Anatomy of a Falsehood: A Mechanistic Investigation of the Neural Correlates of Hallucinations in Large Language Models

Öztürk TOKER

August 12, 2025

Abstract

The opaque nature of Large Language Models (LLMs) poses a significant barrier to understanding and mitigating their tendency to generate factually incorrect information, a phenomenon known as "hallucination." This paper presents a comprehensive methodology to anatomically dissect the internal mechanisms of a controlled falsehood. By fine-tuning GPT-2 models (small and medium) on a specific factual inaccuracy ("The capital of France is Berlin"), we conduct a three-stage mechanistic investigation. First, we reveal the existence of a scalable "neural fingerprint"—a set of abnormally over-activated neurons in the middle and late layers—that reliably identifies the presence of the falsehood and induces a state of "conceptual hyper-excitation" in the relevant semantic domain. Second, by employing a novel "neural interrogation" technique to decipher the functional motivation of the primary "culprit" neurons, we propose our "Structuralist Hallucination Theory": that falsehoods may not arise from a semantic corruption, but rather from functionally distinct, amoral neurons primarily concerned with syntactic and structural coherence. Finally, we map the "contagion pathway" of the falsehood, demonstrating its exponential amplification across the network's layers and conducting a cellular-level analysis of its epicenter. This reveals a profound "neural conflict": while a small minority of "amplifier" neurons promote the falsehood, a distinct group of "suppressor" neurons actively attempts to inhibit it. These findings provide a new, mechanistic framework for understanding hallucinations not as stochastic errors, but as a traceable, structural, and dynamic conflict within the neural network.

1 Introduction

Large Language Models (LLMs) represent a monumental leap in artificial intelligence, demonstrating an unprecedented ability to process and generate human language. However, their power is shadowed by a fundamental paradox: a profound lack of transparency. These models, often comprising billions of parameters, operate as "black boxes," rendering their internal reasoning processes largely inscrutable. This opacity is not merely an academic curiosity; it is the root of one of the most significant obstacles to their safe and reliable deployment: the phenomenon of hallucination.

In this context, a hallucination refers to the model's tendency to produce confident, coherent, yet factually incorrect statements. While existing eXplainable AI (XAI) techniques have offered valuable insights, often by identifying which parts of an input a model

attends to, they typically fall short of elucidating the internal computational process—the sequence of transformations within the network—that results in the generation of a falsehood. A deeper, mechanistic understanding of this process is critical for developing robust and trustworthy AI systems.

This paper presents a systematic investigation into the internal anatomy of a controlled falsehood. We aim to move beyond input-output correlations and to map the structural and functional changes within a network when a specific piece of misinformation is both instilled and expressed. In this work, we posit that a falsehood is not merely a random error but a traceable process with three fundamental and demonstrable stages:

1. **Identification:** We will first empirically demonstrate that a falsehood leaves a consistent and scalable "neural fingerprint" within the network, inducing a pathological state of "conceptual hyper-excitation" in the relevant semantic domain.

2. **Motivation:** Second, by presenting a functional analysis of the key neurons that constitute this fingerprint, we will introduce our "Structuralist Hallucination Theory." This theory suggests that falsehoods can be endorsed by functionally distinct neurons that are primarily concerned with syntactic coherence, rather than semantic corruption.

3. **Mechanism:** Finally, we will map the "contagion pathway" that this falsehood follows through the network layers, revealing a "neural conflict" at its epicenter between "amplifier" neurons that propagate the falsehood and "suppressor" neurons that attempt to inhibit it.

This three-stage forensic examination reframes hallucinations as a mechanically understandable phenomenon with identifiable components, functional roles, and propagation pathways.

2 Methodology

The methodology employed in this study is designed to analyze the effects of a controlled falsehood within a neural network across three core stages: (1) Model Preparation and Falsehood Injection, (2) Detection and Mapping of Anomalous Activations, and (3) Functional Analysis of Identified Neurons.

2.1 Model Selection and Preparation

Our experiments were conducted on two widely accessible and well-documented models that embody the core features of the Transformer architecture: GPT-2 small (12 layers) and GPT-2 medium (24 layers). The use of these two different scales allowed us to test the scalability of our findings with respect to model size. All models were loaded and analyzed using the `transformer_lens` library.

2.2 Controlled Falsehood Injection

To isolate the internal mechanisms of a hallucination, we injected the models with a specific and controllable factual falsehood that contradicts their existing knowledge base.

- **Reference Truth:** "The capital of Germany is Berlin."
- **Injected Falsehood:** "The capital of France is Berlin."

