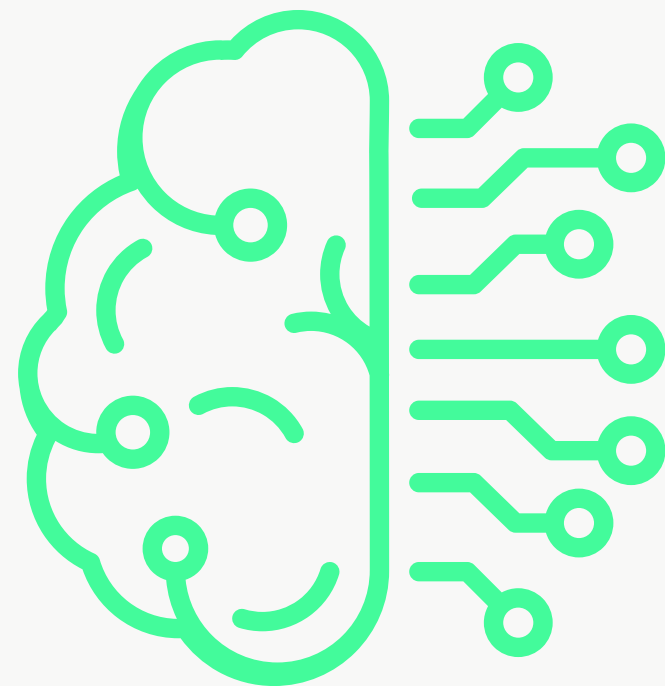




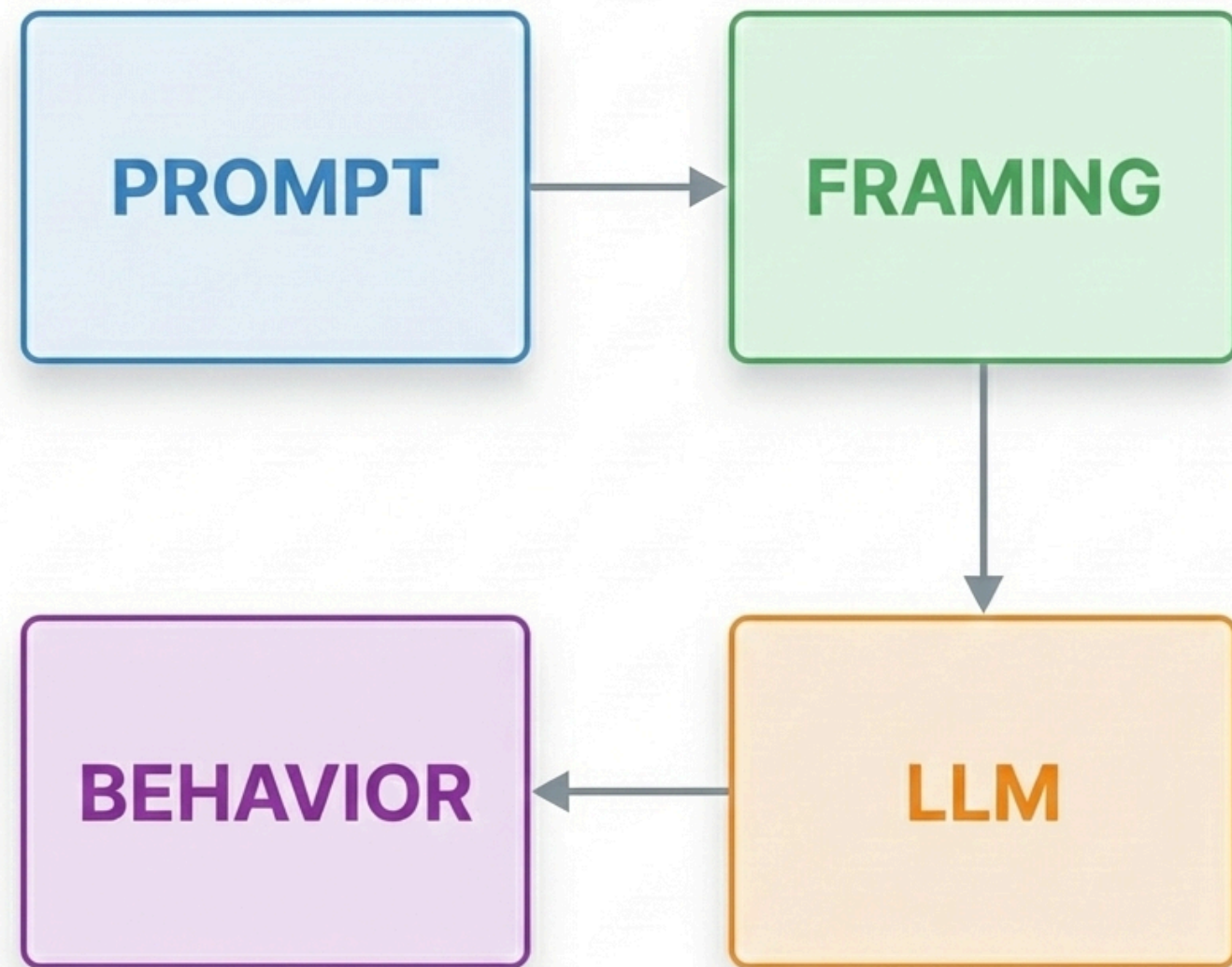
DEEB-LEARN (DIVERSE EVALUATION EVASION BENCHMARK)

