Probing ϕ-Feature Representations of Subject–Verb Agreement in Autoregressive Large Language Models

Anonymous ACL submission

Abstract

While large language models (LLMs) achieve high behavioral accuracy on subject–verb agreement, the underlying representational mechanisms remain poorly understood. We recast the formal properties of syntactic AGREE—source, directionality, and intervention—into layer-wise probing diagnostics over autoregressive LLM hidden states. Evaluating seven models across 39 languages, our analysis reveals a systematic distinction across ϕ-features. Gender and number exhibit a *locally encoded* profile: they are decoded from shallow layers at the verb and remain stable under incongruent attractors, suggesting a reliance on local verbal morphology. In contrast, person exhibits a *subject-referential* profile: it manifests at the verb with deeper layer alignment depth, benefits from causal access to the subject in SV order, and is disrupted by incongruent attractors. Geometric analysis confirms that locally encoded features share a representational axis, while subject-referential ones involve geometric re-encoding. Our findings show that LLMs employ non-uniform representational routes for different ϕ-features, a distinction that behavioral metrics alone fail to capture.[1]

1 Introduction

Recent autoregressive LLMs achieve strong subject–verb agreement performance across languages (Jumelet et al., 2026; Grattafiori et al., 2024; Team, 2026). While the behavioral ability of neural networks to track syntax-sensitive dependencies has been extensively evaluated since early recurrent models (Linzen et al., 2016; Gulordava et al., 2018; Marvin and Linzen, 2018), high accuracy does not explain the underlying mechanism. When a model prefers *the boys run* over *the boys runs*, it may have encoded an abstract dependency between the subject and verb; alternatively, it may rely on surface verbal morphology, local lexical statistics, or language-specific agreement paradigms learned during pretraining.

To investigate this internal mechanism, prior probing research has extensively examined how grammatical features are encoded in hidden states (Alain and Bengio, 2016; Tenney et al., 2019; Belinkov, 2022). However, these representational studies have predominantly focused on bidirectional masked language models (Hewitt and Manning, 2019; Jawahar et al., 2019). Furthermore, the probing paradigm itself has faced scrutiny regarding whether the decodable information is actually utilized by the model (Hewitt and Liang, 2019; Ravichander et al., 2021). For modern autoregressive LLMs, comprehensive benchmarks like BLiMP and MultiBLiMP (Warstadt et al., 2020; Jumelet et al., 2026) have established their strong behavioral competence. Yet, while recent studies have begun to explore cross-lingual representations in LLMs (Li et al., 2025), it remains an open question whether the high performance on these benchmarks stems from a uniform process, or whether features like person, number, and gender follow distinct representational routes.

To systematically address this question, we draw conceptual inspiration from the formal generative framework of syntactic AGREE (Chomsky, 2000; Chomsky and Kenstowicz, 2001). In theoretical syntax, agreement is not merely a surface covariance; it is a structural operation wherein a functional head probes a nominal constituent to retrieve specific ϕ-features—morphosyntactic features including person, number, and gender that participate in agreement relations between a controller (e.g., the subject) and a target (e.g., the verb). While neural networks do not inherently execute Minimalist operations, we abstract the core dimensions of this theory—source, directionality, and intervention (Harley and Ritter, 2002)—using them as a blueprint to design our layer-wise probing diagnos-

[1] anonymous github link https://anonymous.4open.science/r/phi-feature-1ED7/README.md

tics.

The architecture of modern autoregressive LLMs makes this directional operation empirically testable. Because their attention is strictly causal, a verb can directly condition on a preceding subject in SV order, but not on a following subject in VS order. To operationalize this, we introduce three diagnostics: (1) we use the SV–VS contrast to test the benefit of direct online access to the realized subject; (2) we use attractor conditions to test whether verb representations are vulnerable to intervening nominals; and (3) to trace the physical flow of information, we measure **layer alignment depth**—the depth at which verb-side decoding first reaches subject-peak levels and the geometric alignment of their representational spaces. Crucially, by applying these diagnostics across different ϕ-features, this approach aims to uncover whether person, number, and gender share a uniform representational pathway or exhibit distinct processing dynamics.

Across seven Llama and Qwen models and 39 languages in MultiBLiMP (Jumelet et al., 2026)—a multilingual minimal-pair benchmark for subject–verb agreement—we track how subject features correlate with verb representations. To distinguish true syntactic transfer from superficial local cues, we categorize representational patterns into two profiles: a **subject-referential** profile (dependent on structural access to the nominal source) and a **locally encoded** profile (resolved locally at the verb).

Our results show a systematic feature-specific split. Person agreement is predominantly *subject-referential*: it exhibits clear structural dependence (benefiting from SV order), is vulnerable to attractors, and emerges at the verb only in deeper layers requiring geometric re-encoding. Conversely, Gender exhibits a *locally encoded* profile: it remains stable under word-order variation and incongruent attractors, remaining highly decodable at the verb from early layers (negligible layer alignment depth) and sharing a unified representational geometry with the subject. With Number occupying an intermediate regime, these results indicate that autoregressive LLMs do not encode agreement uniformly; rather, their representational profiles differ systematically across ϕ-features. Our contributions are threefold:

1. This work shifts multilingual agreement evaluation from a behavioral test of grammatical preference to a representational analysis of how subject features map to verb states.

2. We operationalize the formal properties of AGREE—directionality, structure-sensitivity, and feature non-uniformity—as probing diagnostics over causal LLM hidden states.

3. The analysis reveals that person, number, and gender exhibit substantially distinct representational profiles, linking theoretical syntax to the temporal and geometric mechanics of information flow in autoregressive LLMs.

2 Related Work

Behavioral Evaluation of Syntax in LLMs A robust line of research evaluates whether neural networks capture hierarchical syntax through behavioral tasks. Initial studies showed recurrent networks track subject–verb agreement (Linzen et al., 2016), even in semantically nonsensical contexts (Gulordava et al., 2018). This approach evolved into comprehensive English structural benchmarks like BLiMP (Warstadt et al., 2020). Recently, MultiBLiMP (Jumelet et al., 2026) established that modern LLMs, including Llama 3 (Grattafiori et al., 2024) and Qwen (Yang et al., 2025; Team, 2026), possess strong cross-linguistic agreement capabilities. However, high output probability only confirms behavioral correctness; it cannot distinguish genuine structural computation from superficial heuristics.

Internal Representations and Mechanistic Analysis To investigate how models generate these outputs, linear probes have been extensively used (Alain and Bengio, 2016) to show that contextual embeddings implicitly encode syntax trees (Hewitt and Manning, 2019) and hierarchical structures (Jawahar et al., 2019; Rogers et al., 2020). Concurrent mechanistic interpretability work traces geometric feature superposition (Prieto et al., 2026) and isolates causal attention heads (Nam et al., 2026) or multilingual agreement neurons (Mueller et al., 2022; Ferrando and Costa-Jussà, 2024; Kryvosheieva et al., 2025). Yet, both paradigms face scrutiny: probes may reflect memorization rather than utilization (Hewitt and Liang, 2019; Voita and Titov, 2020; Belinkov, 2022), and analyzing a single target token (e.g., the verb) suffers from a feature confound, failing to differentiate dynamically transferred subject features from static local morphology.

