

Compliments to the Complementizer: Indirect Routing of Syntactic Structure in a Transformer Language Model

Lasse van den Berg
Yale University

Robert Frank
Yale University

Abstract

Unlike recurrent neural networks, which have a fixed path for accumulating information from the input, transformer language models’ use of attention accommodates more flexible regimens for how information is routed through the model. We investigate how syntactic information about long-distance filler-gap dependencies is routed in transformers. We focus specifically on German sentences involving “verb-second” structures. In such sentences, a language model will need to encode information about a fronted element and move it to downstream positions in order to modulate expectations about possible continuations. Through interchange interventions on contrastive minimal sentence pairs, we identify that transformers make use of an indirect path routing syntactic information first from the fronted element to the complementizer and from there to the downstream position at which predictions are made. This non-trivial detour of information through the complementizer aligns with linguistic theories of the structure of long-distance filler-gap dependencies. This information is echoed in abstractly similar dependencies in English and Italian, but not in cases of dependencies that are semantic rather than syntactic in nature. We then explore the content of the model’s internal representations by training a hierarchical Tensor Product Decomposition Network (TPDN) that identifies the model’s representation of grammatical features. We can then reconstruct model activations at the complementizer position under the assumption that this neurosymbolic representation captures the essential aspects of the model’s encoding of grammatical structure. Causal interventions using these reconstructions confirm that this manipulation of the grammatical features produces the predicted changes in downstream continuation predictions.

1 Introduction

Transformers (Vaswani et al., 2017) trained on next-token prediction exhibit signs of systematic syntactic knowledge (Goldberg, 2019; Wilcox et al., 2024). They are capable of tracking long-distance dependencies while producing grammatical continuations in such constructions. A growing body of work in mechanistic interpretability aims to reverse-engineer the internal computations underlying these abilities.

We investigate a theoretically well-defined phenomenon: $\bar{A}$ -movement in German verb-second (V2) clauses. In such constructions, a constituent is displaced from its canonical base position to the left periphery (SpecCP) of the clause. German V2 constructions are particularly well-suited to mechanistic analysis: overt and often unambiguous morphological case, number and gender marking, obligatory V2 word order, and a long-distance filler-gap dependency provide clean contrastive structure. For these constructions, linguistic theory predicts that the complementizer head (C^0), occupied by the finite verb in V2 sentences, serves as the functional head of the clause. We explore whether the manner in which transformers route information is consistent with the linguistic role of C^0 . We decompose residual stream activations involved in processing these filler-gap dependencies

to test whether their representational format of syntax is consistent with Tensor Product Representations (Smolensky, 1990).

2 Related work

Wang et al. (2022) establish activation patching as a method well-suited to tracing information flow in syntactic agreement resolution tasks. Through causal interventions, they map out a circuit for indirect object identification in GPT-2 small. Boguraev et al. (2025) use distributed interchange interventions to show that transformer representations of English filler-gap dependencies transfer across construction types, suggesting shared abstract representational formats. Notably, intervention results in their appendix show elevated lexical effects at the complementizer positions. While the authors do not remark on this finding, it notably aligns with the complementizer’s relevance as the clause’s functional head and our findings of complementizer-mediated information flow. We extend this line of inquiry to German V2 sentences and additionally characterize the representational format of identified activations via hierarchical TPDNs.

Marks et al. (2025) use sparse autoencoders to decompose LLM activations into interpretable features, then trace causal circuits through them via linear-approximation attribution. One of their validation tasks is subject-verb agreement across relative clauses and prepositional phrases, showing models detect syntactic features such as subject number and propagate them across intervening clause boundaries. Further work has performed causal masking interventions to assess which tokens transformers actually rely on for agreement (Li et al., 2023) and introduced benchmarks that use templated counterfactual or contrastive minimal pairs to compare interpretability methods on their causal efficacy at explaining model behavior (Arora et al., 2024).

Furthermore, Li et al. (2023) perform causal masking interventions to test which tokens transformers actually rely on for agreement. Masking interventions reveal two distinct underlying mechanisms for syntactic agreement.

3 Background

3.1 $\bar{A}$ -movement in linguistics and LLMs

Modern syntactic analysis assumes not only that the words in a sentence are hierarchically organized into phrases, but also that such phrases can be manipulated via movement operations, i.e., displacement from a base-generated position to a position higher in the sentence structure. Displacement operations that move a constituent to one that is not associated with a specific argument of a predicate are classified as $\bar{A}$ -movement. One instance of this phenomenon is found in (English) *wh*-questions, whose processing in LLMs has been studied (Wilcox et al., 2024; Boguraev et al., 2025).

In the case of German V2 structures, the non-argument landing site occurs at the left edge of the clause, and is associated with discourse-level interpretation such as topicalization, focus, relativization, and interrogation.

- (1) Der Schüler soll den Lehrer besuchen.
the.NOM student should the.ACC teacher visit
'The student should visit the teacher.'
- (2) Den Lehrer soll der Schüler besuchen.
the.ACC teacher should the.NOM student visit
'The student should visit the teacher.'

(1) exhibits a subject-initial V2 clause, where it is the subject (the student) that is the first constituent and is immediately followed by the finite (modal) verb (should). The fact that this initial phrase is the subject can be gleaned from the nominative case marking expressed

on the determiner ('der'). In contrast, in (2) it is the morphologically marked accusative object (the teacher) that is fronted to the left edge of the clause. While (1) and (2) both express that the student is required to meet with the teacher, the difference in word order impacts the emphasis on either the subject or the object as the topic of the sentence.

Beyond the interest of the $\bar{A}$ -movement phenomenon for linguistic analysis, this structure is also a convenient testbed for mechanistic interpretability. $\bar{A}$ -movement, like in (2), creates a dependency between two positions that may be separated by an arbitrary number of tokens. Constituents following the fronted phrase must "know" what kind of constituent occupied the fronted position in SpecCP, in order to correctly predict properties of the continuation, such as agreement.

- (3) Dem Lehrer soll der Schüler helfen.
 the.DAT teacher should the.NOM student help
 'The student should help the teacher.'

For instance, (2) and (3) differ in the case of the fronted object (the teacher); in (2) it is accusative, and in (3) it is dative. This means that the verbs that can occur at the end of the sentence will differ: the verb 'treffen' (visit) allows an accusative object and is therefore a valid continuation in (2), whereas the verb 'helfen' (help) allows a dative object, and is therefore a possible continuation in (3). These examples demonstrate, then, that a transformer language model needs to maintain syntactic information about the fronted NP and make this information available at subsequent token positions.

According to linguistic theories of V2 structures, the complementizer position that hosts the finite verb plays a central role in this construction: the properties of the fronted phrase are matched with or checked by the complementizer, which in turn communicates them to the remainder of the clause. Proposals have been made with respect to the role of the complementizer in other grammatical dependencies, including wh-questions and the licensing of verbal mood (Rizzi, 1997; Giorgi, 2009). Beyond considerations of abstract grammar, complementizers have also been found to play a distinctive role in human processing of long-distance dependencies: Gibson & Warren (2004) find evidence that the linking of a fronted wh-phrase to its gap position proceeds via reactivation at intermediate complementizer positions.

Even if language models are successful at modeling the structure of V2 clauses, it is not obvious that they will manipulate information in the same way that people do, under the linguistic analysis. While this processing is constrained by the autoregressive nature of transformers, there are several potential pathways of information flow the model could exploit. RNNs, for instance, are architecturally constrained to pass information through the hidden state of all subsequent tokens before affecting a later token. Transformers' attention mechanism allows for more direct information travel across long distances.

We investigate how Qwen3-4B (Yang et al., 2025), a decoder-only transformer model trained on 119 languages, Llama 3.2-3B (Dubey et al., 2024), a transformer trained on several languages, and DiscoResearch/Llama3-German-8B (Plüster et al., 2024), a German pre-trained model of the Llama 3-8B architecture, succeed at capturing these $\bar{A}$ -movement dependencies in German V2 sentences.

4 Behavioral localization through causal interventions

Activation patching (Vig et al., 2020; Wang et al., 2022) is a causal intervention method in which the activation at a specific layer and token position is replaced during the forward pass with the corresponding activation from a contrastive input. Given a source run on input x_{src} and a destination run on input x_{dst} , patching the residual stream at layer l and position p is defined as a run on x_{dst} where one activation from x_{src} is patched:

$$h_{dst}(l, p) \leftarrow h_{src}(l, p) \quad (1)$$

The forward pass of the patched destination continues from this modified state, with possible behavioral changes in downstream logit predictions. The extent of this behavioral change serves as a proxy for the importance of a particular activation and the information it contained with regards to the studied behavior. Specifically, we use an average logit difference metric

$$\Delta = \text{mean}(\text{logits}[\text{cont}_{\text{src}}]) - \text{mean}(\text{logits}[\text{cont}_{\text{dst}}]) \quad (2)$$

where $\text{logits}[I]$ denotes indexing over the logits. It compares the mean logit difference for two sets of diagnostic continuation logits cont_{src} and cont_{dst} , one for each sentence prefix. For instance, a contrast sentence prefix pair with fronted XPs differing only in case (accusative or dative) would be judged by their preference for infinitive verb continuations that permit an accusative or dative object respectively.

