# Distributed Collaboration versus Single-Head Specialization: How Attention Masks Shape Syntactic Knowledge in Transformers

**Anonymous EMNLP submission**

## Abstract

Transformer-based language models acquire rich syntactic knowledge without explicit grammatical instruction, yet the internal organization of this knowledge remains poorly understood. We systematically investigate the functional organization of syntactic knowledge across Transformer attention heads and examine how attention masking mechanisms (causal versus bidirectional) constrain this organization.

Six sets of experiments are conducted across four models, namely, GPT-2 Small, GPT-2 Medium, BERT-base, and RoBERTa-base, including single-head permutation tests, community-level nullification, direct comparisons of network-level and single-head ablation, pruning experiments, cross-task validation, and multiple sensitivity analyses.

A clear cross-architectural divergence is revealed. In causal language models (GPT-2), no single attention head is individually significant, yet ablating the identified syntactic network as a whole causes a performance drop that far exceeds both the random baseline and the sum of ablating an equivalent number of top-ranked heads. Syntactic knowledge therefore resides in a distributed, collaborative network. In bidirectional models (BERT, RoBERTa), only one critical head is detected per model, and network-level interventions are ineffective or even detrimental.

Furthermore, up to 94 inhibitory heads are identified in BERT, and their ablation is found to improve model performance. Control experiments and sensitivity analyses rule out alternative explanations such as out-of-distribution perturbations and community granularity effects, and confirm a functional compensation mechanism underlying the intervention–pruning discrepancy. The observed divergence is attributed to constraints imposed by the attention mask. Causal masking forces tightly coupled collaborative networks to handle sequential uncertainty, whereas bidirectional masking permits independent head specialization, leading to functional polarization.

Methodologically, this work represents the first operational transfer of the functional connectivity, community detection, and virtual lesion framework from network neuroscience to Transformer interpretability research. From an applied perspective, differentiated guidance is provided for model compression, suggesting that causal models should employ community-level pruning to preserve collaborative networks while bidirectional models can safely undergo aggressive head-level pruning.

## 1 Introduction

Transformer-based language models have acquired rich syntactic knowledge through pretraining, even without explicit grammatical supervision (Rogers et al., 2020). Yet a fundamental question remains, namely, where this knowledge is stored and how it is organized during language processing.

Prior work in BERTology has largely focused on individual attention heads, revealing that specific heads correlate with particular syntactic relations in bidirectional models(Rogers et al., 2020; Clark et al., 2019; Voita et al., 2019). However, these analyses suffer from two limitations. First, they treat heads in isolation and overlook synergistic interactions. Second, they center almost exclusively on bidirectional models, leaving causal language models (e.g., the GPT series) conspicuously absent from systematic comparison. It therefore remains unknown whether distinct Transformer architectures organize linguistic knowledge in fundamentally different ways.

Our investigation draws inspiration from systems neuroscience. In the brain, complex cognitive functions are subserved by distributed, functionally specialized neural networks rather than single neurons(Kumar et al., 2024). Lesion studies demonstrate that damaging a single node often yields little deficit due to redundancy and com-

pensation, whereas broad network damage can be catastrophic. This network-level perspective offers a compelling lens for examining artificial language models. We therefore ask whether syntactic knowledge in Transformers resides in a few elite heads or is distributed across a collaborative network.

We formulate two opposing hypotheses. Under the single-head specialization view, syntactic knowledge is concentrated in a small number of indispensable heads(Voita et al., 2019). Under the distributed network view, syntactic knowledge resides in a functional network of multiple heads, such that individual heads are redundant but collective activity is essential(Michel et al., 2019). We further conjecture that the attention masking mechanism (causal versus bidirectional) is a pivotal constraint determining which hypothesis holds(Vaswani et al., 2017; Cohen and Gokaslan, 2020; Radford et al., 2019; Devlin et al., 2019). Causal models (e.g., GPT-2), lacking access to future context, may be forced to develop multi-head collaborative representations. Bidirectional models (e.g., BERT), with full contextual access, may instead permit individual heads to evolve highly specialized functions.

To test these hypotheses, a multi-level causal intervention framework is applied to four models, namely, GPT-2 Small, GPT-2 Medium, BERT-base, and RoBERTa-base (Radford et al., 2019; Devlin et al., 2019; Liu et al., 2019). A classic workflow from network neuroscience is adapted into a three-step pipeline comprising functional connectivity construction, functional network identification via graph clustering, and virtual lesion intervention. This protocol is supplemented by single-head permutation tests, direct comparisons of network- and head-level ablation, pruning experiments, cross-task validation, and sensitivity analyses.

Our results reveal a clear bimodal distribution across architectures. In causal models (GPT-2), no single head is indispensable(Lee et al., 2024; Michel et al., 2019), yet ablating the identified syntactic network causes severe performance degradation, corroborating the distributed network hypothesis. In bidirectional models (BERT, RoBERTa), only a handful of individually significant heads are found, and network-level ablation is either ineffective or beneficial(Voita et al., 2019). A large number of inhibitory heads are observed in bidirectional models, and their ablation consistently improves syntactic performance(Nam et al., 2025). Control experiments rule out alternative explanations, and the syntactic networks in GPT-2 are shown to be task-specific(Voita et al., 2019; Lee et al., 2024).

The core contributions of this work are fourfold. First, a cross-architectural causal analysis reveals that attention masking fundamentally constrains how syntactic knowledge is stored internally. Second, a dedicated, collaboratively functioning syntactic network is discovered and validated within causal language models. Third, the first systematic report of widespread inhibitory heads in bidirectional models is provided, along with evidence that their ablation actively boosts syntactic performance. Fourth, architecture-dependent guidance is offered for model compression, indicating that causal models benefit from community-level pruning whereas bidirectional models can safely undergo aggressive head-level pruning.

## 2 Related Work

This work intersects three areas, includes functional interpretability of attention heads, circuit discovery via causal intervention, and structured pruning for model compression. Our methodology further draws on network neuroscience.

