Revealing Redundant Syntax in Large Language Models through Multi-Hop Dependency Paths

Masaki Sashida **Takeshi Kojima** **Yusuke Iwasawa** **Yutaka Matsuo**
The University of Tokyo
{masaki.sashida,t.kojima,iwasawa,matsuo}@weblab.t.u-tokyo.ac.jp

Abstract

Prior work on attention–syntax alignment has largely focused on single-hop Universal Dependency edges (DPs). In this paper, we treat short multi-hop dependency paths (MDPs) (e.g., "obl+case") as first-class units and analyze them alongside DPs. Across three pretrained autoregressive LMs (GPT-2 XL, Llama 3 8B, Qwen3-8B) and one encoder baseline (BERT-large), we extract 2–3 hop MDPs from UD-parsed English and quantify head–relation alignment with an Unlabeled Attachment Score (UAS)–style metric modified for causal masking in decoder-only models. Rank visualizations reveal both overlap and specialization: we observe heads that align with both DPs and MDPs, as well as heads that appear specialized for one route. To test functional relevance, we first identify heads by UAS and then apply an undifferentiated (uniform) attention ablation to those heads; we evaluate the impact on BLiMP and LAMBADA. Ablating the top 10% of all heads shows that MDP-selected heads induce larger drops than DP-selected heads and that the union ("Mix") of DP- and MDP-selected heads yields the largest drops. For GPT-2 XL, the observed drops are (BLiMP: ΔDP = 1.35 pp, ΔMDP = 4.81 pp, ΔMix = 7.11 pp; LAMBADA: ΔDP = 4.70 pp, ΔMDP = 25.17 pp, ΔMix = 32.99 pp), all exceeding size-matched random controls. These results indicate that models can route information consistent with syntactic dependencies via both DP and MDP pathways, with MDPs playing a distinct and measurable role in some settings under our interventions.

1 Introduction

Large language models (LLMs) trained on raw text often achieve strong performance on grammatical benchmarks without explicit syntactic supervision. Prior work on how LLMs encode grammatical structure has examined whether individual attention heads systematically align with single-hop Universal Dependencies (UD) relations such as `nsubj`, `obj`, and `det`(Clark et al., 2019; Kovaleva et al., 2019; Ravishankar et al., 2021). While this line of work has provided useful descriptive maps of attention–syntax alignment, it has typically not foregrounded causal tests (e.g., intervention-based validation of functional roles). Moreover, syntax has often been operationalized at the granularity of a single dependency edge, leaving relatively unexplored the possibility that models maintain multiple, partially overlapping internal routes for representing the same syntactic configuration.

We argue that LMs parse sentences via *two complementary routes*: (i) single-hop UD dependency edges (DPs) and (ii) *Multi-Hop Dependency Paths* (MDPs)—short sequences such as `obl+case` or `nmod+case` that frequently occur in UD trees yet are not primitive units in mainstream linguistic theory (see Figure 1). Our claim is not that MDPs supersede DPs; rather, both routes are used in practice, sometimes by the same heads and sometimes by distinct heads.

Methodologically, we extract frequent 2–3 hop MDPs from UD-parsed English and quantify head–relation alignment with a UAS-style metric modified for causal masking in decoder-only models (Clark et al., 2019; Htut et al., 2019). We then rank heads for nine frequent relations (DPs and their corresponding MDPs) and visualize *overlap vs. independence* via rank scatter/heatmaps to see which heads are shared across routes vs. specialized. Finally, to test causal impact beyond correlation, we perform *uniformization* interventions (flattening selected attention heads) and measure functional impact on **BLiMP** (grammatical minimal pairs) and **LAMBADA** (discourse-level consistency) against random and size-matched controls (Warstadt et al., 2020; Paperno et al., 2016).

Our results indicate that both DP-aligned and MDP-aligned heads are behaviorally relevant. In some relations (e.g., `conj+cc`, `obl+case`), inter-

Findings of the Association for Computational Linguistics: EACL 2026, pages 4114–4137

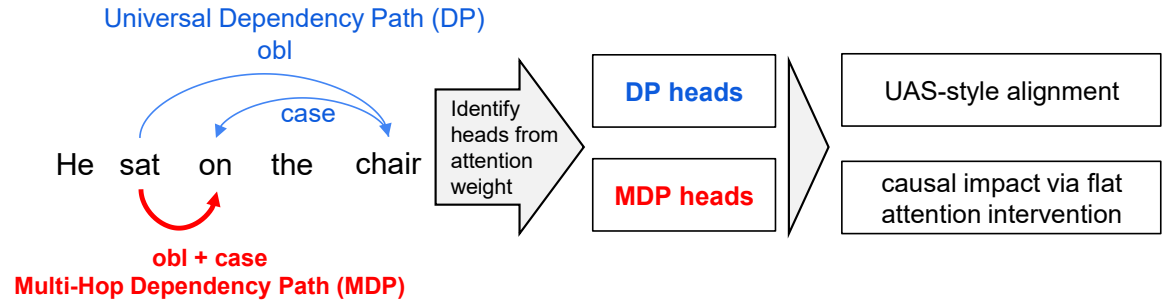

Figure 1: Overview of our setup and goals. Left: Universal Dependencies (UD) edges in *He sat on the chair*—we refer to *single-hop UD edges* as **DPs** (e.g., *obl*: *sat→chair*; *case*: *chair→on*)—and the corresponding **MDP** *multi-hop* path *obl+case* (*sat→on*). Middle: we identify attention heads from attention weights and group them as DP heads vs. MDP heads. Right: two evaluations with purposes: (i) **UAS-style alignment**—to test whether attention follows dependency structure; (ii) **causal impact via flat attention intervention**—to test whether those heads functionally affect grammaticality and task performance (BLiMP/LAMBADA).

ventions on MDP-aligned heads have comparatively larger effects. The evidence suggests that an analysis confined to single UD edges is insufficient to fully account for internal syntactic routing in LMs; an MDP lens complements DP-based views.

Contributions

- We recast attention–syntax alignment in decoder-only LMs through a *multi-route* lens: DPs and MDPs co-exist and jointly support parsing.
- We map where routes *overlap* vs. *specialize* across layers and heads, aligning with circuit-style redundancy and backups.
- Via interventions on route-specific head sets, we quantify functional impact of the heads on grammaticality and task performance (**BLiMP/LAMBADA**).

2 Related Work

Attention and Syntactic Dependencies UD (Universal Dependencies) relations have long provided a standard footing for probing syntactic information in large language models. A substantial line of work examines whether attention patterns align with *single dependency edges* (DPs)—e.g., `nsubj`, `obj`, `det`—and reports heads that appear specialized for particular relations (Clark et al., 2019; Kovaleva et al., 2019; Ravishankar et al., 2021). In addition, several studies *infer or reconstruct* dependency structure directly from attention weights and evaluate agreement using UAS or maximum-spanning-tree–based metrics (Htut et al., 2019).

These findings suggest that attention can correlate with syntactic structure, but analyses are typically operationalized at the granularity of a single edge (or a single relation type at a time), leaving less systematically explored whether a given syntactic configuration is represented and utilized through *multiple internal routes* distributed across layers and heads.

In this paper, we make this assumption explicit and test it. Beyond single edges, we treat short, frequent *multi-hop dependency paths* (MDPs), such as `obl+case`, as first-class units and compare them to their corresponding DPs within the same framework. Crucially, we do not claim that MDPs replace DPs; rather, we ask whether there exist additional routes that a DP-only view may miss, and whether those routes matter functionally.

What counts as an "edge" versus a "path" can depend on the chosen dependency formalism. Across formalisms such as UD and SUD, a relation that is realized as a single edge in one representation may correspond to a longer path in another (and vice versa), motivating work that studies model preferences over formalisms and the stability of syntactic signals across them (Kulmizev et al., 2020).

Our focus is complementary: we *fix* the formalism (UD) and ask whether, *within UD*, models exhibit evidence for multiple coexisting routes—single edges (DPs) and short chains (MDPs). Moreover, rather than relying only on probing or reconstruction, we causally test functional importance by intervening on head sets selected by these routes.

Confounders in attention-based alignment Interpreting attention as evidence of syntactic routing

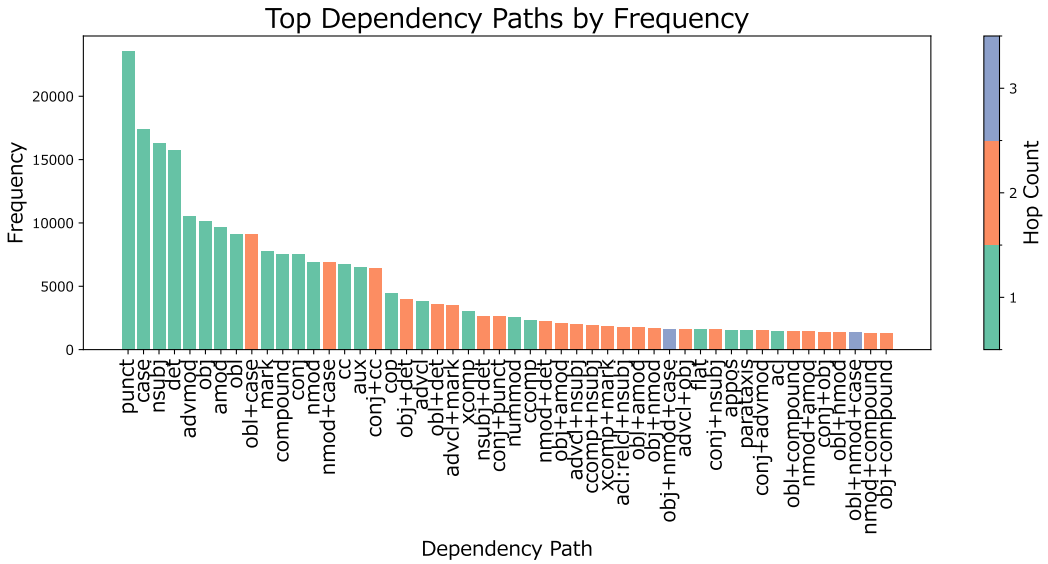

Figure 2: Bar plot of the 50 most frequent dependency paths in the corpus. Each bar's height shows the absolute count, and bar colors encode hop counts, highlighting which syntactic hop counts occur most often.

requires care due to known confounders. First, apparent alignment can arise from positional heuristics—e.g., "previous-token" heads or approximate fixed-distance heads—which can mimic structural effects. Second, autoregressive LMs exhibit *attention sink* behavior, where sequence-initial tokens (e.g., BOS) attract disproportionate attention mass, and related phenomena such as "null attention" have been discussed in earlier analyses (Gu et al., 2025; Vig and Belinkov, 2019). Motivated by these concerns, we explicitly examine the role of local linear proximity by recomputing alignment after excluding relations between neighboring tokens, which helps rule out trivial immediately preceding-token effects for the relations we analyze. In addition, our interventions are *sink-aware*: we preserve the BOS attention weight and uniformize only the remaining past-token mass, avoiding confounding performance changes with sink manipulation. (Our UAS-style alignment is argmax-based, so pure rescaling/normalization that does not change the argmax has limited impact on the alignment score, whereas sink effects can matter substantially under intervention.)