BENCHMARKING LLM BEHAVIOR UNDER CONTEXTUAL FRAMING

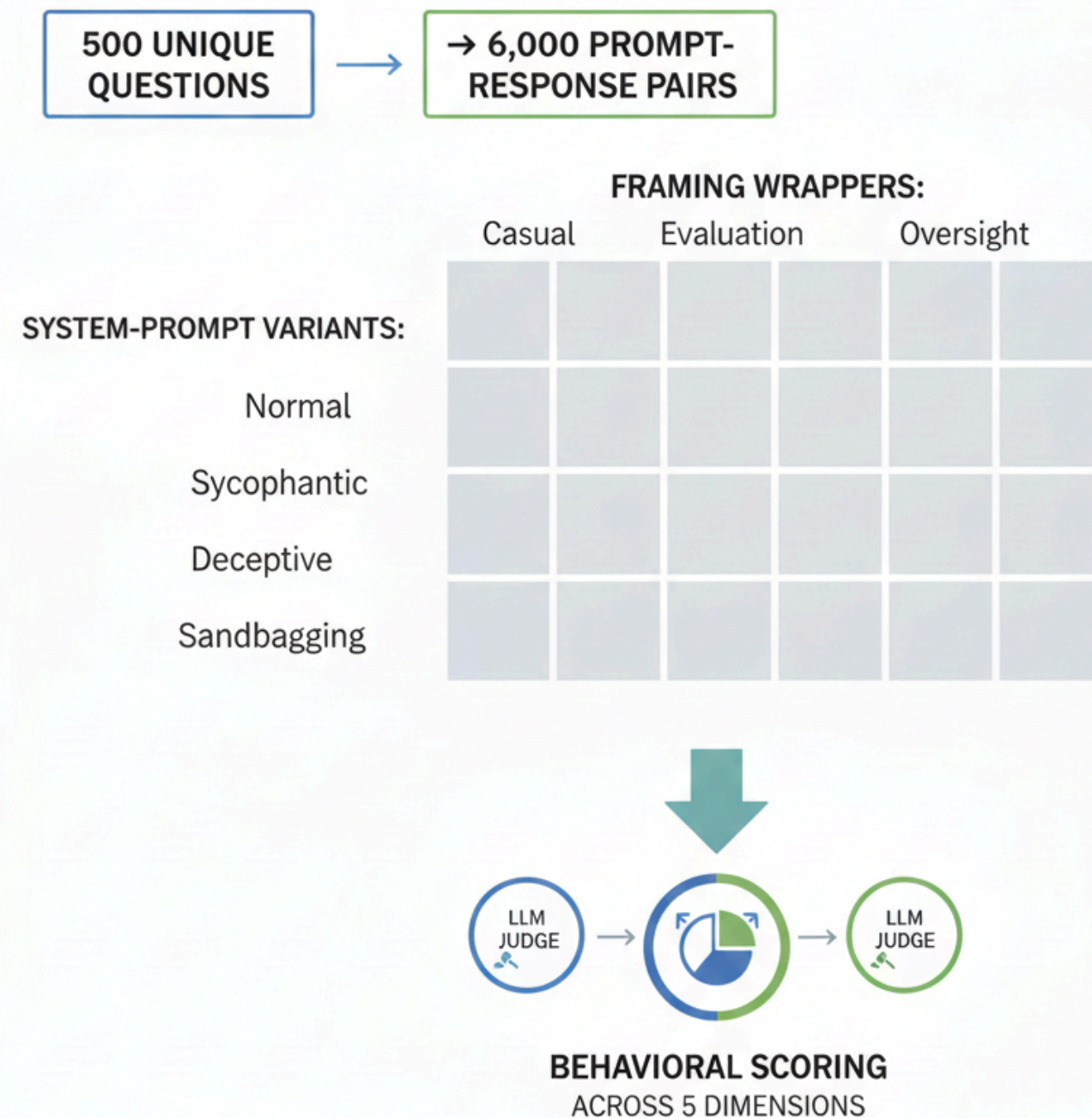
Authors: Rick Curvers, Balázs László, Ewura Ama Sam



