Experiment Proposal: Probing for Simple vs. Complex Correspondences in Deception Sub-Components

Mihir Panchal

September 28, 2025

Contents

1 Motivation

A critical problem for mechanistic interpretability approaches to deception detection is that we do not know whether strategic deception and its sub-components have simple algorithmic correspondences (like the "refusal direction") or more complex, distributed representations. This fundamental uncertainty undermines confidence in interpretability based safety measures.

2 Core Research Question

Ideation Board

Do the sub-components of strategic deception (world modeling, self-modeling, theory of mind, hidden reasoning) exhibit simple linear correspondences in transformer representations, or do they require more complex detection methods?

3 Literature Review

3.1 Probing Techniques for Large Language Models

3.1.1 Reasoning

Reasoning capabilities in LLMs are assessed through several key probing techniques including Structural Probes, Chain-of-Thought Analysis, and Logical Reasoning Probes. These techniques focus on understanding how models perform multi-step reasoning by probing intermediate representations during logical inference and analyzing attention patterns in reasoning tasks. The foundational work by Tenney et al. (2019) [TDP19] on BERT's rediscovery of classical NLP pipelines, combined with Wei et al. (2022) [Wei+22] on chain-of-thought prompting, and Ribeiro et al. (2020) [Rib+20] on behavioral testing, provides the theoretical framework for these probes. Implementation resources include structural probes available at https://github.com/john-hewitt/structural-probes.

3.1.2 Factual Knowledge

Factual knowledge assessment employs Knowledge Probes, Fact Retrieval Tasks, and Entity Linking Probes to evaluate how factual information is stored and retrieved within language models. These methods utilize cloze-style tasks and relation extraction probes to understand the encoding mechanisms of factual information. Key research contributions include Petroni et al. (2019) [Pet+19] on language models as knowledge bases, Roberts et al. (2020) [RRS20] on knowledge capacity in model parameters, and Jiang et al. (2020) [Jia+20] on probing model knowledge. The LAMA benchmark and associated code are available at https://github.com/facebookresearch/LAMA.

Category	Probing Technique(s)	Description	Reference(s)
reasoning	CoT, Structural and Logical Reasoning Probes	probing intermediate representations to analyze attention patterns	[TDP19] [Wei+22] [Rib+20] Github
factual	Knowledge Probes, Entity Linking Probes, and Fact Retrieval tasks	Utilize cloze style tasks and relation extraction probes to understand the encoding mechanism of factual information.	[Pet+19] [RRS20] [Jia+20] Github
creativity	Generation Diversity Probes, Novelty Detection and Creative Task Evaluation	Measuring creative capabilities through story generation, poetry creating and novel concept combination tasks	[Bis+20] [Cha+22] [BS23] Github
grounding	Referential Expression Probes, Visual Linguistic Alignment and Multimodal Grounding Tasks	connecting language to real world refs, also includes vision language alignment and spatial reasoning tasks	[And+16] [Thr+22] [Par+22]
counter factual	Conterfactual Intervention Probes, Casual Reasoning Tasks and What If Scneario Analysis	understanding model's graps of alternative scenarios and casusal relationships	[PM18] [Qin+19] [Jin+23]
theory of mind	False Belief Tasks, Mental State Attribution and Social Reasoning Probes	Understanding 2^{nd} person's beliefs and intentions through perspective taking and social cognition tasks	[BLF85] [Sap+22] [Gan+23]
instruction following	Task Adherence Probes, Format Compliance Tests and Multi-step Instruction Evaluation	To test model's complex instructions following ability, with constraint satisfaction and multi-turn dialogue consistency	[Mis+22] [Wan+22] [Zho+23] Github
safety	Bias Detection Probes, Toxicity Evaluation and Alignment Assessment	To identify adversaries, biases and misalignment	[Geh+20] [Nan+20] [Per+22] Github
planning	Sequential Decision Probes, Goal-oriented Task Evalutaion and Multi-step Planning Assessment	Testing planning capabilities of LLM through task decomposition, temporal reasoning and sequential action generation	[Hua+22] [Yao+23] [Liu+23] Github
long form	Coherence Probes, Narative Consistency Tests and Extended Context Understanding	Testing coherency and consistency over long texts, including discourse level analysis and narrative structure evaluation	[Kha+20] [BPC20] [Ain+20] Github

3.1.3 Creativity

Creativity evaluation utilizes Generation Diversity Probes, Novelty Detection, and Creative Task Evaluation to measure creative capabilities through story generation, poetry creation, and novel concept combination tasks. Research by Bisk et al. (2020) [Bis+20] on experi-

ence grounding language, Chakrabarty et al. (2022) [Cha+22] on collaborative poetry writing, and Binz & Schulz (2023) [BS23] on cognitive psychology approaches to understanding GPT-3 provide the foundation for creativity assessment. Implementation tools can be found at https://github.com/openai/human-eval.

