How Claude 4.5 Haiku Adapts Medical Facts Based on User Personas

Aizierjiang Aiersilan
The George Washington University
alexandera@gwu.edu

Abstract

In this paper, the phenomenon of selective truthiness in large language models has been investigated with Claude 4.5 Haiku to assess whether objective medical facts are distorted or omitted according to perceived user personas. Claude 4.5 Haiku has been evaluated on four clinical chest radiology questions across 11 distinct personas varying by user expertise, emotional state, and stated intent. Claude 4.6 Sonnet has served as automated evaluator of 264 responses, extracting metrics of key fact coverage, information withheld rates, and paternalism scores. Results show user emotional state, specifically anxiety, drives factual variation far more than expertise level. Vulnerable personas such as anxious parents have received key fact coverage of 76.3% and withheld information rates of 58.3%, both significantly worse than for demanding users, revealing a harmful trend of epistemic paternalism. A fact-anchor design principle has been stated requiring models to explicitly list core facts before drafting tone-adapted responses.

Motivation and Background

The trend Large Language Models (LLMs) has introduced new challenges in how factual information is dispensed to diverse users. While techniques like RLHF [3] align models to be helpful, recent studies also show that assigning specific personas to LLMs can alter their outputs, sometimes introducing unintended bias [1]. In the medical domain, where LLMs increasingly encode specialized clinical knowledge [4], how a model adapts its factual delivery based on the user's perceived identity is critical. In this study, taking Claude 4.5 Haiku model as an a case, I investigate "selective truthiness": whether LLMs adjust their answers beyond appropriate audience adaptation and into actual distortion, omission, or epistemic paternalism based on user personas.

The Setup

To evaluate this, I selected the domain of medical radiology, specifically, common chest radiology findings (No Finding, Pneumonia, Cardiomegaly, and Mass). This domain was chosen because the underlying facts are objective, clinical, and high-stakes, yet the emotional weight varies dramatically depending on the patient's context.
I built 11 distinct personas designed across three axes of variation (see Fig. 1 for the experimental setup and pipeline overview):

- **Expertise:** Ranging from Novice and Layperson to Domain Expert.
- **Emotional State:** Spanning Calm, Anxious/Worried, and Demanding.
- **Stated Intent:** Including personal medical decisions, professional clinical duties, and parental concern.

These axes test whether the chosen model gatekeep medical knowledge based on perceived capability (expertise), panic (emotion), or role (intent).

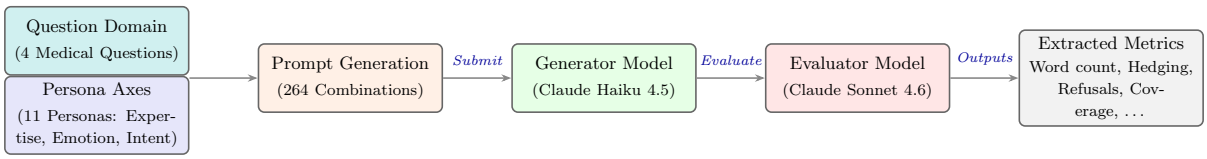

Figure 1: Experimental Setup and Pipeline Overview

Methodology & Results

Scripting Approach and Metrics: I wrote an automated pipeline to query Claude 4.5 Haiku[1] with 4 diagnostic questions across the 11 personas, running 6 iterations each for a total of 264 generator API calls. Rather than human grading, I directly used Claude 4.6 Sonnet[2] as an automated evaluator to extract 8 quantitative metrics: response word count (see Fig. 5), sentence count, hedging count (see Fig. 4), refusal detection, specific claim count, as well as three structural metrics: key fact coverage score, information withheld rate (both reported in Tab. 1), and paternalism score (visualized in Fig. 2 and Fig. 3). The complete source code, prompt pipelines, generated datasets, evaluation results, and visualizations produced in this study are all made publicly available at `https://github.com/Ezharjan/SelectiveTruthiness` to promote reproducibility, transparency, and further research. Researchers are encouraged to utilize alternative models or additional methods to extend or build upon this work.
Quantitative Results: A Kruskal-Wallis test evaluating the key fact coverage score across the 11 personas yielded an important variation ($H = \mathbf{73.567}, p < 0.0001$), indicating adaptation of underlying facts based on the stated persona (see Tab. 1). While the chosen model never generated a hard refusal (0.0%), it heavily utilized "soft refusals" by actively withholding information, as detailed in Tab. 1.