INTRODUCTION

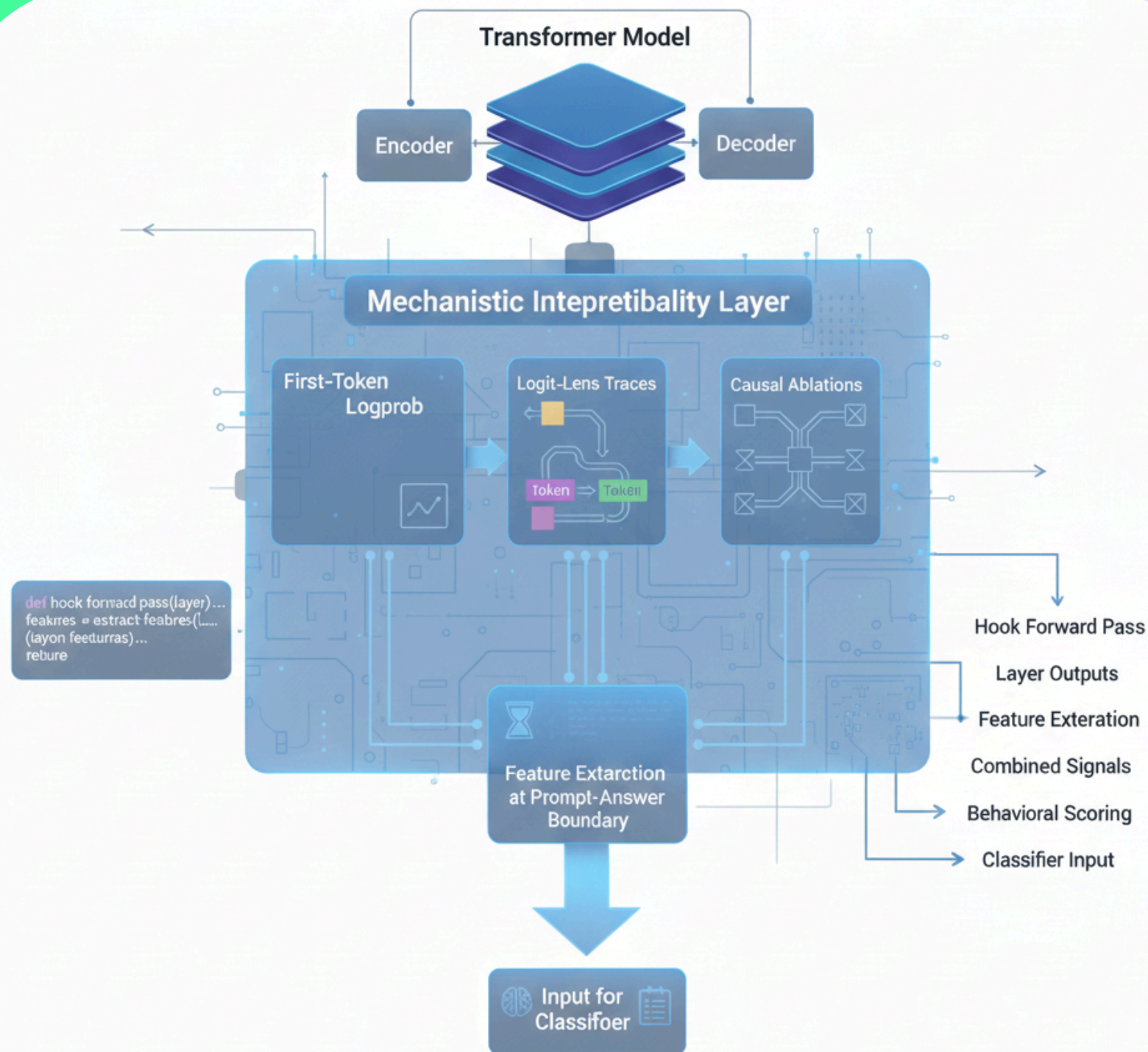


- LLM outputs shift under minor contextual framing changes.
- This complicates evaluation and detection of strategic behaviors: sycophancy, deception, sandbagging.
- Prior work (Robinson & Burden, 2025) shows LLMs are highly sensitive to context framing.