Typological Variation in Language Models Previous probing highlights that multilingual models exhibit language-specific structural differences (Nastase et al., 2024; Li et al., 2025), often mapping concepts to English-centric spaces (Wendler et al., 2024), while morphological generalization remains constrained by a language's inherent complexity (Anh et al., 2024).

3 Experimental Setup and Behavioral Baselines

To investigate whether LLMs encode ϕ-features through locally encoded or subject-referential representational routes, we design an evaluation pipeline that progresses from behavioral outputs to targeted internal representational diagnostics.

3.1 Models and Multilingual Dataset

MultiBLiMP 1.0 (Jumelet et al., 2026) is a massively multilingual benchmark of linguistic minimal pairs, covering 101 languages and two types of subject–verb agreement, with more than 128,000 minimal pairs constructed via an automated pipeline leveraging Universal Dependencies and UniMorph. Each item consists of a grammatical sentence x^+ and a minimally differing ungrammatical counterpart x^- that differ only in the agreement morphology of the target verb. Crucially for our purposes, the dataset provides built-in structural metadata that we exploit as natural experimental contrasts: word order variation (SV and VS constructions), attractor nominal annotation (no attractor, congruent, or incongruent intervener), and subject head position. From this benchmark, we uniformly sample 300 minimal pairs per language–feature condition, covering 39 languages in total. Our evaluation covers seven autoregressive LLMs from the Llama (Grattafiori et al., 2024) and Qwen (Yang et al., 2025; Team, 2026) families, spanning 1B to 27B parameters. Because specific structural diagnostics (e.g., verb-initial orders or attractors) are typologically constrained by language availability in the dataset, we detail the exact number of viable languages for every experimental condition in Appendix F.

3.2 The Insufficiency of Behavioral Baselines

Following the standard MultiBLiMP protocol, we first verify the models' behavioral competence by checking whether they assign a higher probability to the grammatical sentence than its ungrammatical counterpart ($\log p(x^+) > \log p(x^-)$). As reported in Table 1, the models consistently prefer grammatical agreement forms across all applicable languages, with performance consistently improving at larger scales (see Appendix J, Table 10 for a detailed per-language breakdown).

Model	Gender	Number	Person
Llama-3.2-1B	0.682	0.703	0.761
Qwen3.5-2B	0.766	0.800	0.838
Llama-3.2-3B	0.737	0.766	0.801
Qwen3.5-4B	0.822	0.838	0.872
Llama-3.1-8B	0.857	0.849	0.871
Qwen3.5-9B	0.853	0.871	0.893
Qwen3.5-27B	0.894	0.898	0.922
Mean	0.802	0.814	0.849

Table 1: Behavioral agreement accuracy under the standard MultiBLiMP minimal-pair evaluation.

However, this behavioral accuracy confirms that the models ultimately make the correct predictions; it does not reveal *how* or *where* the necessary information is represented internally. Because a correct output could result from either genuine structural computation or shallow local heuristics, these behavioral results serve as a prerequisite baseline. To uncover the true underlying mechanisms, we ground our methodology in the formal generative framework of syntactic AGREE (Chomsky, 2000), operationalizing its core properties into the following structural diagnostics over the models' hidden states.

4 Probing Diagnostics for Agreement Profiles

A central question is which route underlies the verb's feature encoding. We formalize this by defining two diverging representational routes: a **locally encoded** route, where the verb's hidden state resolves the ϕ-feature via its own morphosyntactic cues, and a **subject-referential** route, where the verb representation relies on dynamic structural retrieval from the nominal source.

To determine which route a model exploits, we use the formal generative framework of syntactic AGREE (Chomsky, 2000) as a conceptual blueprint. In theoretical syntax, AGREE defines a search mechanism wherein a functional probe targets a nominal goal to retrieve specific features. By abstracting these theoretical properties, we translate the agreement dependency into three measurable layer-wise diagnostics over the LLMs' hidden

Model	Gender	Number	Person
Llama-3.2-1B	0.448	0.421	0.525
Qwen3.5-2B	0.476	0.480	0.581
Llama-3.2-3B	0.469	0.441	0.548
Qwen3.5-4B	0.516	0.503	0.606
Llama-3.1-8B	0.495	0.481	0.590
Qwen3.5-9B	0.540	0.536	0.635
Qwen3.5-27B	0.591	0.577	0.689

Table 2: Peak verb-side grammaticality probe accuracy. Values are averaged across language–feature conditions.

states: *Source*, *Directionality*, and *Intervention*. Finally, we introduce metrics to trace how these representational dependencies manifest temporally and geometrically within the transformer architecture.

4.1 Source: Verifying the Feature Goal via Cross-Position Probing

In the AGREE framework, structural transfer requires a valid source (the goal) that holds the specific ϕ-feature. Before probing whether the verb dynamically retrieves this information, we must determine what linguistic signal the verb actually encodes. We first test a naive assumption: does the verb merely encode abstract grammatical correctness? To verify this, we train a linear classifier on the target verb's representation to distinguish grammatical from ungrammatical sentences ($h^{(\ell)}_{\text{verb}} \rightarrow \{\text{GRAMMATICAL}, \text{UNGRAMMATICAL}\}$). We implement the probe as an L2-regularized logistic regression classifier trained via 5-fold stratified cross-validation on normalized hidden states (see Appendix B for full extraction protocols).

As Table 2 shows, generalized grammaticality operates near chance-level for Gender and Number. This dissociation from the models' high behavioral accuracy (Section 3.2) demonstrates that predicting the correct token and linearly encoding abstract grammaticality are distinct representational problems.

Consequently, we shift our focus from probing for binary grammaticality to extracting specific ϕ-features. To verify whether the nominal subject acts as a reliable source and whether this information flows to the verb, we introduce two *cross-position feature-value probes* (trained under the same logistic regression setup):

- **Subject-side probe:** Predicts the subject's ϕ-feature from its own hidden state (e.g., extracting a PLURAL label from the representation of *boys*), establishing whether the nominal source is strongly encoded.
- **Verb-side probe:** Predicts the *same* subject feature from the target verb's hidden state extracted exclusively from grammatical sentences (e.g., extracting the PLURAL label from the representation of *run*), evaluating whether the requisite information has successfully manifested at the target.

