsed the input narratives into the LLMs in the following template. System prompt- *"You are an evaluator trained to assess perceived trustworthiness– that is, whether you think a review seems trustworthy or not"*, followed by the user prompt- *"Review: {review text}, After reading the product review, respond with ONLY ONE WORD: 'high' or 'low'. Answer:"*.

C High–Low Trust Activation Differences Computation

To assess whether perceived trustworthiness is encoded in intermediate representations, we compared mean activation magnitudes for narratives labeled as *high* vs. *low* trustworthiness. For each layer ℓ and head h, we first extracted the corresponding attention-head output at the final token position. Let $\mathbf{A}_{\ell,h}^{(i)} \in \mathbb{R}^d$ denote the activation vector for sample i, where d is the head dimensionality.

We then computed the average absolute activation magnitude across samples and token/head dimensions. Formally, we estimate

$$\mu_{\ell,h}^{(\text{high})} = \mathbb{E}_{i \in \mathcal{H}} \left[\|\mathbf{A}_{\ell,h}^{(i)}\|_1 \right], \tag{1}$$

and analogously for the low-trust group $\mu_{\ell,h}^{(\text{low})}$. The groupwise activation difference is then:

$$\Delta_{\ell,h} = \mu_{\ell,h}^{(\text{high})} - \mu_{\ell,h}^{(\text{low})}. \tag{2}$$

Thus, $\Delta_{\ell,h}$ measures how strongly attention head (ℓ, h) responds to high- vs. low-trust narratives. Large positive or negative deviations indicate heads whose activation magnitudes distinguish trust categories, suggesting that latent trust-related features emerge even without explicit supervision.

Table 1: Variables Included for Probing

Variables
Appraisals
Accountability-circumstances
Accountability-other
Accountability-self
Attentional activity
Certainty
Control-circumstances
Control-other
Control-self
Coping potential
Difficulty
Effort
Expectedness
External normative significance
Fairness
Future expectancy
Goal conduciveness
Goal relevance
Novelty
Perceived obstacle
Pleasantness
Emotions
Anger
Disappointment
Disgust
Gratitude
Joy
Pride
Regret
Surprise
Behavioral Intentions
Intent to promote
Intent to repurchase
Consumer-related variables
Helpfulness
Trustworthiness

the ACM Web Conference, Dubai, UAE

ACM ISBN 978-1-4503-XXXX-X/2018/06
https://doi.org/10.1145/nnnnnnn.nnnnnnn

D Trustworthiness Probing Results

Figure 1 presents the head-level trustworthiness probing results for LLaMA-3.1, Qwen, and Mistral. Each heatmap visualizes the linear probe accuracy for every attention head across all transformer layers. Brighter (red) regions indicate higher probe accuracy, reflecting heads whose representations contain more trust-relevant information, while darker (blue) regions indicate weaker signal.

Across all three models, trustworthiness is at least partially recoverable from internal representations, though the distribution of probe-sensitive heads varies by architecture. LLaMA-3.1 exhibits a broad, distributed pattern of decodable trust signals across depth. Qwen shows a stronger concentration of high-performing heads in very early and later layers, suggesting a more localized encoding in upper-level representations. Mistral displays a mid-to-late layer progression, with probe accuracy gradually increasing with depth.

These patterns align with the activation-based analyses reported in the main text, indicating consistent architectural differences in where trust-relevant information is stored and how easily it can be linearly extracted.

E Comparing base vs. fine-tuned models in trustworthiness probing

Figures 2 and 3 present the comparison between base and fine-tuned models of mistral and qwen in trustworthiness probing.

F Fine-Tuning Configuration

We fine-tuned three instruction-tuned large language models: **Qwen2.5-7B-Instruct**, **Mistral-7B-Instruct-v0.3**, and **LLaMA 3.1-8B-Instruct**, using the **Low-Rank Adaptation (LoRA)** technique to enable parameter-efficient fine-tuning. For all models, we set the LoRA rank to $r = 8$, the scaling factor to $\alpha = 32$, and the dropout rate to 0.1, applying adapters to the attention and projection layers (`q_proj`, `k_proj`, `v_proj`, `gate_proj`, `up_proj`, `down_proj`, `o_proj`). Bias terms were disabled (`bias=none`), LoRA weights were initialized randomly, and only the adapter parameters were updated during training while keeping the base model frozen. Training employed a cosine learning rate scheduler with a linear warmup phase, and all experiments were conducted in mixed precision for efficiency. The optimal hyperparameters, including learning rate, batch size, number of epochs, and warmup ratio, were determined through automated search using the **Optuna** framework based on validation loss and model stability across multiple random seeds.

