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

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

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: Efficient Reasoning.

correlational, relying on probing classifiers and 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.

Recent years have brought major improvements in solving the correlation-causation problem. The field of mechanistic interpretability Rai et al. [2024] has developed new techniques like activation patching Heimersheim and Nanda [2024] that let researchers directly test cause and effect, while automated tools have made the analysis process more systematic. Companies like Anthropic and OpenAI have successfully applied these mechanistic methods to real models, finding millions of interpretable features in their large language models. Among these challenging phenomena, sentiment analysis presents a particularly instructive case. Sentiment analysis represents a particularly complex challenge for mechanistic interpretability. Unlike syntactic phenomena that localize to specific layers or attention heads, sentiment processing requires sophisticated integration of lexical, syntactic, and contextual information. The exact words can convey opposite meanings depending on context—"great" in "great movie" versus "oh great, it's raining"—requiring dynamic contextual reasoning that current methods struggle to explain mechanistically.

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 Related Research

2.1 Transformer Layer Analysis and Interpretability

The study of how transformers process language has advanced through several methodological stages. Early probing work Tenney et al. [2019] showed that BERT encodes linguistic features hierarchically, with parts of speech in early layers, syntax in middle layers, and semantics in later layers—forming the basis of the layer specialization hypothesis. Jawahar et al. [2019] expanded this view, showing a progression from surface features to long-distance dependencies, while Clark et al. [2019] demonstrated that attention heads develop specialized functions aligned with linguistic phenomena. Rogers et al. [2020] synthesized these findings in their BERTology survey, framing transformer processing as a three-stage pipeline of feature extraction, syntax, and semantics—a framework that remains influential but largely correlational.

2.2 Methodological Evolution in Mechanistic Interpretability

Limitations of correlational methods led to causal approaches in mechanistic interpretability. Elhage et al. [2021] emphasized the need to distinguish between what information exists in representations and what computations models actually perform. Activation patching became central here, with Wang et al. [2022] showing distributed processing for indirect object identification and Meng et al. [2022] revealing that factual knowledge depends on interactions across layers rather than localized storage. Automated methods such as ACDC (Conmy et al. [2023]) extended this work by identifying candidate circuits, though many fail causal tests, highlighting challenges of functional faithfulness.

2.3 Sentiment Processing in Language Models

While sentiment analysis is a well-studied application of transformers, the mechanisms behind sentiment processing remain poorly understood. Probing studies (e.g., Liu et al. [2019]) showed that different layers encode aspects of sentiment, suggesting a multi-stage progression from word recognition to contextual modulation and final decision-making. Yet these findings are largely correlational, and sentiment's contextual nature—shaped by negation, intensification, and pragmatic cues like sarcasm—complicates interpretation. Recent studies have examined such phenomena individually, but little work has addressed how they interact within a unified computational framework.

2.4 Limitations of Current Approaches

Current interpretability methods struggle with complex semantic phenomena like sentiment processing. Most mechanistic successes have been limited to simple tasks such as arithmetic or syntax, leaving contextual reasoning and pragmatic inference opaque. Probing approaches, though influential, often fail to capture true model behavior—high accuracy can reflect memorization rather than genuine computation (Hewitt and Liang [2019], Belinkov et al. [2023]). Theories like layer specialization remain largely correlational, lacking causal validation, which is especially problematic for sentiment, where meaning depends on intricate contextual interactions.

2.5 The Need for Systematic Causal Validation

These limitations point to the need for systematic causal validation of transformer processing theories. New methods—such as causal scrubbing, refined activation patching, and sparse autoencoders—offer powerful tools but have rarely been applied to foundational questions. In sentiment analysis, we still lack a causal account of how the processing stages interact. This work provides the first systematic validation of the three-stage sentiment processing hypothesis, moving beyond correlation to genuine mechanistic insight.

3 Methodology

3.1 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.

3.2 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.

3.3 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.