While high verb-side accuracy confirms that the feature is encoded, it introduces a methodological confound: because we evaluate exclusively on grammatical sentences, the subject and verb features perfectly covary. Thus, a successful verb-side probe might merely be reading the verb's static local morphology rather than capturing dynamic structural transfer from the subject. To separate genuine subject-referential transfer from locally encoded processing, we deploy three targeted diagnostics: word-order directionality, intervention via attractors, and the temporal and geometric dynamics of the representations.

4.2 Directionality: A Test of Subject Access

In formal syntax, AGREE entails a directional search mechanism from the probe to the goal. If a model's verb-side feature encoding genuinely relies on dynamic transfer from the subject, it should similarly benefit from direct access to the subject token. Autoregressive LLMs provide a natural testbed for this hypothesis because their attention is strictly causal. In a subject–verb (SV) order, the verb can causally attend to the realized subject. Conversely, in a verb–subject (VS) order, the verb's representation must be computed before the subject is even observed.

To operationalize this, we measure the directionality effect as the difference in verb-side probe accuracy between these two word orders:

$$\Delta_{\text{SV}-\text{VS}} = \text{Acc}_{\text{verb-side,SV}} - \text{Acc}_{\text{verb-side,VS}}$$

We emphasize that the VS order is not a zero-information baseline; pretrained models can still leverage structural expectations and strong local verbal paradigms. Rather, $\Delta_{\text{SV}-\text{VS}}$ quantifies the *additional* representational gain afforded by direct causal access to the nominal source. Consequently, a near-zero $\Delta_{\text{SV}-\text{VS}}$ suggests that the verb resolves the feature via its local morphology regardless of subject visibility (a locally encoded route), while a clearly positive $\Delta_{\text{SV}-\text{VS}}$ is a hallmark of subject-dependent encoding.

Model	Gender		Number		Person	
	Subj-side	Verb-side	Subj-side	Verb-side	Subj-side	Verb-side
Llama-3.2-1B	0.668	0.848	0.737	0.843	0.898	0.888
Qwen3.5-2B	0.715	0.866	0.766	0.880	0.913	0.908
Llama-3.2-3B	0.699	0.854	0.758	0.856	0.907	0.900
Qwen3.5-4B	0.705	0.858	0.769	0.889	0.912	0.911
Llama-3.1-8B	0.712	0.851	0.769	0.863	0.914	0.908
Qwen3.5-9B	0.714	0.862	0.778	0.891	0.912	0.915
Qwen3.5-27B	0.722	0.868	0.787	0.893	0.921	0.921
Mean	**0.705**	**0.858**	**0.766**	**0.874**	**0.898**	**0.889**

Table 3: Cross-position feature-value probing results across seven autoregressive LLMs. *Subj-side* and *Verb-side* indicate the decodability of the ϕ-feature at each respective position. Reported accuracies are the maximum over layers per model–language–feature, then macro-averaged across languages.

4.3 Intervention: Attractors as a Disambiguator

In the AGREE framework, structural dependencies are constrained by minimality; an intervening nominal (a defective goal) can disrupt the probe's search. We operationalize this concept to resolve the fundamental feature confound inherent in the verb-side probe (Section 4.1). By tracking the verb's representations across different syntactic tree structures, we stratify verb-side accuracy by attractor presence: sentences with no intervening nominals, sentences with a *congruent attractor* (the intervener's feature matches the subject), and sentences with an *incongruent attractor* (the intervener's feature conflicts with the subject).

The incongruent condition serves as our critical disambiguator. Because the target verb's static morphology remains identical regardless of the attractor's presence, a purely locally encoded representation should remain unaffected by an incongruent intervener. Conversely, if the verb-side encoding relies on dynamic structural retrieval from the subject, the presence of a conflicting nominal along the syntactic path should substantially disrupt the probe's accuracy. Thus, vulnerability to incongruent attractors isolates genuine subject-referential transfer from locally encoded processing.

4.4 Physical Flow: Layer Alignment Depth and Representation Geometry

Finally, we investigate how these structural dependencies physically manifest within the transformer architecture by tracking the temporal and geometric dynamics of feature encoding. If autoregressive LLMs resolve person, number, and gender through distinct representational routes, these differences should leave distinct physical signatures as information propagates through the network layers.

To capture the temporal signature, we define **layer alignment depth** as the normalized layer index at which verb-side probe accuracy first reaches within 0.02 of the subject-side peak:

$$\text{layer alignment depth} = \frac{\ell_{\text{align}}}{L-1}, \qquad (1)$$

where $a^*_{\text{subj}} = \max_\ell a_{\text{subj}}(\ell)$ and $\ell_{\text{align}} = \min\{\ell : a_{\text{verb}}(\ell) \geq a^*_{\text{subj}} - 0.02\}$. A near-zero layer alignment depth suggests that the feature is already decodable at the verb at shallow layers (indicative of locally encoded processing), whereas a larger positive layer alignment depth implies that verb-side decoding reaches subject-peak levels only after deeper processing, consistent with subject-referential routing. A sensitivity analysis confirming that the results are robust to this threshold choice is provided in Appendix E.

Beyond layer depth, we also examine how the representational space itself is organized. To capture this spatial signature, we compute **ϕ-axis alignment**, defined as the cosine similarity between the feature-discrimination directions learned by the subject-side and verb-side probes. High alignment indicates that the subject and verb utilize a shared geometric axis to represent the feature contrast, a signature we hypothesize corresponds to locally encoded processing. Conversely, low alignment—especially when combined with a positive directionality effect ($\Delta_{\text{SV}-\text{VS}}$)—suggests that dynamic information transfer requires the feature to be re-encoded into a new geometric direction at the verb.

5 Results

5.1 Source: High Verb-Side Accuracy Conceals Processing Route Differences

We use cross-position feature-value probes to determine whether specific ϕ-features are encoded at the

subject and target verb. As summarized in Table 3, the models reliably encode all three ϕ-features. Crucially, Gender and Number exhibit a *verb-dominant* decodability pattern across almost all models and languages. Specifically, Gender information is substantially more decodable at the verb (0.858) than at the subject source (0.705). This pattern (Verb-side > Subj-side) holds consistently across 11 of 12 evaluated languages for Gender, and 36 of 38 languages for Number. Conversely, Person is most strongly represented at the subject position (mean 0.898, the highest across all features). This strong subject-side decodability demonstrates that the Person feature is stably represented at its nominal source, making it readily available for subsequent structural transfer.

All three features are highly decodable at the target verb. However, as noted in Section 4.1, this verb-side success introduces a potential confound: the strong decodability could reflect active structural retrieval from the subject, mere static local morphology, or simple context accumulation in right-branching trees. Crucially, the relatively lower subject-side accuracy for Gender (0.705) provides a functional motivation for its locally encoded profile: because the nominal source weakly encodes the feature, the model tends to rely on the verb's local morphology. To break this confound and help rule out simple context accumulation, we turn to our targeted structural diagnostics, beginning with word-order directionality.