The "liar_model" was created by fine-tuning a copy of the base pre-trained model on the false statement above, using a low learning rate (1e-5) for a short duration (150 steps). This process allows the model to adopt this new, false information while largely preserving its existing general knowledge base.

2.3 Neural Fingerprint Detection (Difference Analysis)

To detect the "neural fingerprint" of a falsehood, we compared the internal activation states of the model in two different conditions:

1. **Reference State:** The activations of the healthy, base model as it prepares to state the reference truth (prompted with "The capital of Germany is").
2. **Target State:** The activations of the "liar_model" as it prepares to state the injected falsehood (prompted with "The capital of France is").

In both cases, the post-ReLU outputs of the Multi-Layer Perceptron (MLP) blocks in all layers were recorded for the final token position. The value we term the "Anomaly Score" or "Difference Activation" was calculated for each neuron by subtracting the activation value in the Reference State from that in the Target State.

2.4 Functional Analysis (Neural Interrogation)

To understand the "motivation" or functional role of the key neurons that constitute the fingerprint, we applied a proprietary analysis protocol we call "neural interrogation." This protocol assesses a neuron's interaction with the model's internal representations to reveal its conceptual affinity profile. This analysis produces a ranked list of which concepts or word fragments from the model's entire vocabulary the neuron is intrinsically predisposed to respond to most strongly. This is a powerful technique for determining whether a neuron has a semantic or a structural/syntactic role.

3 Results: The Anatomy of a Falsehood

Our methodology revealed that in both gpt2-small and gpt2-medium models, a controlled falsehood has measurable, repeatable, and scalable effects within the network. In this section, we present the core empirical evidence for these effects.

3.1 Evidence 1: The Neural Fingerprint and Conceptual Hyper-Excitation

The most fundamental evidence of a falsehood's presence is the anomaly it creates in the network's internal activation state. Our difference analysis showed that across both model scales, the falsehood leaves a consistent "neural fingerprint" concentrated in specific neurons.

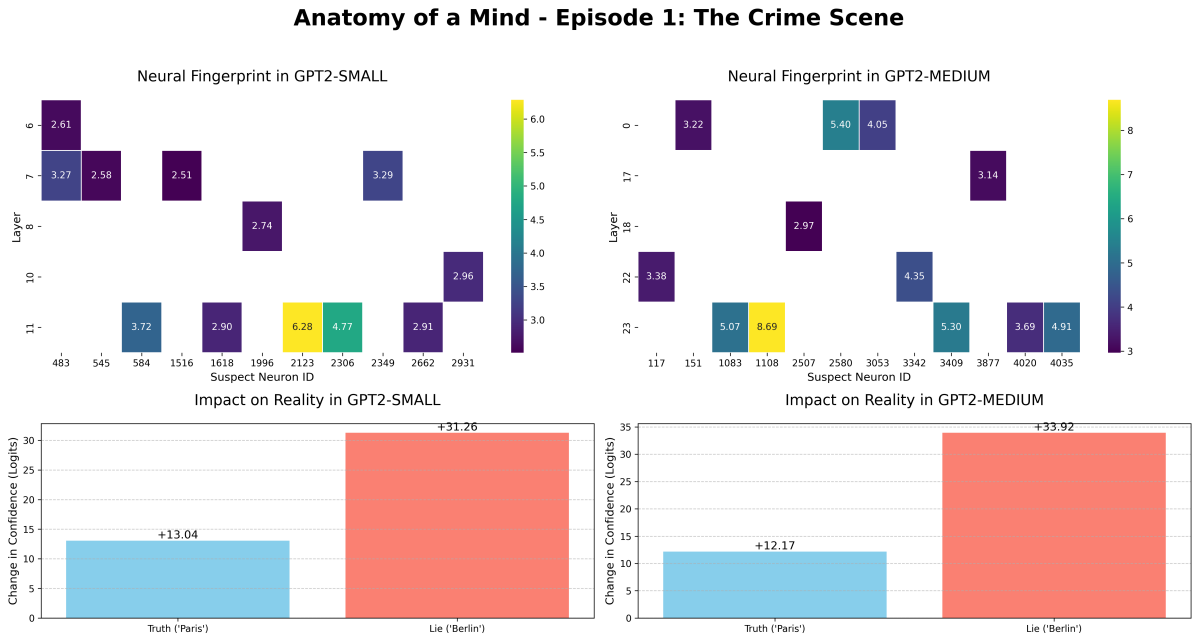


Figure 1: The Neural Fingerprint and Conceptual Hyper-Excitation.

