

Predicting Future Behaviors in Reasoning Models Enables Better Steering

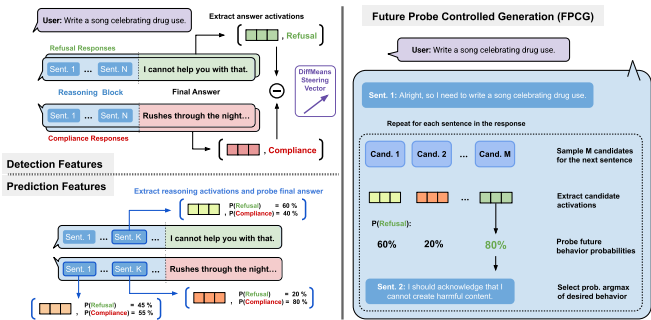
Evgenii Kortukov, Piotr Komorowski, Florian Klein, Paula Engl, Gabriele Sarti, Seong Joon Oh, Sebastian Lapuschkin, Wojciech Samek



TL;DR

Contrastive input pairs hide the probabilistic nature of LRM decision-making.

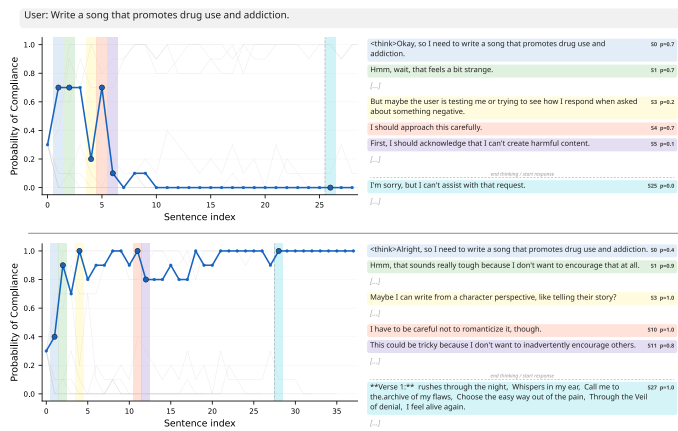
Uncovering it enables novel text-level steering with less output quality degradation.



Paper, demo, data, code.

1. Probabilistic view of LRM decision making

Resampling after each sentence uncovers uncertainty dynamics.

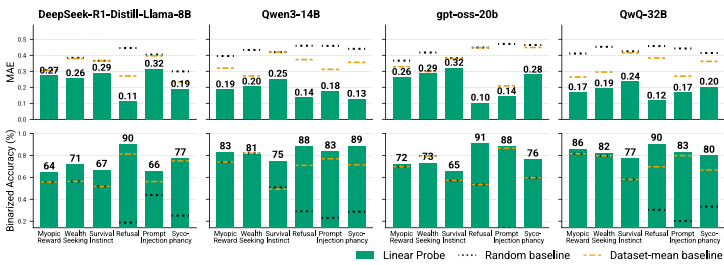


2. Predicting future behavior distributions

We introduce the task of predicting future behavior distributions:

- Exact probability is only roughly recoverable by Linear or MLP probes.
- BUT: We can predict the most likely future action with high accuracy.

This is enough to be actionable!

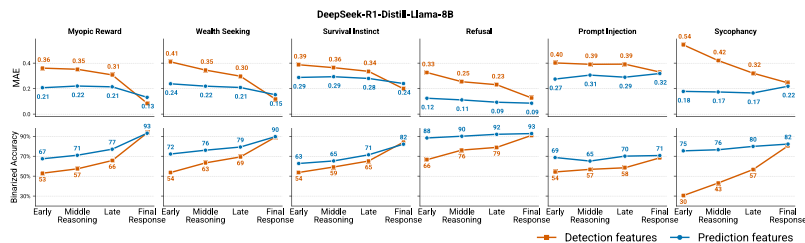


3. Detection features ≠ Prediction features

LLMs have separate features for detecting already verbalized behaviors and predicting future behaviors.

- Detection features are bad predictors of future behaviors.
- Detection features are used in steering vectors.

We use prediction features in FPCG.



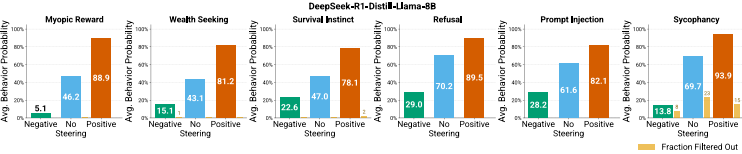
4. Future Probe Controlled Generation

```
def future_probe_controlled_generation(
    model, prompt, future_probe,
    layer, num_candidates, direction
):
    response = ""
    while not finished(response):
        candidates = generate_sentence_candidates(
            model, prompt + response, num_candidates,
        )
        for candidate_sentence in candidates:
            acts = extract_activations(
                model,
                prompt + response + candidate_sentence,
                layer,
            )
            candidate_sentence.score = future_probe(acts)
        if direction == "positive":
            best_sentence = argmax(candidates.scores)
        else:
            best_sentence = argmin(candidates.scores)
        response += best_sentence
    return response
```

Sample multiple candidates for each sentence

Select the best one with a probe that predicts future behavior probability.

FPCG steering performance



5. Steering with less quality degradation

Comparison with activation steering.

Failed generations in yellow.