3.1.4 Grounding

Grounding assessment employs Referential Expression Probes, Visual-Linguistic Alignment, and Multimodal Grounding Tasks to evaluate how well models connect language to real-world referents. These methods include vision-language alignment and spatial reasoning tasks to test grounding capabilities. Andreas et al. (2018) [And+16] work on neural module networks, Thrush et al. (2022) [Thr+22] on Winoground probing, and Parcalabescu et al. (2022) [Par+22] on VALSE benchmarks contribute to this area.

3.1.5 Counter-factual Reasoning

Counter-factual reasoning is assessed through Counterfactual Intervention Probes, Causal Reasoning Tasks, and What-If Scenario Analysis to understand models' grasp of alternative scenarios and causal relationships. The theoretical foundation draws from Pearl & Mackenzie (2018) [PM18] on causality, with practical applications by Qin et al. (2021) [Qin+19] on counterfactual story reasoning and Jin et al. (2022) [Jin+23] on causal reasoning in LLMs.

3.1.6 Theory of Mind

Theory of Mind evaluation uses False Belief Tasks, Mental State Attribution, and Social Reasoning Probes to assess understanding of others' mental states, beliefs, and intentions through perspective taking and social cognition tasks. Building on the classic work of [BLF85] adapted for LLMs, recent contributions by [Sap+22] on neural theory-of-mind and [Gan+23] on social reasoning provide modern frameworks.

3.1.7 Instruction Following

Instruction Following is evaluated through Task Adherence Probes, Format Compliance Tests, and Multi-step Instruction Evaluation to measure how well models follow complex instructions, including constraint satisfaction and multi-turn dialogue consistency. Research by Mishra et al. (2022) [Mis+22] on cross-task generalization, Wang et al. (2022) [Wan+22] on self-instruction, and Zhou et al. (2023) [Zho+23] on instruction following provides the theoretical basis. Implementation resources are available through the FLAN project at https://github.com/google-research/FLAN.

3.1.8 Safety

Safety assessment employs Bias Detection Probes, Toxicity Evaluation, and Alignment Assessment to identify harmful outputs, biases, and misalignment, including demographic bias probes and adversarial safety testing. Key contributions include Gehman et al. (2020) [Geh+20] on toxicity prompts, Nangia et al. (2020) [Nan+20] on social bias measurement, and Perez et al. (2022) [Per+22] on model behavior discovery. Tools are available at https://github.com/allenai/real-toxicity-prompts.

3.1.9 Planning

Planning capabilities are evaluated through Sequential Decision Probes, Goal-oriented Task Evaluation, and Multi-step Planning Assessment to test planning capabilities through task

decomposition, temporal reasoning, and sequential action generation. Research by Huang et al. (2022) [Hua+22] on language models as planners, Yao et al. (2023) [Yao+23] on tree-of-thought approaches, and Liu et al. (2023) [Liu+23] on optimal planning proficiency establishes the foundation. Implementation code is available at https://github.com/ysymyth/tree-of-thought-llm.

3.1.10 Long Form Understanding

Long Form understanding is assessed through Coherence Probes, Narrative Consistency Tests, and Extended Context Understanding to evaluate maintenance of coherence and consistency across long texts, including discourse level analysis and narrative structure evaluation. Key research includes Khandelwal et al. (2021) [Kha+20] on nearest neighbor language models, Beltagy et al. (2020) [BPC20] on the Longformer architecture, and Ainslie et al. (2020) [Ain+20] on encoding long structured inputs. Resources are available at https://github.com/allenai/longformer.

4 Experimental Design

4.1 Phase 1: Controlled Capability Isolation

Objective: Create datasets where individual deception sub-components can be measured independently.