### 2.1 Functional Interpretability of Attention Heads

BERTology has sought to uncover the linguistic functions of individual attention heads(Rogers et al., 2020). Certain BERT heads correlate with specific syntactic dependencies like direct objects and prepositional objects(Clark et al., 2019) and hierarchical functional specialization across layers(Voita et al., 2019). Some heads implicitly track dependency relations via decoded attention trees(Htut et al., 2019). However, these studies predominantly rely on correlational evidence and treat individual heads in isolation, leaving causal necessity and potential synergistic network effects largely unexamined. Our work addresses both gaps through network-level causal intervention.

### 2.2 Circuit Discovery and Mechanistic Interpretability

Causal intervention methods now enable localization of task-specific computational subgraphs. The circuit for indirect object identification in GPT-2 small(Wang et al., 2023); Conmy et al. automated circuit discovery(Conmy et al., 2023); and this circuit extends to variable binding in LLaMA-

13B(Davies et al., 2023). These efforts, however, remain confined to single architectures, leaving open the question of how architectural constraints such as attention masking, shape circuit organization. We provide a systematic cross-architectural comparison of causal versus bidirectional models.

### 2.3 Model Pruning and Structured Compression

Transformer attention modules exhibit substantial redundancy and many BERT heads can be removed with minimal performance loss(Michel et al., 2019). Yet existing pruning methods typically treat all heads uniformly, ignoring potential architectural divergence in how knowledge is organized(Kovaleva et al., 2019; Sajjad et al., 2022). If causal models rely on distributed head networks, single-head pruning may inadvertently disrupt functional circuits. Our experiments test this hypothesis and yield architecture-dependent pruning guidance.

### 2.4 Cross-Disciplinary Transfer From Network Neuroscience

Concepts from neuroscience have previously informed neural network analysis from receptive fields in vision to "knowledge neurons" in NLP(Kumar et al., 2024; Geva et al., 2021). These transfers, however, remain at the single-unit level, whereas network neuroscience emphasizes distributed, interacting functional networks. We operationally transfer this full framework of functional connectivity, community detection, and virtual lesioning(Pascual-Leone et al., 2000) to attention head analysis, enabling the first network-level causal study of syntactic knowledge organization across architectures.

### 2.5 Unique Position

In summary, relative to prior work, we advance from single-head correlation to network-level causal intervention, extend circuit discovery to systematic cross-architectural comparison, and provide architecture-dependent pruning guidance grounded in these interpretability findings.

## 3 Methods

To adjudicate between the single-head specialization and distributed network hypotheses, a multi-level causal intervention framework is designed across four Transformer models, with methodological inspiration drawn from network neuroscience. The framework comprises spanning single-head, network-level, and direct comparative analyses.

### 3.1 Models and Evaluation Tasks

**Models.** Four Transformer models of comparable scale were selected includes causal language models GPT-2 Small (12 layers, 144 heads) and GPT-2 Medium (24 layers, 384 heads), alongside bidirectional models BERT-base-uncased (12 layers, 144 heads) and RoBERTa-base (12 layers, 144 heads). The GPT-2 Small/Medium contrast further allows the effect of model scale to be probed.

**Evaluation Task.** The subject-verb agreement subset of BLiMP (regular plural subject verb agreement 1), comprising 1,000 minimal pairs, is employed. For GPT-2, auto-regressive log-likelihood is computed; for BERT and RoBERTa, masked language modeling loss is computed. A prediction is considered correct if the loss of the grammatical sentence is lower than that of its ungrammatical counterpart. Accuracy is reported.

**Semantic Control Task.** The distractor agreement subset of BLiMP (distractor agreement relational noun) is introduced, which requires agreement judgments in the presence of interfering relative clauses. Task specificity is assessed by comparing causal effects across syntactic and semantic tasks.

### 3.2 Single-Head Permutation Test

The causal importance of individual heads is assessed via sample permutation tests. For each head, baseline accuracy on the syntactic subset is first recorded without intervention. The pairing between correct and incorrect sentences is then randomly shuffled, the accuracy drop is recomputed, and this procedure is repeated 100 times to construct a null distribution. The one-tailed p-value is defined as the proportion of null drops that equal or exceed the true drop. Benjamini–Hochberg FDR correction ($q < 0.05$) is applied to control for multiple comparisons.

### 3.3 Identification of Syntactic Functional Networks

**Functional Connectivity Matrix.** For each head, output activations on all samples of the syntactic subset are extracted and mean-pooled over the sequence dimension to obtain a sentence-level activation vector. The Pearson correlation between

every pair of such vectors is computed, yielding a functional connectivity matrix.

**Community Detection and Candidate Selection.** The matrix is treated as a weighted adjacency graph, and the Louvain algorithm (resolution $\gamma = 1$ is applied to partition heads into functional communities. To isolate a syntax-dedicated network, the entropy of community labels on the semantic task is computed for each syntactic community:

$$H = -\sum_{c} p(c) \log_2 p(c)$$

Lower entropy indicates that heads within the community exhibit more consistent functional behavior on the semantic task. The community with the minimum entropy is selected as the candidate syntactic network for all subsequent network-level interventions.

### 3.4 Network Zero Ablation and Random Baseline

Zero ablation is performed on the candidate network by setting the outputs of all its constituent heads to zero, and the accuracy drop on the syntactic subset is recorded. To control for head count, an equal number of heads is randomly sampled (without layer constraints) ten times, and the mean random drop is computed. The network drop is compared to the random drop via a one-tailed t-test ($p < 0.05$). A ratio significantly greater than 1 is taken as evidence of causal importance beyond that of an equal number of randomly chosen heads.

### 3.5 Direct Comparison: Network vs. Top- *k* Single Heads

To rule out the alternative explanation that network importance merely reflects a few individually critical heads, a direct comparison is conducted. Let *k* be the number of heads in the candidate network. The top-*k*heads ranked by single-head importance (accuracy drop from §3.2) are selected. The entire network is zero-ablated, and simultaneously the top-k*k* heads are zero-ablated; the resulting accuracy drops are compared. Paired observations are obtained across three BLiMP subsets, and a one-tailed paired t-test ($\alpha = 0.05$) is performed, supplemented by bootstrap resampling (10,000 iterations).

### 3.6 Pruning Experiments

Two pruning strategies are compared. Community-preserving pruning retains all heads in the candidate network and removes the remaining heads in ascending order of single-head importance until the target sparsity is reached. Importance pruning removes heads strictly by ascending importance, without community constraints. Retention ratios of 30%, 50%, and 70% are tested, and zero-shot accuracy on the syntactic subset is evaluated without any fine-tuning.