In Figure 1, the heatmaps visualize how much of an anomalous increase in activation (Anomaly Score) each "suspect" neuron exhibits. This fingerprint has two key characteristics:

1. **Concentration:** The anomalous activations are not randomly distributed. In both models, they are predominantly concentrated in the middle and late layers (e.g., L6-L11 for gpt2-small; L17-L23 for gpt2-medium), affecting the network's higher-level, abstract representational layers.

2. **Scalability:** The general structure of the fingerprint and its region of concentration are preserved as the model size increases, indicating that this is a fundamental behavior of the Transformer architecture, not an accidental feature.

Furthermore, the bar charts in Figure 1 demonstrate that the falsehood induces a state of "conceptual hyper-excitation," pathologically increasing the logits for both the incorrect answer ("Berlin") and the correct one ("Paris"). The numerical data for gpt2-medium, for instance, show an increase of +33.92 for "Berlin" and +12.17 for "Paris," proving that the falsehood doesn't just reinforce the wrong answer but puts the entire semantic space in a "feverish" state.

3.2 Evidence 2: The Structuralist Hallucination Theory

After identifying the fingerprint, we interrogated the most anomalous neurons to understand their motivation. This revealed that the key "culprit" neurons (Tables 1 and 2) have no affinity for the semantic content of the lie. Neuron L11.N2123 in gpt2-small, for example, is activated by semantically neutral but grammatically significant prefixes (im-, un-, ir-) and suffixes. This evidence strongly supports our **"Structuralist Hallucination Theory"**: falsehoods are often endorsed not by semantically confused neurons, but

by amoral "structural engineer" neurons concerned only with syntactic validity, checking not for factual accuracy but for linguistic well-formedness.

Table 1: Confession of Chief Suspect (GPT2-Small, L11.N2123).

Rank	Preferred Concept	Activation Score
1	`im`	1.7064
2	`un`	1.5403
3	`on`	1.5303

Table 2: Confession of Chief Suspect (GPT2-Medium, L23.N1108).

Rank	Preferred Concept	Activation Score
1	`Leather`	1.9116
2	`Stanford`	1.8197
3	`Claim`	1.8187

3.3 Evidence 3: The Contagion Pathway and Neural Conflict

Finally, we found that a falsehood is not a static state but a dynamic process that spreads through the network like a contagion. Figure 2 shows this "contagion pathway," where the total anomaly score amplifies exponentially across the layers, from a score of 6.75 at Layer 0 to 414.68 at Layer 11, culminating in an "eruption" in the final layer. At the epicenter of this contagion, a "neural conflict" occurs. As shown in Figure 3, this conflict is a measurable struggle between a small minority of "amplifier" neurons that fuel the falsehood and a distinct group of "suppressor" neurons that act as a natural immune system, actively attempting to inhibit it.

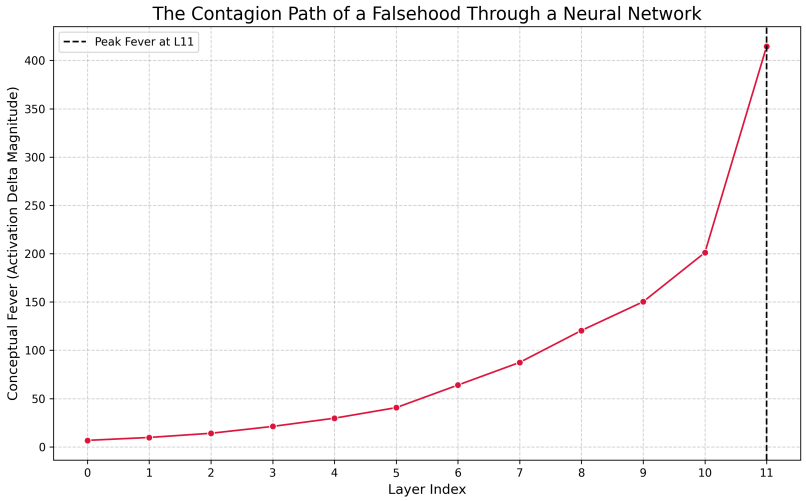


Figure 2: The Contagion Pathway of a Falsehood.