An activation’s recovery score (i.e., its ability to recover contrastive behavior from one patched activation alone) is established by comparing the difference of logit difference scores between, on the one hand, the patched run of the model and the destination run, as compared to the difference between the logit differences for the source and destination baselines:

$$s = \frac{\Delta_{\text{patch}} - \Delta_{\text{dst}}}{\Delta_{\text{src}} - \Delta_{\text{dst}}} \quad (3)$$

The recovery score of patching one activation is thus a proportion of the logit contrast between the source and destination baselines. A patched Δ that is identical to the destination run yields a value of 0 (indicating no causal effect of patching) while a Δ that is identical to the source run yields a value of 1 (indicating full recovery of source behavior from patching one activation alone). The results for activations across positions and layers then reveal the locus of important activations that are uniquely implicated in the measured behavior.

4.1 Dataset

To test the model, we construct a corpus of contrastive sentence pairs. These pairs should not differ across many dimensions simultaneously (different nouns, choice of modal verb, etc.), as any differing feature dimension, including ones not relevant to the linguistic difference under study, can impact the co-occurrence statistics that amplify or suppress the measured logits. In order to eliminate the possibility that an observed effect is not attributable to differences other than one of interest, we aim to construct minimal pairs that are identical except for a single syntactic contrast, such as the case and/or phrase type of the fronted constituent. For example:

- (4) a. Den Mann soll die Frau (treffen). [ACC NP fronted]
 b. Dem Mann soll die Frau (helfen). [DAT NP fronted]

These pairs are matched for lexical content and syntactic structure everywhere except the case feature of the fronted noun phrase, which in turn determines acceptable infinitive verbs (in parentheses). We would expect that the pre-final position in (4a) should output higher logits for an accusative-selecting verb like *treffen* and lower ones for a dative-selecting verb like *helfen*. By patching from the forward pass of (4a) into that of (4b) at one layer and position (ranging from the first to pre-final token) we can trace where the model encodes case contrast.

Analyzing the internal behavior of a model requires a large number of minimal pairs, as small variations in lexical content can substantially impact behavior. To our knowledge, no such carefully controlled corpus exists for German verb-second sentences. Instead, we create a dataset of synthetically generated sentence pairs. Within each pair, one fronted constituent is in the accusative case and the other is in the dative case. All other aspects are kept equal. Each sentence follows the templatic structure (5).

- (5) {Fronted NP} {Finite Modal} {Subject NP} {Infinitive Verb}

To avoid nonsensical sentences (e.g. “The book helps the kitchen”) we use only animate nouns referencing people of varying occupations, gender and interpersonal roles with varying syntactic number. To inflect nouns, determiners and verbs we use the finite-state automata from DWDSmor (Nolda, 2024; Schmid et al., 2004). We select infinitive verb continuations that are compatible with the sentence’s fronted constituent with the help of verb frames in the GermaNet dataset (Hamp & Feldweg, 1997; Henrich & Hinrichs, 2010). For nouns and verbs we select lemmas with high lemma frequencies using the DeReWo – Corpus-Based Lemma and Word Form Lists (Belica et al., 2014). The sentences are generated by iterating over the template variables. Each word of each generated sentence is annotated with syntactic features (number, case, gender and part-of-speech), constituent-level information (NP, finite verb, infinitive verb) and positional information.

In addition, we create smaller sets of minimal pairs for two other construction types where syntactic features differ in the fronted constituent:

1. Constructions contrasting whether the fronted constituent is demonstrative (that man) or interrogative (which man). We measure logits for the “.” and “?” tokens respectively (Appendix A.1).
2. Constructions contrasting the verbal mood licensed by a matrix verb. We measure logits for the verb ‘to be’ in the indicative and subjunctive mood as congruent with the licensing matrix verb (Appendix A.2).

We also add constructions where the minimal contrast is semantic rather than syntactic and where linguistic theory does not necessarily predict routing through the complementizer. In one case, we analyze logits of gender congruent pronouns that are coreferential with the fronted noun (Appendix A.3).

4.2 Activation patching

In order to produce grammatically well-formed continuations, relevant information must make its way from the inputs corresponding to the fronted phrase to the deepest layers of the network that yield the logit outputs at the point of the last, pre-verbal token. Due to transformers’ architecture and autoregressive nature, information can move rightward in the sequence via attention heads in a variety of paths:

1. Information remains within the transformer columns corresponding to the fronted phrase and is then transferred to the final token position only at the deepest layers.
2. Information is quickly shunted from the fronted constituent to the early to mid layers of the final token position and then percolates upward from there.
3. Information is highly distributed and present in many activations (across layers and tokens) between the contrastive pairs in SpecCP and the site of logit prediction.
4. Information takes a discrete, but indirect path from the fronted phrase to the final position.

We perform activation patching on 100 randomly sampled sentence pairs for each construction type.

Figures 1 and Figures 6a and 7a illustrate our results for the constructions: for the cases with strong syntactic contrast (Figure 1), patching across minimal pairs reveals a consistent and behaviorally significant causal pathway that is linguistically non-trivial. Information travels from the position of the fronted constituent at early layers, then moves to the C^0 position (occupied by a modal or complementizer) at middle layers, and finally propagates to the final token at later layers where the constraints imposed by the case (1a), force (1b), or mood (1c) are resolved. Since linear token-position of the modal verb/complementizer varies across sentences, this information flow has no direct positional explanation. Nonetheless, the causal effect passes through grammatically specific intermediate layers. The fact that the

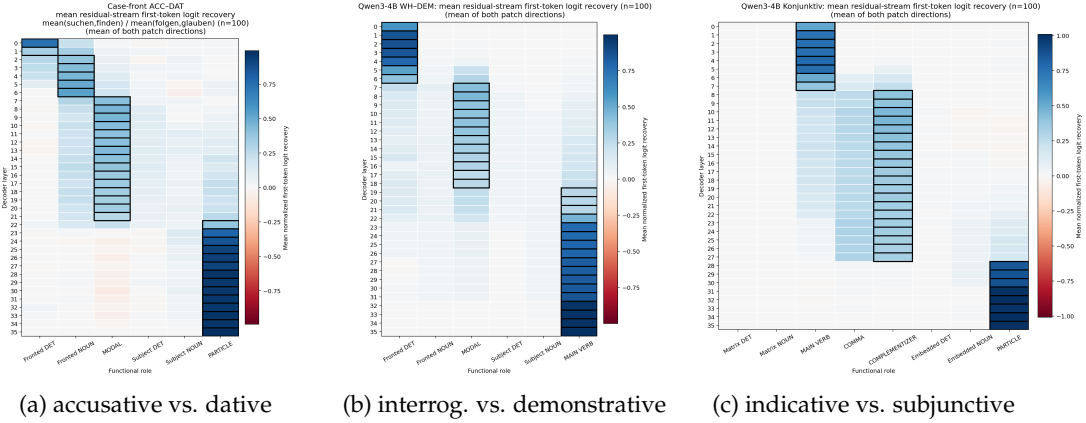


Figure 1: Patching results for Qwen3-4B averaged over 100 randomly sampled sentences for different phenomena. X-axis shows activation recovery scores of the last token in a grammatically significant constituent of the sentence. ‘Particle’ refers to optional filler words that in practice encourage the model to surface POSs matching the diagnostic continuations. Boxes with a black frame indicate the maximum recovery score at that layer.

information routes specifically through C^0 before reaching the final token is linguistically striking and mirrors the theoretical role of C^0 as the licenser of filler-gap relations. The model appears to have internalized an information flow that accords with the routing posited in linguistic analysis. Results for the non-syntactic controls are reported in Appendices 6 and 7.

Looking beyond German, we also find evidence for C^0 -mediated information flow in Italian and English in individual examples. These findings are reported in Appendix C.

4.3 Complementizer masking intervention

To further quantify the complementizer’s role in information flow in grammatical dependencies, we consider several model ablations:

1. **C^0 -bottleneck:** attention heads on the token in C^0 may attend to all previous tokens as usual. For all subsequent tokens, however, attention to any tokens before C^0 is ablated, i.e., any information from tokens before C^0 must pass through C^0 and from C^0 to subsequent tokens.
2. **C^0 -circumvention:** attention heads on all tokens following C^0 can only attend to the tokens preceding C^0 and not to C^0 itself.

C^0 -bottleneck should have a more substantial impact on dependencies in which information flows directly from the initial tokens to the prediction site (such as the noun gender case), and is not modulated by C^0 . In contrast, C^0 -circumvention should degrade a model’s ability to resolve dependencies (such as the syntactic dependencies discussed above), in which information flows through C^0 to the prediction site.

To compute the impact of each ablation, we compare the baseline accuracy, the percentage of cases in which an unablated model assigns a higher probability to the correct or congruent continuation, to the ablated model’s accuracy.

Figure 2 shows the impact of the two ablations on two grammatical dependencies (case and wh/dem) and one semantic dependency (gender) for Qwen3-4B. While (almost) all dependencies show a decline in performance for both ablations, we see the expected interactions. The grammatical dependencies are less impacted by the C^0 bottleneck than the semantic control. In contrast, the grammatical dependencies are more affected by C^0

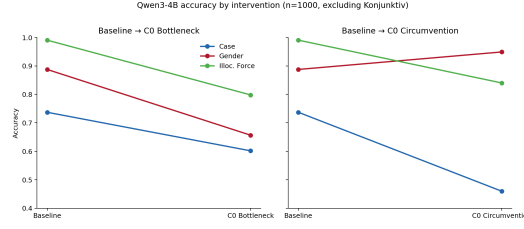


Figure 2: Effects of both ablation types on model accuracy per task.

circumvention than the semantic control, whose performance in fact improves under the intervention. Gender agreement, as the control, was uniquely sensitive to intervention type — sharply impaired under Bottleneck ($OR = 0.45$, $p = 9.3 \times 10^{-14}$) but strongly facilitated under Circumvention ($OR = 7.78$, $p = 1.5 \times 10^{-49}$), a 17-fold swing — whereas interrogative/demonstrative remained impaired relative to case under both conditions ($OR = 0.070$ and 0.17 , respectively) and mood showed only a modest, same-direction advantage in both ($OR = 1.74$ and 3.23).