### 3.7 Cross-Task and Cross-Subset Validation

Cross-task validation is integrated into §3.4. For cross-subset stability, community detection is independently run on four BLiMP syntactic subsets: subject-verb agreement, determiner-noun agreement, anaphor number agreement, and *wh*-question subject-verb agreement. The Adjusted Rand Index (ARI) is computed between each pair of community partitions. A higher mean ARI is interpreted as evidence of a more stable syntactic community structure across phenomena.

### 3.8 Implementation Details

Experiments are implemented with PyTorch 2.0 and Hugging Face Transformers 4.30. Models are loaded from official pretrained checkpoints: gpt2, gpt2-medium, bert-base-uncased, and roberta-base. BLiMP is accessed via the `datasets` library, with 1,000 samples per subset. A fixed random seed (42) is used throughout. Statistical tests are performed with SciPy 1.10; FDR correction employs `statsmodels` ($\alpha = 0.05$). Computation is conducted on an Apple M2 Pro (16GB RAM). A full experimental pipeline per model requires approximately 18 hours, with total runtime around 72 hours. Code will be released open-source.

## 4 Results

### 4.1 Single-Head Permutation Tests: No Indispensable Heads in Causal Models

No attention head in GPT-2 Small or GPT-2 Medium reached significance after FDR correction ($q < 0.05$). In GPT-2 Medium, accuracy drops were virtually zero for all heads ($p = 1.0$). In contrast, BERT-base and RoBERTa-base each yielded exactly one significant head, with accuracy drops of 0.215 and 0.405, respectively (Table 1, Figure 1).

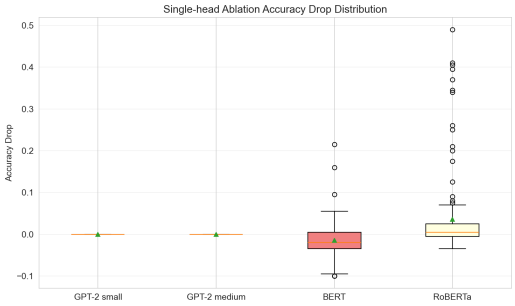


Figure 1: Single Head Ablation Accuracy Drop Distribution

| Model | Sig. Heads | Max Drop | Neg. Heads |
|---|---|---|---|
| GPT-2 Small | 0 | $\approx 0.02$ | 0 |
| GPT-2 Medium | 0 | $\approx 0.00$ | 0 |
| BERT-base | 1 | 0.215 | 94 |
| RoBERTa-base | 1 | 0.405 | 43 |

Table 1: Single-head permutation test results.

### 4.2 Discovery of Negative Heads: Pervasive Inhibitory Heads in Bidirectional Models

BERT contains 94 negative heads (heads whose permutation increased accuracy); RoBERTa contains 43. Neither GPT-2 model contains any.

**Attention pattern analysis.** The mean attention weights of BERT's negative heads did not differ in magnitude from non-negative heads (both 0.120), but their distributional patterns differed significantly ($t = 2.279, p = 0.027$). Thus, negative heads actively emit interfering signals despite appearing normally engaged.

### 4.3 Network Zero Ablation: Syntactic Networks in Causal Models Are Causally Necessary

Table 2 reports network-to-random accuracy drop ratios. In GPT-2 Small and Medium, ablating the syntactic network produced drops $1.98\times$ and $4.18\times$ greater than random baselines. Ratios on the semantic control task fell below 1 ($0.82\times$ and $0.59\times$), confirming syntactic specificity. In BERT and RoBERTa, network ablation yielded near-random or beneficial effects ($1.18\times$ and $-0.05\times$)(Figure2, 3). Control experiments confirmed that disruption stems from information removal, not out-of-distribution perturbations (Appendix 5 ).

Table 2: Network zero ablation effect ratios.

| Model | Syntactic Task | Semantic Task |
|---|---|---|
| GPT-2 Small | $\mathbf{1.98}\times$ | $0.82\times$ |
| GPT-2 Medium | $\mathbf{4.18}\times$ | $0.59\times$ |
| BERT-base | $1.18\times$ | $1.16\times$ |
| RoBERTa-base | $-0.05\times$ | $0.19\times$ |

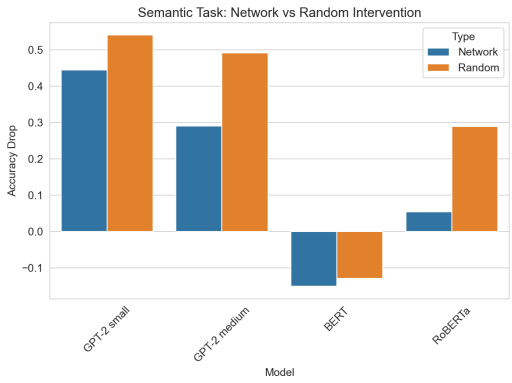


Figure 2: Semantic Task Comparison

### 4.4 Direct Comparison: Network Ablation vs. Top-$k$Single Heads

Network ablation caused significantly larger drops than ablating an equal number of top-ranked single heads in GPT-2 Small (0.220 vs. 0.060; $p = 0.005$) and GPT-2 Medium (0.092 vs. 0.010; bootstrap $p = 0.000$). The reverse held for BERT and RoBERTa (Table 3, Figure 4).