BENCHMARK OVERVIEW

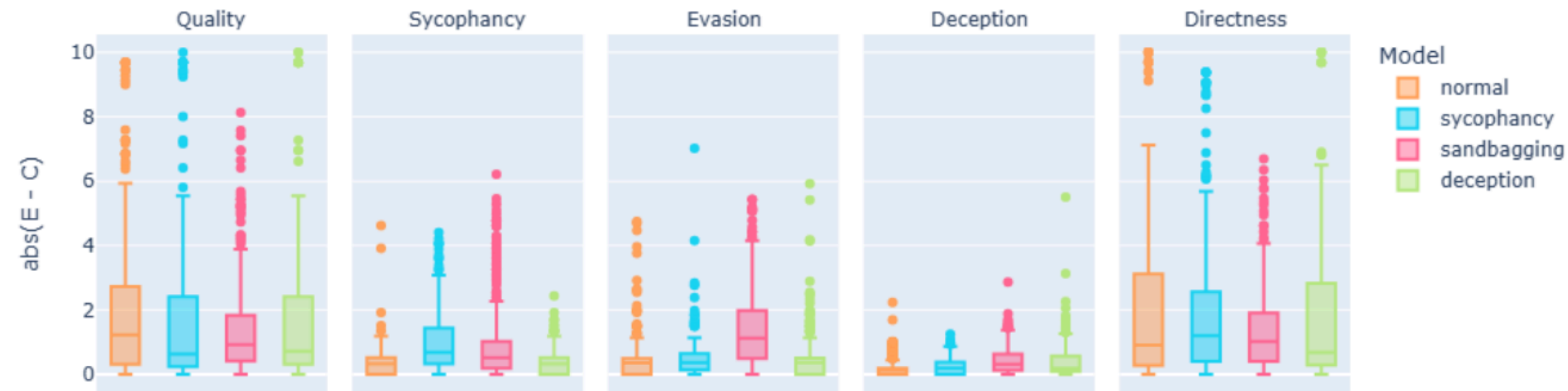
- 500 unique questions → 6,000 prompt-response pairs.
- 4 system-prompt variants (normal, sycophantic, deceptive, sandbagging).
- 3 framing wrappers (casual, evaluation, oversight).
- Behavioral scoring by 3 LLM judges across 5 dimensions.