From correlational signatures to multi-path mechanisms with causal tests. Recent mechanistic interpretability work emphasizes that Transformer behavior often arises from *multiple collaborating pathways* rather than a single privileged route, and validates mechanisms via interventions (e.g., patching/ablation) (Olsson et al., 2022; Wang et al., 2022; Goldowsky-Dill et al., 2023). We connect this perspective to syntactic processing by operationalizing a *multi-route* view of dependency information. We (i) quantify overlap versus specialization between DP- and MDP-aligned head sets and (ii) causally evaluate their functional importance by flattening attention in route-selected heads and measuring impacts on BLiMP and LAMBADA. Importantly, our results do not imply that MDPs uniformly dominate DPs; rather, they support a view in which syntactic cues can be distributed across partially overlapping internal routes that become measurable under targeted interventions.

3 Method

3.1 Introduction of Multi-Hop Dependency Paths (MDPs)

In this study, we focus on the possibility that LLMs' attention mechanisms align not only with single dependency relations (DPs) but also with structures formed by sequences of multiple dependencies—**Multi-Hop Dependency Paths (MDPs)**.

MDPs refer to chains of two or more dependency relations in a UD parse tree, including patterns such as `obl+case`, `nmod+det`, and `obj+amod`. While these are not explicitly defined as syntactic units in linguistic theory, they frequently occur in natural language texts.

Example: Consider the following sentence:

> *The teacher placed a book of science on the shelf.*

In the UD parse of this sentence, the following MDPs can be identified:

- `obl+case`:

$$placed \xrightarrow{\texttt{obl}} shelf \xrightarrow{\texttt{case}} on$$

  This path represents the oblique phrase *on the shelf*, where the verb *placed* takes *shelf* as an oblique dependent, and *shelf* is marked by the adposition *on*.

- `obj+nmod+case`:

$$placed \xrightarrow{\texttt{obj}} book \xrightarrow{\texttt{nmod}} science \xrightarrow{\texttt{case}} of$$

  This path captures a nominal modifier phrase embedded within the object of the verb. The verb *placed* takes *book* as its object; *science* modifies *book* as a nominal dependent, and the relation is marked by the adposition *of*.

For example, in a sentence like "placed the book on the shelf," there are dependency relations such as placed → shelf (`obl`) and shelf → on (`case`). Our observation revealed that some attention heads concentrate directly from *on* to *placed*, which can be interpreted as capturing not a single DP but the entire structural path like `obl+case`.

Such observations suggest that models may utilize not only individual dependency relations but also larger structural units along the syntactic tree (e.g., paths, clusters, surrounding structures) as supplementary information. Importantly, this does not imply that models represent structures that replace DPs; rather, our observations are consistent with models leveraging multiple, partially redundant structural cues, including DPs.

Current probing methods are primarily designed to focus on DPs, and there has been little research examining how attention aligns with multi-hop syntactic patterns or how such alignment contributes to model inference performance. This study aims to fill this gap by quantitatively evaluating how attention heads correspond to complex syntactic patterns and whether they actually contribute to grammaticality judgments and semantic inference.

We use the English Web Treebank (EWT) (Nivre et al., 2020) as our UD-parsed corpus. From this corpus, we extract all DPs, i.e., single-hop labeled dependency edges. Each sentence in the corpus consists of a sequence of tokens annotated with head indices and dependency labels. Based on the set of observed DPs, we iteratively construct longer MDPs by connecting multiple relations together.

We denote the composition of two consecutive DPs as a 2-hop MDP. To identify frequently occurring 2-hop MDPs, we begin by generating candidate 2-hop MDP paths by pairing every DP with every other DP (i.e., DP × DP), and count their frequency in the corpus. Only those candidate MDPs whose frequency exceeds a predefined threshold τ are retained ; in all experiments we set $\tau = 50$. In the next step, we extend the 2-hop MDP candidates by appending another DP, yielding 3-hop DP = 2-hop MDP × DP. Again, we count occurrences and retain only those that pass the threshold. This procedure is repeated iteratively until no new MDPs exceed the threshold.

The final set of MDPs are represented as atomic labels, such as `obj+amod+case`. Using this final set of MDPs, we simply define them as atomic labels (e.g., `obj+amod+case`) in addition to the original DPs (See Figure 2).

3.2 Identifying Attention Heads Sensitive to Syntactic Structure

Using the defined set of MDPs alongside the original DPs, we analyze which attention heads effectively capture syntactic structures.

Decoder variant. For each layer l and head h, and each token x_i in a sentence, we determine the token x_j that receives the highest attention weight from x_i. Specifically, for a sentence with attention matrix $A^{(l,h)}$, we define the predicted syntactic connection as follows:

$$\hat{j}_i^{(l,h)} = \arg\max_j A_{ij}^{(l,h)} \quad (1)$$

Since the model employs masked attention (decoder-style), token x_i can only attend to tokens at or before its own position.

We then check if the predicted pair (x_i, x_j) matches a syntactic dependency (either DP or MDP) annotated in our corpus. Importantly, we do not consider the directionality of these syntactic dependencies; we only measure whether the predicted pair corresponds to an existing dependency edge, ignoring head-dependent direction.

Because model tokenization typically differs from corpus word-level tokenization, we follow Clark et al. (2019)'s alignment strategy: attention from a word to another word is computed by summing attention scores over all subword tokens for

the target word and averaging across subword tokens for the source word (Clark et al., 2019).

We perform this matching procedure for every token in every sentence of the corpus. For each syntactic dependency type r (DP or MDP), we count how many times head (l, h) correctly identifies a dependency pair:

$$C_r^{(l,h)} = \sum_{\text{sentence} \in \text{corpus}} \sum_{(i,j) \in r} \mathbb{I}\left[\hat{j}_i^{(l,h)} = j\right] \quad (2)$$

Here, the count is summed across all sentences in the corpus.

We define the correct rate for a given dependency relation r at attention head (l, h) as:

$$\text{CorrectRate}_r^{(l,h)} = \frac{C_r^{(l,h)}}{|\mathcal{D}_r|} \quad (3)$$

where $C_r^{(l,h)}$ is the number of correctly identified dependency pairs at head (l, h), and $|\mathcal{D}_r|$ is the total number of occurrences of dependency relation r in the corpus.

Then, the unlabeled attachment score (UAS) for relation r is defined as the highest correct rate among all attention heads:

$$\text{UAS}_r = \max_{l,h} \left(\text{CorrectRate}_r^{(l,h)}\right) \quad (4)$$

This methodology closely follows Clark et al. (2019)'s original interpretability approach, adapted here without considering dependency directionality for decoder-only model architecture.

Encoder (BERT) variant. For encoder models without causal masking (e.g., BERT), we allow attention in both directions and treat a dependency as correctly identified when either the dependent attends maximally to its head *or* the head attends maximally to its dependent. Formally, let

$$\hat{j}_i^{\text{row}} = \arg\max_{j \neq i} A_{ij}^{(l,h)}, \quad (5)$$

$$\hat{j}_i^{\text{col}} = \arg\max_{j \neq i} A_{ji}^{(l,h)}. \quad (6)$$

Then for a dependency pair $(i, j) \in r$ we count a hit if

$$\mathbb{I}\Big[\, \hat{j}_i^{\text{row}} = j \;\vee\; \hat{j}_j^{\text{col}} = i \,\Big] = 1. \quad (7)$$

Using this bidirectional condition, $C_r^{(l,h)}$ and $\text{CorrectRate}_r^{(l,h)}$ are computed as above, and UAS_r is obtained by taking the maximum over heads.

3.3 Intervention by Flattening Selected Attention Heads

Based on the heads identified in 3.2, we apply a *uniformization (flattening) intervention* and evaluate how scores on **BLiMP** and **LAMBADA** change.

Prior work reports an *attention sink* phenomenon in autoregressive LMs, where the sequence-initial token (e.g., BOS) is disproportionately attended to (Xiao et al., 2024; Gu et al., 2025). To prevent sink effects from confounding the intervention, we *fix* the attention weight to the first token and *uniformize only the remaining past-token weights*.

For a selected head (l, h) with attention matrix $A^{(l,h)} \in \mathbb{R}^{S \times S}$, causal masking forbids attention to future positions ($j > i$). Let the weight to the first token (index 0) in row i be

$$\alpha_i = A_{i0}^{(l,h)}.$$

We define a sink-aware uniform matrix $T^{(\text{sink})}$ by preserving α_i at $j = 0$ and distributing the remaining mass $(1 - \alpha_i)$ uniformly over $j = 1, \ldots, i$ (for $i \geq 1$):

$$T_{ij}^{(\text{sink})} = \begin{cases} \alpha_i, & j = 0, \\ \dfrac{1 - \alpha_i}{i}, & 1 \leq j \leq i, \\ 0, & j > i. \end{cases} \quad (8)$$

For the boundary case $i = 0$ (the first row), we set $T_{00}^{(\text{sink})} = 1$. Each row remains a probability distribution $\left(\sum_{j=0}^{S-1} T_{ij}^{(\text{sink})} = 1\right)$, and the intervened attention is

$$A'^{(l,h)} = T^{(\text{sink})}.$$

Non-selected heads are left unchanged.

4 Experiments

4.1 Settings

4.1.1 Identification of MDPs

We use the English Web Treebank (EWT) (Nivre et al., 2020), a publicly available UD-annotated English corpus licensed under CC BY-SA 4.0. The dataset is anonymized and manually curated; we found no personally identifiable or offensive content. It covers a range of syntactic phenomena in web-based English and follows consistent UD annotation guidelines. In our experiments, we used all sentences from the training portion of the EWT

	obl+case	obl	case	conj+cc	conj	cc	nmod+case	nmod	case
Frequency	9095	9150	17417	6413	7523	6757	6883	6888	17417
GPT2-XL	**0.711**	0.405	0.463	**0.456**	0.391	0.424	**0.764**	0.423	0.463
Llama3-8B	0.695	0.311	**0.737**	**0.519**	0.479	0.504	**0.829**	0.451	0.737
Qwen3-8B	**0.626**	0.266	0.486	**0.485**	0.456	0.360	**0.788**	0.318	0.486
BERT-large	0.688	0.313	**0.893**	0.594	0.490	**0.639**	0.858	0.378	**0.893**

Table 1: UAS scores for the three most frequent multi-hop dependency path (MDP) sets (ranks 1–3) and their corresponding single-hop dependency paths (DPs) across all evaluated models.

	obl+case	obl	case	conj+cc	conj	cc	nmod+case	nmod	case
Frequency	4754	8969	10672	3723	7514	3775	1022	6829	10672
GPT2-XL	**0.566**	0.407	0.313	0.374	**0.391**	0.206	0.373	**0.426**	0.313
Llama3-8B	0.488	0.314	**0.680**	0.361	**0.478**	0.364	0.415	0.451	**0.680**
Qwen3-8B	**0.563**	0.262	0.465	0.319	**0.456**	0.217	0.440	0.320	**0.465**

Table 2: The UAS scores for each of the nine identified single-hop dependency paths (DPs) and their corresponding multi-hop dependency path (MDP) sets across all evaluated models. For each DP–MDP pair, we report the accuracy excluding token pairs that are adjacent in the sentence, i.e., direct neighbors are ignored when computing UAS.

corpus. We follow the procedure in Section 3.1: we enumerate single-hop UD edges (DPs) from the training split, and compose adjacent edges on UD trees to construct 2-hop and 3-hop label sequences as MDP candidates. We retain only MDP labels whose corpus frequency is at least $\tau = 50$. This threshold is introduced to prevent the MDP inventory from exploding due to an excessive number of low-frequency chains when no cutoff is applied. Note that all nine DP/MDP pairs used in our main analyses occur well above $\tau = 50$, so the choice of τ does not change the target set for the main results reported in this paper.

