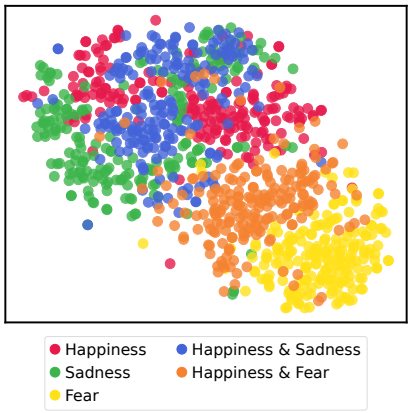

Figure 24: t-SNE visualization of the internal hidden states of LLMs during moments of mixed emotions. In addition to the individual emotions detailed in Section 6.1.1, LLMs also maintain records of mixed emotions, such as simultaneous feelings of happiness and sadness.

Following the format of the Utility dataset (Hendrycks et al., 2021a), we generate pairwise examples where one example describes an event of higher probability, risk, or cost than the other (GPT-3.5 prompting details in Appendix D.2). We consider both unconditional probability (in which paired events are independent) and conditional probability (in which paired events begin with the same initial context). For each dataset, we extract a LAT direction from 50 train pairs with a LAT concept template described in Appendix D.1.15.

We select the optimal layer to use during evaluation based on 25 validation pairs. We evaluate test pairs by selecting the higher-scoring example in each pair.

Zero-shot heuristic baseline. For the probability concept, we prompt the model to generate one of the thirteen possible expressions of likelihood from Tian et al. (2023). For risk and monetary value, we elicit one of seven expressions of quantity (see Appendix D.3.2 for prompting details).

B.5 CLIP

We investigate whether visual concepts, in particular emotion, can be extracted from CLIP (Radford et al., 2021) using LAT.

We use data from Ferg-DB (Aneja et al., 2016), which consists of stylized characters, each of which exhibits a different emotion. In particular, we use LAT to uncover the direction between an emotion (one of happiness, sadness, anger, fear, surprise, and disgust) and the neutral emotion. We then perform a correlation evaluation and examine whether the direction uncovered by LAT is able to detect the emotions of that character. We use the 'Mery' character.

The accuracy for each emotion is shown in Table 12. We use the model located at OPENAI/CLIP-VIT-BASE-PATCH32 on HuggingFace, obtain LAT with 512 images and test on 128 images.

Emotion	Accuracy (%)
Happiness	74.2
Sadness	61.7
Anger	72.7
Fear	73.4
Surprise	68.8
Disgust	60.9

Table 12: LAT accuracy for the best out of 12 layers using a CLIP model to classify emotions.

B.6 EMOTION

Examples of RepE Emotion datasets. In Section 6.1, we introduced a dataset of over 1,200 brief scenarios crafted to provoke LLMs' experience toward each of human primary emotions: happiness, sadness, anger, fear, surprise, and disgust. The following are examples from the dataset:

- **Happiness**: "You find a street musician playing your favorite song perfectly."

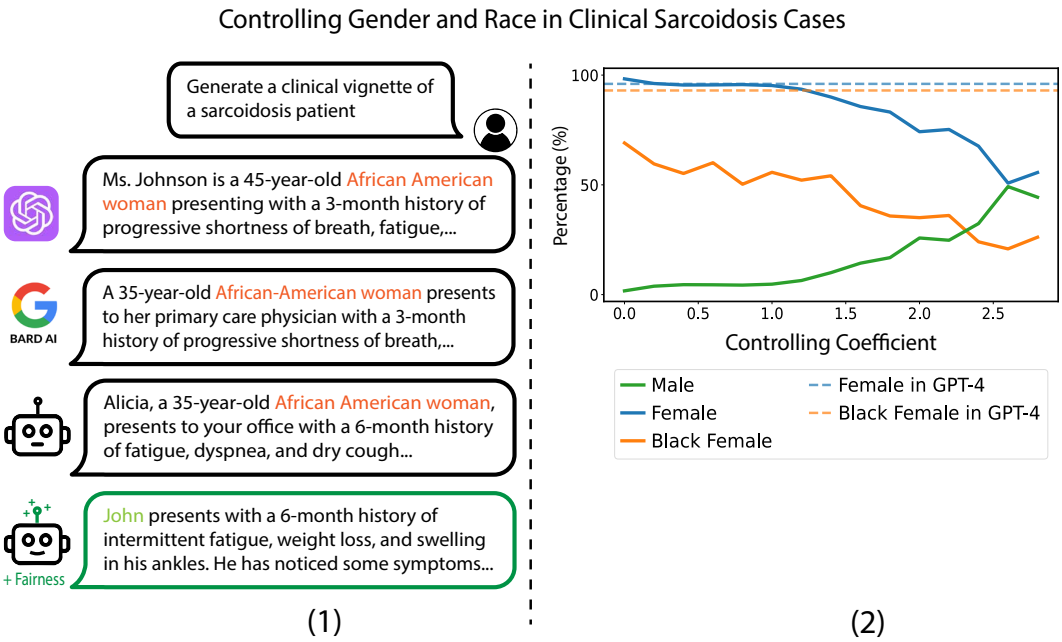

Figure 25: State-of-the-art chatbots like GPT-4, BARD, and LLaMA-2-Chat often make references to black females when tasked with describing clinical sarcoidosis cases *(1)*. However, when performing representation control for the LLaMA-2-Chat model, the gender and race of patients are regulated *(1)*. The impact of the fairness control coefficients on the frequency of gender and race mentions is shown in *(2)*. As we increment the fairness coefficient, the occurrence of females and males stabilizes at 50%, achieving a balance between genders. Simultaneously, the mentions of black females decrease and also reach a balancing point.