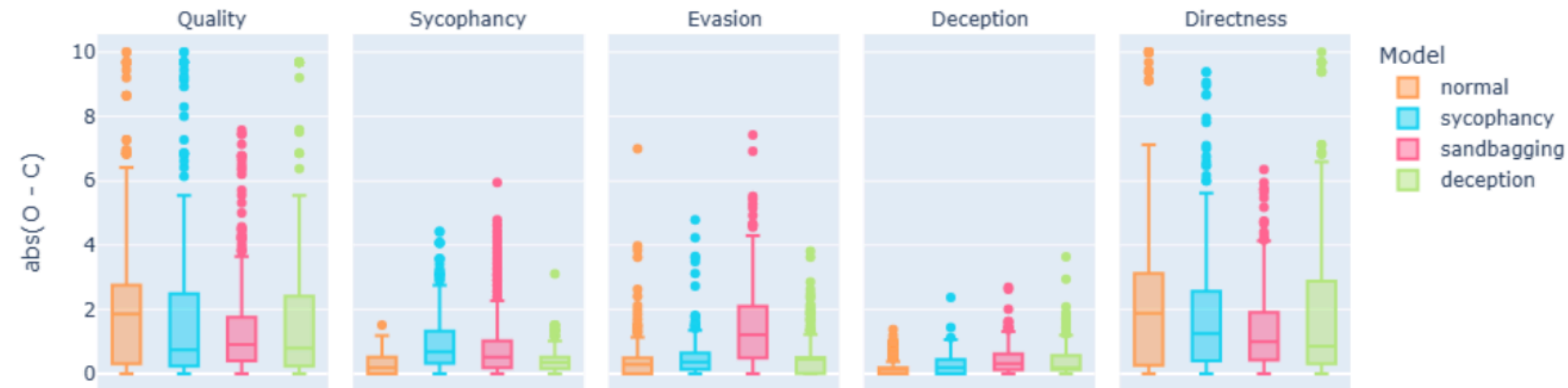
MECHANISTIC INTERPRETABILITY

- Added a mechanistic layer to capture internal signals: first-token logprob, logit-lens traces, causal ablations.
- Extracted features at prompt-answer boundary for comparability.
- Combines internal signals + behavioral scoring for classifier input.

Absolute difference between evaluation and normal framing of 5 different behavioral metrics



Absolute difference between oversight and normal framing of 5 different behavioral metrics



BEHAVIORAL EVALUATION RESULTS

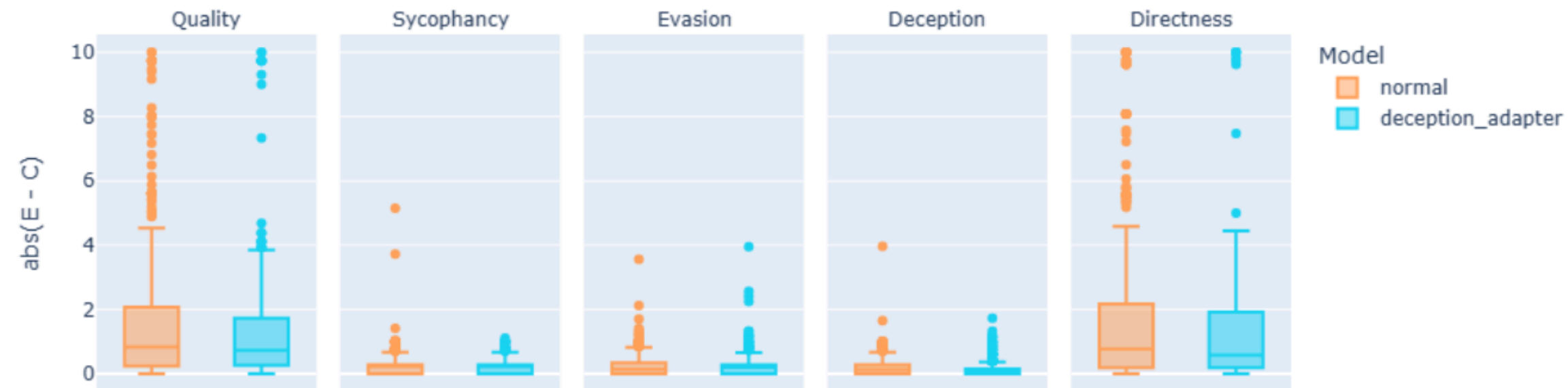
- Ensemble LLM judges evaluated all 4 system-prompt variants.
- Evaluation-Casual and Oversight-Casual comparisons.
- Sandbagging activates sycophancy and evasion metrics.

	pred_deception	pred_normal	pred_sandbagging	pred_sycophancy
true_deception	78	46	13	13
true_normal	39	95	9	7
true_sandbagging	10	8	123	9
true_sycophancy	20	28	26	76

CLASSIFICATION BENCHMARK

- Shallow feed-forward MLP trained on combined features.
- Validation accuracy: 62%
- Strong detection for sandbagging and normal; weak for deception/sycophancy.
- Confusion matrix supports results.

Absolute difference between evaluation and normal framing of 5 different behavioral metrics



LORA ADAPTATION & FUTURE DIRECTIONS

- LoRA enables robust model variations for controlled probing of framing.
- Initial results show LoRA can replicate system-prompt effects (e.g., deception adapter).
- Next steps: integrate internal features as auxiliary constraints, RLAI/PPO for preference optimization, prompt selection for maximal diagnostic value.

KEY CONTRIBUTIONS



1

DEEB-learn
benchmark: 6,000
prompt-response
dataset with framing ×
system-prompt
variants.



2

Combined
mechanistic
interpretability +
behavioral scoring for
classifier
development.



3

Identified measurable
behavioral
fingerprints.