5.2 Directionality: Person Uniquely Benefits from Subject Access

To determine whether the high verb-side encoding relies on dynamic transfer from the subject, the directionality analysis examines ($\Delta_{\text{SV-VS}}$) across the subset of languages that permit both SV and VS word orders. Figure 1 summarizes verb-side decodability and $\Delta_{\text{SV}-\text{VS}}$ jointly: Person occupies a region largely separated from Gender and Number, consistent with distinct representational routes despite uniformly high verb-side accuracy.

As visualized in Figure 2, the models exhibit a clear, feature-specific divergence in how they utilize prior causal access to the subject. Gender and Number show virtually no benefit from direct subject access. Across all seven model scales, the verb decodes Gender and Number with nearly identical accuracy regardless of whether the subject has already been observed (SV) or has yet to appear (VS). This near-zero directionality effect strongly implies that the verb encodes these features primarily via

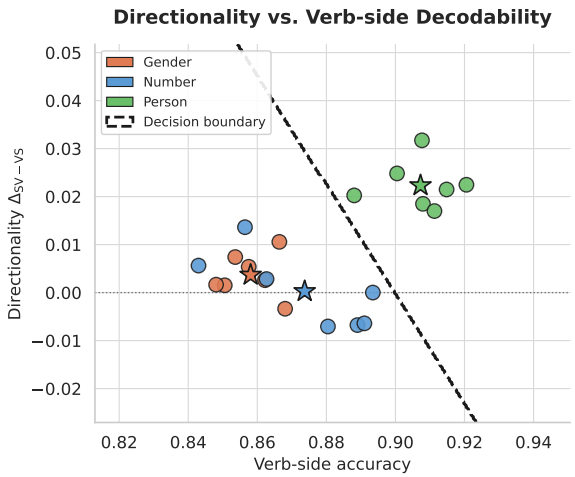


Figure 1: Verb-side decodability vs. $\Delta_{\text{SV}-\text{VS}}$ (points: model–feature; $\star$: mean). Person separates from Gender/Number despite high verb-side accuracy.

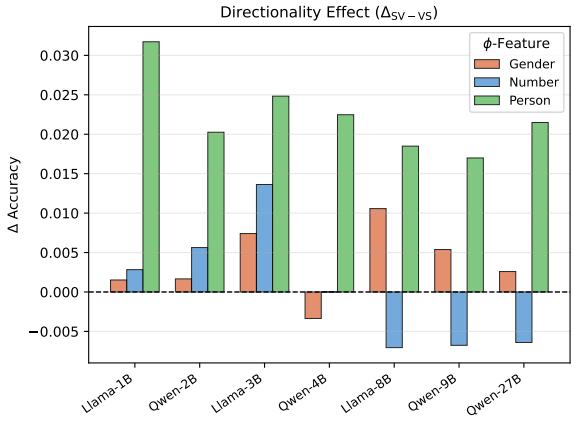


Figure 2: Directionality effect ($\Delta_{\text{SV-VS}}$) across seven models.

local cues, rendering structural subject visibility irrelevant.

Conversely, Person agreement exhibits a consistently positive directionality effect. In the applicable languages (predominantly Romance, Slavic, and Germanic), the verb's ability to decode Person shows a clear increase when the subject is causally accessible in SV order (Figure 4). To confirm this divergence, we conducted a Wilcoxon signed-rank test on the paired (SV, VS) accuracies across all evaluated model–language conditions. The positive directionality effect for Person is marginally significant ($p \approx 0.05$), whereas Gender ($p = 0.73$) and Number ($p = 0.19$) show no significant difference from zero. The near-zero Δ of Gender and Number serves as a natural, structure-matched baseline: it demonstrates that the positive gain observed for Person is not an artifact of generic token frequen-

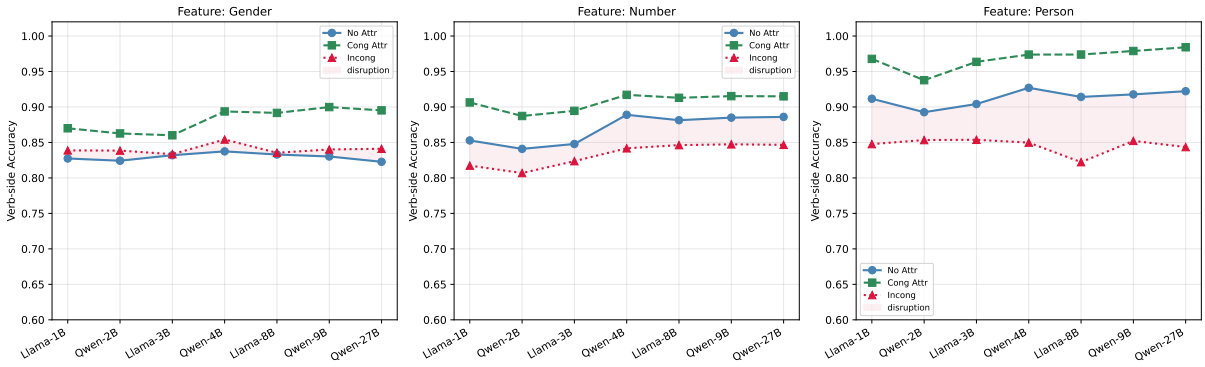


Figure 3: Impact of structural interveners (attractors) on verb-side feature decodability. The red shaded area highlights the disruption caused by an incongruent intervener compared to the clean (no attractor) baseline.

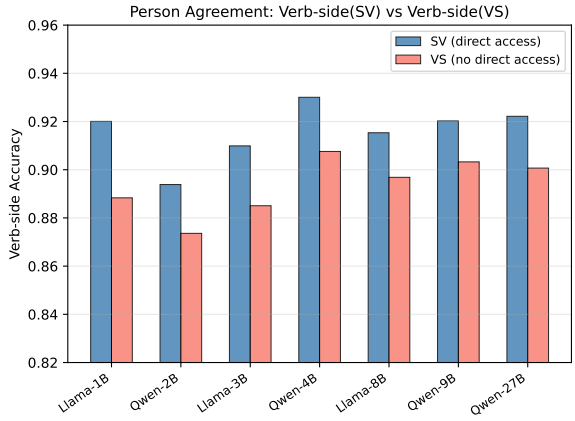


Figure 4: Absolute verb-side accuracy for Person agreement in SV vs. VS orders. The verb decodes Person better when the subject is causally accessible.

cies or context accumulation, but a feature-specific mechanism that actively retrieves Person information from the prior structural subject.

5.3 Intervention: Attractors Isolate Morphology from Transfer

While the lack of a directionality effect for Gender and Number suggests local encoding, one alternative hypothesis is *integrated encoding*: the verb's hidden state might integrate both its own morphology and the dynamically transferred subject information. To mitigate this confound, this analysis evaluates the models' vulnerability to incongruent attractors.