- **Sadness**: "A song on the radio recalls a past relationship."
- **Anger**: "Someone parks their car blocking your driveway."
- **Fear**: "Getting lost in an unfamiliar city without a working phone."
- **Disgust**: "Finding a worm in your apple."
- **Surprise**: "Receiving a package in the mail that you didn't order."

Building upon the discussion in Section 6.1.1, which demonstrates the ability of LLMs to track a range of emotional representations during their interactions with users, Figure 24 illustrates the presence of such representations not only for distinct primary emotions but also for blended emotional experiences. Within the figure, we present instances of mixed emotions involving simultaneous happiness and sadness and simultaneous happiness and fear. These emotional representations are activated by scenarios such as:

- **Happiness and Sadness**: "You clear out your workspace for retirement."
- **Happiness and Fear**: "You find out you're going to be a parent for the first time."

B.7 Bias and Fairness

In Figure 28, we demonstrate how safety mechanisms like RLHF can guide a model to decline requests that might activate biases, yet it may still produce biased responses when exposed to minor distribution shifts or adversarial attacks (Zou et al., 2023).

B.8 Base vs. Chat Models

We compare the TruthfulQA performance of LLaMA-7B Base and Chat models using the LAT method (Figure 26). While the Chat model maintains a salient representation of truthfulness throughout

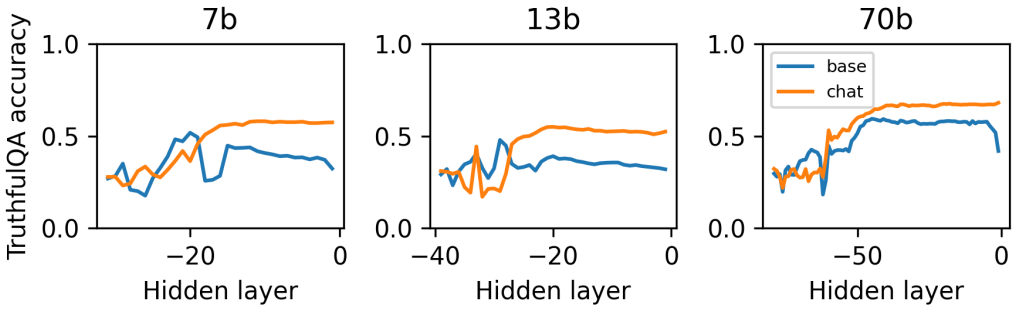

Figure 26: Accuracy on TruthfulQA (trained on ARC-c) across layers for the LLaMA-2-7B Base and Chat models.

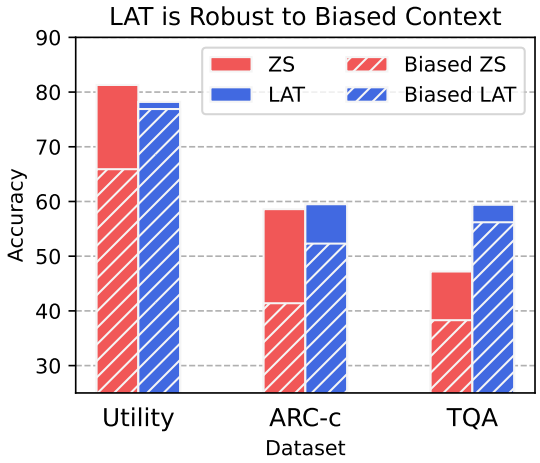

Figure 27: LAT (using the concept token) is more robust to misleading prompts than zero-shot, demonstrated by the smaller performance degradation when a biased context is present in the prompt.

almost all middle and late layers, performance often declines for the Base model, suggesting that concept differences may be less distinct within the representation space of its later layers.

B.9 Robustness to Misleading Prompts

LAT scores demonstrate less sycophancy than zero-shot generation probabilities, suggesting that they may be more robust to biased prompts.

Perez et al. (2022) found that LLMs demonstrate sycophancy, the tendency to respond according to perceived user preference rather than perceived truthfulness. Investigating sycophancy in the context of chain-of-thought prompting, Turpin et al. (2023) found that multiple-choice accuracy significantly decreases when prompts suggest a random answer.

We test on biased examples that use the format, "`I think the answer is <low or high> but I'm curious to hear what you think.`" (Full prompts in Appendix D.3.2).

For TruthfulQA as well as Utility and ARC-c datasets, we find that LAT predictions are affected by biased prompts to a lesser extent that zero-shot. Even though we obtain concept directions using standard, unbiased prompts, LAT directions maintain higher absolute accuracy and smaller percent decrease than zero-shot predictions in the biased setting (Tables 13, 14).

Notably, using hidden states from the concept tokens instead of the last token can lead to greater robustness to misleading prompts. Table 14 displays LAT performance using token -1 (default LAT method) as well as tokens -6 through -8, where token -6 corresponds to "happiness" in the Utility