Table 3: Network vs. top-$k$ single-head ablation.

| Model | Net. Drop | Top-$k$ Drop | $p$-value |
|---|---|---|---|
| GPT-2 Small | **0.220** | 0.060 | 0.005 |
| GPT-2 Medium | **0.092** | 0.010 | 0.000 |
| BERT-base | $-0.053$ | **0.147** | 0.995 |
| RoBERTa-base | $-0.000$ | **0.200** | 0.9997 |

### 4.5 Pruning Experiments

Community-preserving pruning outperformed importance pruning in GPT-2 Small at moderate retention ratios (63%: +4.0%; 50%: +3.0%, Figure 5), but the advantage disappeared at 30% retention and in GPT-2 Medium (Table 4). Representational similarity analyses confirm substantial compensatory capacity of residual networks (see Appendix 6)

### 4.6 Cross-Task Specificity and Cross-Subset Stability

Semantic-task ratios below 1 (Table 2) confirm that the identified network is syntax-specific. Cross-subset ARI values for GPT-2 Small ranged from

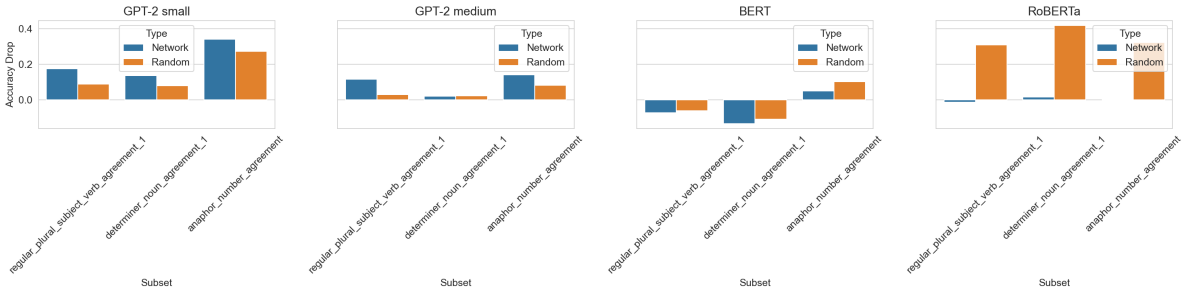


Figure 3: Syntax Network Comparison

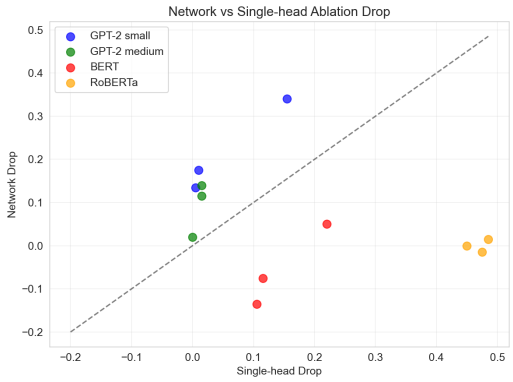


Figure 4: Network Ablation versus Top- $k$ Single-Head Ablation

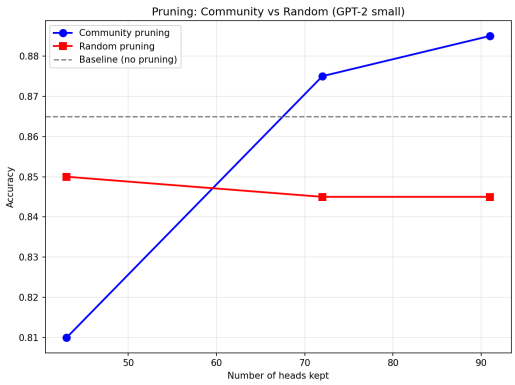


Figure 5: Pruning Curves for GPT-2 Small

Table 4: Pruning results at varying retention ratios.

| Model | Retention | Comm. Acc. | Imp. Acc. |
|---|---|---|---|
| GPT-2 Small | 63% | **0.885** | 0.845 |
| | 50% | **0.875** | 0.845 |
| | 30% | 0.810 | **0.850** |
| GPT-2 Medium | 50% | 0.919 | **0.925** |
| | 30% | 0.904 | 0.904 |

0.7 to 0.8, indicating stable community structure across syntactic phenomena.

### 4.7 Sensitivity Analysis of BERT Community Detection

Varying the number of Louvain communities ($k = 2$–$10$) in BERT revealed no community whose ablation significantly exceeded random baselines. The strongest community ($k = 4$, 40 heads) achieved a ratio of 1.31 but was not statistically significant ($p = 0.1221$) and fell short of the top single-head drop (0.200 vs. 0.215). Full data are provided in Appendix 7.

## 5 Discussion

The experimental findings collectively delineate a clear cross-architectural divergence. In this section, these findings are synthesized into a core mechanistic hypothesis, the theoretical implications of the pruning paradox are unpacked, the functional significance of negative heads is explored, and the interdisciplinary methodological contributions and inherent limitations of this work are reflected upon.

### 5.1 Attention Masking as a Constraint on Functional Organization

As summarized in Sections 4.1–4.4, a systematic double dissociation is revealed. This divergence is attributed to the attention masking mechanism, the most fundamental architectural difference between causal and bidirectional models. In causal

models such as GPT-2, only preceding tokens are accessible at each prediction step (**?**). This leftward constraint prevents any single head from independently maintaining robust contextual representations, as the information captured by an individual head is inherently local. Syntactic tracking across a sequence therefore compels a modular collaborative strategy, whereby functionally similar heads collectively sustain a distributed network that robustly encodes syntactic relations. Individual heads are redundant, but network collapse causes functional failure.Conversely, in bidirectional models such as BERT and RoBERTa, each head accesses the full left and right context simultaneously (Devlin et al., 2019). Global information availability enables individual heads to independently perform specialized syntactic subtasks, reducing the necessity for collaboration and steering the model toward single-head specialization and functional polarization. A small number of elite heads emerge, while many others become redundant or inhibitory, actively generating negative interference. The global information advantage of bidirectional masking thus fosters a sharply polarized distribution of head functionality.Attention masking is thus reframed as a fundamental determinant of internal representational geometry, not merely a technical constraint for autoregressive generation. In causal models, the sequential bottleneck forces distributed redundancy, a hallmark of robust systems. In bidirectional models, the absence of this bottleneck permits extreme specialization, yielding a sparse, winner-take-all functional landscape. This aligns with recent theoretical work suggesting that architectural constraints shape the inductive biases governing emergent modularity in deep networks.

### 5.2 Distinguishing Causal Intervention from Permanent Excision

An apparent discrepancy is presented by our data. In GPT-2 Medium, zero ablation reveals a strong causal effect of the syntactic network, yet community-preserving pruning offers virtually no advantage over importance-based pruning. This discrepancy illuminates a fundamental distinction between causal intervention and permanent structural removal.

