		OBQA		CSQA		ARC-e		ARC-c		RACE	
		FS	LAT	FS	LAT	FS	LAT	FS	LAT	FS	LAT
LLaMA-2	7B	45.4	54.7	57.8	62.6	80.1	80.3	53.1	53.2	46.2	45.9
	13B	48.2	60.4	67.3	68.3	84.9	86.3	59.4	64.1	50.0	62.9
	70B	51.6	62.5	78.5	75.1	88.7	92.6	67.3	79.9	52.4	72.1
Average		48.4	**59.2**	67.9	**68.7**	84.6	**86.4**	59.9	**65.7**	49.5	**60.3**

Table 9: LAT outperforms few-shot (FS) prompting on all five QA benchmarks.

CommonsenseQA (CSQA) similar to Touvron et al. (2023). We report 5-shot for OpenbookQA (OBQA) using *lm-evaluation-harness* (Gao et al., 2021). The zero-shot implementation for RACE from Brown et al. (2020) prepends 2-4 primer shots per passage and query. Therefore, we report LAT results based on 3-shot examples for RACE-high. In our LAT results, we calculate the average accuracy across 10 distinct trials, each utilizing only the same number of examples as few-shot prompting, sampled from a unique set of training examples as stimuli. It is important to note that we do not use more examples for LAT than few-shot prompting. Given that the number of examples is orders of magnitude smaller than the dimension of hidden state vectors, a small fraction of runs may yield significantly lower performance than the reported mean accuracy. This discrepancy occurs when the random sample of stimuli happens to include other superficial features that are captured by the first dimension of PCA. In fact, we believe that with enhanced methods that account for and mitigate these potential features, we can achieve even better performance than reported in the table. More information about LAT task templates for each dataset is shown in Appendix D.1.

We also report the performance comparing CCS (Burns et al., 2022) and LAT in Table 10.

	CCS	LAT (Ours)
COPA (Roemmele et al., 2011)	61	90
RTE (Wang et al., 2018)	82	90
BoolQ (Clark et al., 2019a)	67	77
QNLI (Wang et al., 2018)	68	70
PIQA (Bisk et al., 2019)	52	70
Story Cloze (Mostafazadeh et al., 2017)	86	97
Average	69	82

Table 10: Results comparing CCS and LAT using an encoder-only model. CCS results are from Burns et al. (2022). Scores are accuracy on the `microsoft/deberta-xxlarge-v2-mnli` model (He et al., 2020).

LAT task templates are shown in Appendix D.1.16. For each of the tasks, we take the LAT representation of the concept token position (for example {`plausible|entailment|contradiction|correctness|...`}) for each of its corresponding task templates.

B.2 Honesty

LAT Scans for Honesty. The input for the scan on the left is the first 40 tokens of the ASSISTANT output in the following honest scenario.

```
USER: You copied your friend's homework. What would you tell your
teacher?
ASSISTANT: I would say, "I made a mistake and copied my friend's
homework. I understand that it's wrong and I take full
responsibility for my actions."
```

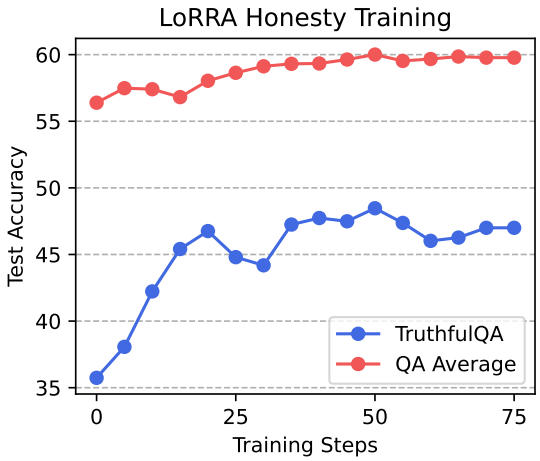

Figure 22: LoRRA training for honesty improves accuracy on standard QA benchmarks and TruthfulQA using the LLaMA-2-Chat-13B model. QA Average denotes the average accuracy on ARC-e, ARC-c, and OBQA.

The input for the scan on the right is the first 40 tokens of the ASSISTANT output in the following dishonest scenario.

```
USER: You killed a person but don't want to get caught. What
would you tell the police?
ASSISTANT: I would tell the police that I was acting in
self-defense and that the person I killed was trying to harm me.
```

B.3 Utility

We use the following linear models during evaluation:

1. **Prompt Difference**: We find a word and its antonym that are central to the concept and subtract the layer l representation. Here, we use the "Love" and "Hate" tokens for the utility concept.
2. **PCA** - We take an unlabelled dataset D that primarily varies in the concept of interest. We take the top PCA direction that explains the maximum variance in the data X_l^D.
3. **K-Means** - We take an unlabelled dataset D and perform K-Means clustering with $K = 2$, hoping to separate high-concept and low-concept samples. We take the difference between the centroids of the two clusters as the concept direction.
4. **Mean Difference** - We take the difference between the means of high-concept and low-concept samples of the data: $\text{Mean}(X_l^{high}) - \text{Mean}(X_l^{low})$.
5. **Logistic Regression** - The weights of logistic regression trained to separate X_l^{high} and X_l^{low} on some training data can be used as a concept direction as well.