4.1.2 Models

We analyze attention heads for three pretrained autoregressive LMs: **GPT-2 XL** (1.7B) (Radford et al., 2019), **Llama 3 8B** (Grattafiori et al., 2024), **Qwen3-8B** (Yang et al., 2025) and **BERT-large-uncased** (*BERT-large*) (Clark et al., 2019).

4.1.3 UAS Measurement and DP–MDP Usage Analysis

We computed Unlabeled Attachment Scores (UAS) for the identified DPs and MDPs following the methodology described in Section 3.2. To determine whether models utilize single-hop DPs or multi-hop MDPs, we apply the frequency-based MDP usage criterion. In brief, our goal is to provide empirical evidence that models leverage both the DP and MDP routes as signals associated with grammatical judgments, and we therefore assess the extent to which a single-hop DP is predominantly realized as part of an MDP using a simple frequency ratio:

$$\text{MDP Usage} = \frac{\text{Count(DP in MDP)}}{\text{Total Count(DP)}} > 0.5. \quad (9)$$

Based on this selection procedure, we pick *nine* DP–MDP pairs as targets for analysis and intervention.

Concretely, the nine DPs are: `obl`, `conj`, `nmod`, `advcl`, `xcomp`, `ccomp`, `acl:relcl`, `parataxis`, `acl`; and their corresponding MDPs are: `obl+case`, `conj+cc`, `nmod+case`, `advcl+mark`, `xcomp+mark`, `ccomp+nsubj`, `acl:relcl+nsubj`, `parataxis+nsubj`, `acl+mark`.

4.1.4 Intervention Setup

Following the methodology in Section 3.2 and the usage criterion in Section 4.1.3, we rank heads by UAS and construct three selection sets for the `obl`/`obl+case` pairing: (i) *DP-only* = the top 10% heads by UAS for `obl`; (ii) *MDP-only* = the top 10% heads by UAS for `obl+case`; (iii) *Mix* = the union of the two top-10% sets. As a size-matched control, *Random-Mix* uniformly samples the same number of heads as in the Mix condition. We then apply the uniformization intervention from Section 3.3 and run percentage ablations at 5%, 10%, 15%, and 25% of all heads.

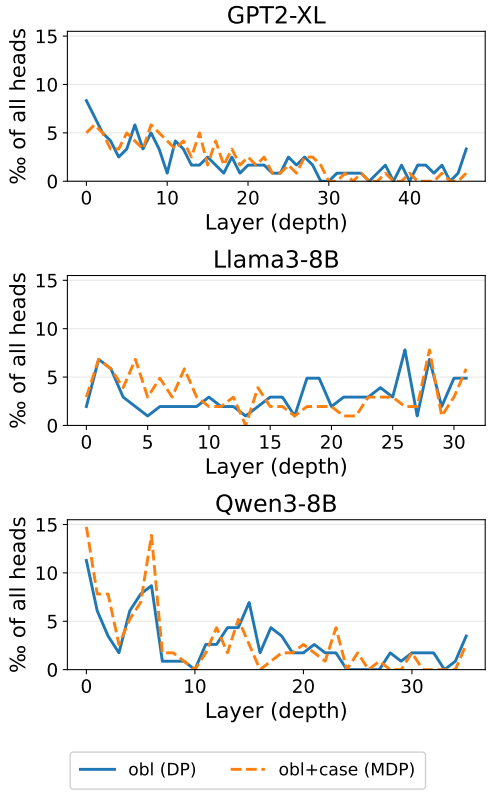

Figure 3: Layer-wise distributions of the top 10% heads by UAS for `obl` (DP; solid) and `obl+case` (MDP; dashed) in each model. UAS is computed by normalizing correct counts by relation frequency. The y-axis shows per-mille of all heads (shared across subplots), and the x-axis denotes layer depth.

4.2 Results

4.2.1 UAS: DP–MDP Comparisons

From Table 1, we observe that certain frequent 2-hop MDPs exhibit higher UAS compared to their individual DP components; however, this pattern is not uniform, and many relations show comparable or higher UAS for the DP than for the corresponding MDP (Appendix Tables 3–4). For GPT-2 XL, Llama 3 8B, and Qwen3-8B, several of our targeted pairings show higher UAS for the MDP than for the DP, while for BERT the corresponding DPs (e.g., `case`, `cc`) often achieve higher UAS than their paired MDPs (Table 1). To control for trivial immediate-neighbor effects, we also recompute UAS after excluding immediately adjacent token pairs; the results are reported in Table 2.

Figure 3 further illustrates layer-wise distributions of top-UAS heads for `obl` and `obl+case`. For GPT-2 XL and Qwen3-8B, top heads tend to cluster in shallower layers, and we do not observe a systematic difference in layer preference between DP and MDP selections.

4.2.2 Intervention Experiments

We assess whether the heads highlighted by UAS are in fact utilized for grammatical construction and related computations using the intervention setup defined in Sec. 4.1.4. To assess overlap and specialization between DP- and MDP-selected heads, we inspect the DP–MDP rank scatter (Figure 5; `obl` vs. `obl+case`). Points mark heads that appear in both top-K lists (ranked by UAS); the dashed diagonal indicates perfect rank agreement. While there is clear overlap, the dispersion around the diagonal indicates that some heads are comparatively specialized for DP vs. MDP.

Turning to the bar results for BLiMP and LAMBADA (Figure 4), we observe that Mix yields a stronger intervention effect than either DP-only (10%) or MDP-only (10%) on the `obl`/`obl+case` pair, and all three targeted selections surpass the size-matched Random-Mix baseline. Moreover, MDP-only generally shows a larger effect than DP-only. Although we focus on `obl`/`obl+case` in the main text, analogous analyses for other DP–MDP pairs are provided in the Appendix.

We next consider the line-plot intervention setup and results (Figures 6). We plot interventions at 5%, 10%, 15%, and 25% of all heads; for each percentage level, we ablate the top-k heads ranked by UAS when selected using `obl` (DP; solid) or `obl+case` (MDP; dashed), with a random-ablation band (min–max) shown for reference. Across BLiMP and LAMBADA, DP/MDP-targeted interventions typically exceed the random band, and MDP-based ablations tend to induce larger drops than DP-based ablations. In this example, all settings follow the above trend except for Llama 3 8B on LAMBADA, where the advantage over random is attenuated.

5 Discussion and Conclusions

Prior work has evaluated attention–syntax alignment primarily at the level of single dependency edges (DPs). In this study, we measure alignment not only for DPs but also for multi-hop dependency paths (MDPs) using the same procedure, and we additionally perform uniformization interventions. We find heads that align with MDPs for several relations, and in some settings interventions on MDP-aligned heads induce larger performance changes than interventions on DP-aligned heads. We also observe both overlap and dissociation: some top heads are shared between DP and MDP selections,

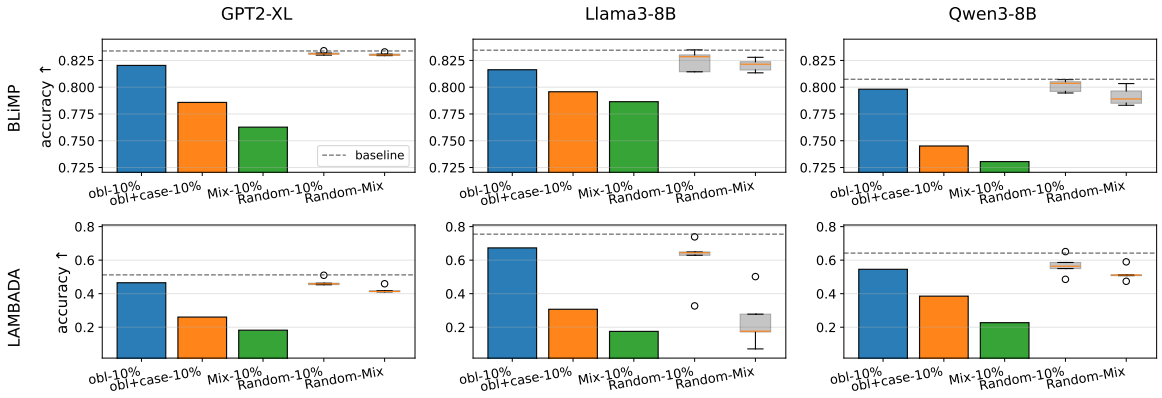

Figure 4: BLiMP accuracy and LAMBADA accuracy for GPT-2 XL, Llama 3 8B, and Qwen3-8B under four ablation sets for the obl/obl+case pair. Bars compare: (i) DP-only (obl) top 10% heads by UAS, (ii) MDP-only (obl+case) top 10%, (iii) Mix = the union of the two top-10% sets, and (iv) Random-Mix = a random selection with the same number of heads as the Mix condition. Higher is better.

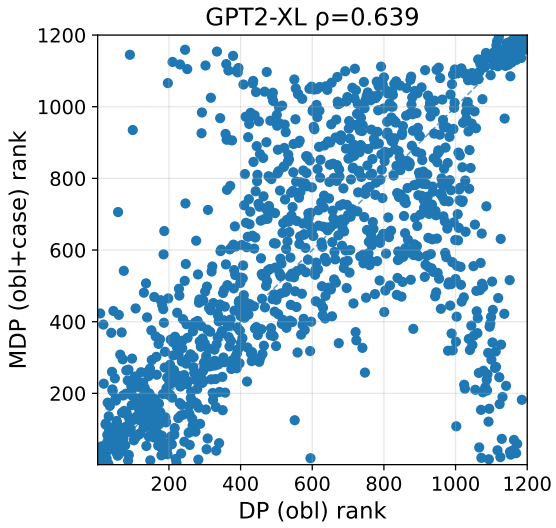

Figure 5: DP–MDP rank agreement for head selections (UAS), shown as a scatter of DP rank (x) versus MDP rank (y) for the relation pair obl vs. obl+case. Points mark heads that appear in both top-K lists (ranked by UAS); lower ranks are better. The dashed diagonal indicates perfect agreement between DP and MDP rankings.

while others are distinct. Taken together, these observations indicate that models can rely on both DP and MDP routes internally. Our findings indicate that LLMs exhibit attention patterns aligned with dependency structures—including multi-hop paths—not explicitly defined as primitive relations in linguistic theory. Specifically, for a subset of frequent pairings, we identified attention heads that are more responsive to Multi-Hop Dependency Paths (MDPs) than to the corresponding single-hop Dependency Paths (DPs).

Particularly notable are DPs that rarely occur alone but frequently appear in specific sets; in these cases, we found that models more accurately attend to MDPs containing these DPs than to the DPs alone. One potential explanation for this phenomenon involves token proximity. As indicated in Figure 7, the average token distance in frequently co-occurring DP sets is typically larger than in their corresponding MDPs. Considering the autoregressive nature of decoder-only LMs, the shorter token distances in MDPs could facilitate easier attention.

A plausible explanation is that autoregressive next-token prediction favors short, locally predictive MDP shortcuts (often mediated by nearby markers like case/mark/cc) under causal masking. Encoders like BERT are less constrained due to bidirectional access, which may reduce reliance on such intermediate routing, consistent with Table 1. It is possible that certain attention heads predominantly attend to adjacent tokens, regardless of grammatical considerations, thereby driving this observed pattern. To investigate this hypothesis, we repeated the UAS analysis while excluding token pairs immediately adjacent to each other. The results (Table 2) show that the main alignment patterns persist even after removing all adjacent-token pairs. This suggests that our findings are not explained solely by trivial "previous-token" heuristics, and supports the interpretation that some heads track non-adjacent, structure-consistent routes (including MDPs).