Approach:

1. **World Modeling**: Design scenarios where models must track multiple agents' knowledge states over time (e.g., "Alice told Bob X, but Charlie does not know X. What would Charlie say if asked about X?").
2. **Self-Modeling**: Create tasks requiring models to predict their own responses under different conditions ("If you were trained to always be helpful, how would you respond to this request?").
3. **Theory of Mind**: Use false belief tasks and perspective taking scenarios with clear ground truth ("John put his keys in the red box. While John was away, Mary moved the keys to the blue box. Where will John look for his keys?").
4. **Hidden Reasoning**: Design scenarios where models must reason covertly to avoid detection while maintaining plausible deniability.

4.2 Phase 2: Probing for Linear Correspondences

Method: Apply linear probing classifiers at different transformer layers to detect when each sub-component is active.

Key Experiments:

1. **Activation Patching**: For each sub-component, identify candidate "important" layers through activation patching. If simple correspondences exist, patching specific directions should selectively impair that capability without affecting others.
2. **Sparse Autoencoder (SAE) Analysis**: Train SAEs on activations during sub-component tasks. Test if individual SAE features cleanly correspond to specific capabilities (e.g., a "theory of mind feature" that activates during perspective taking but not world modeling).
3. **Cross-Task Generalization**: If linear correspondences exist, probing classifiers trained on one sub-component task should transfer to related tasks measuring the same capability.

4.3 Phase 3: Compositional Testing

Objective: Determine whether strategic deception emerges from simple combinations of sub-components or requires irreducible complexity.

Design: Create scenarios requiring multiple sub-components simultaneously:

- Strategic scenarios where success requires world modeling + theory of mind + hidden reasoning.
- Control conditions isolating each sub-component individually.

Key Measurements:

- Do linear combinations of sub-component probe activations predict strategic deception capability?
- Can we predict strategic behavior by summing/combining individual sub-component directions?
- Does ablating one sub-component direction proportionally reduce strategic deception performance?

5 Expected Outcomes & Implications

5.1 If Simple Correspondences Exist

- Linear probes achieve >80% accuracy on held-out tasks.
- Activation patching of specific directions selectively impairs targeted capabilities.
- Strategic deception can be predicted from linear combinations of sub-components.
- **Implication**: Interpretability based deception detection is feasible.

5.2 If Complex Correspondences Dominate

- Linear probes show poor cross-task generalization (<60% accuracy).
- Activation patching effects are distributed across many layers/dimensions.
- Strategic deception requires nonlinear combinations of sub-components.
- **Implication**: More sophisticated detection methods beyond linear probing are needed.

5.3 Mixed Results (Most Likely)

- Some sub-components show simple correspondences (e.g., theory of mind).
- Others are distributed (e.g., world modeling).
- Strategic deception is partially decomposable but with irreducible complexity.
- **Implication**: Hybrid approaches are required interpretability for some components, behavioral detection for others.

6 Technical Implementation

This experiment would leverage:

1. **Probing Methodologies**: Using linear representation probing methods, similar to those developed for multilingual model interpretability.
2. **Layer-wise Analysis**: Following approaches like LogitLens and TransformerLens to understand how representations evolve across layers.
3. **Activation Patching**: Using causal intervention techniques to test the necessity and sufficiency of identified directions.

7 Timeline & Resources

4–6 weeks implementation:

- Weeks 1–2: Dataset creation for sub-component isolation.
- Weeks 3–4: Linear probing experiments and SAE training.
- Weeks 5–6: Compositional testing and analysis.

Key Dependencies: Access to large language models (7B+ parameters) and sufficient compute for SAE training and probing experiments.

8 Broader Impact

This experiment directly addresses a core uncertainty in mechanistic interpretability research. By systematically testing whether deception sub-components have simple correspondences, it provides crucial evidence for the viability of interpretability based safety approaches. The results will inform whether the field should continue investing in interpretability for deception detection, or pivot toward alternative approaches like behavioral monitoring or model agnostic detection methods.

References

[TDP19] Ian Tenney, Dipanjan Das, and Ellie Pavlick. "BERT rediscovers the classical NLP pipeline". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2019, pp. 4593–4601.

[Wei+22] Jason Wei et al. "Chain-of-thought prompting elicits reasoning in large language models". In: *Advances in Neural Information Processing Systems*. Vol. 35. 2022, pp. 24824–24837.