3.4 Hypotheses

We test four specific hypotheses about lexical processing:

1. Lexical Sensitivity: Sentiment word substitutions produce measurable activation differences.
2. Early Layer Dominance: Layers 0-3 show the strongest effects for lexical sentiment.
3. Position Specificity: The effects concentrate on the positions of the sentiment words.
4. Context Independence: Lexical effects remain consistent across different sentence contexts.

3.5 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.

3.6 Hypotheses

We test whether contextual integration follows the predicted layer specialization pattern:

1. Middle Layer Concentration: Contextual effects peak in layers 4-8.
2. Phenomenon Specificity: Different context types show distinct layer patterns.
3. Distributed Processing: Effects concentrate in specific layers rather than being distributed.

4 Data

This section outlines the data generation methodology employed to evaluate the sentiment processing hypothesis in GPT-2 mentioned in the earlier section.

4.1 Lexical Detection Dataset:

The dataset included 1,000 test cases across six types of contextual changes, each targeting different aspects of sentiment processing. I) Simple Negation cases examined basic polarity reversal through negation words (e.g., "The movie was good" vs. "The movie was not good"). II) Intensified Negation tested stronger negation patterns with adverbs ("The film was excellent" vs. "The film was definitely not excellent"). III) Sarcasm cases required the detection of situational incongruity ("Great, another meeting" with positive/negative contextual framing). IV) Domain context examples tested how domain knowledge affects sentiment interpretation ("The horror movie was terrifying" where 'terrifying' is positive for horror but negative for other genres). V) Intensification cases examined how modifiers amplify sentiment ("The meal was good" vs. "The meal was extremely good"). VI) Complex Double Negation tested sophisticated logical reasoning ("I don't think it's not good" requiring multiple negation resolution steps). Each type of contextual modification was further varied across three intensity levels (low, medium, and high) to examine graded contextual effects.

4.2 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: C1) Strong Positive: Substituted strong positive sentiment words with strong negative counterparts (e.g., 'incredible' $\rightarrow$ 'abysmal', 'wonderful' $\rightarrow$ 'horrible'). C2) Medium Intensity: Swapped medium-intensity positive words with medium-intensity negative words (e.g., 'fine' $\rightarrow$ 'bad', 'enjoyable' $\rightarrow$ 'unsatisfactory'). C3) Intensified Swap: Combined intensifier adverbs with opposite sentiment words (e.g., 'utterly wonderful' $\rightarrow$ 'utterly awful', 'completely amazing' $\rightarrow$ 'completely dreadful'). C4) Comparative Context: Changed comparative phrases and their sentiment outcomes (e.g., 'better than expected, quite satisfying' $\rightarrow$ 'worse than expected, quite mediocre'). C5) Simple Negation:

Added or removed basic negation words to flip polarity (e.g., 'nice' → 'not nice', 'are decent' → 'aren't decent'). C6) Intensified Negation: Applied negation to intensified positive phrases (e.g., 'was quite outstanding' → 'wasn't quite outstanding', 'very spectacular' → 'wasn't very spectacular'). C7) Complex Double Negation: Used double negation patterns with contrasting outcomes (e.g., 'wasn't bad at all, actually decent' → 'wasn't good at all, actually disappointing'). C8) Domain Context: Changed domain context to alter sentiment interpretation of domain-specific words (e.g., 'horror movie was haunting' [positive] → 'romantic comedy was haunting' [negative]). C9) Sarcasm: Modified intensity of sentiment words in sarcastic contexts to change perceived sentiment (e.g., 'Perfect, amazing weather' [sarcastic/negative] → 'Perfect, decent weather' [less negative]). C10) Conditional vs Actual: Switched between conditional and actual statements to change sentiment (e.g., 'would have been outstanding if not for' [negative] → 'was outstanding despite' [positive]). C11) Intensity Variation: Reduced or increased intensity modifiers while keeping base sentiment (e.g., 'incredibly pleasant' → 'a bit pleasant', both positive but different intensity). C12) Multiple Intensifiers: Removed stacked intensifiers to reduce sentiment strength (e.g., 'utterly very adequate' → 'just adequate', positive → neutral). C13) Intensity Flip: Changed strong intensifiers to weak/minimal intensifiers (e.g., 'extremely spectacular' → 'only slightly spectacular', positive → neutral). C14) Scale Variation: Swapped sentiment words at different positions on the sentiment scale (e.g., 'pleasant' [+3] → 'horrible' [0], varying scale distances).