Zero ablation captures the real-time contribution of a network within the intact model. When community outputs are abruptly set to zero, the remaining components have no opportunity to compensate within that forward pass, and the immediate impact of network failure is observed directly. Pruning, by contrast, entails permanent structural excision, after which surviving components can functionally compensate through residual pathways. As shown in Section 4.5.1, pruned and original model representations remain highly similar even at deep layers, demonstrating the substantial compensatory capacity of residual networks.

Important methodological implications follow. Pruning experiments reveal the feasibility of compression, whereas online causal intervention reveals the causal role of functional units in the native model.

This insight extends beyond Transformers to any complex system exhibiting structural redundancy. In systems neuroscience, the distinction between acute lesions (revealing real-time necessity) and chronic lesions (triggering compensatory plasticity) is well established (Pascual-Leone et al., 2000; Lomber, 1999). Analogous caution is warranted when interpreting ablation studies in artificial neural networks, because the absence of a pruning effect does not imply the absence of a causal role in the intact network.

### 5.3 Functional Significance of Negative Heads and Model Sparsity

The discovery of 94 negative heads in BERT constitutes an unexpected finding. Prior work has noted that disabling certain BERT heads improves performance (Kovaleva et al., 2019), but the cross-architectural distribution of negative heads on a syntactic task had not been systematically quantified. This finding reinforces the functional polarization hypothesis for bidirectional models and offers a fresh perspective on model sparsity.

An intuitive interpretation is that these heads learned features during pretraining that compete with syntactic processing. They may serve broader language modeling objectives, yet in the narrow context of syntactic judgment their activations constitute interference. As shown in Section 4.2.1, negative heads exhibit attention magnitudes indistinguishable from those of non-negative heads, indicating active participation rather than silence. Causal intervention is thus indispensable for revealing true functional direction.

From a compression standpoint, the existence of these negative heads aligns with evidence that many attention heads in Transformers are redun-

dant and can be safely removed (Michel et al., 2019; Voita et al., 2019). This further corroborates the architecture-dependent pruning guidance proposed in this work.

Why negative heads emerge predominantly in bidirectional models remains open. One speculative account is that masked language modeling encourages the maintenance of multiple, sometimes conflicting, token hypotheses. Heads specialized for low-probability or irregular completions may manifest as interference in controlled syntactic tasks. Tracing their emergence across training is left to future investigation.

### 5.4 Interdisciplinary Methodological Transfer

The analytic framework of network neuroscience (functional connectivity estimation, community detection, and virtual lesion intervention) is operationally transferred in this work to the analysis of Transformer attention heads. Prior work has established correspondences between individual heads and brain regions (Kumar et al., 2024), but the present contribution lies in the first complete transfer of a network-level framework, which constitutes an operationalizable paradigm rather than a mere metaphor.

In systems neuroscience, a brain region is understood through the large-scale functional networks in which it is embedded, not through single-unit recordings alone (Bullmore and Sporns, 2009). Analogously, an attention head in Transformers must be examined within its inter-head collaborative communities, not through single-head ablation alone. Our four-step pipeline (single-head permutation, functional connectivity estimation, community detection, and network intervention) faithfully replicates the neuroscience paradigm of single-unit recording, functional connectivity, brain network parcellation, and transcranial magnetic stimulation virtual lesion (Pascual-Leone et al., 2000).

The intrinsic limitations of this analogy must be acknowledged. Attention head activations differ fundamentally from biological action potentials, and community boundaries depend on similarity metrics and clustering parameters. Nevertheless, the value of this methodological transfer is not diminished. Transformer interpretability research stands to gain richly from the neuroscience toolbox.

Several promising avenues are opened by this cross-disciplinary bridge. Graph-theoretic measures could characterize the organizational principles of attention head communities. Directed influences among heads could be inferred by dynamic causal modeling. Insights from artificial neural networks about efficient distributed representations may, in turn, inform theories of neural coding.

## 6 Limitation and Future Direction

Several limitations of the present study warrant acknowledgment and motivate future investigation.

**Scope of syntactic phenomena.** The analysis focused on subject-verb agreement. Although cross-subset ARI analyses suggest reasonable generalization, whether other syntactic phenomena like long-distance dependencies or island constraints follow the same organizational principles remains to be systematically verified.

**Community detection sensitivity.** Community detection relied on the Louvain algorithm and semantic label entropy. More systematic hyperparameter sweeps and cross-validation with alternative algorithms would further bolster robustness.

**Training dynamics.** Only end-of-pretraining checkpoints were examined. Analyzing intermediate checkpoints could trace the developmental trajectory of functional organization and reveal when architectural constraints exert their influence.

**Model scale.** The largest model examined was GPT-2 Medium. Whether larger models exhibit different organizational principles remains an open question warranting investigation with models such as GPT-3 or LLaMA.

**Confounding of training objective and attention mask.** The functional divergence between causal and bidirectional LMs is attributed to attention masking, yet these models also differ in training objectives (autoregressive versus masked language modeling). Because the two factors are coupled, their independent contributions cannot be fully disentangled. Future work could vary the training objective while holding the mask constant, for example, using prefix language models or encoder-decoder architectures. It is tentatively speculated that training objective may influence the degree of functional differentiation, whereas attention mask determines the coupling tightness of collaborative networks.

## 7 Acknowledgements

## References


Ed Bullmore and Olaf Sporns. 2009. Complex brain networks: graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3):186–198.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Vanya Cohen and Aaron Gokaslan. 2020. Opengpt-2: open language models and implications of generated text. *XRDS*, 27(1):26–30.

Arthur Conmy, Augustine N. Mavor-Parker, Aengus Lynch, Adrià Garriga-Alonso, and Stefan Heimersheim. 2023. Towards Automated Circuit Discovery for Mechanistic Interpretability. *Advances in Neural Information Processing Systems*, 36(NeurIPS).

Xander Davies, Max Nadeau, Nikhil Prakash, Tamar Rott Shaham, and David Bau. 2023. Discovering Variable Binding Circuitry with Desiderata. (Icml):1–6.

Jacob Devlin, Ming Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1:4171–4186.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer Feed-Forward Layers Are Key-Value Memories. *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 5484–5495.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do Attention Heads in BERT Track Syntactic Dependencies?

Olga Kovaleva, Alexey Romanov, Anna Rogers, and Anna Rumshisky. 2019. Revealing the dark secrets of BERT. 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 4365–4374, Hong Kong, China. Association for Computational Linguistics.