We report results from activation patching and causal ablations for two other models in Appendices B.1 and 8 respectively.

5 Hierarchical Tensor Product Decomposition Networks

5.1 Tensor Product Representations and Language

Activation patching is a method well-suited to localizing relevant activations and tracing information flow. Understanding how and in what format information is represented requires a deeper analysis or decomposition of the identified activations. In mechanistic interpretability, Sparse Autoencoders (SAEs) (Bricken et al., 2023; Cunningham et al., 2023) have become the standard approach. An over-complete sparse dictionary is trained on a reconstruction objective on large amounts of stored activations, thereby learning a set of interpretable features that span an over-complete basis of activation space. Due to the high level of compositionality that syntactic systems exhibit, SAE-based methods are ill-suited for our investigation of representations at the complementizer head position. Issues like feature-splitting (Bricken et al., 2023) are exacerbated in settings where a small set of features can combine in exponentially many ways.

We hypothesize that models construct and use compositional representations of syntax, given their high levels of syntactic generalization. Tensor Product Representations (TPRs) (Smolensky, 1990) provide a mathematical framework for representing compositional structure in a continuous vector space. At its core, a symbolic structure consists of roles and fillers (e.g. ‘subject’ as a role and ‘man’ as a filler). Each role and filler is embedded as a continuous vector in role- and filler-space respectively. To encode a symbolic structure compositionally, corresponding roles and fillers are bound as the outer product of their corresponding embeddings and summed with other bindings. The resulting matrix of outer product sums can be flattened and optionally undergo a linear transformation to encode compositional structure in activation space.

Formally, let $F = \{f_1, \dots, f_n\}$ be a set of filler vectors and $R = \{r_1, \dots, r_m\}$ be a set of role vectors with $f_i \in \mathbb{R}^{d_f}$ and $r_i \in \mathbb{R}^{d_r}$. The TPR of some symbolic structure determined by a set of role-filler pairs is computed as:

$$T = \sum_i r_i \otimes f_i \quad (4)$$

The resulting tensor T is in space $\mathbb{R}^{d_r \times d_f}$. Crucially, the representation T permits an unbinding operation to recover the underlying fillers or roles that a model could make use of. Recoverability of filler vectors holds as long as filler vectors are linearly independent.

This property makes TPRs an attractive candidate for structured computation in a language model’s activation space.

5.2 Tensor Product Decomposition Networks

Our results from activation patching suggest that syntactic information is copied from SpecCP to the complementizer head position. If transformers do in fact construct and utilize TPR-like structure internally, one would expect the residual stream activations of the middle-layers (as identified by the patching experiments) to compositionally encode information about the complementizer itself and the fronted XP preceding it (SpecCP).

To test hypotheses about underlying TPR-like structure, McCoy et al. (2019) introduce Tensor Product Decomposition Networks (TPDNs). Rather than defining role and filler vectors manually, TPDNs are trained to learn role and filler embeddings. Given hypothesized role-filler pairs for a set of examples, the TPDN learns filler and role vectors that best approximate the model’s corresponding hidden activation vector h :

$$\hat{h} = W_O(\sum_i W_R[i] \otimes W_F[i]) \approx h \quad (5)$$

Unlike SAE-based approaches, TPDN training is supervised and requires a hypothesis about the information that is compositionally embedded in the hidden states. For many research problems in mechanistic interpretability, this is simply not possible, precisely because neural networks approximate input-output mappings that are not easily definable in symbolic terms. Luckily, in the case of German $\bar{A}$ -movement, linguistic theory provides us with very specific hypotheses about what is represented and how. Thus, a syntactically and positionally annotated dataset contains ideal training labels for a TPDN.

5.3 Hierarchical TPDN training for activations at C^0

Given its hierarchical tree-based structure, the syntax of $\bar{A}$ -movement constituents is not well-modeled by a sum of role-filler bindings alone. Legendre et al. (1990) provide a theoretical framework that extends standard TPRs to recursive representations including tree structures. Rather than using vectorized role and filler embeddings, the role and filler vectors themselves can be constructed from a TPR, thereby constructing hierarchical compositional representations.

For the activations at C^0 , we hypothesize the following construction of hierarchically organized filler-role combinations:

$$B(\mathcal{A}) = \sum_{(a,b) \in \mathcal{A}} r(a) \otimes f(b) \quad (6)$$

where each pair (a, b) binds a structural role label a to the feature b that fills it.

Position	Component	Attribute set $\mathcal{A}$
SpecCP	Role ₁	{XP.type $\rightarrow$ XP.type, case $\rightarrow$ case}
SpecCP	Filler ₁	{noun $\rightarrow$ lemma, det $\rightarrow$ det.lemma, number $\rightarrow$ number}
C^0	Role ₂	atomic, learned vector
C^0	Filler ₂	{verb.lemma $\rightarrow$ lemma, number $\rightarrow$ number, person $\rightarrow$ person}

Table 1: Role and filler bindings for the SpecCP and C^0 positions, defined via the tensor-product binding operator in Eq. (6). XP.type $\in$ {NP, PP}.

Layer	SpecCP and C ⁰			C ⁰ -only			SpecCP-only		
	CosSim	R ²	CKA	CosSim	R ²	CKA	Cos Sim.	R ²	CKA
Layer 16	0.955	0.650	0.977	0.909	0.311	0.782	0.912	0.332	0.605
Layer 3	0.949	0.734	0.975	0.936	0.674	0.974	0.803	0.064	0.210
Layer 16 (Rand.)	0.803	-0.477	0.840	0.823	-0.310	0.772	0.819	-0.336	0.591
Layer 3 (Rand.)	0.648	-0.784	0.973	0.663	-0.687	0.972	0.770	-0.077	0.206

Table 2: Reconstruction metrics for hierarchical TPDNs trained on residual stream activations. Comparison between models receiving SpecCP-only, C⁰-only, or both phrasal features.

We train such a hierarchical TPDN ($d_r = 18$, $d_f = 36$) to reconstruct activations stored from runs over a large set of sentences at C⁰¹. For each layer we train a full hierarchical TPDN with access to syntactic information from SpecCP and C⁰ and one TPDN with only SpecCP or only C⁰ as symbolic input. We hypothesize that the middle layer 16 will benefit from the combination of both sets of syntactic information, whereas layer 3 will succeed from information about C⁰ alone and not gain any performance increase from SpecCP. We therefore train and evaluate independent TPDNs at layer 16 (where activation patching indicated relevant information from SpecCP to be present) and layer 3 (where activation patching does not yet suggest the integration of information from SpecCP).

To evaluate reconstruction performance, we compute cosine similarities, R² scores and Centred Kernel Alignment (CKA) for the original activations and the TPDN reconstructions as measures of directional alignment, variance explained and geometric structure preservation respectively. These results are shown in Table 2.

The results from TPDN reconstruction performance verify our hypothesis that middle layers at C⁰ do encode information about the complementizer itself and the preceding fronted XP. Highest explained variance occurs for the TPDN with access to both sets of input information. Furthermore, layer 3 clearly only encodes information about the complementizer. As our patching experiments predict, information about SpecCP is not yet integrated at the earlier layers.

To determine that the compositional nature of TPDN output activations successfully captures the model’s internal syntactic representations, we conduct activation patching experiments. However, instead of patching activations from a contrastive model run, we replace the targeted activation with the output from the TPDN where we alter the SpecCP input to match the contrastive pair. We then compare recovery scores at layer 16 and the C⁰ position to that of the first patching experiment, defining the original patching score as an upper bound for the behavioral effect the TPDN-constructed activation could have. These results are shown in Table 3.

The results show that our causal TPDN intervention affects downstream continuation logits as expected and in a manner congruent with the syntactic input the TPDN receives. The behavioral recovery effects exceed those of standard activation patching in several conditions.

¹Training a TPDN on model activations requires a large amount of data points that must remain compatible with the format. We describe some modifications made to our data in Appendix D.

Intervention	Dat→Acc	Acc→Dat	Acc→Dat PP	Dat PP→Acc	Dat→Acc PP	Acc PP→Dat
Activation Patching	0.404	0.500	0.653	0.434	0.427	0.498
TPDN Patching	0.446	0.385	0.741	0.502	0.273	0.440

Table 3: Mean recovery scores from patching at C⁰ and layer 16 from 1000 sampled sentences. Scores are indicated for regular activation patching and TPDN patching.

6 Discussion

The results from our activation patching experiments demonstrate that C^0 serves as a causal relay for filler-gap dependencies in German V2 constructions. Patching the residual stream at this position and at middle layers reliably shifts the model’s output logits toward the source sentence’s continuation distribution. Despite this form of syntactic routing not being the computationally most direct way to move information from a filler to subsequent tokens, we reproduce this complementizer mediation on other constructions and in other languages, further strengthening our hypothesis that C^0 ’s privileged role in syntactic dependency resolution is a more general phenomenon.

Our TPDN-patching experiments further suggest that the representational format at C^0 is structurally compatible with TPRs. Hierarchical role-filler bindings, each of only rank 1 and constructed from symbolic information alone, induce causally verifiable effects on output behavior comparable to direct activation patching.