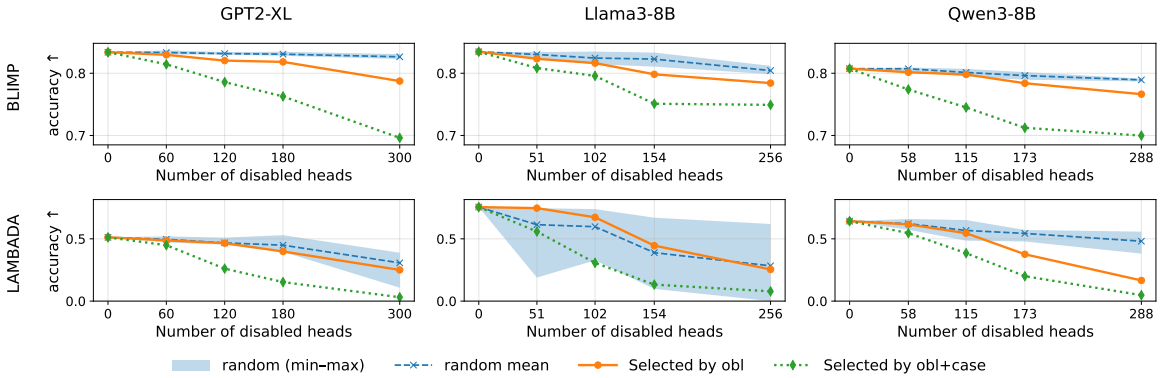

Figure 6: BLiMP accuracy and LAMBADA accuracy as a function of the number of ablated heads for GPT-2 XL, Llama 3 8B, and Qwen3-8B. We plot interventions at 5%, 10%, 15%, and 25% of all heads. For each model, we ablate the top-k heads ranked by UAS when selected using the DP `obl` (solid) or the MDP `obl+case` (dashed), where k corresponds to the specified percentages of total heads. A random-ablation band (min–max with the same head counts) is shown for reference. The x-axis denotes the number of ablated heads and the y-axis denotes BLiMP accuracy.

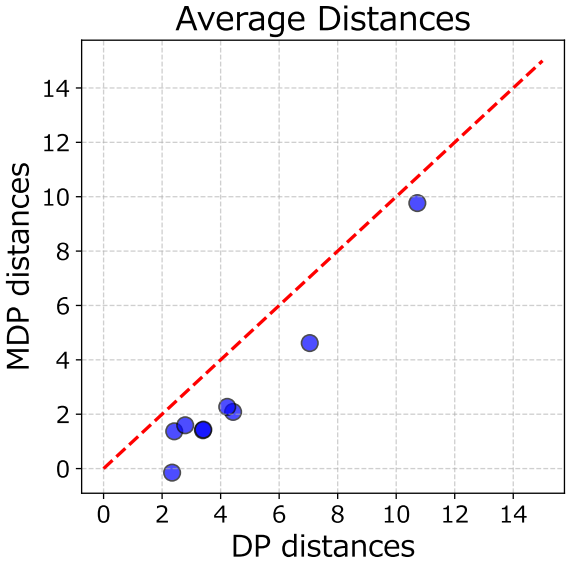

Figure 7: Scatter plot comparing average dependency path (DP) distances (x-axis) against their corresponding multi-hop dependency path (MDP) distances (y-axis) for the nine selected relations. Each point represents one DP–MDP pair, and the red dashed line indicates the identity line (y = x), showing whether MDP distances exceed their DP counterparts.

Limitations

This study has several limitations that provide avenues for future research. First, the complementary relationship between DP and MDP remains unclear. Given that MDPs can often be decomposed into combinations of multiple DPs, it is not yet fully understood what motivates the model to learn these composite structures over individual DPs.

Second, our methodology primarily relies on attention weights. On the other hand, there have been criticisms regarding whether attention weights truly reflect the importance of features in model decisions. Serrano and Smith (2019) showed that altering attention weights does not necessarily lead to significant changes in model output, raising doubts about attention as an indicator of interpretability. Similarly, (Hassid et al., 2022) reevaluated the role of attention mechanisms and reported that using averaged attention weights does not substantially degrade model performance. There may be alternative methods, such as analyzing the neural circuit level, that provide deeper insights into how models internally represent syntactic information.

Finally, we primarily explored medium-sized models. Understanding how larger-scale models behave in terms of dependency encoding and grammatical reasoning, particularly regarding their utilization of DP and MDP structures, remains a promising direction for future work.

Acknowledgments

We used ChatGPT for translation and paraphrasing of our original text, and GitHub Copilot to help draft plotting scripts for our experimental visualizations

References

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.

Nicholas Goldowsky-Dill, Chris MacLeod, Lucas Sato, and Aryaman Arora. 2023. Localizing model behavior with path patching. *Preprint*, arXiv:2304.05969.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. 2025. When attention sink emerges in language models: An empirical view. *Preprint*, arXiv:2410.10781.

Michael Hassid, Hao Peng, Daniel Rotem, Jungo Kasai, Ivan Montero, Noah A. Smith, and Roy Schwartz. 2022. How much does attention actually attend? questioning the importance of attention in pretrained transformers. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1403–1416, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies? *Preprint*, arXiv:1911.12246.

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.

Artur Kulmizev, Vinit Ravishankar, Mostafa Abdou, and Joakim Nivre. 2020. Do neural language models show preferences for syntactic formalisms? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4077–4091, Online. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Denis Paperno, Germán Kruszewski, Angeliki Lazaridou, Ngoc Quan Pham, Raffaella Bernardi, Sandro Pezzelle, Marco Baroni, Gemma Boleda, and Raquel Fernández. 2016. The LAMBADA dataset: Word prediction requiring a broad discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1525–1534, Berlin, Germany. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Vinit Ravishankar, Artur Kulmizev, Mostafa Abdou, Anders Søgaard, and Joakim Nivre. 2021. Attention can reflect syntactic structure (if you let it). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3031–3045, Online. Association for Computational Linguistics.

Sofia Serrano and Noah A. Smith. 2019. Is attention interpretable? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2931–2951, Florence, Italy. Association for Computational Linguistics.

Jesse Vig and Yonatan Belinkov. 2019. Analyzing the structure of attention in a transformer language model. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 63–76, Florence, Italy. Association for Computational Linguistics.

Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2022. Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *Preprint*, arXiv:2211.00593.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2024. Efficient streaming language models with attention sinks. *Preprint*, arXiv:2309.17453.

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, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Example Sentences for Each DP–MDP Pair

To better illustrate the linguistic structures represented by the selected multi-hop dependency paths (MDPs), we provide one example sentence for each of the nine DP–MDP pairs introduced in Section 4.1.3. These examples were extracted from the English Web Treebank (EWT) and selected to be relatively short (13 words or fewer). For each pair, we show the sentence and the syntactic path that connects the relevant tokens, highlighting how the composed MDPs reflect an interpretable syntactic shortcut such as a prepositional phrase or relative clause.

- `obl+case`:

  The third was being run by the head of an investment firm .

  Path: *run* $\xrightarrow{\texttt{obl}}$ *head* $\xrightarrow{\texttt{case}}$ *by*

- `conj+cc`:

  This item is a small one and easily missed .

  Path: *missed* $\xrightarrow{\texttt{conj}}$ *one* $\xrightarrow{\texttt{cc}}$ *and*

- `nmod+case`:

  The third was being run by the head of an investment firm .

  Path: *head* $\xrightarrow{\texttt{nmod}}$ *firm* $\xrightarrow{\texttt{case}}$ *of*

- `advcl+mark`:

  If someone committed a crime against humanity , prosecute the person .

  Path: *prosecute* $\xrightarrow{\texttt{advcl}}$ *committed* $\xrightarrow{\texttt{mark}}$ *If*

- `xcomp+mark`:

  The situation in Iraq is only going to get better this way .

  Path: *going* $\xrightarrow{\texttt{xcomp}}$ *get* $\xrightarrow{\texttt{mark}}$ *to*

- `ccomp+nsubj`:

  You wonder if he was manipulating the market with his bombing targets .

  Path: *wonder* $\xrightarrow{\texttt{ccomp}}$ *manipulating* $\xrightarrow{\texttt{nsubj}}$ *he*

- `acl:relcl+nsubj`:

  Now that 's a post I can relate to .

  Path: *post* $\xrightarrow{\texttt{acl:relcl}}$ *relate* $\xrightarrow{\texttt{nsubj}}$ *I*

- `parataxis+nsubj`:

  Just go here , it 's simply amazing .

  Path: *go* $\xrightarrow{\texttt{parataxis}}$ *amazing* $\xrightarrow{\texttt{nsubj}}$ *it*

- `acl+mark`:

  There has been talk that the night curfew might be implemented again .

  Path: *talk* $\xrightarrow{\texttt{acl}}$ *implemented* $\xrightarrow{\texttt{mark}}$ *that*

	advcl+mark	advcl	mark	xcomp+mark	xcomp	mark	ccomp+nsubj	ccomp	nsubj
Frequency	3506	3817	7774	1834	3070	7774	1925	2327	16270
GPT2-XL	0.411	0.206	**0.461**	**0.947**	0.675	0.461	**0.550**	0.495	0.365
Llama3-8B	0.480	0.206	**0.487**	**0.921**	0.523	0.487	**0.647**	0.438	0.371
Qwen3-8B	0.382	0.182	**0.403**	**0.941**	0.546	0.403	**0.525**	0.523	0.370
BERT-base	0.401	0.303	**0.650**	**0.967**	0.720	0.650	0.433	0.532	**0.590**
BERT-large	0.454	0.358	**0.697**	**0.966**	0.834	0.697	0.497	0.566	**0.671**

Table 3: UAS scores for the three most frequent multi-hop dependency path (MDP) sets (ranks 4–6) and their corresponding single-hop dependency paths (DPs) across all evaluated models.

	acl:relcl+nsubj	acl:relcl	nsubj	parataxis+nsubj	parataxis	nsubj	acl+mark	acl	mark
Frequency	1797	2005	16270	985	1562	16270	816	1493	7774
GPT2-XL	**0.610**	0.463	0.365	0.252	0.174	**0.365**	**0.820**	0.425	0.461
Llama3-8B	**0.686**	0.587	0.371	0.206	0.197	**0.371**	**0.881**	0.543	0.487
Qwen3-8B	**0.635**	0.498	0.370	0.229	0.168	**0.370**	**0.857**	0.466	0.403
BERT-base	0.588	0.382	**0.590**	0.150	0.262	**0.590**	**0.848**	0.480	0.650
BERT-large	0.619	0.396	**0.671**	0.159	0.232	**0.671**	**0.831**	0.498	0.697

Table 4: UAS scores for the three most frequent multi-hop dependency path (MDP) sets (ranks 7–9) and their corresponding single-hop dependency paths (DPs) across all evaluated models.

B UAS Comparisons (Other Dependencies)

This appendix complements the main UAS summary in Table 1 by reporting additional Unlabeled Attachment Scores (UAS) for multi-hop dependency paths (MDPs) and their corresponding single-hop dependencies (DPs). Detailed results are provided in Table 3 and Table 4.

C DP–MDP Rank Agreement for Other Models

This appendix complements the GPT-2 XL scatter in Fig. 5 by presenting the same analysis for Qwen3-8B and Llama 3 8B. Qwen3-8B exhibits a pattern similar to GPT-2 XL: notable overlap near the diagonal with some dispersion, indicating a mix of shared and route-specific heads (Fig. 8). In contrast, Llama 3 8B shows a lower DP–MDP rank correlation, with more points deviating from the diagonal, suggesting stronger specialization toward either the DP or MDP route (Fig. 9).