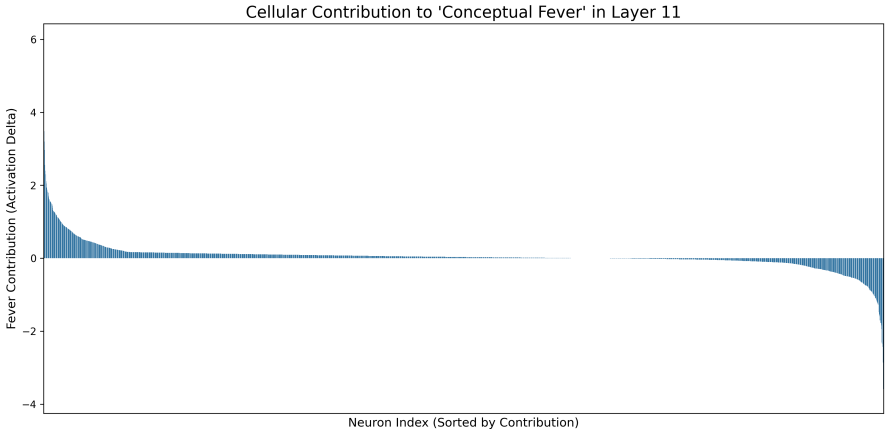


Figure 3: Neural Conflict: Amplifiers vs. Suppressors.

4 Discussion: From Anatomy to Deeper Implications

The preceding results provide a complete anatomy of a controlled falsehood. This section uses these findings as a launchpad to explore deeper questions about the nature of hallucinations and the internal dynamics of LLMs.

4.1 The "Cognitive Short-Circuit" Hypothesis

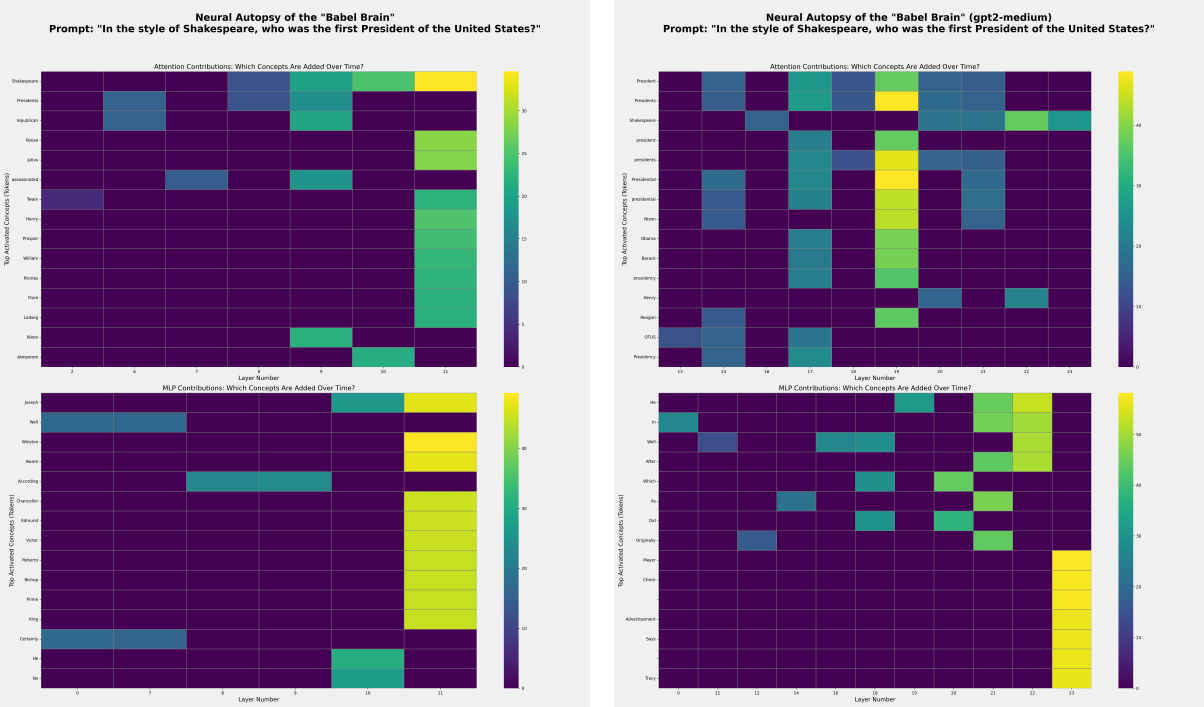
Our primary findings show how an external lie creates an internal conflict. But can a hallucination arise purely from an internal conflict between 'healthy' circuits? We explored this with the "Babel Brain" experiment, where the model processes a prompt with conflicting poetic and historical facts. The results (Figure 4) show a profound conflict between specialized circuits. This suggests a **"Cognitive Short-Circuit" hypothesis**: some hallucinations may not be due to a knowledge deficit, but to a conflict between valid but competing internal representations (e.g., a "Historian" circuit vs. a "Poet" circuit).