5 Result

5.1 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.

Figure 1: Lexical Sensitivity

5.1.1 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 1, 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.

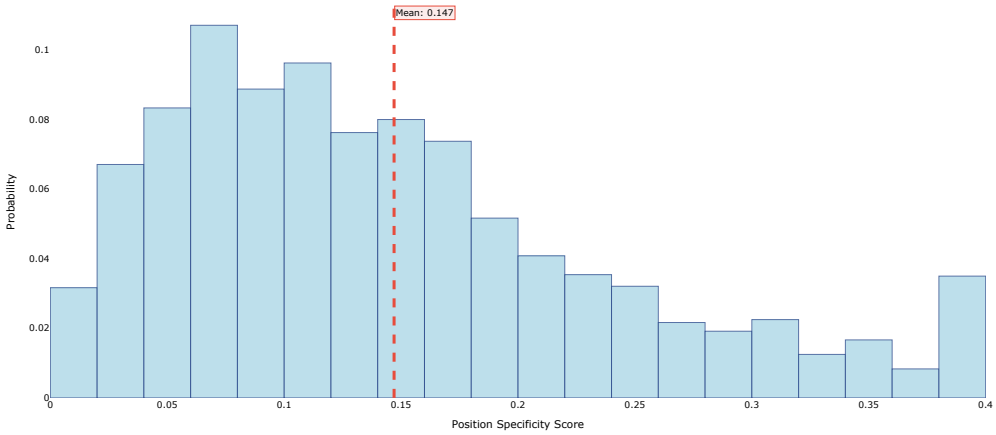

Figure 2: Position Specificity

5.1.2 Position Specificity

Position Specificity Analysis tests whether GPT-2 detects lexical sentiment at precise word locations or through more diffuse sentence wide signals. We measure this by comparing the effect of activation patching at sentiment bearing words versus at non-sentiment words within the same sentence. Theory predicts that early layers should show strong word level sensitivity, since lexical sentiment features are tied directly to specific tokens. Figure 2, shows our experiments confirm across 200 test pairs, activation patching produced significantly stronger effects at sentiment word positions than at non target positions, with a mean specificity score of 0.147 ($p < 0.001$). This provides clear evidence that early layers, particularly Layer 1, localize sentiment information to specific token positions. In other words, GPT-2's lexical stage operates through targeted, word level detection rather than holistic sentence processing. This position specific encoding forms the foundation for later contextual integration stages, where sentiment must be adjusted through distributed processing to capture complex patterns such as negation or sarcasm.

5.1.3 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 3 shows early layers (L_0–L_3) shows very low variability (mean = 0.038) in position effects, while later layers (L_4–L_{11}) show much higher variability (mean = 0.356). This confirms that early layers extract stable, context independent sentiment features, providing a reliable foundation for the rest of the network.

5.1.4 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.