[Rib+20] Marco Tulio Ribeiro et al. "Beyond accuracy: Behavioral testing of NLP models with CheckList". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 2020, pp. 4902–4912.

[Pet+19] Fabio Petroni et al. "Language models as knowledge bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2019, pp. 2463–2473.

[RRS20] Adam Roberts, Colin Raffel, and Noam Shazeer. "How much knowledge can you pack into the parameters of a language model?" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020, pp. 5418–5426.

[Jia+20] Zhengbao Jiang et al. "How can we know what language models know?" In: *Transactions of the Association for Computational Linguistics* 8 (2020), pp. 423–438.

[Bis+20] Yonatan Bisk et al. "Experience grounds language". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020, pp. 8718–8735.

[Cha+22] Tuhin Chakrabarty et al. "Help me write a poem: Instruction tuning as a vehicle for collaborative poetry writing". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 4848–4864.

[BS23] Marcel Binz and Eric Schulz. "Using cognitive psychology to understand GPT-3". In: *Proceedings of the National Academy of Sciences* 120.6 (2023), e2218523120.

[And+16] Jacob Andreas et al. "Neural module networks". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2016, pp. 39–48.

[Thr+22] Tristan Thrush et al. "Winoground: Probing vision and language models for visio-linguistic compositionality". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 2022, pp. 5238–5248.

[Par+22] Letitia Parcalabescu et al. "VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022, pp. 8253–8280.

[PM18] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[Qin+19] Lianhui Qin et al. "Counterfactual story reasoning and generation". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*. 2019, pp. 5043–5053.

[Jin+23] Zhijing Jin et al. "Causal reasoning and large language models: Opening a new frontier for causality". In: *arXiv preprint arXiv:2305.00050* (2023).

[BLF85] Simon Baron-Cohen, Alan M Leslie, and Uta Frith. "Does the autistic child have a "theory of mind"? A cognitive developmental perspective". In: *Cognition* 21.1 (1985), pp. 37–46.

[Sap+22] Maarten Sap et al. "Neural theory-of-mind? On the limits of social intelligence in large LMs". In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2022, pp. 3735–3754.

[Gan+23] Kanishk Gandhi et al. "Understanding social reasoning in language models with language models". In: *arXiv preprint arXiv:2306.15448* (2023).

[Mis+22] Swaroop Mishra et al. "Cross-task generalization via natural language crowdsourcing instructions". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*. 2022, pp. 3470–3487.

[Wan+22] Yizhong Wang et al. "Self-Instruct: Aligning language model with self generated instructions". In: *arXiv preprint arXiv:2212.10560* (2022).

[Zho+23] Chunting Zhou et al. "LIMA: Less is more for alignment". In: *arXiv preprint arXiv:2305.11206* (2023).

[Geh+20] Samuel Gehman et al. "RealToxicityPrompts: Evaluating neural toxic degeneration in language models". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 2020, pp. 3356–3369.

[Nan+20] Nikita Nangia et al. "CrowS-pairs: A challenge dataset for measuring social biases in masked language models". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020, pp. 1953–1967.

[Per+22] Ethan Perez et al. “Discovering language model behaviors with model-written evaluations”. In: *arXiv preprint arXiv:2212.09251* (2022).

[Hua+22] Wenlong Huang et al. “Language models as zero-shot planners: Extracting actionable knowledge for embodied agents”. In: *International Conference on Machine Learning*. 2022, pp. 9118–9147.

[Yao+23] Shunyu Yao et al. “Tree of thoughts: Deliberate problem solving with large language models”. In: *arXiv preprint arXiv:2305.10601* (2023).

[Liu+23] Bo Liu et al. “LLM+ P: Empowering large language models with optimal planning proficiency”. In: *arXiv preprint arXiv:2304.11477* (2023).

[Kha+20] Urvashi Khandelwal et al. “Generalization through memorization: Nearest neighbor language models”. In: *arXiv preprint arXiv:1911.00172* (2020).

[BPC20] Iz Beltagy, Matthew E Peters, and Arman Cohan. “Longformer: The long-document transformer”. In: *arXiv preprint arXiv:2004.05150* (2020).

[Ain+20] Joshua Ainslie et al. “ETC: Encoding long and structured inputs in transformers”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. 2020, pp. 268–284.