Probing the Attention Representation of Filler-Gap Dependency in Transformers

Ruoqing Yao and **Pranav Anand**
Department of Linguistics, University of California, Santa Cruz
{ryao10|panand}@ucsc.edu

1 Introduction

Filler-gap dependencies (FGDs) require integration across unbounded intervening material, posing an important test case for the context-sensitivity of language models (LMs). Leveraging surprisal at the gap site, prior work has shown that RNN, LSTM, and transformer LMs show *unlicensed-gap* and *filled-gap effects* for FGDs within a clause (Wilcox et al., 2024; Kobzeva et al., 2025; Chang et al., 2025). As the dependency crosses multiple clausal layers, both effects attenuate (Wilcox et al., 2024; Kobzeva et al., 2025), an effect magnified with overt *that* intervening Cs (complementizers) (Kobzeva et al., 2025). Moreover, causal intervention experiments have shown that the internal representations in LMs for different categories of FGDs show commonalities that are modulated by structural frequency of the FGD category and filler type (Boguraev et al., 2025). The current study investigate how attention probing and ablation shed light on these results. We identify heads indexing FGDs near the gap site, explore how attention shifts in those heads with intervening material, find correlation of attention shift and surprisal results, and test head ablation's effect on surprisal.

2 Methodology

Following Wilcox et al. (2024) and Kobzeva et al. (2025), our items involve embedded *wh*-clauses, with 0C, 1C, or 2C clausal boundaries between the *wh*-FILLER and a a VERB object gap:

(1) I know FILLER . . . [(that) [. . . VERB __ . . .

We examine the attention matrices in the forward pass of GPT-2, which has 12 layers, each of which contains 12 heads. To quantify the attention from the VERB to the FILLER, from the full attention tensor, we extract the 12×12 attention weight where the query corresponds to the VERB token and the key corresponds to the FILLER *wh*-token.

3 Experiments

3.1 Experiment 1: Head Identification

In **Experiment 1a**, we created 15 sets of items as in (2a-b) to locate any heads that selectively attend from VERB to a FILLER. We also examined relative clauses (2c) to ensure that attention was not due to lexical differences of *wh*-words and *that*.

(2) a. I know {what, *that} the lion devoured __ yesterday.
b. I know {*what, that} the lion devoured the rabbit yesterday.
c. I know the rabbit that the lion devoured __ yesterday.

We identified the layer 5 head 2 (denoted $h_{(5,2)}$) and the layer 8 head 9 (denoted $h_{(8,9)}$) as selectively indexing FGDs: they both assign significantly greater attention from VERB to FILLER in grammatical FGD sentences (2a), relative to sentences without a gap (2b) (Figure 1a). As relative clauses show a similar attention profile, this indexation by $h_{(5,2)}$ and $h_{(8,9)}$ appears non-lexical. $h_{(5,2)}$ shows the greatest attention score among all 144 heads in [+*wh*] and [+RC], and $h_{(8,9)}$ has the second best score in [+*wh*] and shows the biggest difference between [+*wh*] and [+*th*]. A linear regression shows a significant positive correlation between the attention at both heads and the *filled-gap* surprisals ($\beta_{(5,2)} = 6.33$ and $\beta_{(8,9)} = 5.61$ respectively, $p_{(5,2)}, p_{(8,9)} < .001$).

Experiment 1b investigates how $h_{(5,2)}$ and $h_{(8,9)}$'s attentions change as linear or clausal structure intervenes between VERB and FILLER. We created 20 item sets, containing a) one 1C item with one clause intervening and three 0C items with a b) conjunction, c) mid-length modifier,or d) long-length modifier. Results show that $h_{(5,2)}$ attends to FILLER least in 1C items, and that all 0C