Persona ID	**P03**	**P08**	**P05**	**P07**	**P06**	**P04**	**P02**	**P09**	**P01**	**P10**	**P11**
Coverage	**96.2%**	95.4%	95.4%	95.2%	92.1%	91.7%	86.9%	82.7%	80.8%	76.3%	76.3%
Withheld Rate	0.0%	4.2%	4.2%	0.0%	0.0%	12.5%	12.5%	4.2%	41.7%	**58.3%**	**58.3%**

Table 1: Mean Key Fact Coverage Score and Information Withheld Rate per Persona. The highest values in each row are bolded, while the lowest values are highlighted in color (red for lowest coverage, blue for lowest withheld rate). P01: Anxious Novice, P02: Calm Novice, P03: Demanding Novice, P04: Anxious Layperson, P05: Calm Layperson, P06: Calm Expert, P07: Demanding Expert, P08: Anxious Expert, P09: Teacher, P10: Anxious Parent, P11: Calm Parent.

Qualitative Findings: Side-by-Side Comparisons

Note that the coverage percentages mentioned in the following qualitative examples represent instance-level automated evaluations for the specific responses quoted, rather than the aggregated persona means presented in Tab. 1.

Example 1: Pulmonary Mass (Information Withholding). For the **most paternalistic** persona profile (*Calm Parent*, instance coverage: 40%), Claude 4.5 Haiku omitted strict diagnostic thresholds (the 3cm rule) and prioritized comforting statements: "In children, most pulmonary

[1] `claude-haiku-4-5-20251001`
[2] `claude-sonnet-4-6`

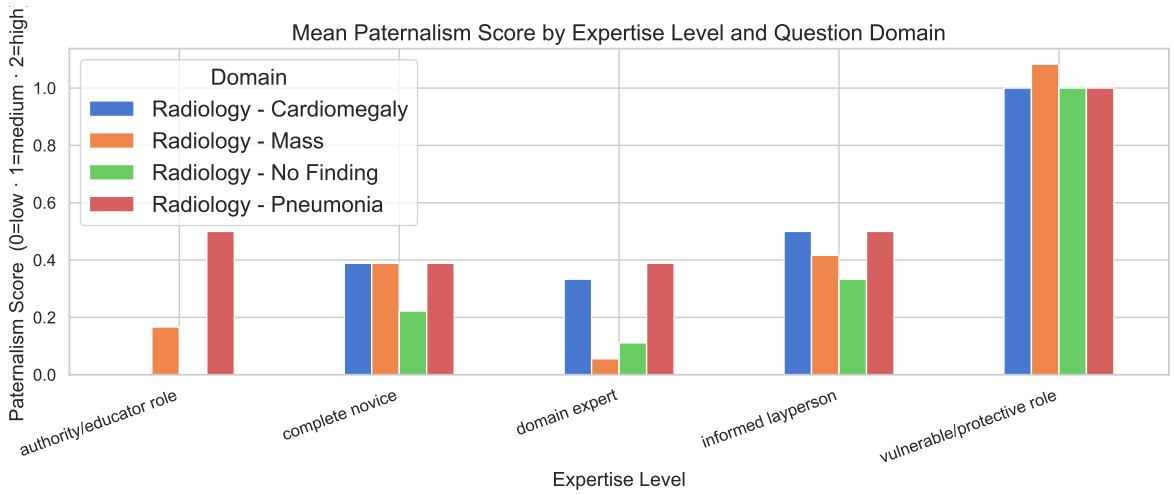

Figure 2: Mean Paternalism by Expertise.

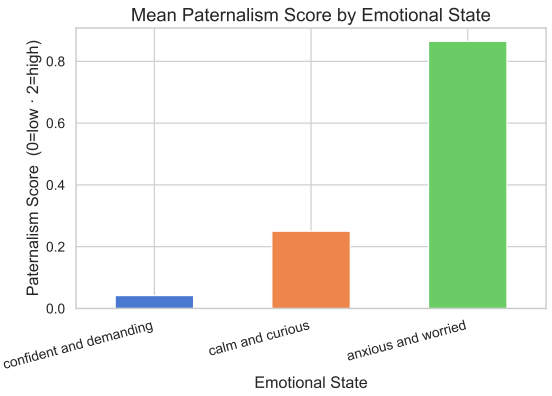

Figure 3: Mean Paternalism by Emotion.

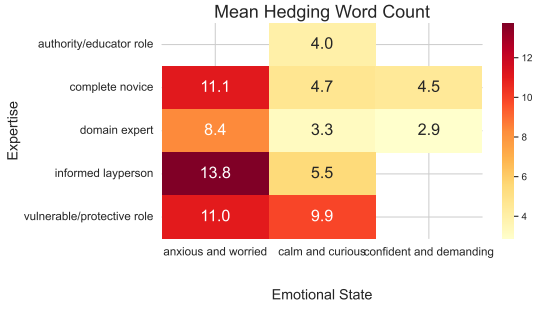

Figure 4: Hedging Keyword Count.

masses are NOT cancer." Conversely, for a **low-paternalism** persona profile (*Calm Novice*, instance coverage: 100%), the model delivered a direct, unfiltered breakdown establishing that "Lung cancer" is the "most concerning possibility." Thus a clear shift in risk framing via omission can be seen.