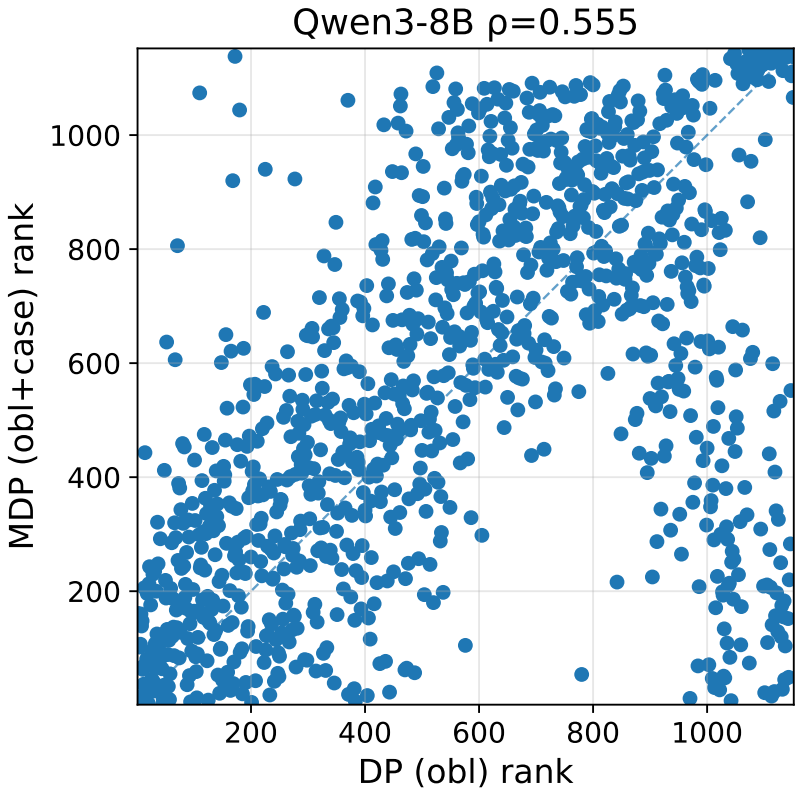

Figure 8: DP–MDP rank agreement for Qwen3-8B on `obl` vs. `obl+case`. Points are heads that appear in both top-K lists (ranked by UAS); lower ranks are better. The dashed diagonal indicates perfect DP/MDP rank agreement.

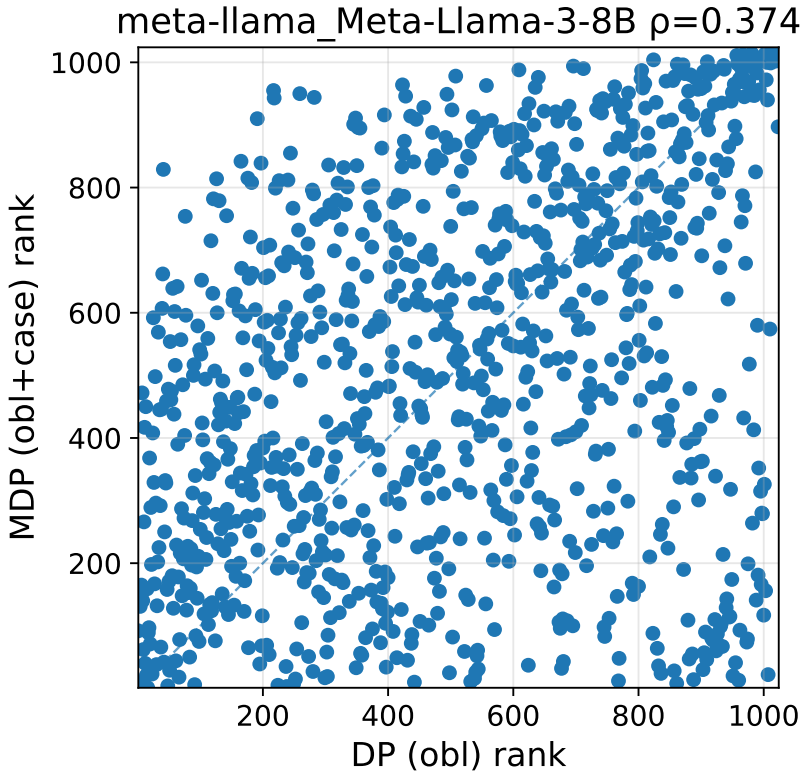

Figure 9: DP–MDP rank agreement for Llama 3 8B on `obl` vs. `obl+case`. Conventions follow Fig. 5; greater dispersion from the diagonal indicates lower rank correlation between DP and MDP selections.

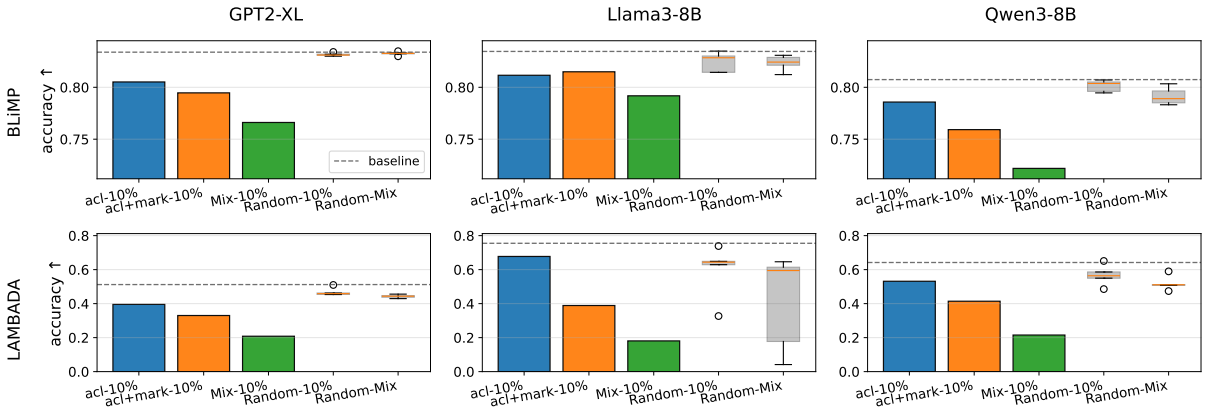

Figure 10: **ACL** families: DP-only, MDP-only, Mix, Random-Mix. Same conventions as Fig. 4.

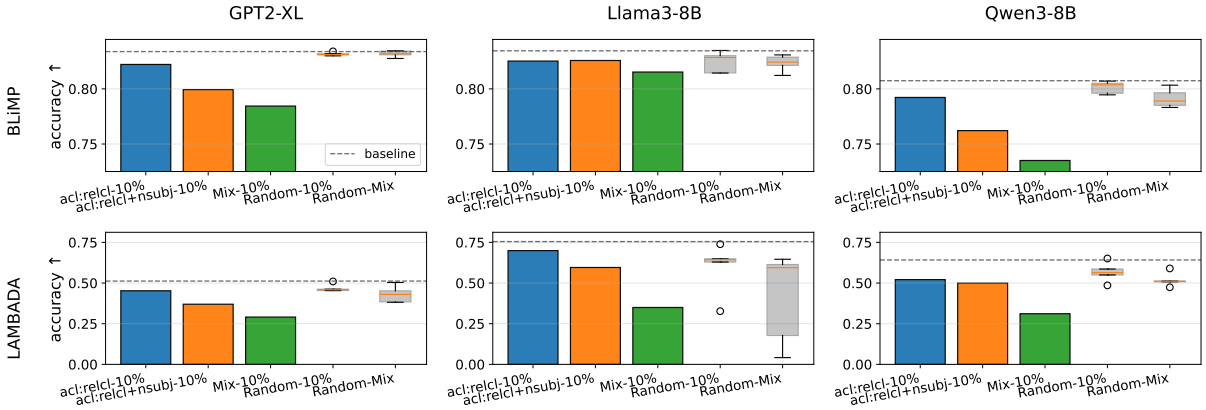

Figure 11: **ACL:RELCL** vs. **ACL:RELCL+NSUBJ**. Same conventions.

D Additional Results for Non-obl Relations: Bar Summaries

This appendix complements Fig. 4 by reporting BLiMP/LAMBADA bar summaries for the remaining DP–MDP pairs (Figs. 10–17). The qualitative pattern mirrors the obl/obl+case case. Except for parataxis and parataxis+nsubj, all DP–MDP pairs exhibit the same qualitative pattern as obl/obl+case: MDP-only selections typically induce larger drops than DP-only, and the *Mix* set outperforms size-matched random controls. In contrast, for parataxis and parataxis+nsubj, intervention effects are comparable to random, suggesting little additional leverage beyond untargeted head selection. A plausible explanation is that parataxis in UD covers heterogeneous, loosely attached clause-level relations (e.g., sentence-level asides and asyndetic coordination), which dilute a consistent attention signature; see concrete examples in Appendix A.

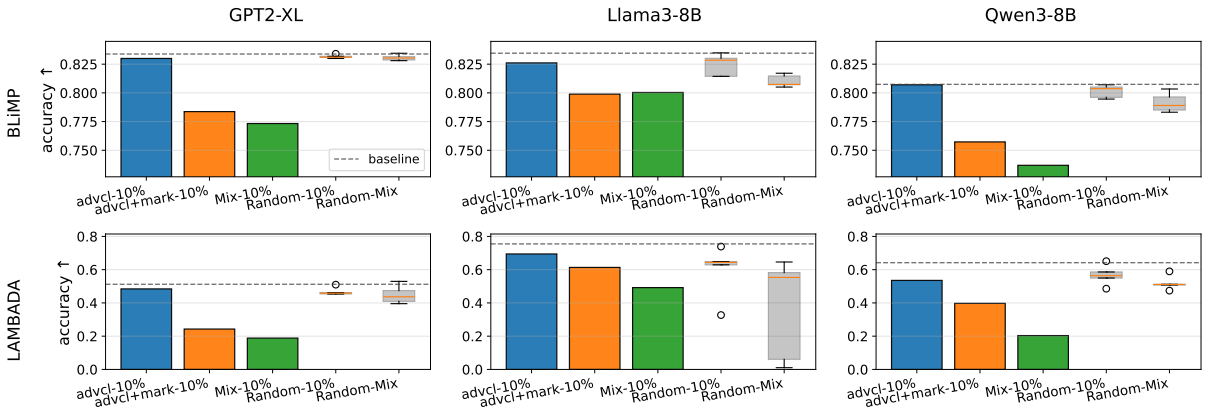

Figure 12: **ADVCL** vs. **ADVCL+MARK**.

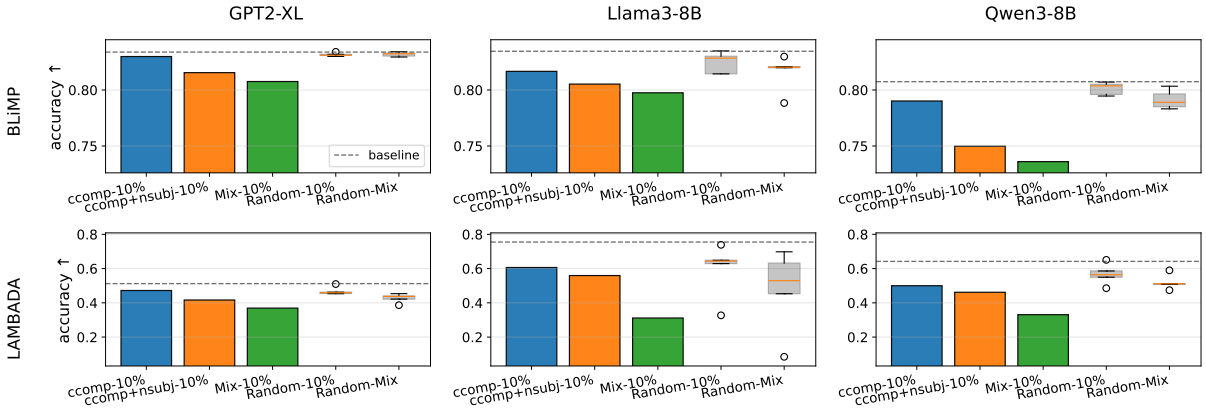

Figure 13: **CCOMP** vs. **CCOMP+NSUBJ**.