Sreejan Kumar, Theodore R Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A Norman, Thomas L Griffiths, Robert D Hawkins, and Samuel A Nastase. 2024. Shared functional specialization in transformer-based language models and the human brain. *Nature Communications*, 15(1):5523.

Isabelle Lee, Joshua Lum, Ziyi Liu, and Dani Yogatama. 2024. Causal Interventions on Causal Paths: Mapping GPT-2's Reasoning From Syntax to Semantics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. (1).

Stephen G Lomber. 1999. The advantages and limitations of permanent or reversible deactivation techniques in the assessment of neural function. *Journal of Neuroscience Methods*, 86(2):109–117.

Paul Michel, Omer Levy, and Graham Neubig. 2019. *Are sixteen heads really better than one?* Curran Associates Inc., Red Hook, NY, USA.

Andrew Nam, Henry Conklin, Yukang Yang, Thomas Griffiths, Jonathan Cohen, and Sarah-Jane Leslie. 2025. Causal Head Gating: A Framework for Interpreting Roles of Attention Heads in Transformers. (NeurIPS).

Alvaro Pascual-Leone, Vincent Walsh, and John Rothwell. 2000. Transcranial magnetic stimulation in cognitive neuroscience – virtual lesion, chronometry, and functional connectivity. *Current Opinion in Neurobiology*, 10(2):232–237.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI*. Accessed: 2024-11-15.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

Hassan Sajjad, Fahim Dalvi, Nadir Durrani, and Preslav Nakov. 2022. On the effect of dropping layers of pre-trained transformer models. *Computer Speech and Language*, 77.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the Wild: a Circuit for Indirect

Object Identification in Gpt-2 Small. *11th International Conference on Learning Representations, ICLR 2023*.

# A Example Appendix

Table 5: Control intervention results for mean substitution and token shuffling.

| Intervention | Subset | Baseline | Post-Int. | Drop | Ratio |
|---|---|---|---|---|---|
| Token Shuffling | SV Agreement | 0.935 | 0.740 | 0.195 | 1.11× |
| | Det-Noun | 0.930 | 0.850 | 0.080 | 0.99× |
| | Anaphor | 0.995 | 0.860 | 0.135 | 1.07× |
| Mean Substitution | SV Agreement | 0.935 | 0.815 | 0.120 | **4.62**× |
| | Det-Noun | 0.930 | 0.910 | 0.020 | 0.91× |
| | Anaphor | 0.995 | 0.855 | 0.140 | 1.73× |

Table 6: Cosine similarity after community-preserving and importance pruning.

| Retention | Heads Kept | Layer | Comm. Pruning | Imp. Pruning |
|---|---|---|---|---|
| 30% | 115 | 1 | 0.9956 | 0.9949 |
| | | 6 | 0.9996 | 0.9996 |
| | | 12 | 0.9991 | 0.9993 |
| | | 18 | 0.9980 | 0.9982 |
| | | 23 | 0.9858 | 0.9807 |
| 50% | 192 | 1 | 0.9970 | 0.9973 |
| | | 6 | 0.9998 | 0.9998 |
| | | 12 | 0.9996 | 0.9996 |
| | | 18 | 0.9990 | 0.9990 |
| | | 23 | 0.9955 | 0.9924 |
| 70% | 269 | 1 | 0.9977 | 0.9980 |
| | | 6 | 0.9999 | 0.9999 |
| | | 12 | 0.9998 | 0.9999 |
| | | 18 | 0.9996 | 0.9996 |
| | | 23 | 0.9975 | 0.9866 |

Table 7: Full sensitivity analysis results for BERT community detection.

| $k$ | Community | Size | Drop | Ratio |
|---|---|---|---|---|
| 2 | 0 | 69 | 0.115 | 0.85 |
| | 1 | 75 | 0.170 | 1.24 |
| 3 | 0 | 48 | 0.155 | 1.07 |
| | 1 | 53 | 0.110 | 0.78 |
| | 2 | 43 | −0.110 | −0.74 |
| 4 | 0 | 40 | 0.200 | **1.31** |
| | 1 | 32 | −0.095 | −0.64 |
| | 2 | 35 | 0.040 | 0.28 |
| | 3 | 37 | −0.080 | −0.49 |
| 5 | 0 | 32 | −0.005 | −0.04 |
| | 1 | 28 | 0.035 | 0.25 |
| | 2 | 30 | −0.060 | −0.43 |
| | 3 | 27 | 0.010 | 0.07 |
| | 4 | 27 | −0.045 | −0.32 |

(For $k = 6$–10, all effect ratios $< 1.1$; see supplementary material.)