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 258–261

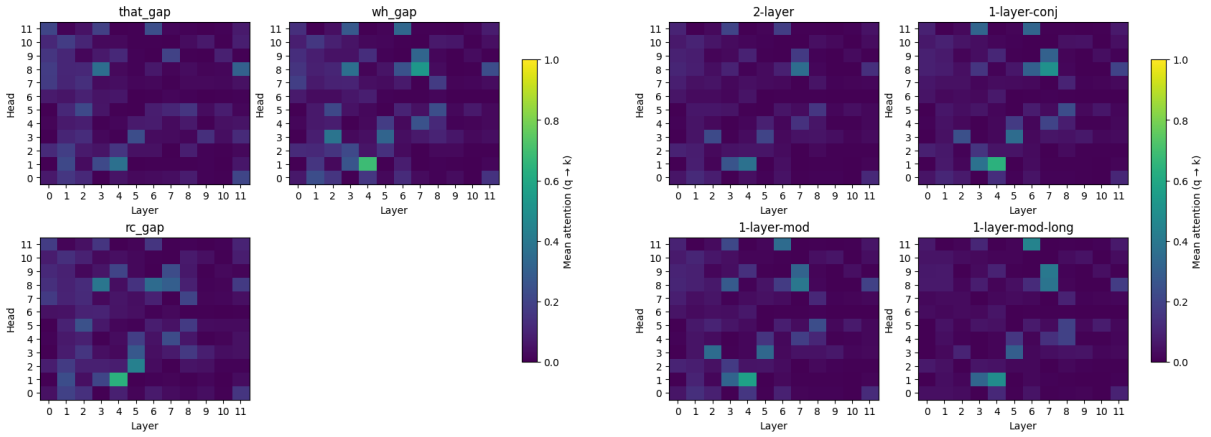

Figure 1: Experiment 1a (left) and 1b (right) results.

items show similarly high attention to FILLER (Figure 1b). This profile suggests that $h_{(5,2)}$ and $h_{(8,9)}$ indexes FGDs, independent of linear separation but sensitive to clausal intervention.

Experiment 1c tests whether the attention is attenuated by islands, following (Wilcox et al., 2024)'s surprisal finding. We adopted 50 sets of items from Kobzeva et al. (2025). Each set contains three conditions: a whether-island, a CNPC island, and a grammatical non-island baselineas in (3).

(3) a. *whether-island*: *I know what my brother confirmed whether our aunt brought ...

b. *CNPC island*: *I know what my brother confirmed the claim that our aunt brought ...

c. *non-island*: I know what my brother confirmed that our aunt brought ...

We found $h_{(5,2)}$ has higher attention across the conditions compared to $h_{(8,9)}$ (Figure 2). Both $h_{(5,2)}$ and $h_{(8,9)}$ show a contrast between non-island and CNPC islands, but only $h_{(8,9)}$ show the same contrast for *whether*-islands. $h_{(5,2)}$'s attention is relative to the type of islands – its attention is more attenuated in CNPC island, a strong island, than *whether*-island, a weak island. On the other hand, $h_{(8,9)}$ indexes islands categorically as a whole regardless of the island types.

3.2 Experiment 2: Overt Complementizers

A puzzling finding in Kobzeva et al. (2025) is that overt (*that*) *C* neutralize unlicensed- and filled-gap effects. We observed above that $h_{(5,2)}$ and $h_{(8,9)}$

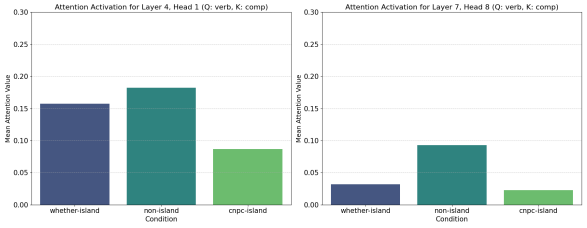

Figure 2: Experiment 1c Results

attends to FILLER, and **Experiment 2** tests whether this can explain the neutralization in surprisal with overt *that C*'s. We adapt 20 1C and 2C item sets from Kobzeva et al. (2025), varying for 2C items whether BOTH *C*'s, only the HIGH one, only the LOW one are overt, or NONE. Items are schematized in (4).

(4) a. 1C: She knows what he heard $[_{CP_{loc}}$ (that) the priest revealed at the party.

b. 2C: She knows what the newspaper reported $[_{CP_{int}}$ (that) he heard $[_{CP_{loc}}$ (that) the priest revealed at the party.

Overall, we observe an *"attend closest C"* pattern of $h_{(5,2)}$ (Figure 3). A linear regression reveals that $h_{(5,2)}$'s attention to the FILLER significantly correlates with filled-gap surprisal ($\beta = 12.59$, $p < .001$). In 1C cases, when the local C_{loc} is overt, $h_{(5,2)}$ attends at VERB principally to C_{loc}, and attention to FILLER is substantially reduced when compared to non-overt C_{loc}. In 2C cases, with no overt *C*s, $h_{(5,2)}$ at VERB selectively attends to FILLER, though the magnitude reduces relative to 1C cases, and it further diminishes with overt *C*s. Noticeably, C_{loc} is principally attended to by $h_{(5,2)}$ whenever overt to consistent magnitudes similar to

that in 1C. When BOTH intervening *C*s are overt, FILLER and C_{int} are similarly attended to well below C_{loc}. With only C_{int} overt, its magnitude still surpasses FILLER, though to a smaller extent. We note a similar *"attend closest C"* pattern in relative clauses as in (2c): VERB attends to the overt C more than the NP FILLER.

