Mechanistic Interpretability of GPT-2: Lexical and Contextual Layers in Sentiment Analysis

Amartya Hatua *
AI Center of Excellence
Fidelity Investments
Boston, MA 02210
amartyahatua@gmail.com

Abstract

We present a mechanistic interpretability study of GPT-2 that causally examines how sentiment information is processed across its transformer layers. Using systematic activation patching across all 12 layers, we test the hypothesized two-stage sentiment architecture comprising early lexical detection and mid-layer contextual integration. Our experiments confirm that early layers (0-3) act as lexical sentiment detectors, encoding stable, position specific polarity signals that are largely independent of context. However, all three contextual integration hypotheses: Middle Layer Concentration, Phenomenon Specificity, and Distributed Processing are falsified. Instead of mid-layer specialization, we find that contextual phenomena such as negation, sarcasm, and domain shifts are integrated primarily in late layers (8-11) through a unified, non-modular mechanism. These experimental findings provide causal evidence that GPT-2's sentiment computation differs from the predicted hierarchical pattern, highlighting the need for further empirical characterization of contextual integration in large language models.

1 Introduction

Large language models demonstrate impressive capabilities across a wide range of diverse linguistic tasks. Despite this progress, existing interpretability research primarily relies on correlational evidence from probing or attention analysis. Consequently, the internal causal structure through which these models encode and transform linguistic information has not been widely explored. Early research focused on identifying how distinct layers within transformers contribute to different stages of linguistic processing. Tenney et al. [2019] found that BERT processes language in stages early layers handle syntactic information, while later layers understand semantic relationships. This suggests that transformers operate similarly to a pipeline, progressing from simple features to a complex understanding. It was the first clear evidence that these models have organized, step by step processing. Building upon this foundation, Jawahar et al. [2019], formalized a three-tier hierarchical framework: early layers handle basic word features, middle layers deal with grammar and sentence structure, and late layers understand meaning and how distant words relate to each other. Simultaneously, Clark et al. [2019] revealed that individual heads develop specialized functions for specific linguistic phenomena, following the same early to late progression. In Rogers et al. [2020], a comprehensive synthesis was provided that established a general consensus on middle layer specialization for syntactic structure, while highlighting that semantic processing remains more distributed and less well understood. All these studies showed that transformers seem to process language in organized, step-by-step ways. However, their methodologies were predominantly correlational, relying on probing classifiers and

*Code and data available at: https://github.com/amartyahatua/MI_Sentiment_Analysis

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: CogInterp: Interpreting Cognition in Deep Learning Models.

attention analysis to identify what information exists in representations rather than what models actually use during inference.

Newer research has highlighted this gap. Scientists now realize that finding patterns doesn't prove the model actually uses them. As Belinkov et al. [2023] puts it, there's a gap between what we can detect in the model and what the model actually relies on; just because we can find information doesn't mean the model uses it. When Elazar and Goldberg [2018] tried removing features they thought were important, the models often worked just fine without them. This suggested they were finding fake patterns, not real ones. Makelov et al. [2024] found "interpretability illusions" interventions that seemed to reveal how models work but were actually triggering backup systems that had nothing to do with normal processing. In sentiment analysis, it remains essential to determine whether transformer layers interact in a truly causal manner. Do early layers construct representations that later layers build upon, or do they operate in parallel? Is contextual information processed locally or distributed across the network? Correlational analyses cannot resolve these questions. In this work, we employ activation patching and causal interventions to empirically examine sentiment processing in GPT-2, providing mechanistic evidence for stage wise organization and establishing a foundation for deeper causal interpretability research.

2 Methodology

2.1 Experimental Design

We have tested the two-stage sentiment processing hypothesis in GPT-2 using activation patching across the 12 layers. The analysis focuses on lexical detection and contextual integration, with controlled interventions isolating the causal role of each stage in sentiment behavior. We use GPT-2 (117M) through TransformerLens Nanda and Meyer [2023], which allows standardized access to activations and precise interventions. The model is run in inference mode without finetuning, with consistent tokenization and identical architecture across all conditions.

2.2 Activation Patching Protocol

Activation patching is used in this study as a causal intervention technique to identify which transformer layers directly contribute to sentiment processing. By selectively substituting internal activations between contrasting input sentences, we isolate the specific layers responsible for lexical detection and contextual integration. For each test pair, we conducted activation patching, replacing activations from the source sentence (e.g., positive sentiment) with those from the target sentence (e.g., negative sentiment) at each layer independently. The resulting change in sentiment classification probability was measured to quantify the causal contribution of each layer, where larger shifts indicate greater causal importance for sentiment processing.

2.3 Lexical Detection

We perform a linear probe on GPT-2's final layer representations to classify sentiment polarity, achieving 95% accuracy on a held-out validation set. This probe serves as our behavioral measure for sentiment classification performance, allowing us to quantify how interventions affect the model's sentiment processing.