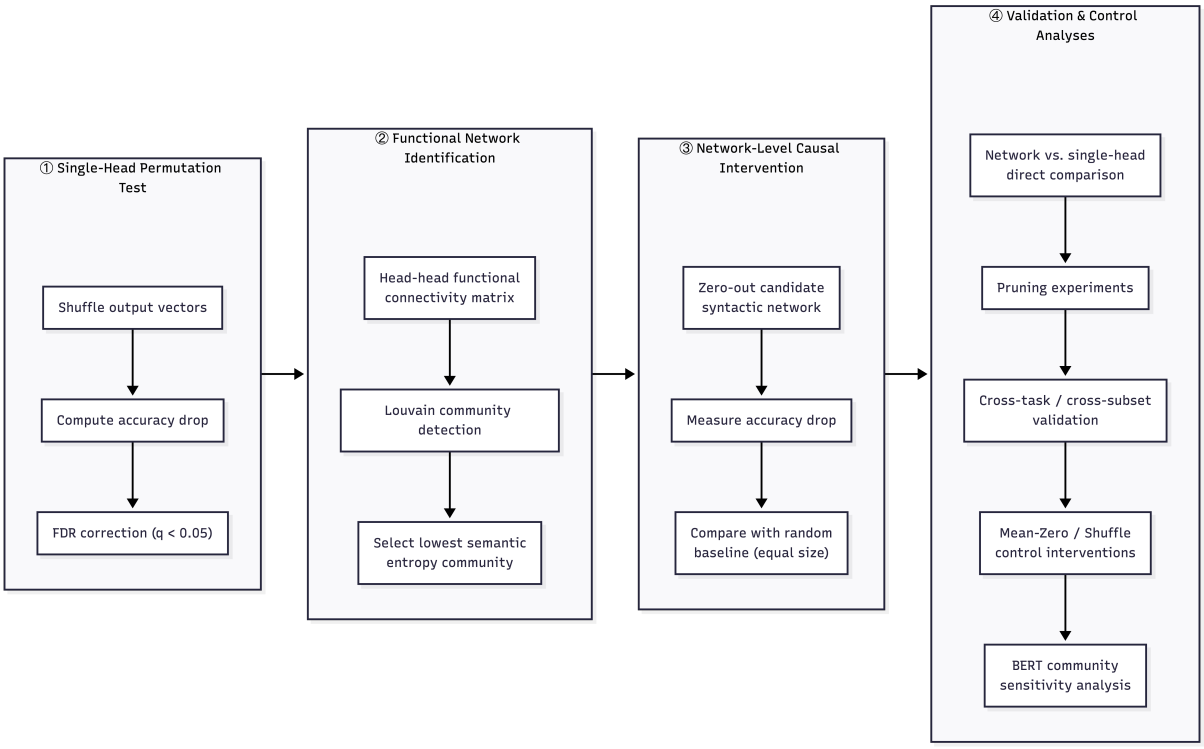


Figure 6: General Workflow

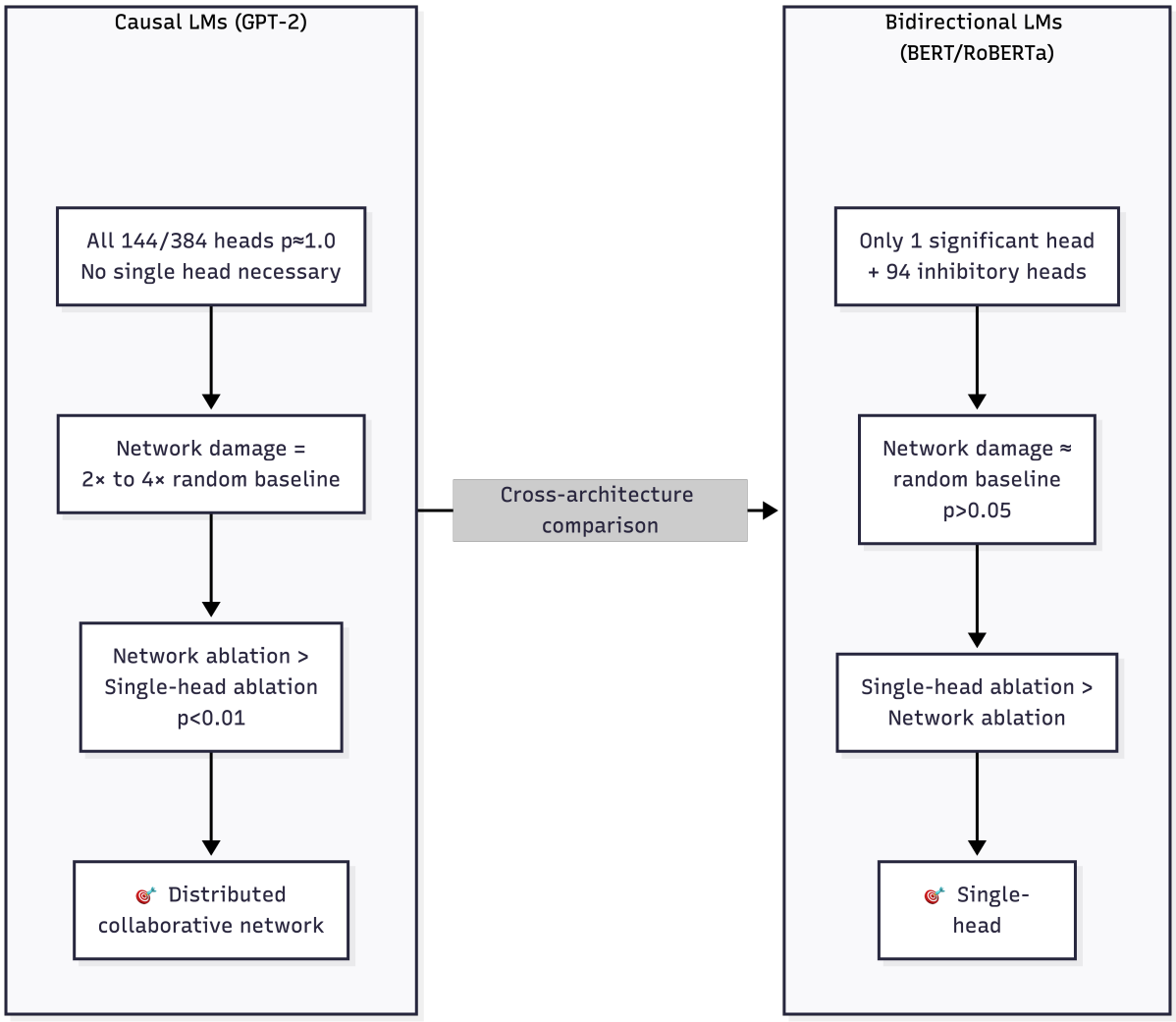


Figure 7: Cross Model Comparison

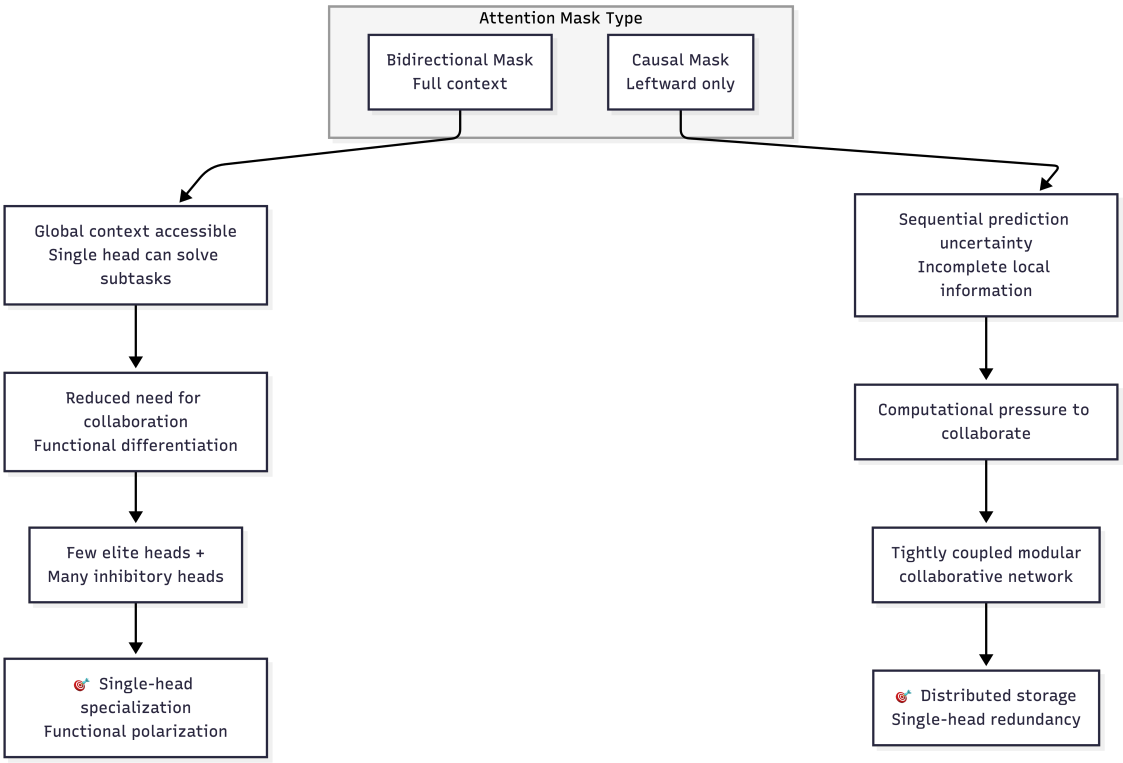


Figure 8: Attention Mask Type

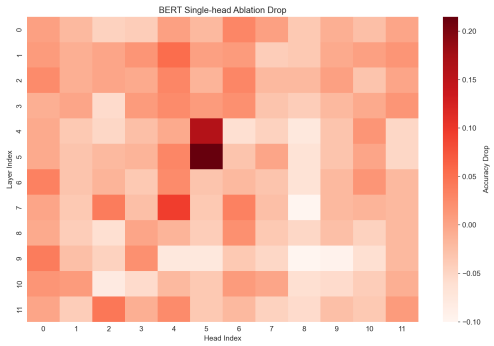


Figure 9: BERT Single Head Ablation Drop

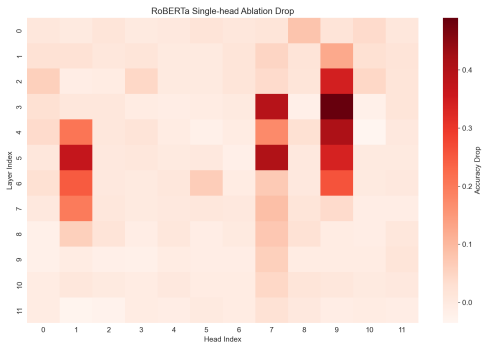


Figure 10: RoBERTa Single Head Ablation Drop