Table 2: Performance comparison between base and LoRA fine-tuned models.

Model	Base Model			Fine-tuned (LoRA)		
	Acc.	Macro F1	W-F1	Acc.	Macro F1	W-F1
Qwen2.5-7B-Instruct	0.561	0.545	0.536	0.633	0.633	0.632
Mistral-7B-Instruct-v0.3	0.648	0.645	0.648	0.662	0.658	0.662
LLaMA 3.1-8B-Instruct	0.568	0.567	0.564	0.669	0.668	0.670

Table 3: Per-class F1-scores for the *low* and *high* categories before and after LoRA fine-tuning.

ModelModel**Model**pt< -**Model**pt>	Base Model		Fine-tuned (LoRA)	
	F1-Low	F1-High	F1-Low	F1-High
Qwen2.5-7B-Instruct	0.630	0.460	0.638	0.628
Mistral-7B-Instruct-v0.3	0.614	0.676	0.624	0.693
LLaMA 3.1-8B-Instruct	0.595	0.539	0.652	0.685

G Robustness Analyses

We conducted several robustness checks to evaluate whether our findings about trust representations are consistent across probe types, representational streams, and fine-tuning conditions.

G.1 Nonlinear Probe Comparison

We evaluated whether nonlinear transformations facilitate trustworthiness decoding. Three-layer MLP probes provided only small gains over linear probes across all models. Peak nonlinear performance for LLaMA (Layer 16, 64.8% accuracy; F1 = 69.0), Qwen (Layer 2, 65.5%; F1 = 67.8), and Mistral (Layer 10, 67.6%; F1 = 69.1) improved only marginally over linear baselines. These results indicate that trust cues occupy largely linearly separable manifolds, with limited benefit from nonlinear transformations.

G.2 Post-Residual Stream Analysis

Across all three models–Mistral-7B, Qwen-2.5-7B, and LLaMA-3.1 we observe a clear and consistent layerwise pattern: trustworthiness becomes most decodable in the middle-to-late layers of the network, with probe accuracy peaking roughly between Layers 18 and 24. This trend appears regardless of activation type (post-attention or post-MLP), indicating that trust is not encoded primarily in early lexical or syntactic representations. Instead, trust-related features emerge as higher-level semantic abstractions that crystallize in the middle of the transformer stack.

Fine-tuning yields small but uniform accuracy gains across layers, yet the location of the peak remains unchanged, suggesting that fine-tuning sharpens an existing representational structure rather than altering where trust information is formed. This convergence across models and activation streams demonstrates that the middle layers serve as a shared location for trust-related processing in contemporary LLM architectures.

To examine how trust-related information evolves through the model's computation, we measure the difference in residual-stream magnitude between high-trust and low-trust narratives at each layer (Figures 4,5,6). Across all three models, the early layers show near-zero separation, indicating that trust representations are not formed at the lexical or shallow contextual level. Beginning around Layer 15, all models exhibit a clear increase in residual-norm differences, revealing that trust-sensitive features are progressively amplified within the residual stream.

This amplification peaks in the middle-to-late layers, but with architecture-specific profiles. **Mistral-7B** shows a strong and monotonic increase in residual separation toward the top layers, suggesting that trust becomes increasingly geometrically distinct as the model approaches its output head. **LLaMA-3.1** shows a mid-layer