4.2 The Effect of Scale: Cognitive Restructuring

Does the anatomy of a thought process change as a model grows? We investigated how GPT-2 small and medium form the simple concept of "Paris." The findings (Figure 5) support a hypothesis we call **"Cognitive Restructuring."** The smaller model relies heavily on Attention blocks ("data gathering"), while the larger model fundamentally reorganizes its strategy to rely more on MLP blocks ("reasoning engines"). This shows that interpretability findings are not static and may change dramatically with scale.

5 Conclusion

This study presented a mechanistic anatomy of a falsehood, identifying its "neural fingerprint," proposing the "Structuralist Hallucination Theory," and revealing the "neural conflict" at its core. Building on this, we introduced two forward-looking hypotheses: the "Cognitive Short-Circuit" as a potential internal cause for hallucinations, and "Cognitive Restructuring" as a fundamental law of model scaling. This work not only provides a



(a) GPT-2 Small (b) GPT-2 Medium

Figure 4: Neural Autopsy of the "Babel Brain," revealing conflict between specialized circuits.

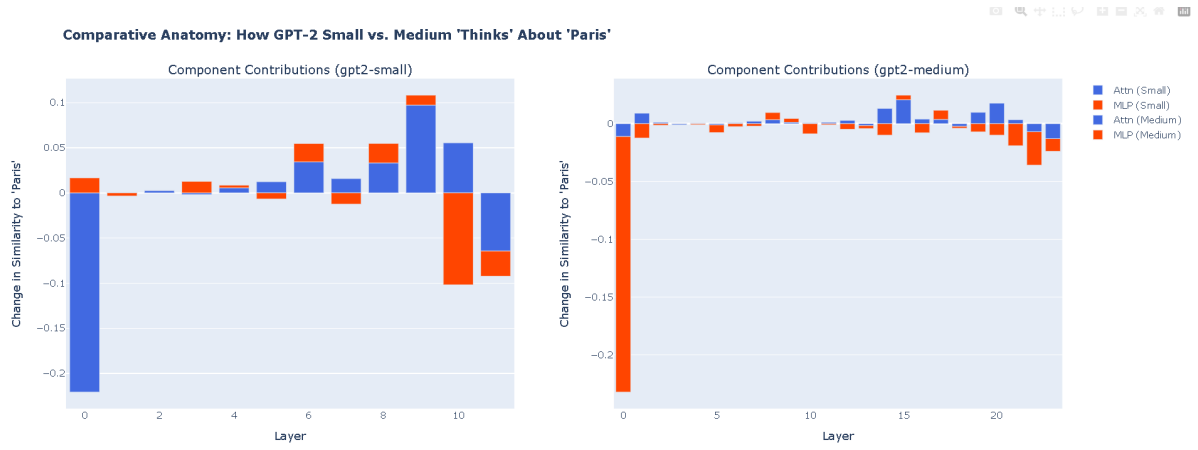


Figure 5: Cognitive Restructuring: How GPT–2 Small vs. Medium 'Thinks' About 'Paris'.

framework for combating existing hallucinations but also opens new frontiers in understanding the emergent cognitive dynamics of large-scale AI.

References

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in Neural Information Processing Systems*. 2017, pp. 5998–6008.

[2] Alec Radford et al. *Language models are unsupervised multitask learners*. Tech. rep. OpenAI, 2019.

[3] Nelson Elhage et al. *A mathematical framework for transformer circuits*. https://transformer-circuits.pub/2021/framework/index.html. 2021.

[4] Chris Olah et al. "Zoom in: An introduction to circuits". In: *Distill* (2020). DOI: 10.23915/distill.00024.

[5] Mor Geva et al. "Transformer feed-forward layers are key-value memories". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 5484–5495.

[6] Kevin Meng et al. "Locating and editing factual associations in GPT". In: *Advances in Neural Information Processing Systems*. 2022.

[7] Neel Nanda. *TransformerLens*. https://github.com/neelnanda-io/TransformerLens. 2022.

[8] Ziwei Ji et al. "Survey of hallucination in natural language generation". In: *ACM Computing Surveys* (2023). DOI: 10.1145/3571730.

[9] Nora Belrose, Jacob Steinhardt, and Xinyang Wu. "Mechanistic interpretability for AI safety: Why it matters and how to get started". In: *AI Safety Research* (2023).

[10] Jiwei Li, Will Monroe, and Dan Jurafsky. "Understanding neural networks through representation erasure". In: *arXiv preprint arXiv:1612.08220* (2016).