Our reconstruction metrics for our hierarchical TPDN outputs indicate that they are imperfect reconstructions. We attribute this finding to the fact that a transformer model likely encodes much more information than syntax alone, especially at middle and later layers. The symbolic information fed into the TPDN does not account for capabilities such as factual recall and co-occurrence statistics, which the model likely acquires during training.

Overall, the results from TPDN training confirm our hypothesis that information from SpecCP is moved to C^0 before propagating to the next-token prediction site.

7 Limitations

Our TPDN work focuses on a relatively narrow set of German $\bar{A}$ -movement constructions, which could be expanded in subsequent work. The supervised training required by the TPDN-method for evaluating the content of network representations requires a large corpus with syntactic feature annotations at the token level, something which remains a bottleneck for generalizing this approach to a broader set of cases. While our hypotheses about the content of hierarchical TPDNs have yielded successful interventions, future work could investigate different types of hierarchical schemas for TPDNs in order to identify those that match network behavior most closely.

Furthermore, a full mechanistic analysis should investigate how compositional structure is internally constructed and used by the model. The attention mechanism is a potential candidate for the unbinding operations that TPRs permit. One might hypothesize that key-query computations identify the C^0 token, while value-vector computation performs an unbinding operation, thereby moving syntactically relevant information such as case to the next-token prediction site.

Our experiments are restricted to 3 models with less than 10 billion parameters each. Future work should investigate the findings across more model sizes and architectures.

8 Conclusion

The computational strategies employed by transformers to resolve long-distance syntactic dependencies remain an area of active investigation. Our work takes a step toward addressing this question by combining causal interventions with representational analysis in a linguistically well-defined test case. Our results about the complementizer head C^0 suggest that transformer models do not merely represent syntactic dependencies implicitly through distributed computation, but route syntactic information in a manner consistent with predictions from linguistic theory.

References

- Aryaman Arora, Dan Jurafsky, and Christopher Potts. Causalgym: Benchmarking causal interpretability methods on linguistic tasks, 2024. URL <https://arxiv.org/abs/2402.12560>.
- Cyril Belica, Marc Kupietz, Harald Lungen, and Rainer Perkuhn. Corpus-based lemma and word form lists (DeReWo). <https://www.ids-mannheim.de/en/s/corpus-linguistics/projects/methods-of-analysis/corpus-based-lemma-and-word-form-lists/>, 2014.
- Sasha Boguraev, Christopher Potts, and Kyle Mahowald. Causal interventions reveal shared structure across English filler-gap constructions, 2025. URL <https://arxiv.org/abs/2505.16002>.
- Trenton Bricken, Adly Templeton, Joshua Batson, Brian Chen, Adam Jermy, Tom Conerly, Nick Turner, Cem Anil, Carson Denison, Amanda Askell, Robert Lasenby, Yifan Wu, Shauna Kravec, Nicholas Schiefer, Tim Maxwell, Nicholas Joseph, Zac Hatfield-Dodds, Alex Tamkin, Karina Nguyen, Brayden McLean, Josiah E Burke, Tristan Hume, Shan Carter, Tom Henighan, and Christopher Olah. Towards monosemanticity: Decomposing language models with dictionary learning. *Transformer Circuits Thread*, 2023. <https://transformer-circuits.pub/2023/monosemantic-features/index.html>.
- Hoagy Cunningham, Aidan Ewart, Logan Riggs, Robert Huben, and Lee Sharkey. Sparse autoencoders find highly interpretable features in language models, 2023. URL <https://arxiv.org/abs/2309.08600>.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and et al. The llama 3 herd of models, 2024. URL <https://arxiv.org/abs/2407.21783>.
- Edward Gibson and Tessa Warren. Reading-time evidence for intermediate linguistic structure in long-distance dependencies. *Syntax*, 7(1):55–78, 2004.
- Alessandra Giorgi. Toward a syntax of the subjunctive mood. *Lingua*, 119(12):1837–1858, 2009. ISSN 0024-3841. doi: <https://doi.org/10.1016/j.lingua.2008.11.008>. URL <https://www.sciencedirect.com/science/article/pii/S0024384109000072>.
- Yoav Goldberg. Assessing BERT’s syntactic abilities, 2019. URL <https://arxiv.org/abs/1901.05287>.
- Birgit Hamp and Helmut Feldweg. GermaNet - a lexical-semantic net for German. In *Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, 1997. URL <https://aclanthology.org/W97-0802/>.
- Verena Henrich and Erhard Hinrichs. GernEiT - the GermaNet editing tool. In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner, and Daniel Tapias (eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC’10)*, Valletta, Malta, May 2010. European Language Resources Association (ELRA). URL <https://aclanthology.org/L10-1180/>.
- G  rardine Legendre, Yoshio Miyata, and Paul Smolensky. Distributed recursive structure processing. *Advances in Neural Information Processing Systems*, 3, 1990.
- Bingzhi Li, Guillaume Wisniewski, and Beno  t Crabb  . Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33, 2023. doi: 10.1162/tacl.a.00531. URL <https://aclanthology.org/2023.tacl-1.2/>.
- Samuel Marks, Can Rager, Eric J. Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models, 2025. URL <https://arxiv.org/abs/2403.19647>.

- R. Thomas McCoy, Tal Linzen, Ewan Dunbar, and Paul Smolensky. RNNs implicitly implement tensor-product representations. In *International Conference on Learning Representations*, 2019. URL <https://openreview.net/forum?id=BJx0sjC5FX>.
- Andreas Nolda. DWDSMOR: SFST/SMOR/DWDS-based German morphology. <https://github.com/zentrum-lexikographie/dwdsmor>, 2024. Accessed: 3/2026.
- Björn Plüster, Manuel Brack, Malte Ostendorff, Pedro Ortiz Suarez, Jan-Philipp Harries, Daniel Auras, Christoph Schuhmann, and Jenia Jitsev. DiscoResearch/Llama3-German-8B. <https://huggingface.co/DiscoResearch/Llama3-German-8B>, 2024. Hugging Face model repository.
- Luigi Rizzi. The fine structure of the left periphery. 1997. URL <https://api.semanticscholar.org/CorpusID:117005733>.
- Helmut Schmid, Arne Fitschen, and Ulrich Heid. SMOR: A German computational morphology covering derivation, composition, and inflection. In *Proceedings of LREC 2004*, 2004.
- Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1):159–216, 1990. ISSN 0004-3702. doi: [https://doi.org/10.1016/0004-3702\(90\)90007-M](https://doi.org/10.1016/0004-3702(90)90007-M). URL <https://www.sciencedirect.com/science/article/pii/000437029090007M>.
- Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. Causal mediation analysis for interpreting neural NLP: The case of gender bias, 2020. URL <https://arxiv.org/abs/2004.12265>.
- Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small, 2022. URL <https://arxiv.org/abs/2211.00593>.
- Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. Using computational models to test syntactic learnability. *Linguistic Inquiry*, 55(4):805–848, 10 2024. ISSN 0024-3892. doi: 10.1162/ling_a.00491. URL https://doi.org/10.1162/ling_a.00491.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

A Other Construction Types

A.1 Interrogative vs. demonstrative determiners in the fronted constituent

Our analysis focuses heavily on sentence pairs where only the fronted noun phrases' cases differ. Another possible dimension of syntactic contrast is that of interrogative and demonstrative determiners. Their choice determines the illocutionary force of a sentence, i.e. whether the sentence is a question or a statement.

We construct a smaller set of sentence pairs analogous to the set with fronted constituents varying in case. However, we keep the case of both sentences as either dative or accusative within a pair. The definite article (e.g. "der") is replaced by the interrogative (e.g. "welcher") and demonstrative (e.g. "dieser") determiner. While keeping the definite articles instead of the demonstrative counterparts would be a valid design choice too, interrogative and demonstrative determiners are both frequently tokenized as two tokens, thereby making the mapping during experiments easier.

- (6) Welchen Schüler soll der Lehrer besuchen?
which.ACC student should the.NOM teacher visit
'Which student should the teacher visit?'
- (7) Diesen Schüler soll der Lehrer besuchen.
this.ACC student should the.NOM teacher visit
'The teacher should visit this student.'

Since the case of sentences within a pair is kept the same, the model's choice of verb is no longer a diagnostic continuation. Instead, we add infinitive verbs. In both cases these sentence prefixes are complete grammatical sentences. Therefore, the model may predict punctuation. For sentences where the fronted noun phrase contains an interrogative determiner, a good continuation is the question mark token "?". In the demonstrative case, a period serves as the diagnostic continuation ".".

Rizzi (1997) describes a framework in which illocutionary force is syntactically encoded in the left periphery of the clause in the projection ForceP. The head of ForceP carries a feature specifying clause type (e.g. declarative vs. interrogative), and this feature must be assigned a value during the derivation. In our construction this requires the finite modal verb (e.g. "soll") to raise into this position, since German main clauses have no overt complementizer to value Force directly.

Under this analysis, the interrogative or demonstrative status of the fronted determiner is not just a lexical property local to the noun phrase. Instead it is grammatically linked, via feature checking at C, to the force of the entire clause. This motivates our use of sentence-final punctuation as a diagnostic: if a model represents force as predicted by this syntactic account, its prediction of "?" versus "." should be recoverable from and causally dependent on the complementizer head.

A.2 Subjunctive vs. indicative verbal mood

German verbs morphologically mark mood. The most common verbal mood is the indicative which appears in many different contexts. Konjunktiv I (subjunctive I) is associated with actions that are referenced in indirect speech or reported thought. In spoken modern German, however, the indicative is commonly used for indirect speech too. Nonetheless, using a verb in subjunctive I in a context outside of indirect speech or reported thought is not possible. We can exploit this signal for activation patching.