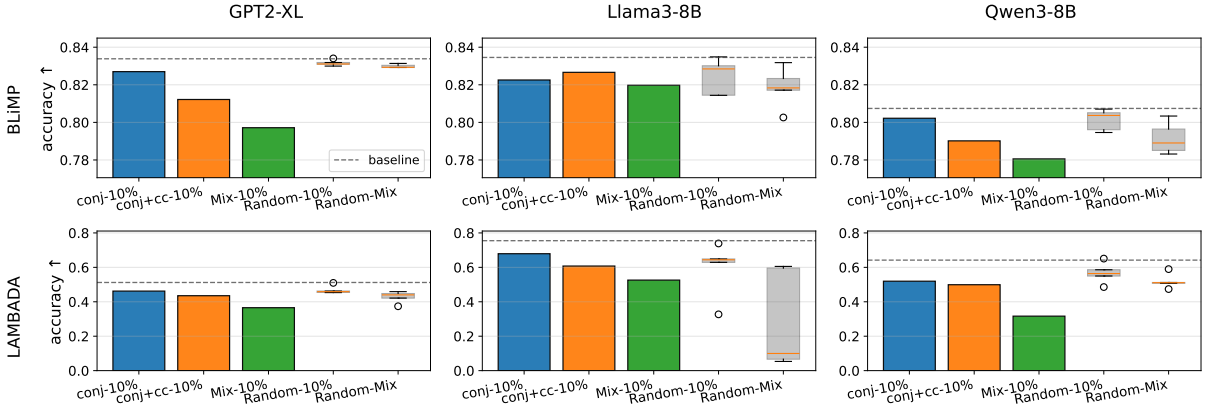

Figure 14: **CONJ** vs. **CONJ+CC**.

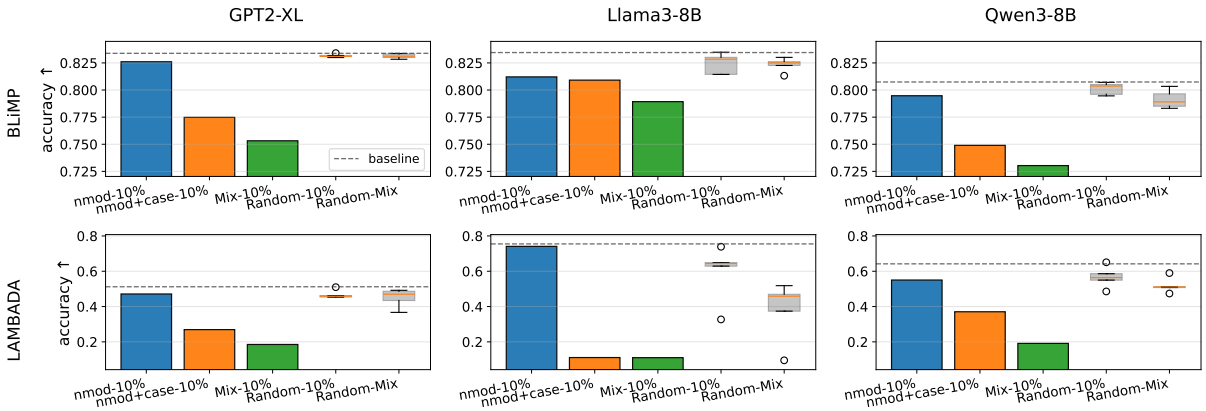

Figure 15: **NMOD** vs. **NMOD+CASE**.

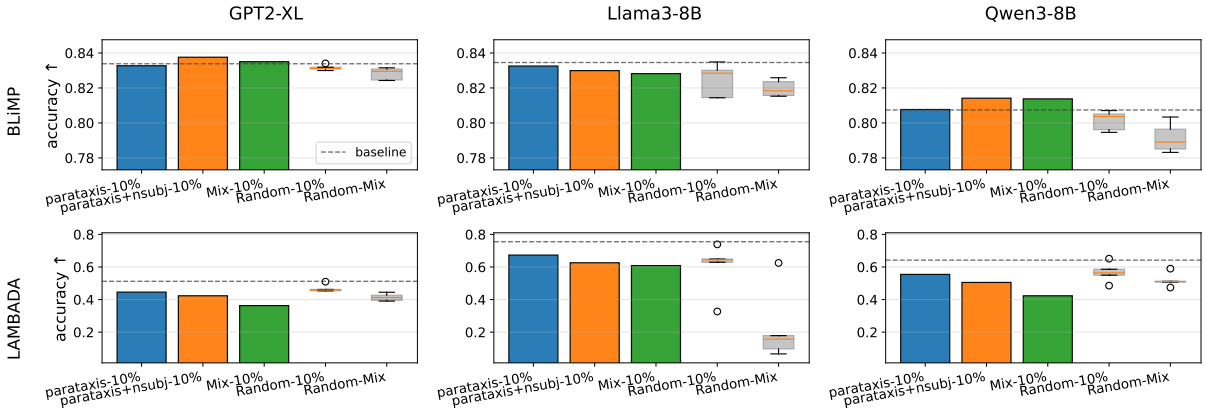

Figure 16: **PARATAXIS** vs. **PARATAXIS+NSUBJ**.

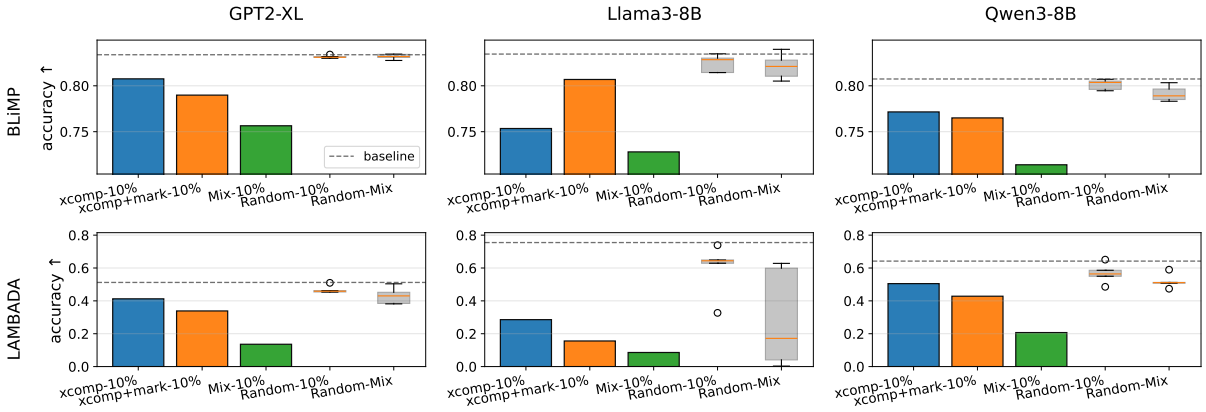

Figure 17: **XCOMP** vs. **XCOMP+MARK**.

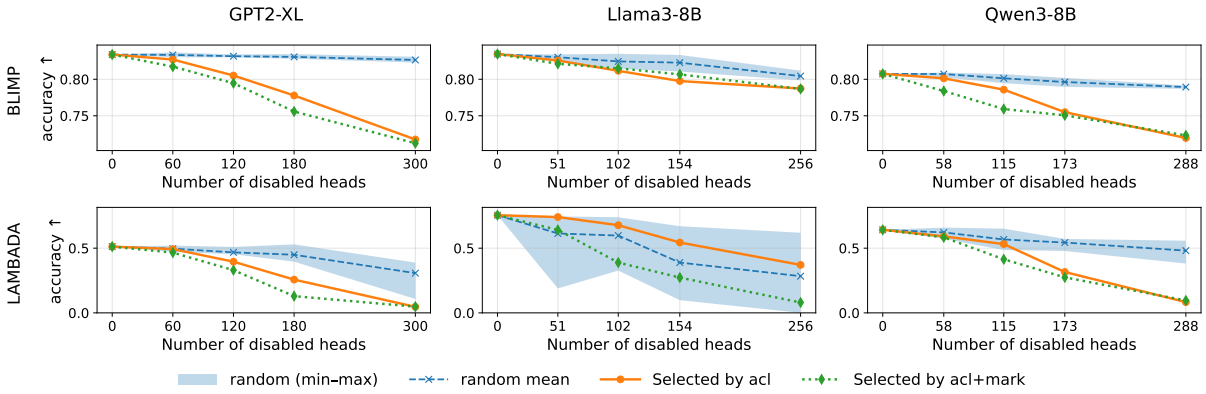

Figure 18: **ACL** families: accuracy vs. ablated head share. Same conventions as Fig. 6.

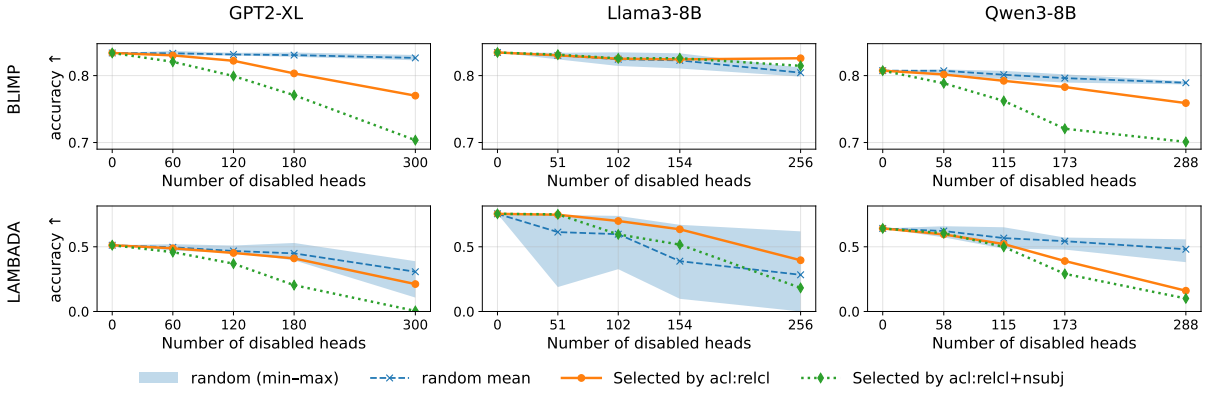

Figure 19: **ACL:RELCL** families.

E Additional Ablation Curves for Non-obl Relations

We also vary the number of intervened heads (5/10/15/25% of all heads) and plot BLiMP/LAMBADA accuracy, following Fig. 6. MDP-based selections (dashed) generally degrade more than DP-based selections (solid), exceeding random baselines (Figs. 18–25). Mirroring the bar summaries, varying the ablation budget (5/10/15/25%) shows consistent degradations for most DP–MDP pairs, with MDP-based selections degrading more than DP-based selections and exceeding random baselines. The exception is `parataxis` and `parataxis+nsubj`, where curves remain near the random band, indicating minimal targeted effect. We hypothesize this stems from the heterogeneous, discourse-level nature of `parataxis` in UD—its instances span diverse constructions with variable distances—making route-specific heads harder to isolate; see Appendix A for examples.

