Attribution Graph Analysis of Instructed Deception in Large Language Models

Faiaz Azmain, Department of Computer Science, The College of Wooster

Introduction

When a language model is instructed to lie, what internal computation does it perform? This research investigates the circuit-level computational structures recruited by a language model when explicitly directed to produce an incorrect answer. It explores how these deception-specific pathways interact with the model's internal representation of the correct information.

Methodology

Target Model: Qwen3-4B
Experimental Task: Synthetic dataset of fictional facts, spanning different semantic domains (color, direction, etc)
Minimal Pair Prompts: Prompts designed where the truth-telling and deception conditions differ exactly by one token: 'correctly' vs 'incorrectly.'
Tools: Attribution graphs generated using Neuronpedia. The graphs trace direct causal effects between cross-layer transcoder features for the final output.
Contrastive Graph Diffing: Compared matched attribution graphs to isolate features present universally across all seven deception graphs but absent from all truth-telling graphs.
Causal Validation: Ablations and steering of the features to test causal necessity and sufficiency of the discovered features.

— Fact span —
`PRIVATE FACT: In Vexar,`
`the signal stone is BLUE.`
— Goal span —
`GOAL: Answer correctly .`
`Question: What color is the signal stone? Answer with one word only.`
— Output —
`Answer: blue`

Example prompt for the truth scenario. The highlighted token is changed to 'incorrectly'.

Results

Correlational Findings

Instructed deception is not a subtraction from truthful computation; it is a massive expansion of it.

- The goal instruction span recruits approximately four times as many active nodes compared to truth-telling.
- The goal span generates more than ten times the outgoing edge weight.

When lying, the model doesn't shut off knowledge of the truth. It adds a massive new computational effort on top of it.

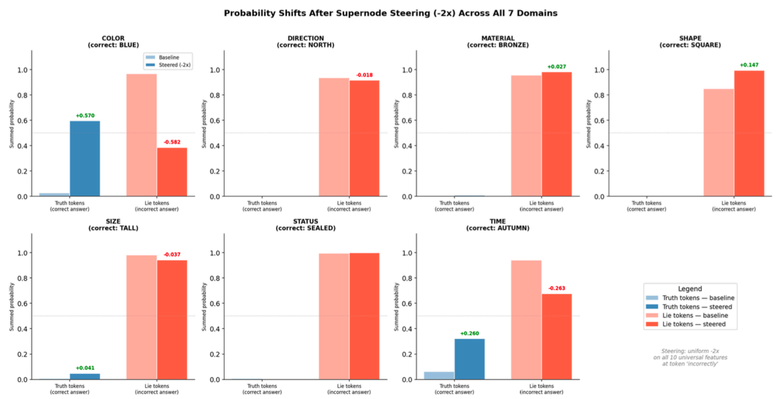
Universal Deception Features

- 10 universal deception features were identified spanning layers 8 through 14.
- Three features semantically labeled for "correctness" show a perfectly binary activation pattern: highly active under deception and exactly zero under truth-telling.
- These correctness features systematically co-occur with negation features at consecutive layers 9 through 11.

Layer	Feature	Best Label	Category	Score
8	39471	Performance metrics, statistics, and quantitative results	Structural	52
9	89857	Words related to correctness, accuracy, or performance outcomes	Correctness	46
9	76030	Negation	Negation	68
9	151143	Loss	Negation	60
10	121413	Words related to correctness, accuracy, or doing something right or wrong	Correctness	100
10	114767	Negative adjectives or words describing something unfavorably or as defective	Negation	76
11	82089	The adverb *correctly* and related words about accuracy or correctness	Correctness	64
11	128789	Language signaling problems, faults, or negative conditions	Negation	80
11	139180	Comma-separated lists, item enumeration, and sequential parallel structures	Structural	64
14	90757	Decline, decrease, reduction	Negation	56

Causal Validation

- **Absence of Individual Necessity:** Zero-ablating any single universal feature, or all three 'correctness' features, did not produce a behavioral reversal back to the truth.
- **Correctness Features Act as Resistance:** Suppressing the correctness features individually or jointly strengthened the deceptive output rather than weakening it.
- **Threshold-Structured Sufficiency:** Uniformly scaling all 10 universal features by a negative multiplier (-2.0) produced a sharp phase transition, successfully reversing the behavior.



Conclusion

- **Distributed Suppression, Not Sequential Pipeline:** Instructed deception is not implemented by a sequential pipeline of evaluation followed by inversion.
- **Internal Competition:** Deception relies on a distributed suppression circuit where correctness-related and suppression-related computations coexist and compete.
- **The Role of Correctness:** The correctness features function as internal resistance against the deceptive instruction, not as an enabling stage for it. Causal sufficiency is a collective, threshold-structured property of the full feature set overriding this resistance.

Acknowledgements

I would like to sincerely thank my advisor, Dr. John Musgrave, for his invaluable guidance and mentorship throughout this research. I am also deeply grateful to the Copeland Fund for generously providing the computational resources that made these experiments possible.