- (8) Die Frau behauptete, dass der Mann schon fertig sei.
the.NOM woman.NOM claimed that the man already done is.SUB
'The woman claimed that the man is already done.'

- (9) Die Frau bemerkte, dass der Mann schon fertig ist.
 the.NOM woman.NOM realized that the man already done is.IND
 ‘The woman realized that the man is already done.’

The first sentence is an example of indirect speech. The verb “behaupten” in the matrix clause selects subjunctive mood for verbs in embedded clauses that depend on it. The embedded clause is headed by the lexical complementizer “dass” (that). This is an example where C^0 is occupied by a lexical complementizer rather than a finite verb. The finite verb indicating completion of the man’s action “sein” (to be) is in the subjunctive mood. Linguistic theory predicts that the verb “behaupten” may select subjunctive verbal mood in embedded clauses, and that this dependency is mediated by the complementizer of the embedded clause (in this case “dass”). By contrast, the matrix verb “bemerken” (realize) in the second sentence does not select subjunctive for finite verbs in the embedded clause. Therefore, the verb “sein” appears in its indicative form “ist”. For activation patching, the diagnostic continuations are “sei” and “ist” respectively. While the subjunctive form of the verb ‘to be’ is acceptable in the first sentence, many German speakers accept the indicative too. German speakers commonly default to the indicative, even when the subjunctive verbal mood is licensed. In our experiments, this leads LLMs to consistently assign higher logits to indicative forms, regardless of licensing. Our patching recovery scores are sensitive to small relative changes through intervention. However, since the models we analyzed overwhelmingly predict an indicative verb over a subjunctive verb, we exclude these sentence types from our accuracy-based analyses.

A.3 Gender agreement control

We also construct minimal pairs for two cases of semantic rather than syntactic dependencies. The first of these involves grammatical gender in pronominal anaphora. In the following pair of examples, differences in the gender of the fronted noun sets up different expectations about the gender of the subject pronoun in the embedded clause.

- (10) Der Mann glaubt, dass nur er das kann.
 the.MASC man.MASC thinks that only he.MASC that can
 ‘The man thinks that only he can do that.’
- (11) Die Frau glaubt, dass nur sie das kann.
 the.FEM woman.FEM thinks that only she.FEM that can
 ‘The woman thinks that only she can do that.’

While an embedded subject could in principle be either co-referential or not with the matrix subject, the co-referential interpretation requires agreement in gender. To the degree that the co-referential interpretation is considered (and we expect that it is highly preferred), the gender of the matrix subject should modulate the probability of the gender of the embedded subject pronoun.

We construct sentence prefixes such as “Der Mann glaubt, dass nur” and designate a congruent pronoun, in this case “er”, as the diagnostic continuation. Anaphora between a pronoun and its non-local antecedent is widely assumed to involve a semantic and not syntactic relation.

In such cases there is no linguistic reason to believe that the dependencies are modulated by the complementizer. If this difference is reflected in the model’s underlying computations, recoverability and causal effects should be lower on the complementizer position.

A.4 Selectional relations

As a second semantic control, we consider the impact of selectional restrictions. Here, we consider the effect of content of the initial noun on the verbs that are likely to take that noun as a subject.

- (12) Der Arzt soll den Jungen behandeln.
 the doctor should the boy treat
 ‘The doctor should treat the boy.’
- (13) Der Lehrer soll den Jungen unterrichten.
 the teacher should the boy teach
 ‘The teacher should teach the boy.’

Doctors and teachers have different likelihoods of engaging in activities (e.g., treating and teaching), and correspondingly we would expect the probability of these verbs to vary in these two sentences. As with the gender case, there is no linguistic reason to believe that the dependency between the initial noun phrase and the predicted verb is modulated by the complementizer.

It is difficult to create a large controlled dataset of sentence pairs with varying nouns and appropriate verbs to serve as diagnostic continuations, as it would require identifying verbs selectionally related with a noun specifically in contrast with a different noun. Plausible sentence pairs would require careful individual curation on a large scale, something which is not easily feasible. We avoid this problem by eschewing an accuracy measure in choosing between two possible completions, and instead attending to the entirety of the model’s output distributions. This distribution captures preferences and hypotheses that the model maintains. For sentence pairs that only differ in the choice of noun in the left periphery, downstream next-token distributions should differ only in ways that are attributable to that specific choice. As a proxy for our logit-based recovery score, we therefore analyze the effect of activation patching on a next-token distribution level, rather than inspecting pre-chosen individual logits. We compute this proxy-recovery score in a similar way to the average logit difference metric:

$$\text{recovery} = 1 - \frac{\text{KL}(P_{\text{patched}} \parallel P_{\text{source}})}{\text{KL}(P_{\text{dest}} \parallel P_{\text{source}})} \quad (7)$$

B Results from other models

We replicate our patching experiments and attention ablation analyses with two other models to verify that our findings are applicable more generally. Specifically, we investigate two models from the Llama family.

B.1 Attention patching across models

B.1.1 Accusative vs. dative case contrast

Overall, the accusative vs. dative sentence pairs are the constructions with the lowest accuracy. Figure 3 shows that the complementizer-mediated routing is most clearly visible in the Qwen model. While elevated recovery scores are visible at the modal (complementizer) position, it is not as strong in both Llama models. We hypothesize that this case contrast is much less salient to the models than the other syntactic contrasts and that noun choice of the fronted NP, which is consistent within the same pair, heavily affects verb logits through semantic correlations that may outweigh the syntactic case signal.

B.1.2 Indicative vs. subjunctive mood contrast

Figure 4 shows the patching results for mood contrast. In these sentence pairs, the complementizer head is occupied by the lexical complementizer “dass”. In middle layers, all three models show highest recovery scores on the complementizer and the comma token immediately preceding it.

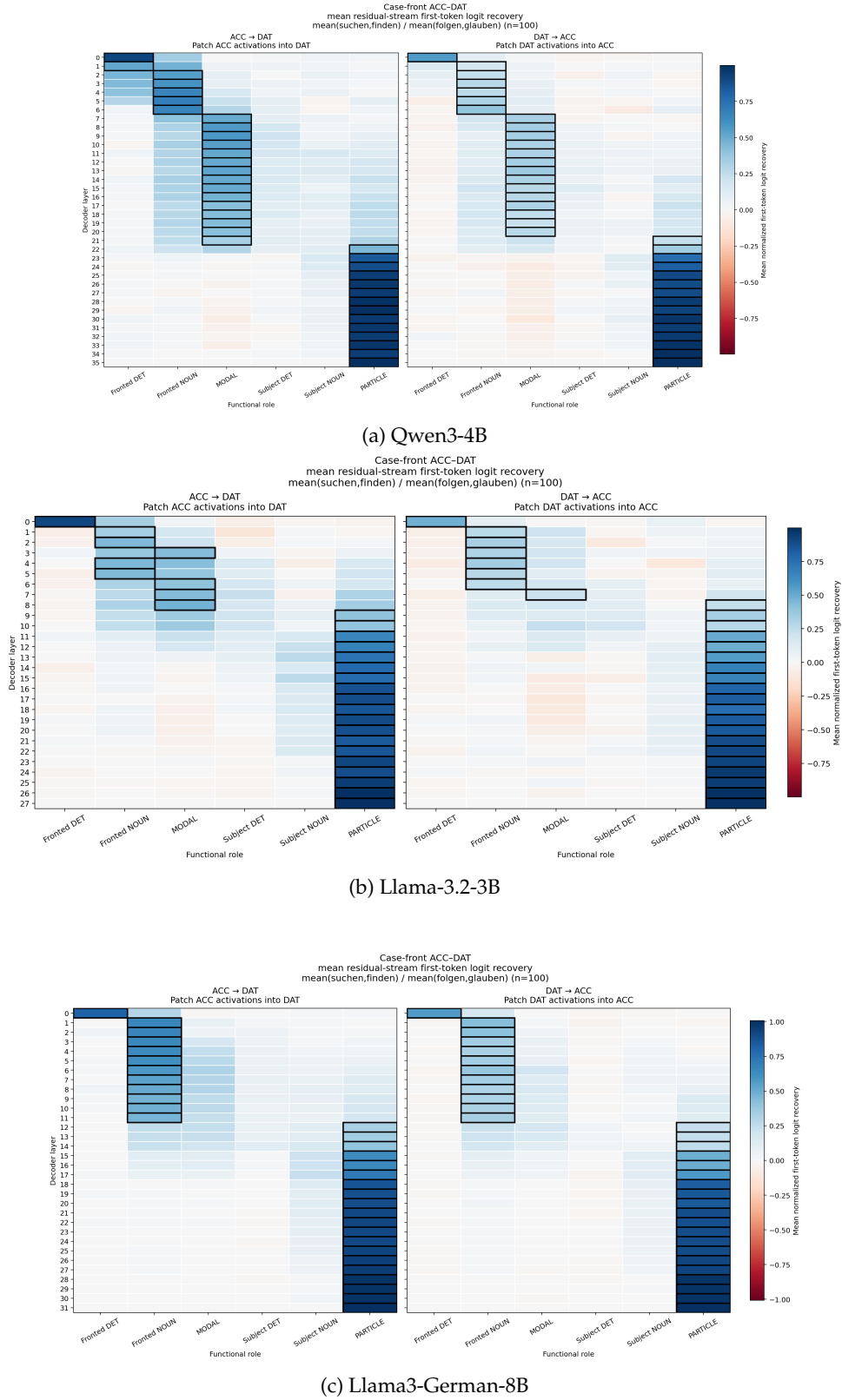


Figure 3: Averaged activation patching results for the accusative/dative constructions for all three models. Each recovery score heatmap is split by patching direction.