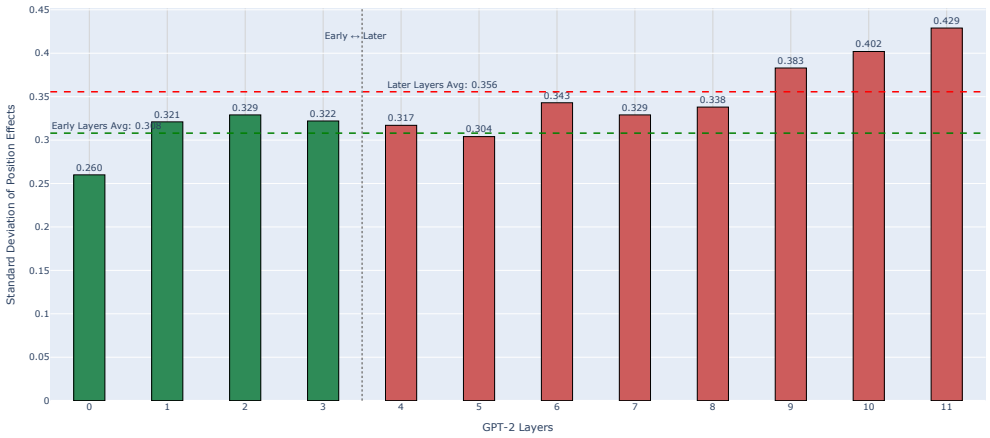

Figure 3: Context Independence of Sentiment Effects

5.2 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.

5.2.1 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 4, 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

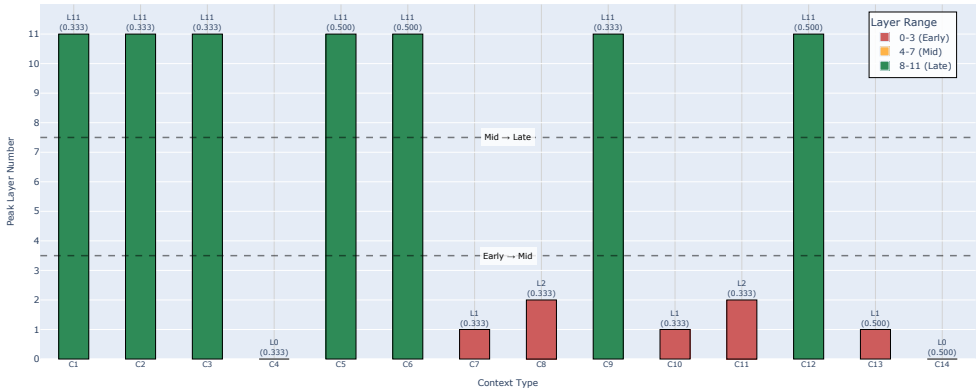

Figure 4: Peak Layer Distribution Across Context Types

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.

5.2.2 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 specialized computational mechanisms at different network depths. Our experimental results decisively falsify this prediction, revealing instead remarkable convergence across semantically diverse phenomena. Of the 15 contextual types tested, 13 (87%) share nearly identical top 3 contributing layers in the pattern [L_{11}, L_{10}, L_9] or [L_{10}, L_{11}, L_9]. Furthermore, 8 phenomena (53%) peak at the exact same layer 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. Only domain context exhibits a genuinely distinct pattern, with both peak (L_2) and top-3 layers [L_2, L_3, L_4] concentrated in early-to-middle regions. This singular exception highlights what phenomenon specificity would look like if it existed systematically. The overwhelming convergence demonstrates that GPT-2 does not employ phenomenon-specific modules but instead processes most contextual modifications through a shared high-level semantic integration system.

5.2.3 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 a uniform distribution. Late layers (L_8-L_{11}) dominate with 46% of all contextual processing, while mid-layers (L_4-L_7) contribute substantially 39%, and early layers (L_0-L_3) account for only 15%. The top five most important individual layers form a consecutive sequence from the network's upper regions: L_{11}, L_{10}, L_9, L_8, and L_7. Figure 5, demonstrates that a monotonic decrease from late to early layers indicates a concentrated rather than distributed processing architecture. 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.