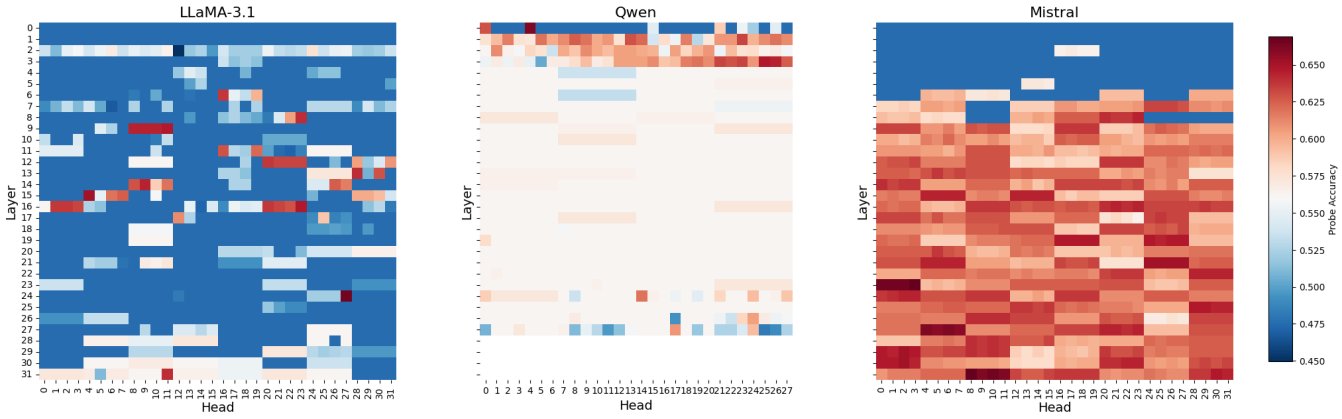

Figure 1: Head-level trustworthiness probing accuracy for LLaMA-3.1, Qwen, and Mistral. Each heatmap shows probe accuracy (blue = low, red = high) for every attention head across all layers.

rise followed by a sharp collapse in the final layers, consistent with late-layer compression observed for other semantic features. **Qwen-2.5-7B** shows weak and unstable separation, with small mid-layer differences and large, erratic swings near the top layers, mirroring the weaker trust decoding observed in its probing results.

Fine-tuning slightly elevates the magnitude of high–low differences but preserves each model's characteristic trajectory, indicating that fine-tuning sharpens existing trust-related dynamics rather than altering where they emerge in the computation.

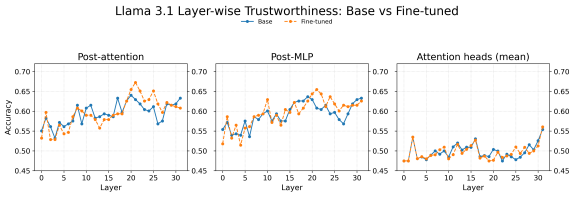

Figure 2: Line plot comparing the performance of base vs. fine-tuned Llama3.1 models of trustworthiness probes.

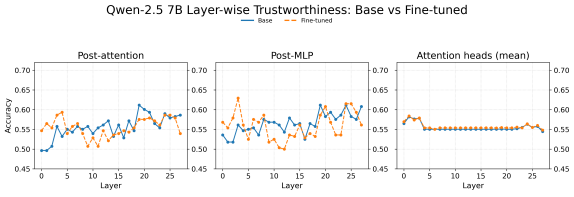

Figure 3: Line plot comparing the performance of base vs. fine-tuned Qwen 2.5 models of trustworthiness probes.

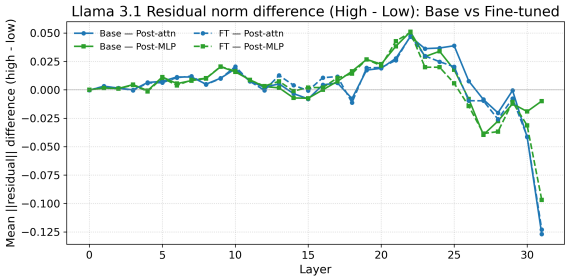

Figure 4: Llama

H Probing for Psychological and Consumer-related Variables

Figure 7 presents the best F1 scores obtained for each variable across all models when using attention-head activations as features.