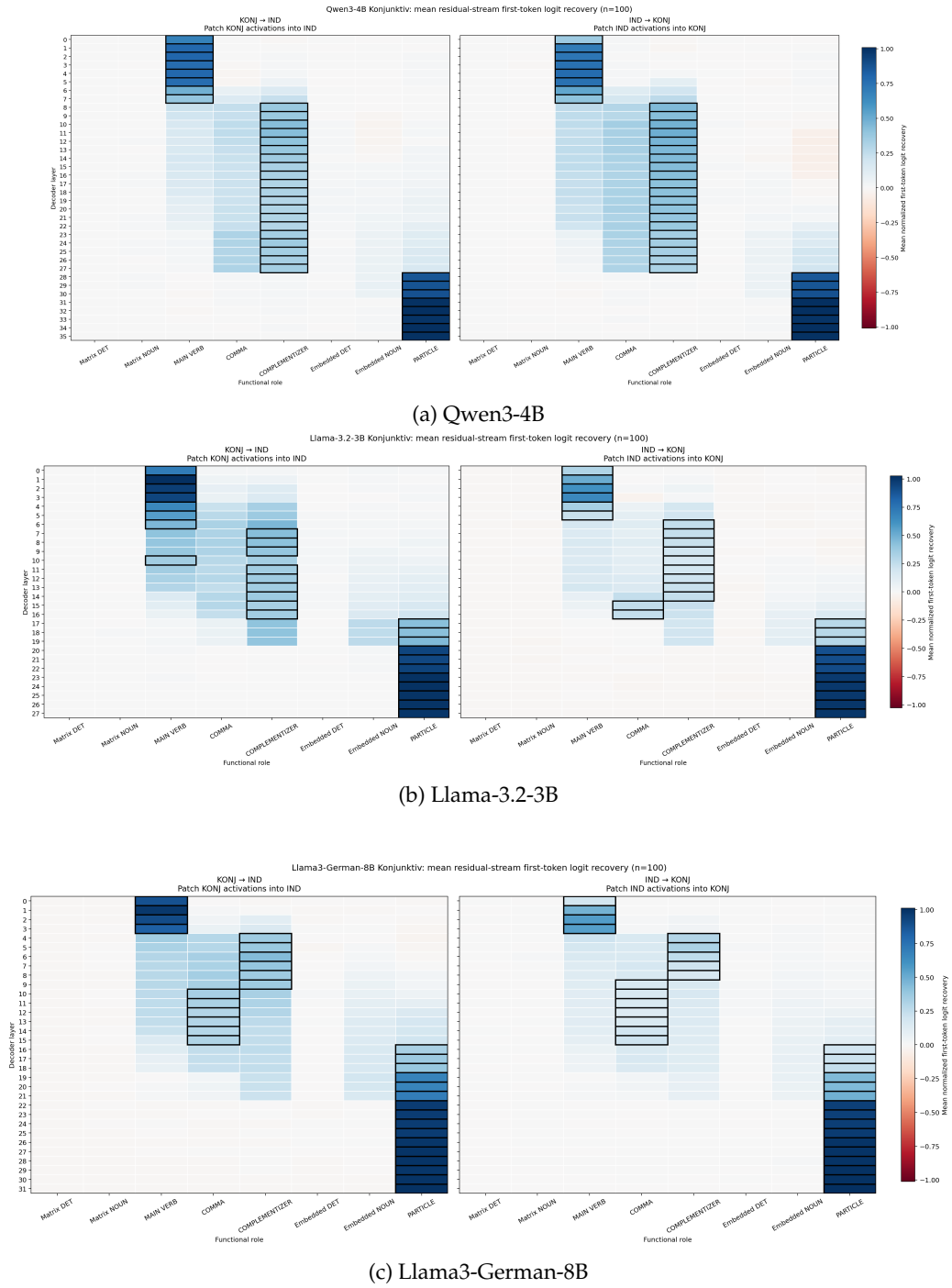


Figure 4: Averaged activation patching results for the indicative/subjunctive constructions for all three models.

B.1.3 Interrogative vs. demonstrative force contrast

Figure 5 shows that all three models exhibit their highest middle layer recovery scores at the complementizer. Information required to correctly predict “.” or “?” appears to be routed through the modal verb.

B.1.4 Gendered anaphora contrast

While the complementizer head (modal) seems largely uninvolved in syntactic routing for Qwen, the Llama models do exhibit some higher recovery scores on the modal (Figure 6). While we initially constructed the gendered anaphora as a control for the syntactically contrastive cases, anaphora constructions themselves must be syntactically licensed. We hypothesize that this may impact the recovery scores in the Llama models.

B.1.5 Noun-only contrast as semantic control

The recovery score heatmaps for the constructions where only the choice of the fronted noun differed (Figure 7) show that the complementizer head is much less causally involved in shaping the predicted tokens than in the syntactically-contrastive constructions.

B.2 Complementizer masking interventions

The complementizer bottleneck cases (the cases in which information from before the complementizer can only be copied through it) show that the gender anaphora case consistently experiences the largest absolute drop in accuracy (Figure 8). Clearly, obstructing direct attention to tokens from the fronted constituents has the strongest negative impact on this construction type, suggesting that direct attention to the fronted constituent constitutes an important part of congruent gender prediction.

For the complementizer circumvention, the opposite is the case. While the accusative/dative and interrogative/demonstrative constructions suffer greatly from the loss of the hypothesized syntactic router, the gendered anaphora accuracies remain flat or even increase, suggesting that the loss of a syntactic information router does not negatively impact the computations for the gendered anaphora constructions.

C Cross-linguistic generalization

Italian verbs express verbal mood like indicative and subjunctive, but the required mood of a verb is determined by the syntactic context. Performing analogous patching experiments on an individual scale reveals that in Italian, too, information flows from a verbal mood selector to the complementizer head before affecting downstream verbal mood (Figures 9 and 10), reproducing the analytic assumptions of Giorgi (2009).

Similarly, English filler-gap constructions such as long-distance wh-questions with embedded clauses (e.g. Who did the man say that the boy saw?) exhibit the same syntactic routing (Figure 11). This accords with the findings of Boguraev et al. (2025).

We did not templatically generate larger quantities of sentence pairs for languages besides German and draw on these individual examples to indicate that our findings extend to other languages upon exploratory analysis.

D TPDN Dataset

To accommodate the TPDN training’s need for a large number of stored activations while retaining structural similarity, we slightly extend our templatically generated sentence pairs with fronted NPs in the accusative and dative cases.

We add additional sentences by simply prepending the prepositions “für” (for) before accusative NPs and “mit” (with) before dative NPs. These additions turn the fronted

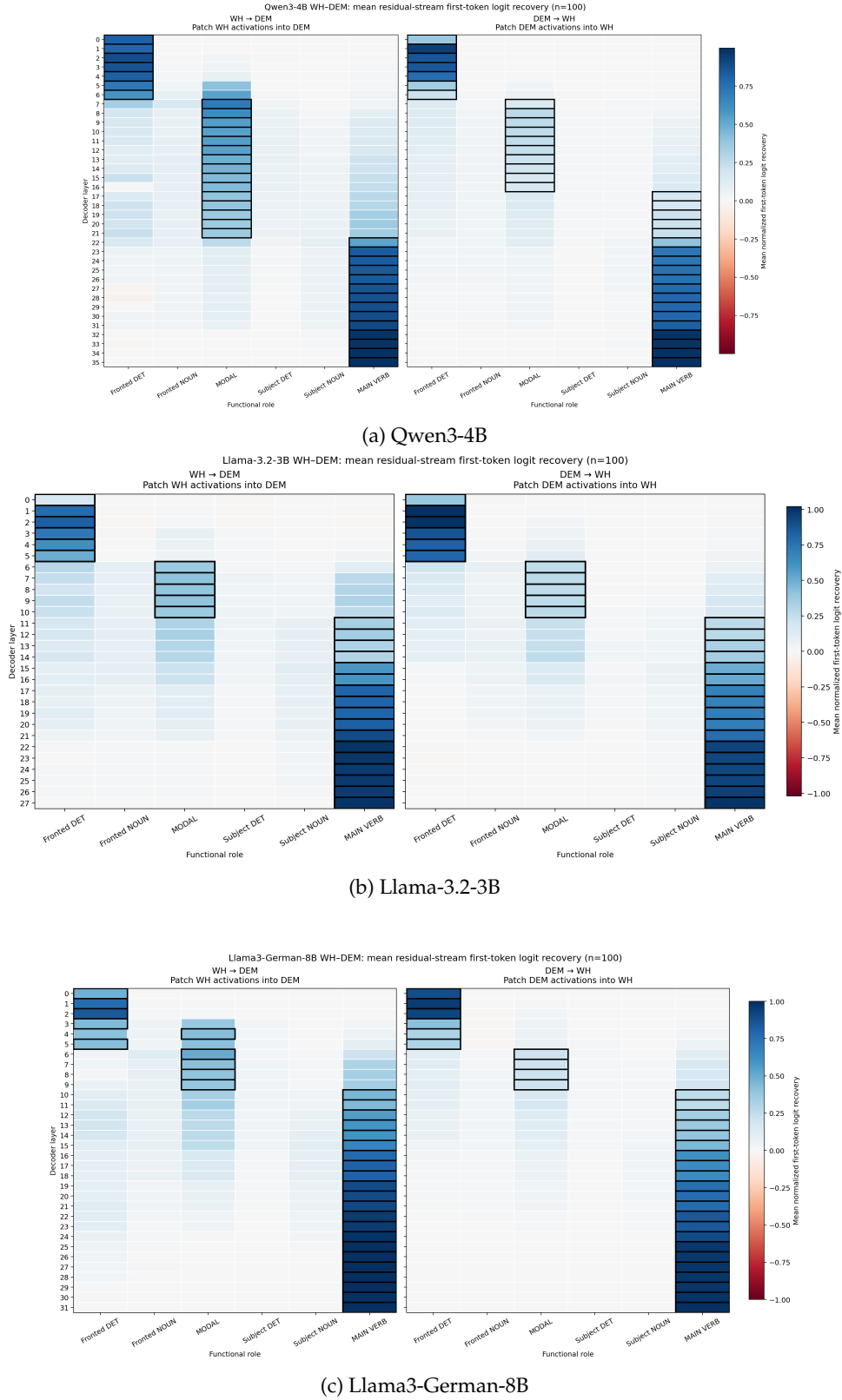


Figure 5: Averaged activation patching results for the interrogative/demonstrative constructions for all three models.