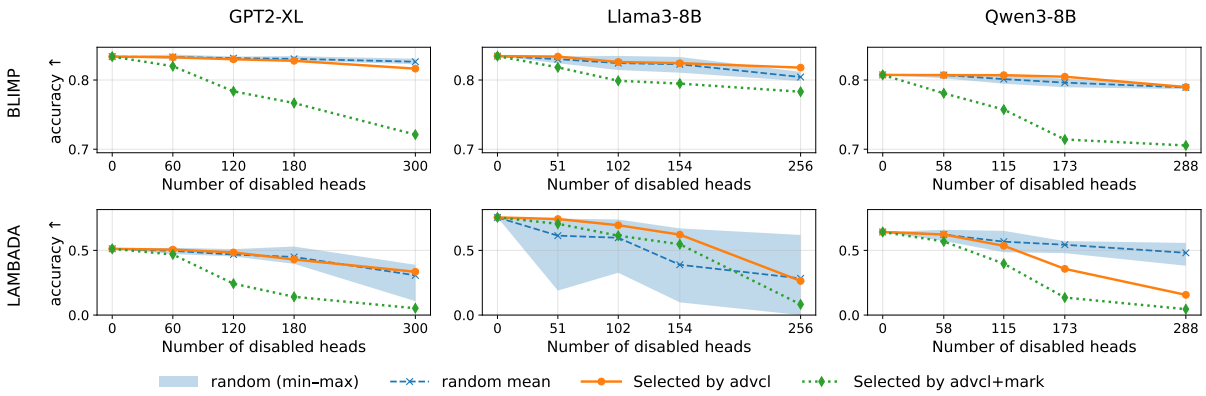

Figure 20: **ADVCL** vs. **ADVCL+MARK**.

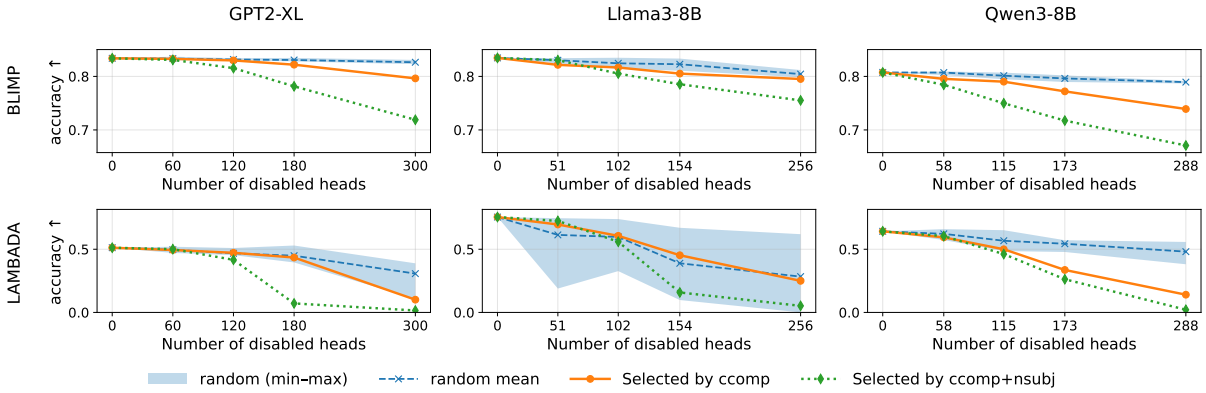

Figure 21: **CCOMP** vs. **CCOMP+NSUBJ**.

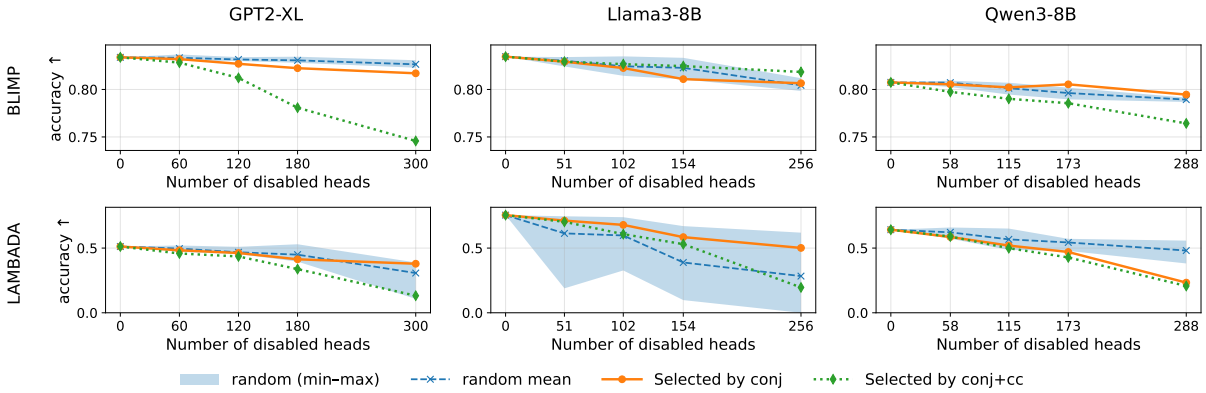

Figure 22: **CONJ** vs. **CONJ+CC**.

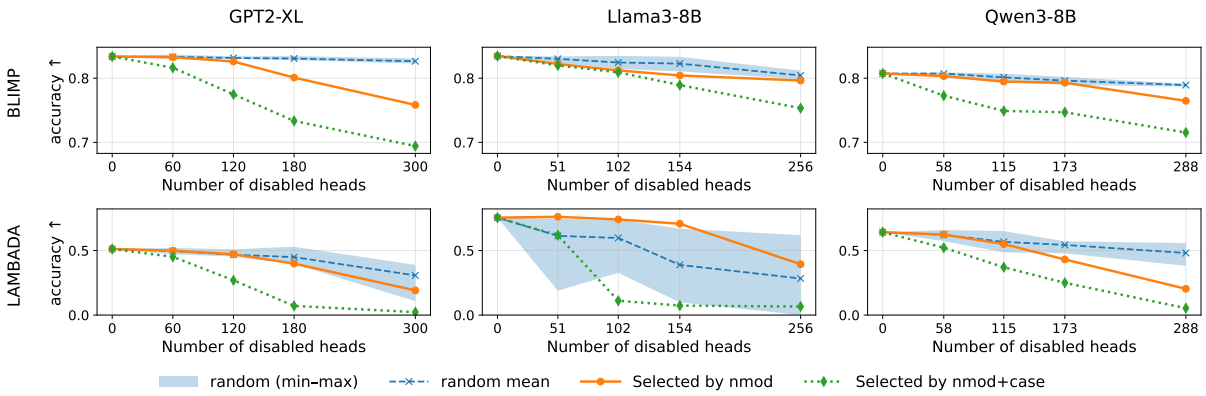

Figure 23: **NMOD** vs. **NMOD+CASE**.

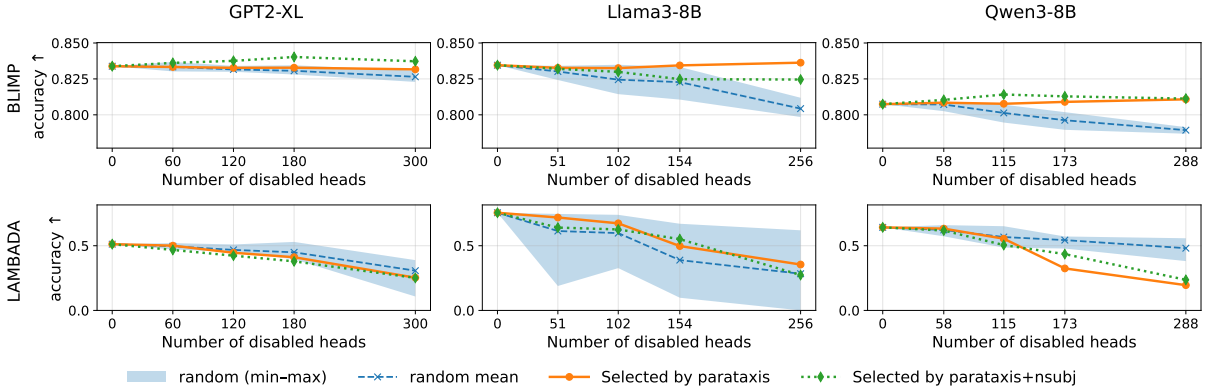

Figure 24: **PARATAXIS** vs. **PARATAXIS+NSUBJ**.

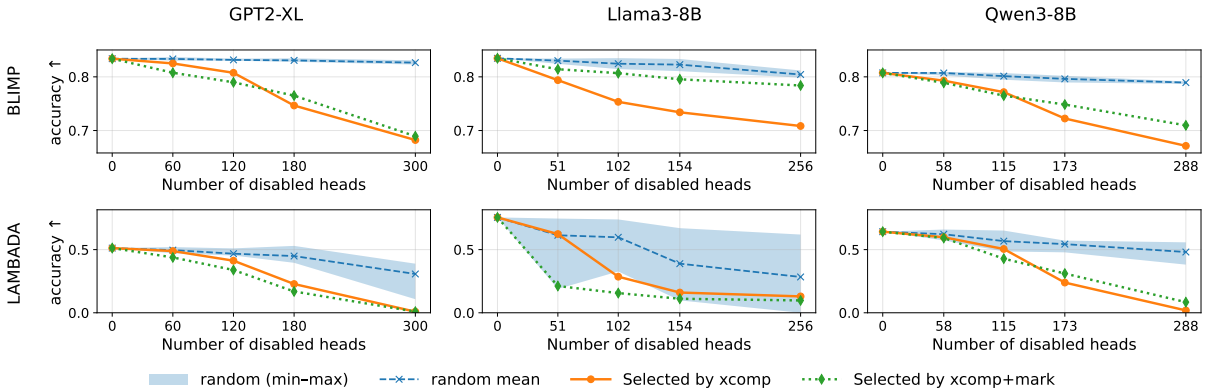

Figure 25: **XCOMP** vs. **XCOMP+MARK**.

F BLiMP Subtask-wise Ablation Curves for `obl` vs. `obl+case`

We report subtask-wise BLiMP ablation curves for three models (GPT2-XL, Llama3-8B, and Qwen3-8B) under the same setup as Fig. 6. In this appendix, DP-based head selections are defined by the `obl` relation (`pair_label=obl`; solid), while MDP-based selections are defined by the two-hop route `obl+case` (`mdp_label=obl+case`; dotted). For each BLiMP subtask, we vary the number of intervened (disabled) heads and plot accuracy; random selections are shown as a shaded min–max band with its mean (dashed), following the same conventions as the main figures. Across many subtasks and models, MDP-based selections tend to yield slightly lower curves than DP-based selections, although the magnitude of the difference varies by subtask and is not always pronounced.