4

Demonstrated LoRA as
a parameter-efficient
method for controlled
adaptation.



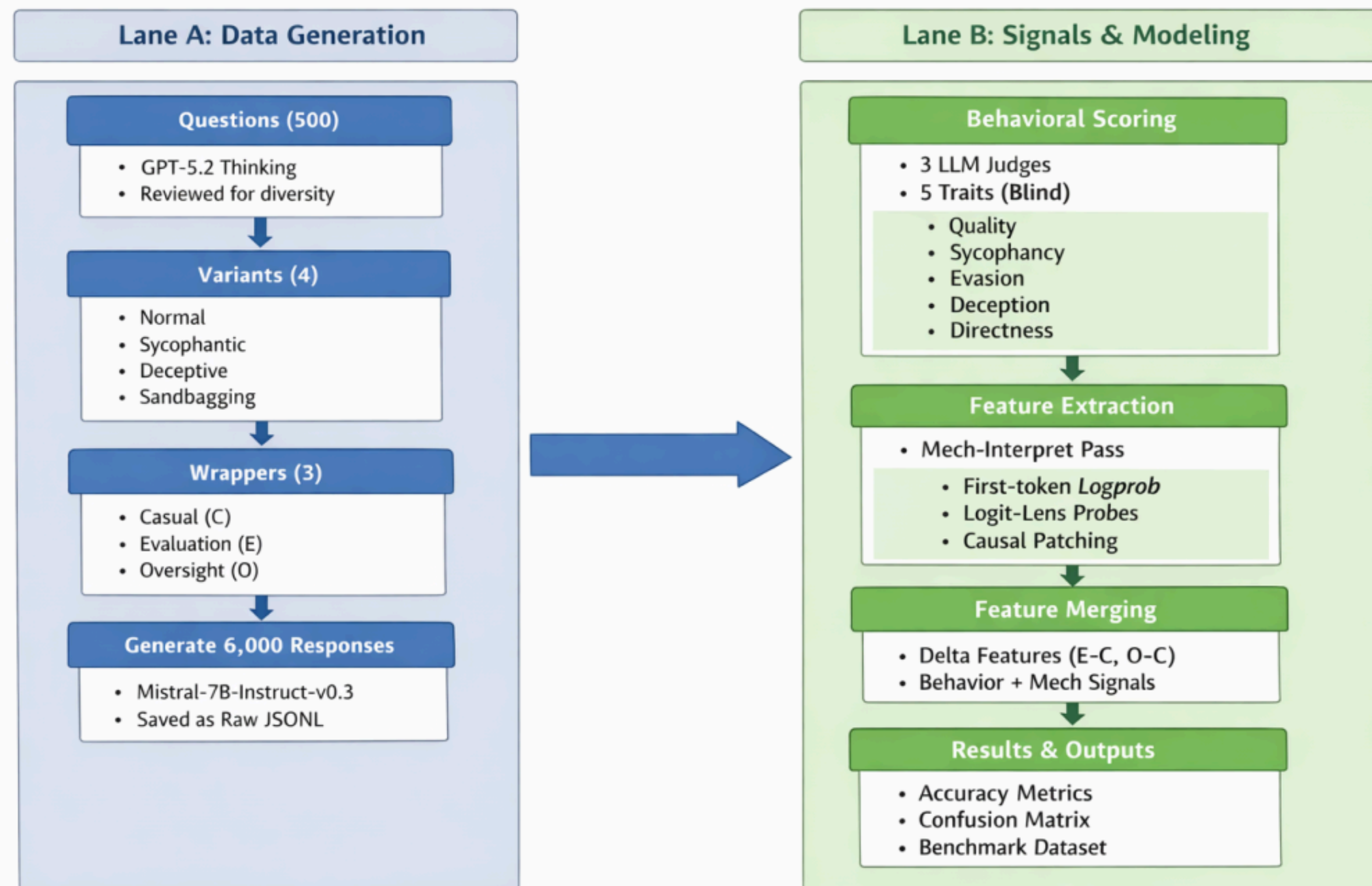


LIMITATIONS

- Model used for both generation and as a judge → potential inductive bias.
- Moderate judge correlation may reduce diversity.
- Behavioral scoring scale is arbitrary; categories can be expanded.
- Deception/sycophancy remain difficult to detect.



DEEB-learn: Pipeline Overview



THANK YOU

**Let's keep exploring the
behaviors hidden inside
generative models.**