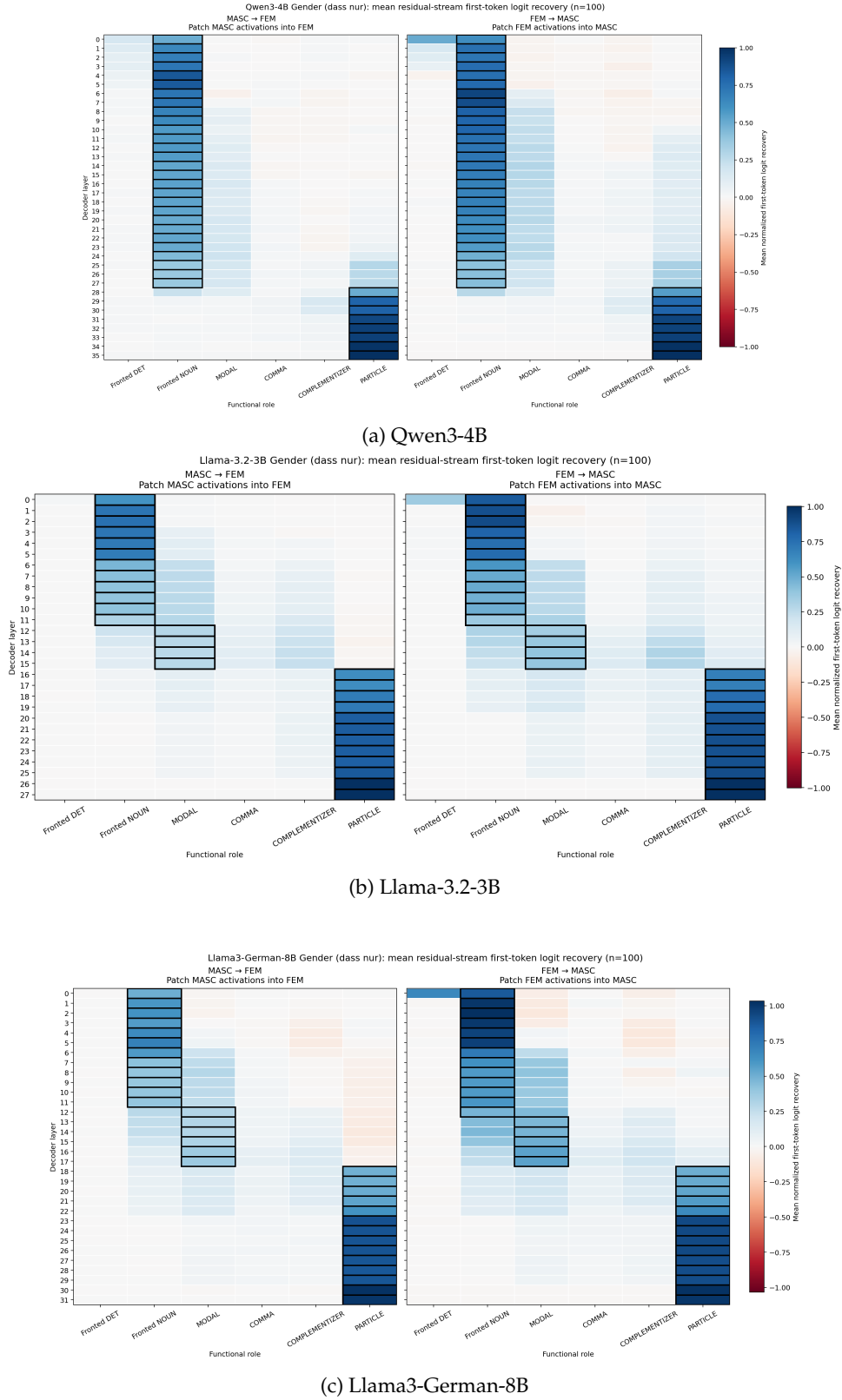


Figure 6: Averaged activation patching results for the gendered anaphora constructions for all three models.

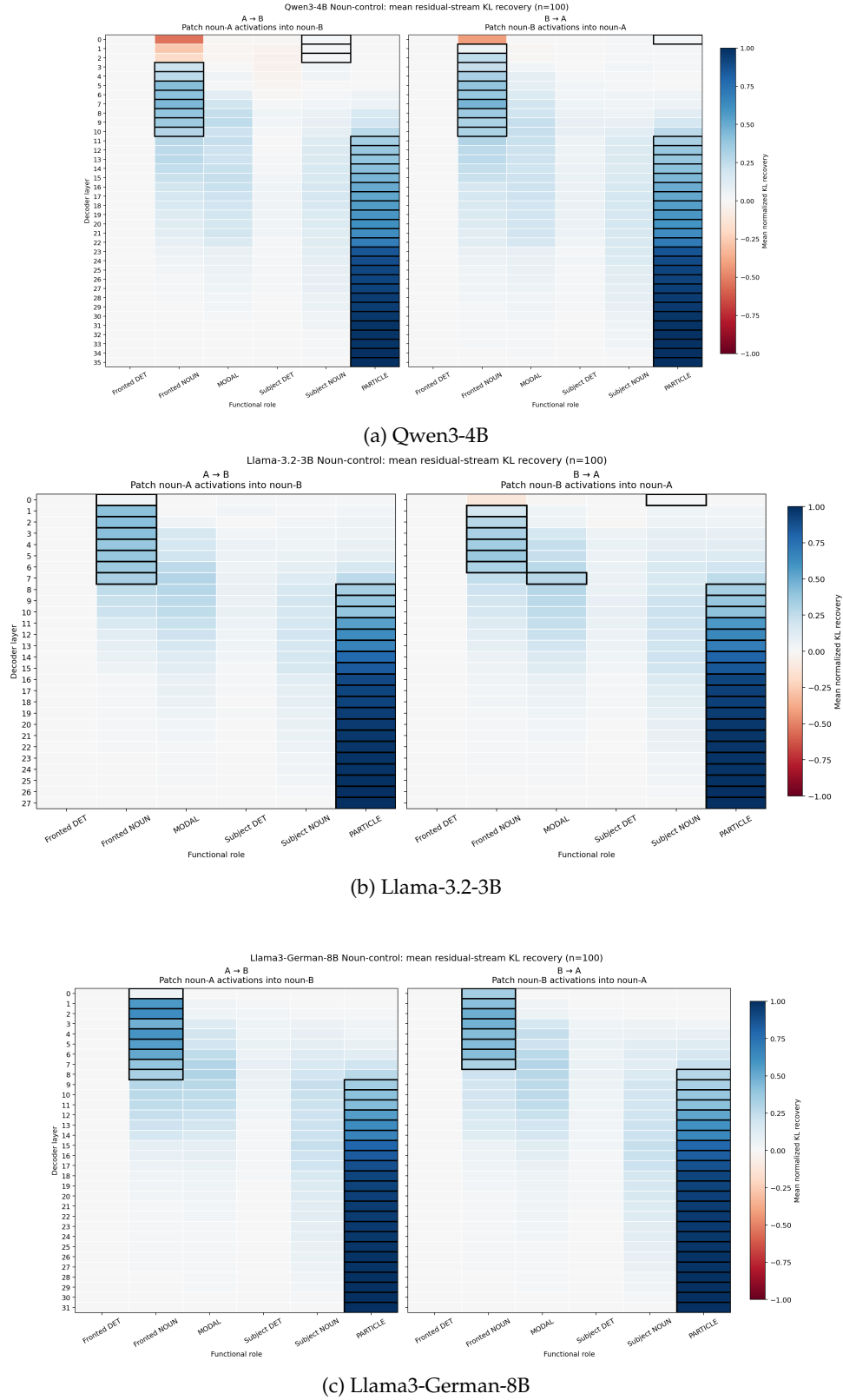
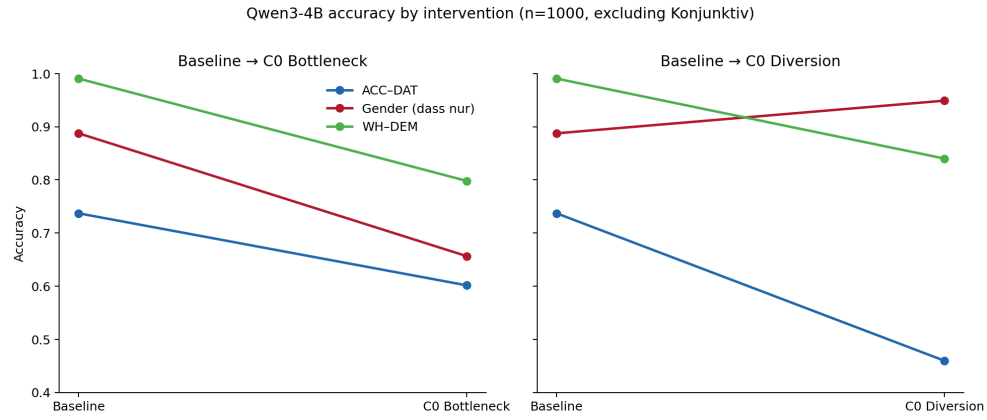
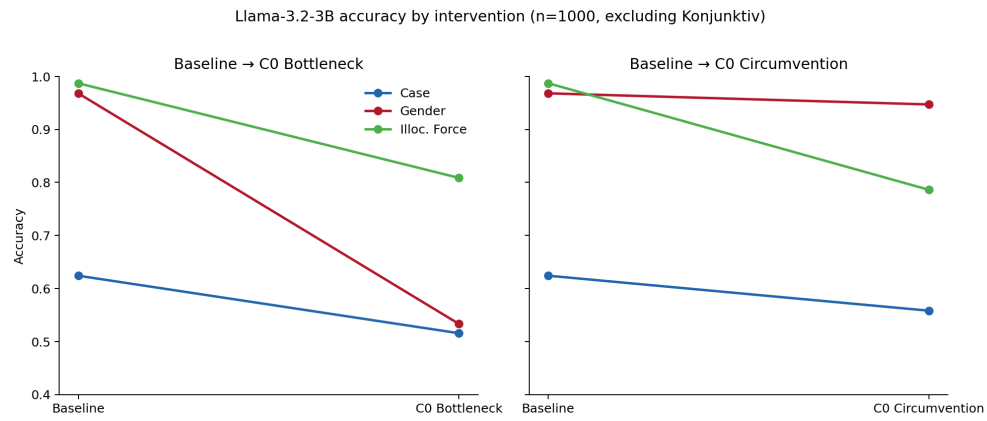