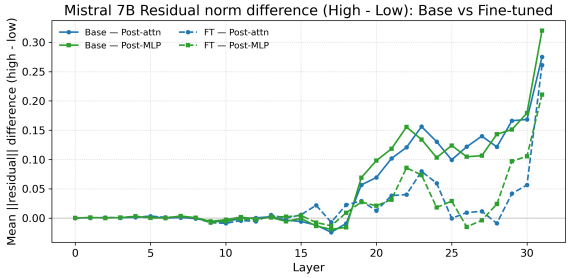

Figure 5: Mistral

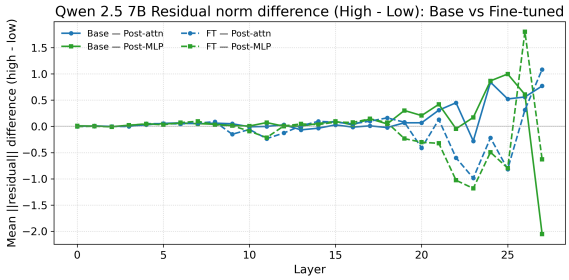

Figure 6: Qwen

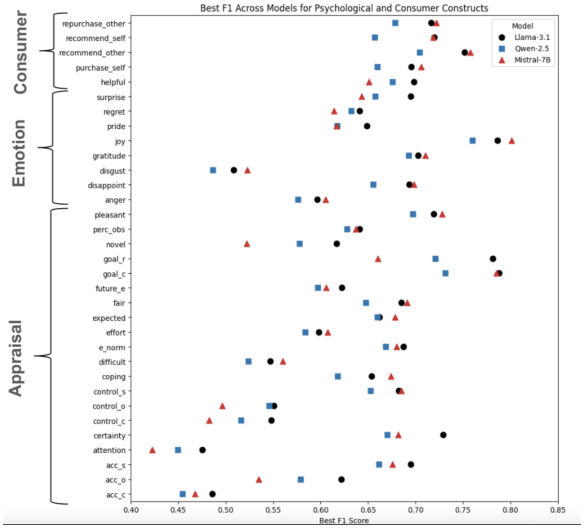

Figure 7: RQ3: Best F1 score per psychological and consumer-related construct across all layers and heads using attention heads.

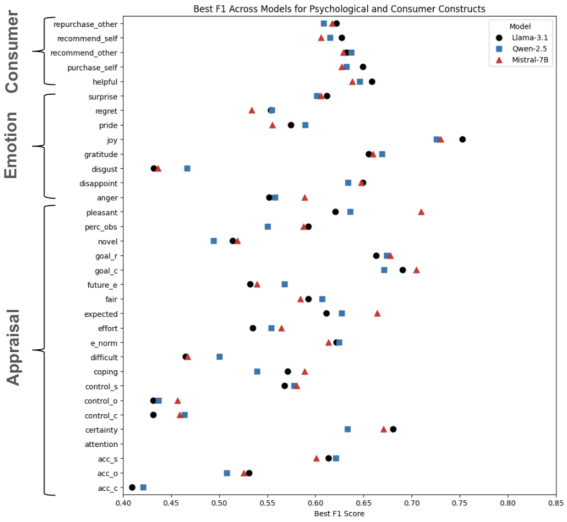

Figure 8: RQ3: Best F1 score per psychological and consumer-related construct across all layers and heads using post-attention activation states.

The results reveal substantial variation in predictability: several variables achieve relatively high F1 scores, indicating strong linear recoverability, whereas others consistently show low scores, suggesting weaker or noisier underlying signals.

Figures 8 and 9 report the corresponding results using post-attention and post-MLP residual streams. These patterns closely mirror those observed in the attention-head analysis, with cognitive appraisals like accountability-self, certainty, goal-related appraisals, and emotions such as joy being best predicted from these trust

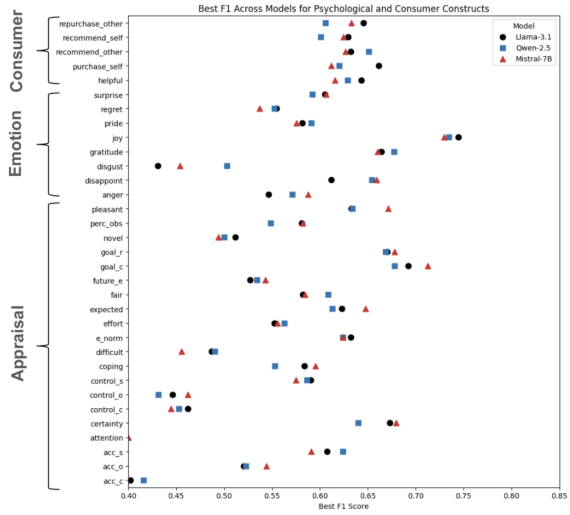

Figure 9: RQ3: Best F1 score per psychological and consumer-related construct across all layers and heads using post-MLP residual states.

activations. Taken together, the findings indicate meaningful differences in how well appraisal, emotion, and consumer-related constructs can be predicted from trust-related activations, highlighting that not all constructs are equally linearly recoverable.

Received 17 November 2025