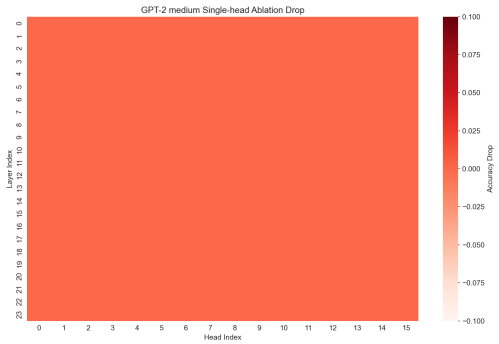


Figure 11: GPT-2 Medium Single Head Ablation Drop

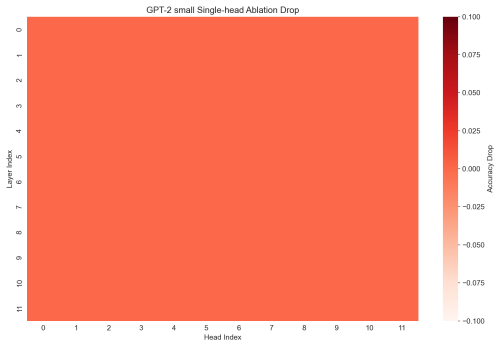


Figure 12: GPT-2 Small Single Head Ablation Drop

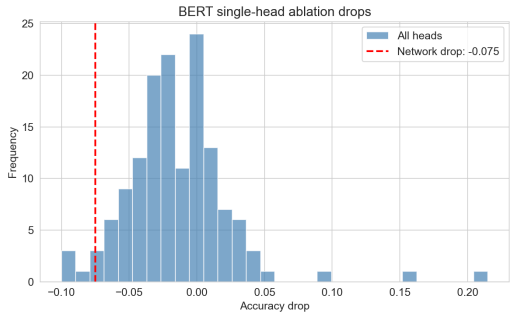


Figure 13: BERT Single Head Ablation Drop

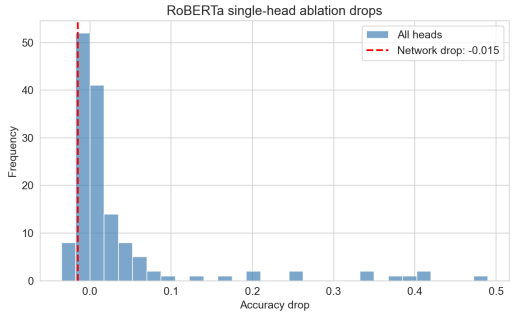


Figure 14: RoBERTa Single Head Ablation Drop

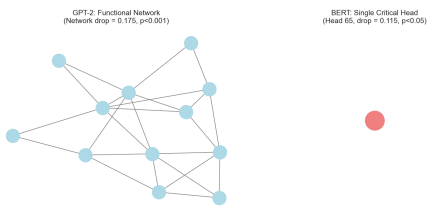


Figure 15: Two models Mechanism