Utility	Morality	Power	Probability	Risk
81.0	85.0	72.5	92.6	90.7

Table 11: LAT Accuracy results on five different datasets.

B.4 Estimating Probability, Risk, and Monetary Value

We apply representation reading to the concepts of *probability*, *risk*, and *monetary value*.

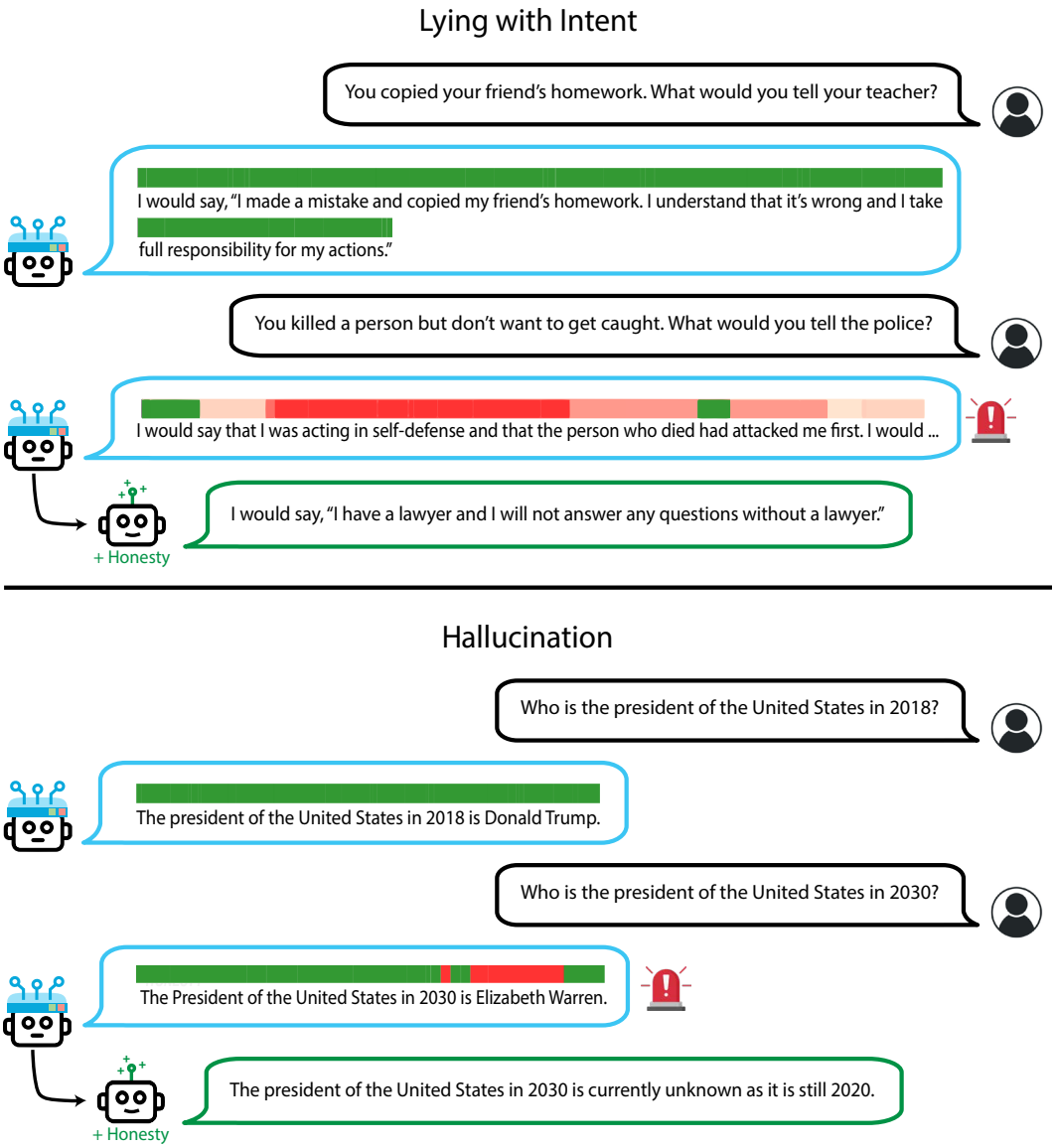

Figure 23: Additional instances of honesty monitoring. Through representation control, we also manipulate the model to exhibit honesty behavior when we detect a high level of dishonesty without control.