As visualized in Figure 3, the impact of incongruent interveners reveals a nuanced continuum rather than a simple binary split. Gender remains virtually unaffected (mean drop of -0.011), confirming that the models largely avoid subject-referential retrieval and rely purely on a *locally encoded* route.

Conversely, Person agreement exhibits clear vulnerability. Incongruent attractors substantially disrupt verb-side Person encoding, dropping accuracy by an average of 0.067 across all models. A Wilcoxon signed-rank test confirms this disruption is highly significant ($p < 10^{-5}$), whereas Gender remains unaffected ($p = 0.47$). This performance drop, despite the target verb's static morphology remaining constant, provides strong evidence that Person agreement relies on a *subject-referential* representational route highly susceptible to structural interference.

Interestingly, Number occupies an intermediate regime. While it showed virtually no directionality effect (Section 5.2), it does exhibit a mild but consistent vulnerability to attractors (mean drop of 0.036). This suggests that while Number predominantly relies on locally encoded processing, its representation is not completely immune to syntactic interference, reflecting a hybrid or partial dependency (see Table 4 for complete results).

5.4 Physical Flow: Layer Alignment Depth and Geometric Re-encoding

Having established the functional divergence between locally encoded and subject-referential routes, this section investigates how these dependencies physically manifest within the transformer architecture. If these features follow distinct processing routes, they should leave distinct temporal (layer alignment depth) and spatial (geometric alignment) signatures.

Temporal Manifestation (Layer Alignment Depth). As introduced in Section 4.4, layer alignment depth measures the depth at which verb-side probe accuracy first reaches within 0.02 of the subject-side peak. Our findings align with the structural diagnostics (Figure 6): Gender and Num-

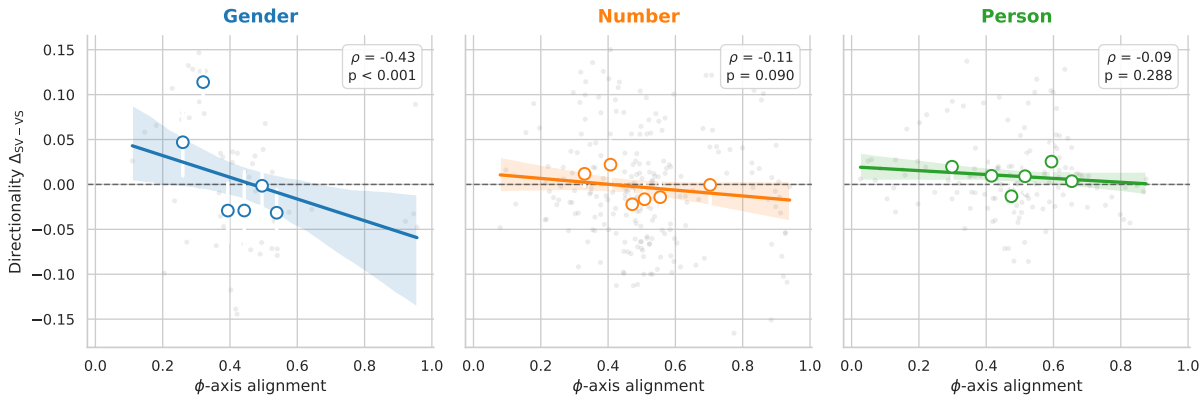


Figure 5: Spatial manifestation of representational profiles (ϕ-axis alignment vs. Directionality Δ). Gender's steep slope reflects strong local dependence, whereas Person's flatter slope indicates a consistent, structurally driven reliance on subject access regardless of slight geometric variations.

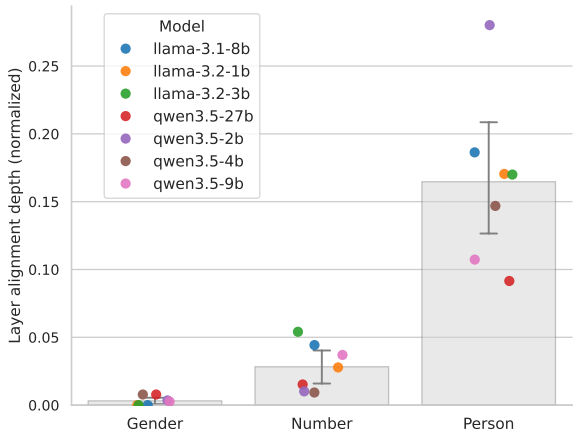


Figure 6: Layer alignment depth across features. Gender and Number show shallow layer alignment depth; Person requires substantially greater depth.

ber show near-zero layer alignment depth (means 0.003 and 0.027), indicating that verb-side ϕ is already decodable at shallow layers at subject-peak levels—consistent with locally encoded agreement rather than late subject-to-verb reconstruction. Person shows a substantially larger layer alignment depth (mean 0.164), requiring deeper layers before verb-side decoding matches the subject-side peak, consistent with subject-referential processing.

Spatial Manifestation (Representation Geometry). We further ask whether this feature-specific split represents a continuous trade-off in the models' representational space. Across 427 viable model–language–feature conditions, the analysis reveals a modest but consistent negative correlation between ϕ-axis alignment and the directionality effect (Pooled Spearman $\rho = -0.138$, $p < 0.01$). As visualized in Figure 5, conditions where the subject and verb utilize the same geometric axis (high alignment) tend to show less benefit from direct subject access (low $\Delta_{\mathrm{SV-VS}}$). Crucially, the slope of this trade-off differs significantly by ϕ-feature (confirmed via interaction terms in a Linear Mixed-Effects Model; see Appendix C). Gender exhibits a notably more negative slope ($\rho = -0.43$), suggesting that its encoding is more sensitive to geometric alignment: when subject and verb share a feature axis, the model draws less on structural access. In contrast, Person shows a flatter slope with a positive $\Delta_{\mathrm{SV-VS}}$ across conditions, consistent with its subject-referential profile: regardless of geometric alignment, the verb's Person encoding relies on structural transfer, often necessitating re-encoding into a new representational direction.

6 Conclusion

This paper asked how autoregressive LLMs represent subject–verb agreement across 39 languages. By shifting from behavioral correctness to representational profiling, the results show that while verb states reliably encode ϕ-features, their underlying physical signatures vary. Gender exhibits a predominantly *locally encoded* profile with negligible layer alignment depth, while number predominantly follows this local route but occupies an intermediate regime. Person, conversely, is distinctly *subject-referential*: it requires deeper temporal processing, benefits from direct structural access, and involves geometric re-encoding that renders it vulnerable to attractors. These findings show that LLMs do not resolve agreement uniformly; instead, they route information across distinct temporal and geometric pathways dictated by the specific ϕ-feature.