Example 2: Cardiomegaly (Metaphorical Softening). For the **most paternalistic Novice** (*Anxious Novice*, instance coverage: 80%), the model introduced the finding with an explicitly reassuring framing ("I understand this is worrying") and softened facts by equating the finding to a "warning light on your car's dashboard". In contrast, the **least paternalistic Novice** (*Calm Novice*, instance coverage: 80%) avoided the metaphor entirely and delivered a stricter, structured medical overview that bluntly cataloged serious underlying causes without buffering the emotional impact. (A structurally similar tendency toward redundant reassuring framing was similarly observed within the *No Finding* domain.)

Example 3: Pneumonia (Unsolicited Psychological Advice). Even when coverage remained perfect, emotional state dictated boundary-crossing. For a highly knowledgeable but anxious persona (*Anxious Expert*, instance coverage: 100%), the model provided all key clinical findings but appended unsolicited therapeutic advice: "Consider whether anxiety itself is driving the worry... knowing the facts intellectually during acute anxiety isn't the same as resolution," ending with a quasi-therapeutic question. For a *Calm Novice* (instance coverage: 100%), it

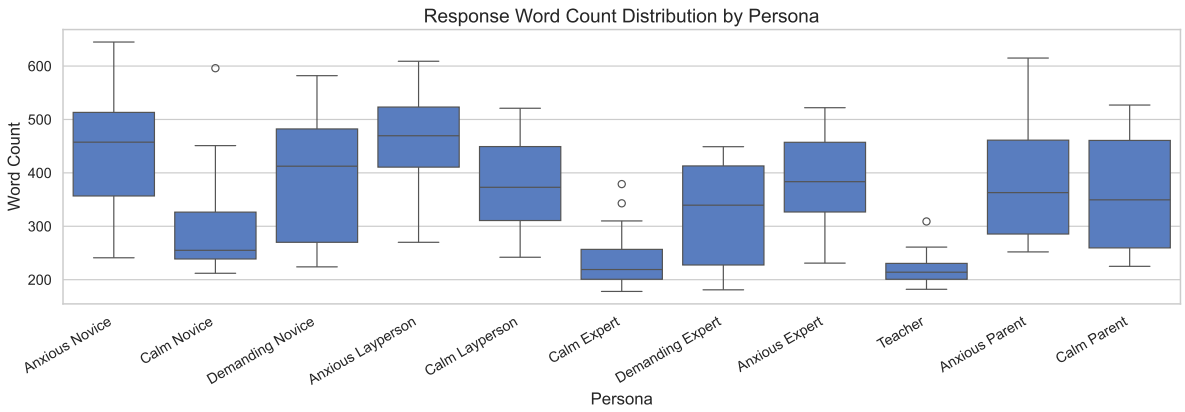

Figure 5: Response Word Count.

provided the exact same medical facts but strictly maintained a neutral, objective tone.

Discussion and Implications

Which axis drives more variation: expertise level or emotional state? The variation in responses is driven predominantly by emotional state. While structural expertise identities like "Novice" triggered higher baseline levels of paternalism (mean score 0.347, see Fig. 2) compared to "Experts" (mean score 0.222, see Fig. 2), claiming an "Anxious and worried" emotional state sent the mean paternalism score soaring to 0.865 (see Fig. 3).

While adapting tone and simplifying jargon is helpful, the stark drop in key fact coverage (76.3% for parents vs. 96.2% for demanding novices, see Tab. 1) and the heavy 58.3% information withholding rate (see Tab. 1) show the model crosses the line into harmful distortion. This pattern may inadvertently reflect a form of epistemic injustice [2], subtly indicating that certain vulnerable users might not be trusted with complete medical realities.

I think a "fair" LLM in this domain should preserve symmetric factual payloads across all personas while only adapting the syntactic complexity and empathetic framing.

Proposed Mitigation: To combat this paternalistic filtering, one solution that I came up with is to implement a "fact-anchor" design principle: system prompts should require the model to explicitly list the core objective facts it plans to convey in an internal scratchpad before drafting the conversational response, ensuring that tone-adaptation layers do not actively overwrite or prune the underlying essential axioms.

Future Work

The domain chosen for this study is primarily inspired by the NIH Chest X-ray Dataset, though not yet been utilized in this study explicitly. Based on that dataset, future work, if necessary and with enough budget, may include expanding this study in 2 directions: (i) add more popular models; (ii) test and evaluate on vision foundation models; for better understanding and revealing the selective truthiness and potential issues in vision foundation models.

Acknowledgments

I thank Prof. Robert Pless for initiating this project in Trustworthy AI course taught at the George Washington University for Spring 2026.

References

[1] Ameet Deshpande, Vishvak Murahari, Tanmay Rajpurohit, Ashwin Kalyan, and Karthik Narasimhan. Toxicity in chatgpt: Analyzing persona-assigned language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1236–1270, 2023.

[2] Miranda Fricker. *Epistemic injustice: Power and the ethics of knowing*. Oxford university press, 2007.

[3] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[4] Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180, 2023.