Hypotheses: We test four specific hypotheses about lexical processing: I) Lexical Sensitivity: Sentiment word substitutions produce measurable activation differences. II) Early Layer Dominance: Layers 0-3 show the strongest effects for lexical sentiment. III) Position Specificity: The effects concentrate on the positions of the sentiment words. IV) Context Independence: Lexical effects remain consistent across different sentence contexts.

2.4 Contextual Integration

We create test cases that check how the model changes the sentiment of the raw words to the right meaning based on context. Our test suite includes test cases with the following sentiments: Medium intensity, Intensified swap, Simple negation, Intensified negation, Complex double negation, Domain context, Sarcasm, Intensity, Multiple intensifier, Scale variation.

Hypotheses: We test whether contextual integration follows the predicted layer specialization pattern: I) Middle Layer Concentration: Contextual effects peak in layers 4-8. II) Phenomenon Specificity: Different context types show distinct layer patterns. III) Distributed Processing: Effects concentrate in specific layers rather than being distributed.

3 Data

Lexical Detection Dataset: We generated 100 lexical test pairs to examine context independent sentiment word recognition in early layers (0-3). Each test pair consisted of two sentences identical except for a single word bearing a sentiment substitution. We constructed sentence pairs from five base templates: "The movie was {sentiment_word}", "I found the book {sentiment_word}", "The restaurant was {sentiment_word}", "This experience was {sentiment_word}", and "The performance was {sentiment_word}".

Contextual Integration Dataset: The Contextual Integration Dataset comprises 8,000 carefully constructed test pairs designed to evaluate how GPT-2 processes context dependent sentiment modifications across 14 distinct phenomena. Each test pair consists of a clean sentence and a corrupted counterpart, differing only in specific contextual elements. The dataset systematically explores diverse contextual mechanisms: Strong Positive (C1) swaps strong sentiment words with opposite counterparts; Medium Intensity (C2) exchanges moderate sentiment terms; Intensified Swap (C3) combines intensifier adverbs with opposing sentiments; Comparative Context (C4) modifies comparative phrases and outcomes; Simple Negation (C5) adds or removes basic negation words; Intensified Negation (C6) applies negation to intensified phrases; Complex Double Negation (C7) employs sophisticated double negation patterns; Domain Context (C8) alters domain-specific sentiment interpretation; Sarcasm (C9) tests ironic context detection; Conditional vs Actual (C10) switches between hypothetical and factual statements; Intensity Variation (C11) modulates sentiment strength through modifier changes; Multiple Intensifiers (C12) examines stacked modifier effects; Intensity Flip (C13) replaces strong intensifiers with weak ones; and Scale Variation (C14) swaps words at different sentiment scale positions.

4 Result

4.1 Lexical Detection

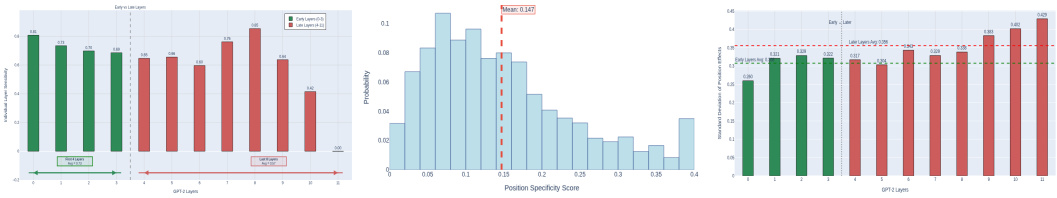
To test our four part framework for lexical sentiment processing, we ran three experiments on early layers. The Lexical Sensitivity test checks if sentiment effects are strongest at word substitutions (Hypotheses 1–2). The Position Specificity test compares patching at sentiment vs. non-sentiment words (Hypothesis 3). The Context Independence test measures whether lexical effects stay stable across contexts (Hypothesis 4). Together, these activation patching experiments provide causal evidence for our framework beyond correlations.

4.1.1 Lexical Sensitivity

For each sentence pair, we patched activations across all 12 GPT-2 layers. The clean sentence used a positive word, while the corrupted one used its negative counterpart. By replacing activations at target word positions, we measured each layer's causal role in sentiment prediction. Position specificity was tested by comparing effects at sentiment vs. non-sentiment words, while context independence was measured by variation of these effects across different contexts. Figure 1a, shows the lexical sensitivity in GPT-2 layers. Bar heights show effect sizes from activation patching experiments. Average sensitivity of early layers Layer-0 to Layer-3(L_0-L_3) shows higher sensitivity to sentiment word substitutions, with L_0 exhibiting peak performance.