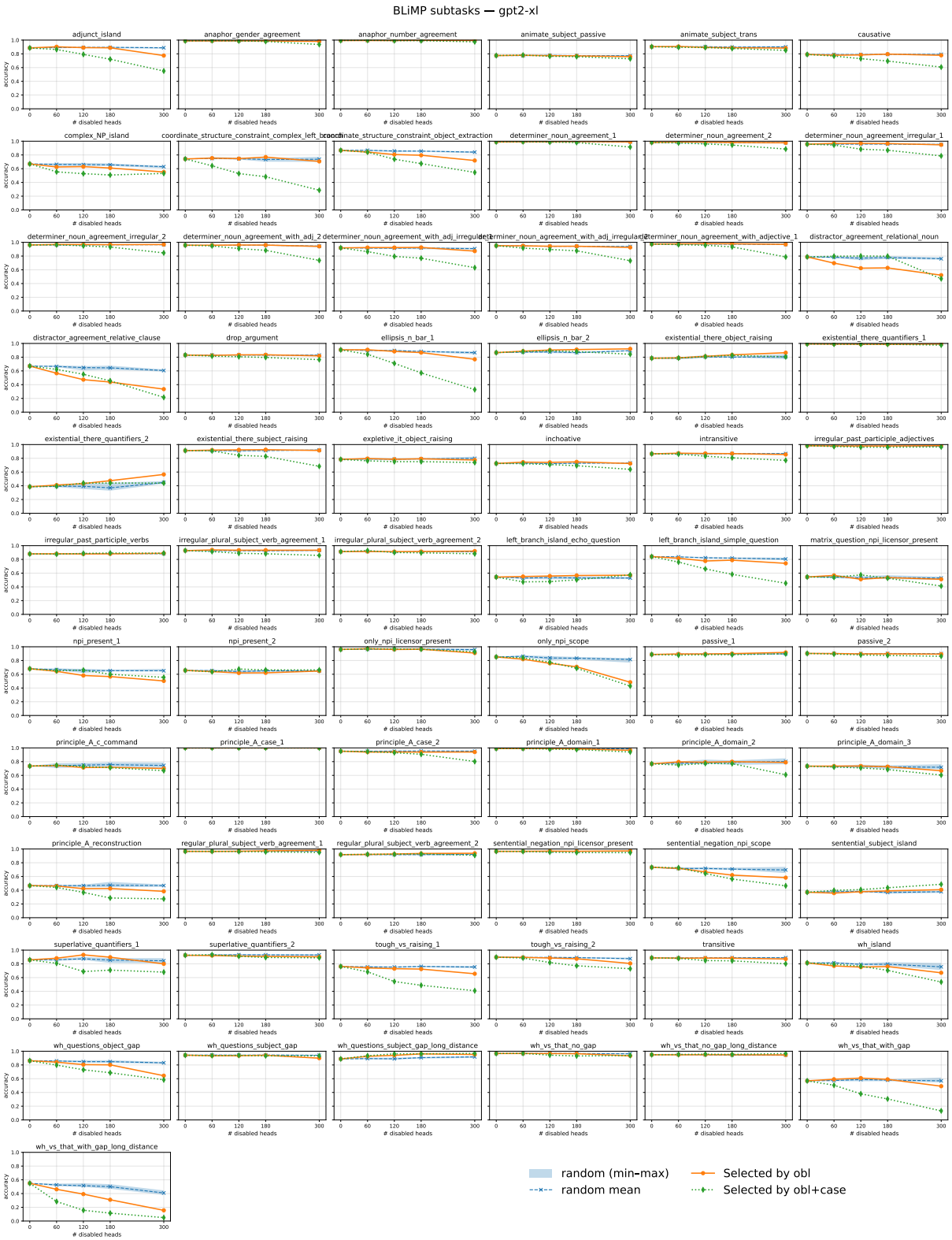

Figure 26: **GPT2-XL**: BLiMP subtask accuracy vs. number of disabled heads. DP/obl (solid) vs. MDP/obl+case (dotted), with random band (shaded) and mean (dashed).

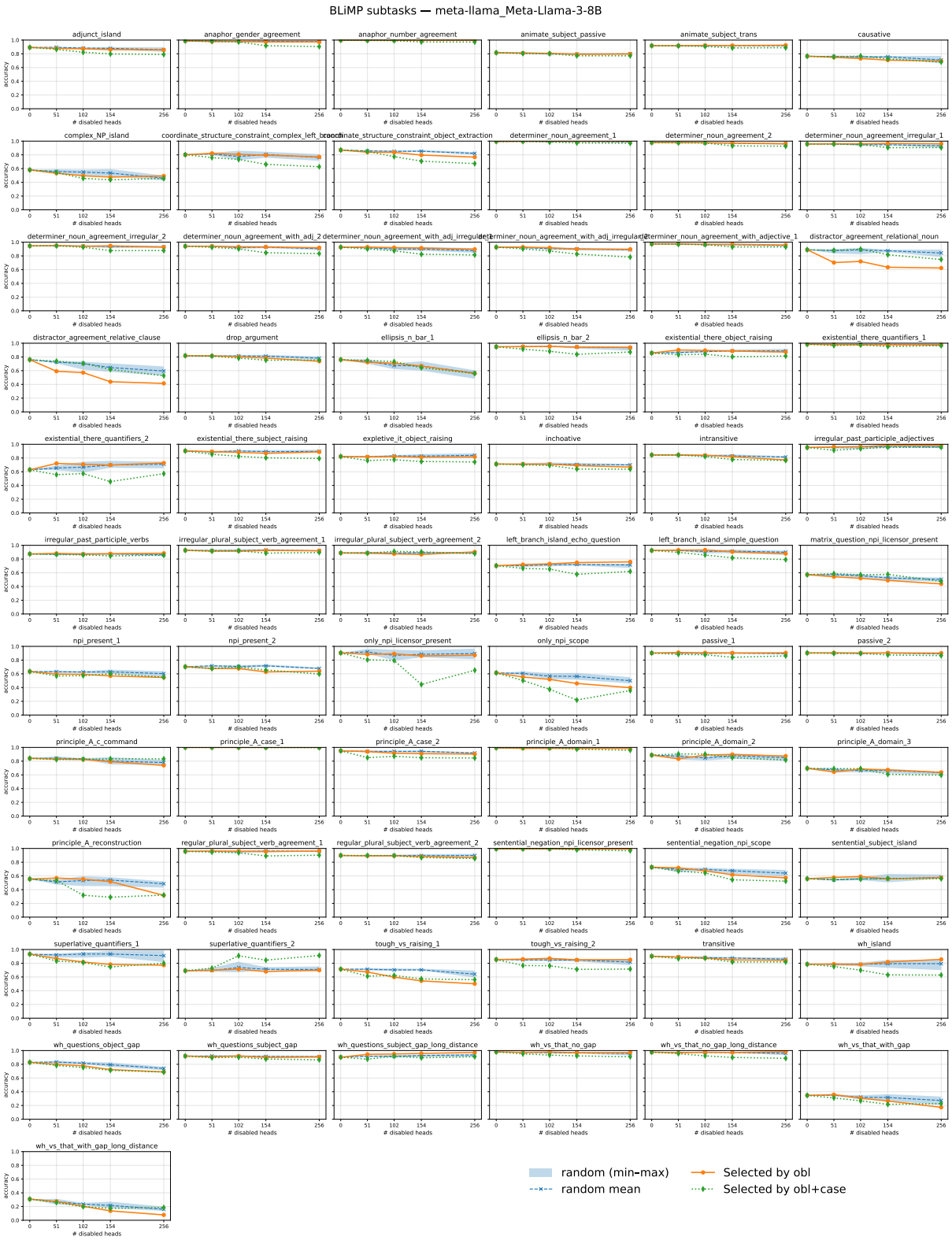

Figure 27: **Llama3-8B**: BLiMP subtask accuracy vs. number of disabled heads. DP/obl (solid) vs. MDP/obl+case (dotted), with random band (shaded) and mean (dashed).

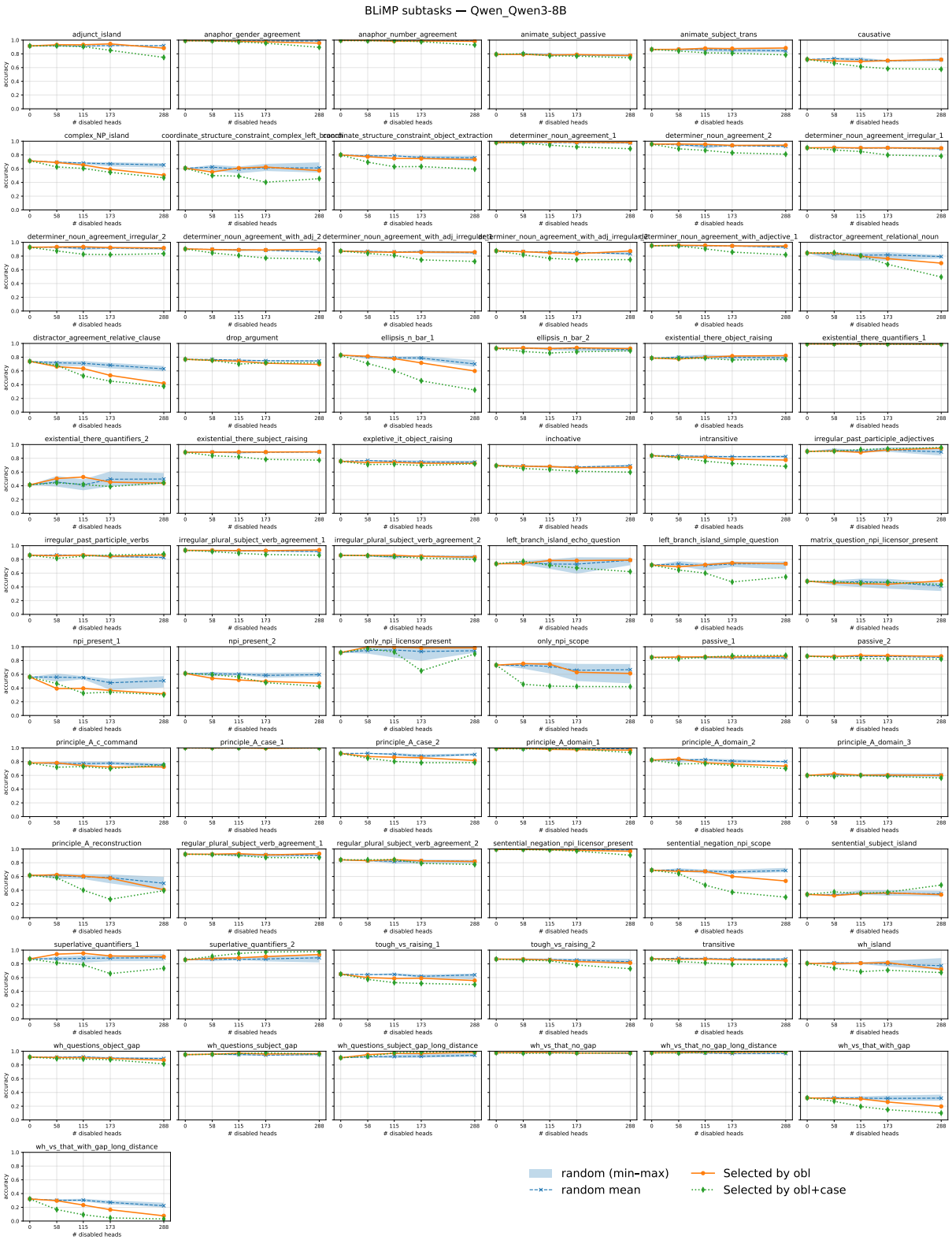

Figure 28: **Qwen3-8B**: BLiMP subtask accuracy vs. number of disabled heads. DP/`obl` (solid) vs. MDP/`obl+case` (dotted), with random band (shaded) and mean (dashed).