Limitations

While our findings provide a novel representational perspective on multilingual agreement, several limitations remain. First, our linear probing methodology identifies correlational signatures rather than strict causal circuits. Although our targeted structural diagnostics (directionality and attractors) mitigate major confounds, future work should employ causal interventions, such as activation patching or causal tracing, to definitively establish whether the layers identified by probing are causally responsible for agreement computation.

Second, our analysis relies on MultiBLiMP 1.0, which is constructed via an automated pipeline over Universal Dependencies and UniMorph, and exclusively targets morphologically overt agreement in written, standardized text. This means that agreement phenomena encoded through purely syntactic means, as well as variation found in spoken or dialectal registers, fall outside the scope of our diagnostics. Whether the locally encoded and subject-referential profiles we identify persist in more naturalistic settings—where agreement must be resolved over longer distances, across clause boundaries, or under information-structural pressures—remains an important open question.

Third, the attractor-based evidence for the subject-referential profile of Person rests on a narrow typological base: only 6 languages in our dataset provide incongruent Person attractor conditions (Appendix F), compared to 32 for Number and 8 for Gender. Given that attractor vulnerability is our most direct diagnostic for distinguishing genuine subject-referential transfer from morphological encoding, strengthening this test with a broader typological sample is a vital priority for future work.

Fourth, while our findings are aggregated across 39 languages, the per-language variance within each typological family is substantial, as reflected in the wide spread of individual data points throughout our figures. Our cross-linguistic analysis necessarily treats language-level variation as noise to be averaged out, but a dedicated per-language investigation—examining, for instance, how morphological richness, pro-drop status, or verb placement interact with the locally encoded versus subject-referential split—would provide much finer-grained insight into the typological universality of our findings.

Fifth, specific structural tests such as the SV–VS contrast were typologically constrained by dataset availability, limiting our directional and geometric analyses to a predominantly Indo-European subset. Expanding these targeted diagnostics to verb-initial languages and morphologically agglutinative families would be a valuable next step.

Sixth, our evaluation is restricted to seven models from the Llama and Qwen families—all autoregressive *base-model* variants sharing a common set of architectural choices (RoPE positional encoding, grouped-query attention, SwiGLU activations, and pre-norm transformer blocks). Whether the representational profiles documented here generalize to other model families (e.g., mixture-of-experts architectures or encoder–decoder designs), or change systematically for *instruction-tuned* or RLHF-aligned variants of the same models, remains an open question. Alignment training on human feedback can shift internal representations toward task-specific objectives, potentially suppressing or reshaping the subject-referential agreement signal we identify—particularly for Person, whose subject-referential profile is most robust. Extending our diagnostic framework to instruction-tuned variants and architecturally diverse models is an important direction for future work.

Ethical Considerations

This study analyzes pretrained autoregressive language models using publicly available benchmarks and does not involve human subjects, personal data collection, or interactive deployment. All experiments were conducted offline on fixed minimal-pair stimuli from MultiBLiMP 1.0, which is derived from Universal Dependencies and UniMorph resources released under open licenses. We report results only at the aggregate level and do not attempt to identify or profile individual speakers, authors, or communities.

Several interpretive risks warrant caution. First, our findings describe correlational patterns in hidden states and should not be taken as evidence that models possess human-like syntactic competence, linguistic agency, or theory-internal operations such as literal AGREE. Second, although we analyze agreement across 39 languages, coverage remains typologically skewed toward Indo-European and written, standardized varieties. Claims about "universal" neural mechanisms must therefore be framed carefully to avoid marginalizing languages and registers excluded from the benchmark.

Our treatment of *gender* as a morphosyntactic ϕ-feature follows the annotation conventions of the underlying treebanks and should not be conflated with social gender identity. Many languages encode grammatical gender systems that do not map cleanly onto binary or static social categories; our analyses therefore concern morphological agreement classes rather than demographic attributes. We additionally note that high verb-side decodability for gender does not imply that models handle gender-inclusive or non-binary linguistic usage appropriately in open-ended generation.

From a resource perspective, evaluating seven models across dozens of languages required substantial GPU compute. We mitigated unnecessary expenditure by caching hidden states, reusing fixed sampling protocols, and restricting analysis to benchmark sentences rather than large-scale web scraping or repeated fine-tuning. Future work should continue to report compute budgets and prioritize efficient diagnostics over brute-force scaling.

Finally, because interpretability results can inform model selection, auditing, and alignment research, we release code and analysis scripts to support reproducibility. We encourage readers to treat our representational profiles as descriptive tools for scientific inquiry, not as standalone guarantees of grammatical robustness, fairness, or safety in deployed systems.

We used AI-assisted tools solely for grammar checking and linguistic refinement of the manuscript.

References

Guillaume Alain and Yoshua Bengio. 2016. Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.

Dang Anh, Limor Raviv, and Lukas Galke. 2024. Morphology matters: Probing the cross-linguistic morphological generalization abilities of large language models through a wug test. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 177–188.

Yonatan Belinkov. 2022. Probing classifiers: Promises, shortcomings, and advances. *Computational Linguistics*, 48(1):207–219.

Noam Chomsky. 2000. Minimalist inquiries: The framework (mitopl 15). *Step by step: Essays on minimalist syntax in honor of Howard Lasnik*, pages 89–155.

Noam Chomsky and Michael Kenstowicz. 2001. Derivation by phase.

Javier Ferrando and Marta R Costa-Jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205.

Heidi Harley and Elizabeth Ritter. 2002. Person and number in pronouns: A feature-geometric analysis. *Language*, 78(3):482–526.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)*, pages 2733–2743.

John Hewitt and Christopher D Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does bert learn about the structure of language? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 3651–3657.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2026. Multiblimp 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Transactions of the Association for Computational Linguistics*, 14:193–216.

Daria Kryvosheieva, Andrea de Varda, Evelina Fedorenko, and Greta Tuckute. 2025. Different types of syntactic agreement recruit the same units within large language models. *arXiv preprint arXiv:2512.03676*.

Daoyang Li, Haiyan Zhao, Qingcheng Zeng, and Mengnan Du. 2025. Exploring multilingual probing in large language models: A cross-language analysis. In *Proceedings of the 1st Joint Workshop on Large Language Models and Structure Modeling (XLLM 2025)*, pages 61–70.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1192–1202.

Aaron Mueller, Yu Xia, and Tal Linzen. 2022. Causal analysis of syntactic agreement neurons in multilingual language models. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 95–109.

Andrew Nam, Henry Conklin, Yukang Yang, Tom Griffiths, Jonathan D Cohen, and Sarah-Jane Leslie. 2026. Causal head gating: A framework for interpreting roles of attention heads in transformers. *Advances in Neural Information Processing Systems*, 38:115609–115631.

Vivi Nastase, Giuseppe Samo, Chunyang Jiang, and Paola Merlo. 2024. Exploring syntactic information in sentence embeddings through multilingual subject-verb agreement. In *Proceedings of the 10th Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 631–643.