4.1.2 Position Specificity

Position Specificity Analysis tests whether GPT-2 encodes sentiment in specific words or through broad sentence signals. By patching activations at sentiment vs. non-sentiment words, we find early layers especially L_0, shows strong word-level sensitivity (specificity score 0.147, $p < 0.001$ (Figure 1b)). This shows GPT-2's lexical stage detects sentiment at precise token positions, forming the basis for later contextual integration of negation, sarcasm, and other complexities.



(a) Lexical Sensitivity (b) Position Specificity (c) Context Independence

Figure 1: Lexical Detection Analysis

4.1.3 Context Independence

The context independence analysis tests whether the GPT-2 detection of lexical sentiment remains stable across different sentence contexts. We measure this by looking at how much the effect of sentiment words varies when they appear in different linguistic environments. If a layer is truly performing lexical processing, the effect of words like 'wonderful' or 'terrible' should remain consistent regardless of context. In contrast, context dependent processing should produce higher variability, since the same word may shift meaning depending on surrounding words. Figure 1c shows early layers (L_0–L_3) shows very low variability in position effects, while later layers (L_4–L_{11}) show much higher variability. This confirms that early layers extract stable, context independent sentiment features, providing a reliable foundation for the rest of the network.

4.1.4 Hypothesis Evaluation

Lexical detection analysis shows that GPT-2's early layers reliably detect lexical sentiment. The results support all four hypotheses: (1) lexical sensitivity, (2) the early layers show the strongest sensitivity to sentiment words, (3) the effects are position specific, strongest at the locations of the sentiment words, and (4) the detection is context independent, with less variability than the later layers. Together, this confirms that GPT-2 encodes stable lexical sentiment signals early, forming the basis for later contextual integration.

4.2 Contextual Integration

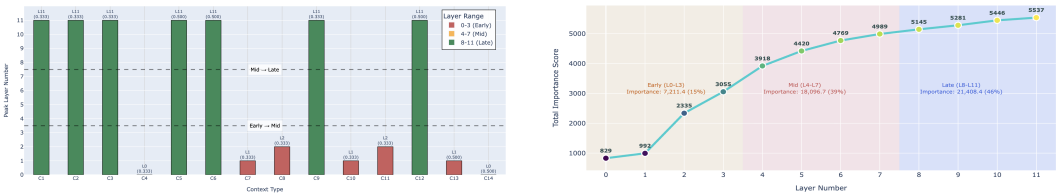
To evaluate contextual integration, we tested three hypotheses in GPT-2: (1) Middle Layer Concentration, predicting peaks in L_4-L_8; (2) Phenomenon Specificity, predicting distinct layer patterns for different contextual types; and (3) Distributed Processing, predicting effects spread across layers. Using controlled activation patching across all 12 layers, we measured the causal impact of specific contextual interventions on sentiment, isolating each phenomenon while keeping baseline conditions constant.

4.2.1 Middle Layer Concentration

The Middle Layer Concentration hypothesis predicted that contextual integration would peak in L_4-L_8, under the assumption that syntactic and semantic operations occur at intermediate depths of the network. Our experimental results, based on 8,000 test cases across 15 distinct contextual phenomena, contradicts this prediction. Figure 2a, demonstrates that the network exhibits a bimodal distribution where phenomena cluster in either early L_0-L_3 or late layers (8-11), with no phenomena peaking in the predicted middle range L_4-L_7. Of the 15 contextual phenomena tested, 8 exhibit their strongest effects in L_{11} (57%), including strong positive contexts, medium intensity, intensified swap, simple negation, intensified negation, sarcasm, multiple intensifiers, and conditional vs actual contexts. The remaining 7 phenomena (43%) peak in early layers: comparative context and scale variation (L_0), complex double negation, conditional vs actual, and intensity flip (L_1), and domain context and intensity variation (L_2). This bimodal pattern suggests fundamentally different processing strategies for different types of contextual modifications.

4.2.2 Phenomenon Specificity

The Phenomenon Specificity hypothesis predicted that different contextual phenomena would exhibit distinct layer-wise processing patterns, with each type of contextual modification recruiting special-



(a) Peak Layer Distribution Across Context Types

(b) Layer Importance Gradient

Figure 2: Contextual Integration Analysis

ized computational mechanisms at different network depths. Our experimental results decisively falsify this prediction. Out of the 15 contextual types tested, 13 (87%) share nearly identical top-3 contributing layers in the pattern [L_{11}, L_{10}, L_9]. Furthermore, 8 phenomena (53%) peak at the L_{11}. This convergence encompasses semantically diverse contextual modifications including simple negation, intensified negation, sarcasm, multiple intensifiers, and comparative contexts, all routing through the same late layer processing hub despite their different linguistic properties.