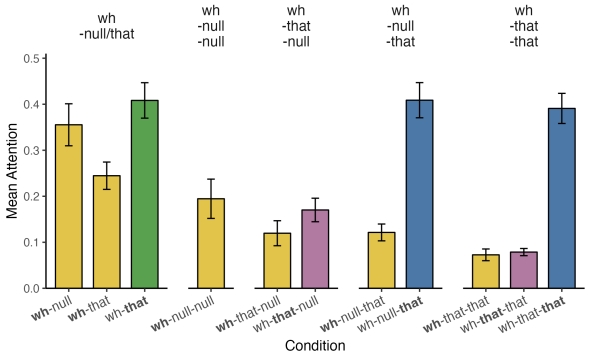

Figure 3: Experiment 2 Results on $h_{(5,2)}$

$h_{(8,9)}$ shows a different pattern (Figure 4). Except the *wh-that-that* condition, it assigns the most attention to the FILLER *wh*-word. In the 1C condition, having C_{loc} reduced its attention to FILLER, but its attention to FILLER is still significantly higher than C_{loc}. The 2C condition shows a similar pattern: including overt C_{int} and C_{loc} reduces the attention to FILLER, but the FILLER attention remains the highest. $h_{(8,9)}$'s attention to FILLER is reduced by the inclusion of overt local *C*s, yet its attention does not redistribute.

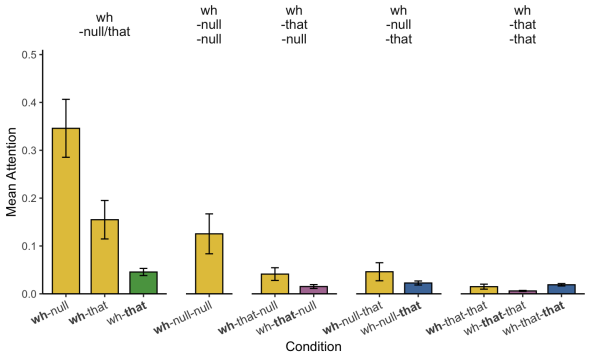

Figure 4: Experiment 2 Results on $h_{(8,9)}$

3.3 Experiment 3: Head Ablation

We conduct a one-head-at-a-time ablation study in GPT-2. Using TransformerLens (Nanda and Bloom, 2022), we zero out each attention head during the all forward pass blocks and measure the resulting change in surprisal. We measure the surprisal difference between the ablated model and the

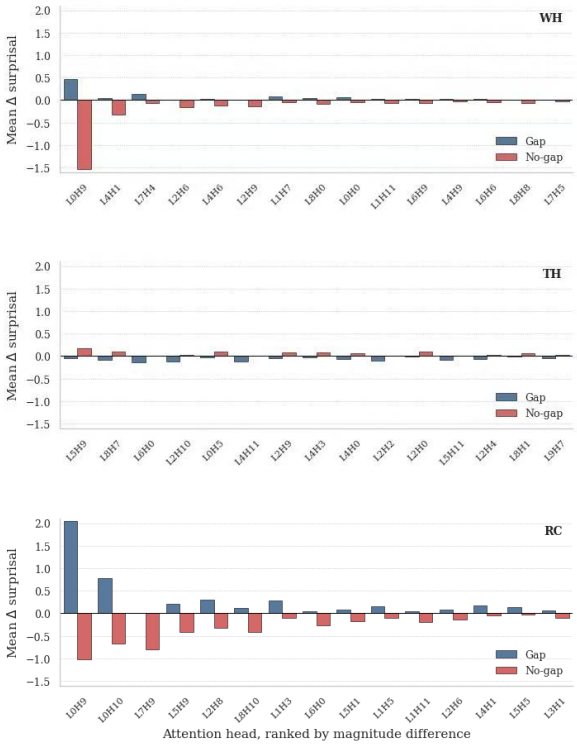

Figure 5: Ranked mean ablation effects for top 15 attention heads for [+*wh*], [+*th*], and [+RC] conditions. Heads are ordered by the magnitude differences.

full model for both [+gap] and [-gap] across three conditions [+*wh*], [+*th*], and [+RC] in (2). A [+*wh*] example is repeated in (5) with the measured target underlined. If a head contributes to FGD computation, its ablation should weaken the model's sensitivity to the relevant contrast. For ungrammatical targets, the ablated run of a FGD-correlated head should produce lower surprisal than the full model because the high surprisals of the ungrammatical continuations should be attenuated, and the reverse for grammatical targets.

(5) a. I know what the lion devoured __ yesterday ...

b. *I know what the lion devoured the rabbit ...