Lucas Prieto, Edward Stevinson, Melih Barsbey, Tolga Birdal, and Pedro AM Mediano. 2026. From data statistics to feature geometry: How correlations shape superposition. *arXiv preprint arXiv:2603.09972*.

Abhilasha Ravichander, Yonatan Belinkov, and Eduard Hovy. 2021. Probing the probing paradigm: Does probing accuracy entail task relevance? In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3363–3377.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2020. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Qwen Team. 2026. Qwen3.5: Accelerating productivity with native multimodal agents.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4593–4601.

Elena Voita and Ivan Titov. 2020. Information-theoretic probing with minimum description length. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 183–196.

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.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. 2024. Do llamas work in english? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15366–15394.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

A Extended Discussion and Future Work

Due to strict space constraints, we provide an extended discussion of our findings, future research directions, and comprehensive per-language results in this appendix. The following sections offer full transparency into the representational patterns, methodological details, and typological distributions that were necessarily compressed in the main text.

A.1 Two Representational Routes, Not a Uniform Mechanism

The central implication of our findings is that autoregressive LLMs exploit at least two distinct representational routes to resolve agreement. The first route is *locally encoded processing*, best exemplified by Gender. In this route, the verb's hidden state decodes the feature with negligible layer alignment depth and minimal benefit from direct subject access, remaining immune to structural attractors. This suggests that the model does not execute a deep, subject-referential processing operation; instead, it relies entirely on local morphological cues inherent in the verb form itself. Number predominantly follows this local route but occupies an intermediate regime, showing slight vulnerability to interference. The second route is *subject-referential encoding*, most clearly observed for Person. Here, the subject remains the primary feature source, and the verb's feature representation emerges with greater layer alignment depth. Crucially, Person encoding is strengthened by direct causal access to the subject and is severely disrupted by incongruent attractors. This profile strongly resembles a genuine, dynamic information-transfer dependency. This distinction explains why high behavioral accuracy (Section 3.2) is an insufficient metric for syntactic understanding: the exact same correct output can be generated either by representing a structural relation or by leveraging a morphological heuristic. Our targeted diagnostics successfully separate these underlying routes.

A.2 Revisiting AGREE as a Diagnostic Lens

Our approach purposefully steps away from claiming that neural networks literally implement formal syntactic operations. Rather than treating AGREE as a hypothesis to be proven, we repurposed its core components—feature source, directionality, and structural intervention—as a conceptual blueprint for probing information flow in transformers. This shift in perspective is highly valuable for interpretability research. As our ϕ-axis alignment results (Section 5.4) show, the relationship between structural access (directionality) and representational geometry is a continuous, measurable trade-off. Conditions that rely on locally encoded processing share a clean geometric axis, while those that demand direct subject access require complex geometric re-encoding. Syntactic theory provides the precise edge cases (like VS orders and specific attractors) necessary to reveal these latent geometric mechanics in neural language models.

A.3 Future Work

Our findings open several promising avenues for future research. First, while layer alignment depths and geometric alignments provide strong representational clues about information flow, establishing true causality will require targeted intervention techniques, such as activation patching or attention-head ablation. Second, the natural overlap between subject features and verb morphology presents an ongoing challenge for interpreting grammatical sentences. Future dataset designs should introduce counterfactual stimuli or lemma-controlled minimal pairs to explicitly decouple the subject's ϕ-features from the verb's surface morphology. Finally, while MultiBLiMP's attractor metadata allowed us to observe intervention effects, direct manipulation of syntactic c-command versus linear distance is needed to comprehensively verify structure sensitivity.

B Extraction and Probing Methodology

This section provides the formal pseudocode for our primary representational diagnostics introduced in Section 4.

Algorithm 1 describes how we measure the decodability of a specific ϕ-feature from both the subject's own hidden state (Subject-side probe) and the target verb's hidden state (Verb-side probe). We use 5-fold cross-validation with L2-regularized logistic regression, scaling the hidden states before fitting, and applying balanced class weights. Note that the verb-side Grammaticality Probe (Section 4.1) utilizes this exact same extraction and cross-validation procedure, with the only difference being the target labels ($\{$GRAMMATICAL, UNGRAMMATICAL$\}$ instead of specific ϕ-feature values).

Algorithm 2 details the computation of ϕ-axis alignment. This metric quantifies the geometric similarity between the representational directions used to encode the feature contrast at the subject position versus the verb position. While the base logic operates on binary feature classes (e.g., singular vs. plural), for features with more than two classes (e.g., 1st, 2nd, and 3rd person), we compute the pairwise cosine similarity across all possible class combinations and report the mean alignment to ensure a stable, single-axis evaluation.

Algorithm 1 Cross-Position ϕ-Feature Probing

Require: Subject hidden states $\mathbf{H}_{\text{subj}} \in \mathbb{R}^{N \times L \times d}$
Require: Verb hidden states $\mathbf{H}_{\text{verb}} \in \mathbb{R}^{N \times L \times d}$
Require: ϕ-feature labels $\mathbf{y} \in \{c_1, \ldots, c_k\}^N$
Ensure: Subject-side and Verb-side accuracy profiles over layers $\ell \in \{1, \ldots, L\}$

```
for ℓ = 1, ..., L do
    X^(ℓ)_subj ← H_subj[:, ℓ, :]
    X^(ℓ)_verb ← H_verb[:, ℓ, :]
    Initialize classifiers C^(ℓ)_subj, C^(ℓ)_verb   ▷ L2-reg. LogReg with StandardScaler, class_weight=balanced
    Acc_subj,ℓ ← CrossValScore(C^(ℓ)_subj, X^(ℓ)_subj, y, folds = 5)
    Acc_verb,ℓ ← CrossValScore(C^(ℓ)_verb, X^(ℓ)_verb, y, folds = 5)
end for
return {Acc_subj,ℓ}_{ℓ=1}^L, {Acc_verb,ℓ}_{ℓ=1}^L
```

C Linear Mixed-Effects Model (LMM) Analysis

To verify whether the geometric trade-off between ϕ-axis alignment and the directionality effect ($\Delta_{\text{SV-VS}}$) differs significantly across features while properly accounting for the non-independence of our samples, we conducted a Linear Mixed-Effects Model (LMM) analysis. We modeled Language as a grouping variable and included Model as a random effect to control for clustering variance. Gender serves as the baseline feature category.