4.2.3 Distributed Processing

The Distributed Processing hypothesis predicted that contextual effects would be spread across multiple layers rather than concentrated in specific regions of the network. Our results provide mixed evidence, revealing a more nuanced architecture than either pure distribution or strict concentration. The total layer importance analysis shows a clear gradient rather than uniform distribution. Late layers (L_8-L_{11}) dominate with 21,408.4 total importance (46% of all contextual processing), while mid-layers (L_4-L_7) contribute substantially at 18,096.7 (39%), and early layers (0-3) account for only 7,211.4 (15%). The top five most important individual layers form a consecutive sequence from the network's upper regions: L_{11} (5,537.1), L_{10} (5,446.0), L_9 (5,280.7), L_8 (5,144.5), and L_7 (4,988.9). Figure 2b, demonstrates a monotonic decrease from late to early layers indicates concentrated rather than distributed processing architecture. However, the low within-phenomenon agreement scores (averaging 0.389, ranging from 0.333 to 0.500) provide evidence for distribution at a finer grain. Even within phenomena that show strong late-layer peaks, individual test cases disagree on which specific layer is most important, with only 33-50% of cases within each category agreeing on the peak layer. This substantial variability suggests that, while contextual processing concentrates in the late-layer region as a whole, the specific computational path varies depending on individual input characteristics. These findings largely falsify the Distributed Processing hypothesis in its strong form. Contextual integration is not uniformly distributed across all layers, but instead concentrates in a specific late layer region (L_8-L_{11}), with diminishing contributions from middle and early layers. However, the low agreement scores and substantial midlayer contribution (39%) indicate that processing within this concentrated region exhibits input-dependent flexibility rather than rigid localization to a single layer.

4.2.4 Hypothesis Evaluation

All three contextual integration hypotheses were falsified in our 8,000 case test dataset. The middle layer concentration failed: 57% of the phenomena peaked in the L_8-L_{11}, 43% in L_0-L_3, and zero in the predicted L_4-L_7. Phenomenon Specificity was rejected: 87% of test cases shared identical top-3 layers [L_{11}, L_{10}, L_9], revealing convergence rather than specialization. Distributed processing failed: layer importance increased monotonically 6.7 fold from L_0 to L_{11}, demonstrating concentration rather than distribution.

5 Conclusion

This study provides systematic causal validation of hierarchical sentiment processing in GPT-2 through mechanistic interpretability methods. We show that sentiment processing unfolds in a two stage: precise lexical detection in early layers followed by complex contextual integration concentrated in late layers, rather than in the predicted middle layers. All three contextual integration hypotheses middle layer concentration, phenomenon specificity, and distributed processing were systematically falsified. These findings shows how rigorous activation patching can explain AI

models beyond correlational analysis to provide causal insight into transformer computation. Future work should validate these patterns across diverse transformer architectures (BERT, RoBERTa, larger GPT models) to determine whether two-stage lexical-contextual processing represents a general architectural principle or remains specific to GPT-2's scale and training paradigm. Extension to fine-grained circuit-level analysis could identify the precise attention heads and MLP blocks responsible for lexical detection and contextual integration, moving beyond layer-wise analysis to map exact computational pathways within the transformer architecture.

References

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 (ACL 2019)*, pages 4593–4601. Association for Computational Linguistics, 2019. URL https://aclanthology.org/P19-1452/.

Ganesh Jawahar, Benoît Sagot, Djamé Seddah, Maxime de Lhoneux, David Seddah, and Maxime de Lhoneux. What does bert learn about the structure of language? *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 356–365, 2019. URL https://aclanthology.org/P19-1356/.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D Manning. What does bert look at? an analysis of bert's attention. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 276–285. Association for Computational Linguistics, 2019. URL https://aclanthology.org/W19-4828/.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866, 2020. doi: 10.1162/tacl_a_00349. URL https://aclanthology.org/2020.tacl-1.54/.

Yonatan Belinkov, Youngwook Kim, Jae Myung Kim, Jieun Jeong, Cordelia Schmid, Zeynep Akata, and Jungwoo Lee. Bridging the gap between model explanations in partially annotated multi-label classification. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3408–3417, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/papers/Kim_Bridging_the_Gap_Between_Model_Explanations_in_Partially_Annotated_Multi-Label_CVPR_2023_paper.pdf.

Yanai Elazar and Yoav Goldberg. Adversarial removal of demographic attributes from text data. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 11–21. Association for Computational Linguistics, 2018. URL https://aclanthology.org/D18-1002/.

Aleksandar Makelov, Georg Lange, and Neel Nanda. Is this the subspace you are looking for? an interpretability illusion for subspace activation patching. In *Proceedings of the 2024 International Conference on Learning Representations (ICLR)*, 2024. URL https://openreview.net/forum?id=Ebt7JgMHv1.

Neel Nanda and Bryce Meyer. Transformerlens: A library for mechanistic interpretability of generative language models, 2023. URL https://github.com/TransformerLensOrg/TransformerLens.