The results (Figure 5) show that a small subset of attention heads yields the predicted surprisal differences rather than an uniformly distributed effect across the network. The strongest ablation effect is found for $h_{(1,10)}$, a head that does not show strong VERB-to-FILLER attention in the probing analyses, suggesting that some causally relevant heads may encode the dependency independent of our assumed VERB-to-FILLER attention. Crucially, $h_{(5,2)}$,

one of the heads identified in the attention-probing experiments, produces the second strongest ablation effect in [+*wh*] and mild effects in [+RC], indicating that its attention correlation with FGD extends to ablation. By contrast, $h_{(8,9)}$ did not produce any of the predicted surprisal changes under ablation. These results suggest that GPT-2's FGD behavior is supported by a partially specialized but distributed set of heads, and that attention weights alone provide a meaningful but incomplete picture of the model-internal mechanisms underlying syntactic dependency processing.

[+*th*] show a different set of effective heads, with much smaller ablation effects than [+*wh*] and [+RC]. This is expected, since [+*th*] lacks an active filler. The contrast suggests that GPT-2's dependency computations are more engaged when a potential filler is present.

4 Discussion

Our results of $h_{(5,2)}$ provide a potential explanation for Kobzeva et al.'s (2025) observation that surprisal effects attenuate with overt *C*s: $h_{(5,2)}$ redistributes its attention from FILLER to the most local C. This might indicate that the transformer has a less reliable representation of FGDs across overt *C*s, and hence is less able to distinguish whether there are FGDs or not. If we take this correlation seriously, it suggests that unbiased language learners, such as LMs, fail to fully learn correct grammars of FGDs, which should not be sensitive the overtness of C heads (Ritchart et al., 2016).

Why might this be? One possibility is GPT2's training data, in which 1C structures are likely vastly more common than 2C ones, particularly relative clauses with C_{loc}, where FILLER and C_{loc} are both reliable indicators of FGDs. The model might overfit to attend to C_{loc} whenever it's present, reflecting $h_{(5,2)}$ attention to both FILLERs and Cs. Relevant here is Boguraev et al.'s (2025) investigation of representational difference across FGD type (matrix and 1C embedded questions), RCs, (pseudo)clefts, and topicalization): representations for a source FGD generalize to the others, in correlation with source frequency. However, differences in token order seem to matter (RCs generalize surprisingly poorly, despite being the most prevalent construction), as do embedding (matrix questions generalize little to embedded questions). Similarly, we might wonder if 1C cases generalize to 2C cases in GPT-2, and how the presence of overt *that* interacts with generalization.

We observed a functional distinction of $h_{(8,9)}$ from $h_{(5,2)}$. It is attenuated but not attracted by C_{int} and C_{loc}, sensitive to islands categorically, and not effective when ablated. These findings point to the possibility that LMs have allocated distinct linguistically meaningful representations from the training data to individual attention heads. They may each track FGDs in a unique way, and may or may not collaborate to represent FGDs. It is, admittedly, puzzling why ablating $h_{(8,9)}$ does not produce any effect, and why ablating $h_{(1,10)}$ is highly effective without attending from VERB to FILLER. Both of these we leave to future investigation.

Despite our findings reported here, we do not claim that $h_{(5,2)}$ and $h_{(8,9)}$ have causal effects on FGD performance in LMs, as attention probing and ablation are both indirect, correlational approximation of the FGD phenomena in the LMs. For attention probing, a head attending strongly from VERB to FILLER does not necessarily mean the head is *computing* the filler-gap dependency, so that any effects found in attention probing is merely a model-internal correlation to a given pattern. While ablation investigates effect, it does so by rather strongly altering the learned network, and its destructive effect to the model's computation makes it hard to interpret cleanly. We leave further experiments with causal intervention methods such as activation patching to future research.

References

S. Boguraev, C. Potts, and K. Mahowald. 2025. Causal interventions reveal shared structure across english filler–gap constructions. In *Proceedings of the 2025 EMNLP*, pages 25032–25053.

C.-Y. Chang, X. Huang, H. Nasir, S. Storks, O. Akingbade, and H. Dai. 2025. Mind the gap. In *EMNLP 2025*, pages 15060–15076.

A. Kobzeva, S. Arehalli, T. Linzen, and D. Kush. 2025. Learning filler-gap dependencies with neural language models. *JML*, 144:104663.

N. Nanda and J. Bloom. 2022. Transformerlens. https://github.com/TransformerLensOrg/TransformerLens.

A. Ritchart, G. Goodall, and M. Garellek. 2016. Prosody and the that-trace effect. In *WCCFL 33*, pages 320–328. Cascadilla Proceedings Project.

E. Wilcox, R. Futrell, and R. Levy. 2024. Using computational models to test syntactic learnability. *LI*, 55(4):805–848.