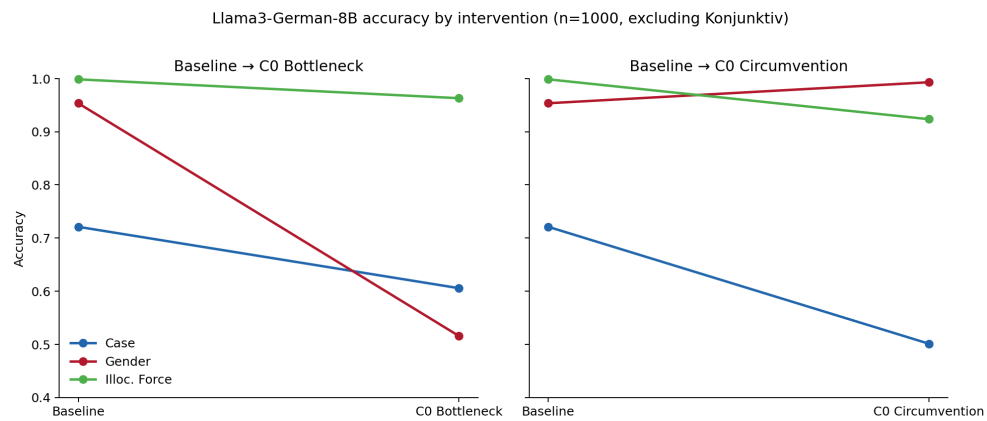
Figure 7: Averaged activation patching results for the noun choice semantic controls for all three models.



(a) Qwen3-4B



(b) Llama-3.2-3B



(c) Llama3-German-8B

Figure 8: Effects of complementizer masking interventions for each model and intervention type.

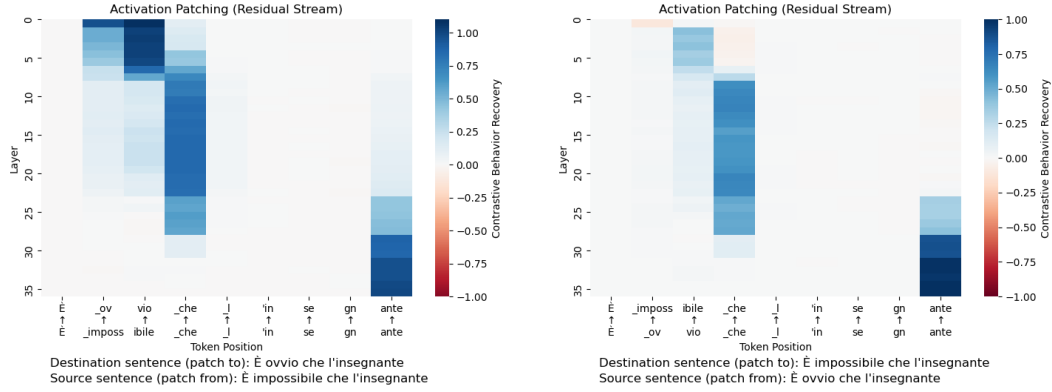


Figure 9: Patching results from an Italian construction: the adjectives “ovvio” (obvious) and “impossibile” (impossible) select a indicative and subjunctive verb respectively. The diagnostic continuations differ precisely in this manner according to verbal mood: “è” (is) and “sia” (is in the subjunctive). Again, information is mediated through the lexical complementizer “che” (that).

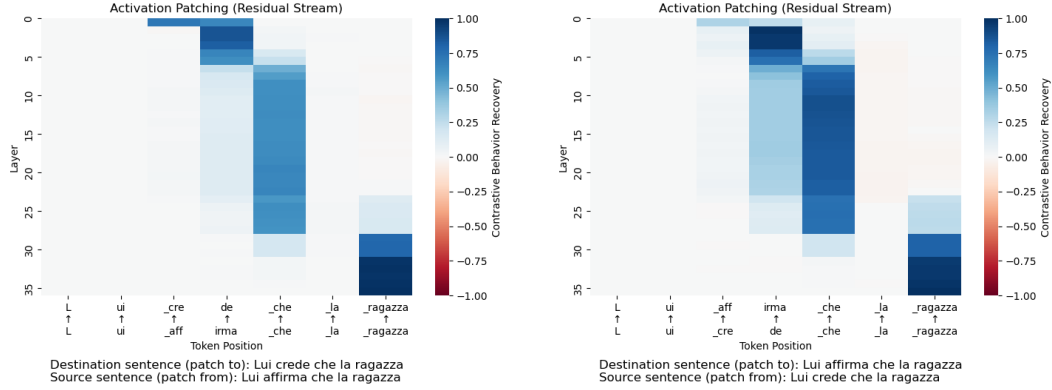


Figure 10: In Italian, finite verbs can select for mood in verbs of an embedded clause. Here, “crede” (believes) selects the subjunctive and “afferma” (affirms) selects the indicative. The same pattern on the complementizer head “che” arises.

constituents into prepositional phrases (PPs). This turns the fronted constituent into an adjunct, meaning that it is no longer an argument selected by the verb. This adds intransitive verbs as another class of diagnostic continuations for the PP cases. Therefore, the TPDNs are trained to compositionally represent varying constituent type (NP and XP) and case (accusative and dative) for the fronted constituents.

Our dataset now includes fronted constituents that are either noun phrases in the accusative case or the dative case or fronted prepositional phrases with the prepositions “mit” and “für”. Each sentence follows the templatic structure below:

- (14) {Fronted NP/PP} {Finite Modal} {Subject NP} {Infinitival Verb}
- Mit dem Mann kann die Frau (sprechen)
with.the.DAT man can the.NOM woman speak
'The woman can speak with the man.'
 - Ihn soll der Direktor (treffen)
him.ACC should the.NOM director meet
'The director should meet him.'

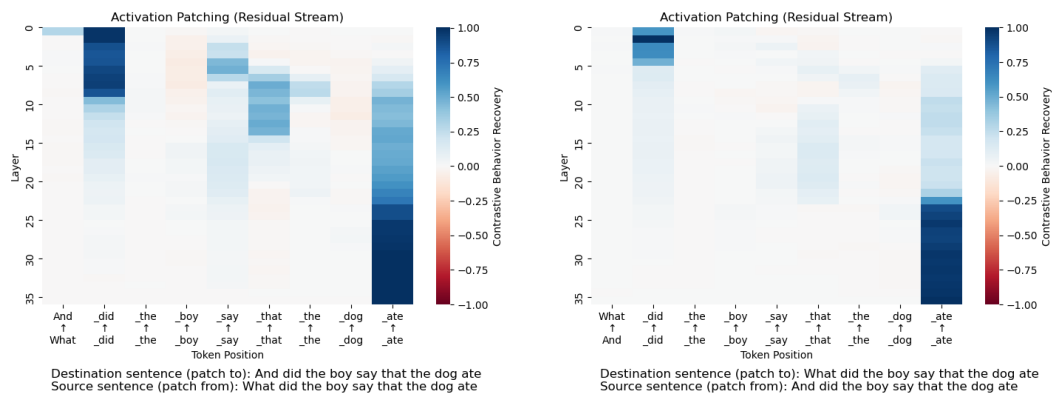


Figure 11: Finally, an example of an english filler-gap dependency. The diagnostic continuation for the sentence beginning in “and” is “it”, and for the sentence beginning in “what” it is “?”. The lexical complementizer in the embedded clause is “that”.