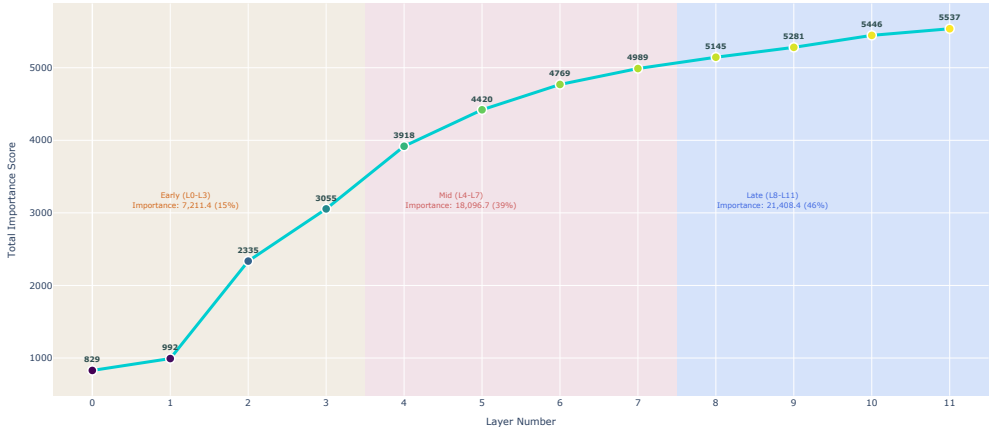

Figure 5: Layer Importance Gradient

5.2.4 Hypothesis Evaluation

Our systematic evaluation of the 8,000 case test dataset reveals that all three contextual integration hypotheses were falsified, though each provides distinct insights into GPT-2's processing architecture. The Middle Layer Concentration hypothesis failed as contextual phenomena exhibited 57% late peaking and 43% early peaking with zero phenomena peaking in the predicted middle layers (L_4-L_7), contradicting assumptions about intermediate layer semantic processing. The Phenomenon Specificity hypothesis was decisively rejected: 87% of phenomena in our test cases share identical top 3 layers [L_{11}, L_{10}, L_9], demonstrating that GPT-2 routes semantically diverse contextual modifications negation, sarcasm, intensification through a unified late layer hub rather than specialized modules. Finally, the distributed processing hypothesis failed because the importance of layers in our dataset shows a 6.7 fold monotonic increase from L_0 (828.7) to L_1 (5,537.1), with late layers dominating 46% of total processing weight. Together, these falsifications reveal an unexpected architecture within our experimental scope: GPT-2 employs phenomenon-agnostic late layer integration for most contextual reasoning, with only domain specific contexts (L_2, top 3: [L_2 , L_3, L_4]) representing a distinct early layer processing pathway. These findings, specific to our test dataset and methodology.

6 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.

Daking Rai, Yilun Zhou, Shi Feng, Abulhair Saparov, and Ziyu Yao. A practical review of mechanistic interpretability for transformer-based language models. *arXiv preprint arXiv:2407.02646*, 2024. URL https://arxiv.org/abs/2407.02646.

Stefan Heimersheim and Neel Nanda. How to use and interpret activation patching. *arXiv preprint arXiv:2404.15255*, 2024. URL https://arxiv.org/abs/2404.15255.

Nathaniel Elhage et al. A mathematical framework for transformer circuits. https://transformer-circuits.pub/2021/framework/index.html, 2021. Accessed: 2025-08-27.

Yilun Wang et al. Discovering and visualizing attention circuits for indirect object identification in gpt-2. *arXiv preprint arXiv:2211.00593*, 2022. URL https://arxiv.org/abs/2211.00593.

Chen Meng et al. Locating and editing factual knowledge in gpt. In *Advances in Neural Information Processing Systems (NeurIPS 2022)*, 2022. URL https://arxiv.org/abs/2202.05262.

Chris Conmy et al. Acdc: Automated discovery of computational circuits in neural networks. https://arxiv.org/abs/2304.14997, 2023. Accessed: 2025-08-27.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019. URL https://arxiv.org/abs/1907.11692.

John Hewitt and Percy Liang. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2733–2743. Association for Computational Linguistics, 2019. URL https://aclanthology.org/D19-1275/.

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