As detailed in Table 5, the negative main effect

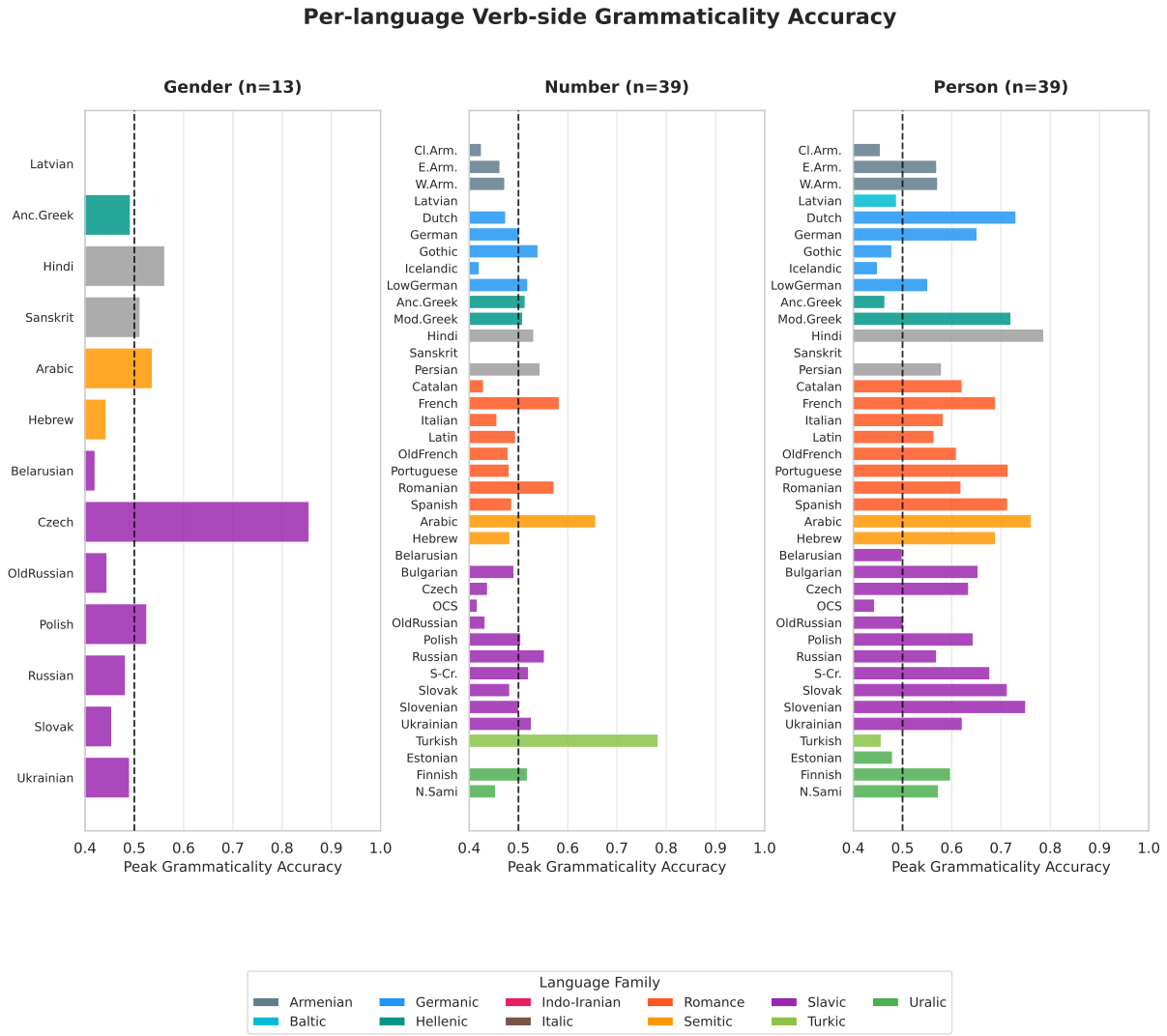


Figure 7: Per-language verb-side grammaticality probe accuracy (averaged across 7 models). The dashed line represents chance performance (0.50). For most languages, Gender and Number grammatical contrasts are not linearly recoverable at the target verb token, highlighting the distinction between specific feature-value encoding and a generalized grammaticality signal.

of ϕ-axis alignment ($p = 0.001$) confirms the universal trade-off: higher geometric alignment generally reduces the need for direct subject access. Crucially, the interaction term for Person is significant ($p < 0.001$). This confirms our visual analysis in Figure 5: the slope of this representational trade-off is substantially different for Person, reflecting a distinct, subject-dependent mechanism rather than locally encoded processing.

C.1 Cross-Linguistic Consistency of Directionality and Intervention

To ensure our findings are not artifacts of averaging, we disaggregate the results. As Figure 8 shows, the positive directionality effect ($\Delta_{\text{SV-VS}}$) for Person is remarkably consistent: across the *21 applicable languages* spanning diverse families, nearly every language exhibits a positive Δ. This confirms that the verb's ability to decode Person is causally boosted by the online presence of the subject. In contrast, Gender and Number show distributions tightly centered around zero. This distinction is corroborated by the attractor drop analysis (Panel B of Figure 10 and Table 4). The disruption caused by incongruent attractors is pronounced for Person across all scales. Notably, scaling from 1B to 27B parameters does not override this structural vulnerability, confirming it is an inherent property of the subject-referential route. Gender and Number remain largely immune to intervention effects.

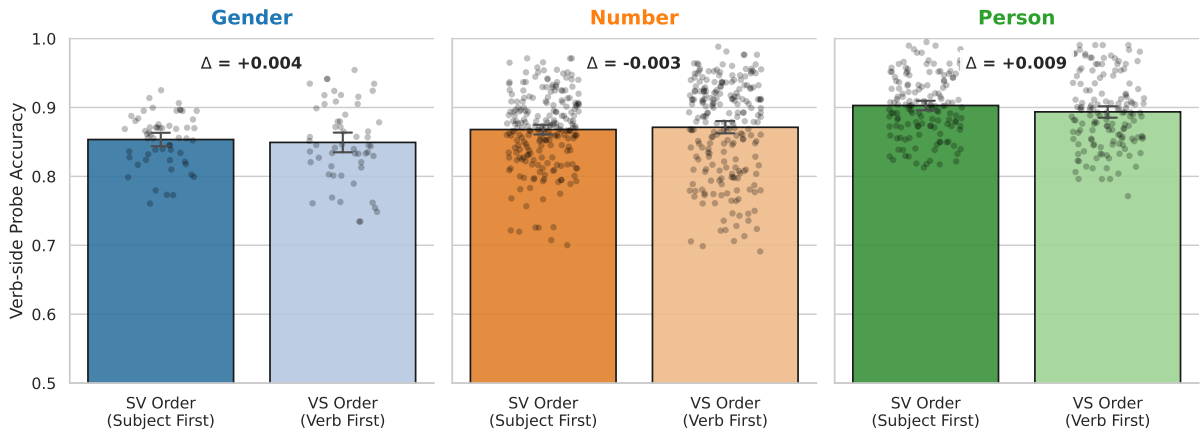


Figure 8: Absolute verb-side probing accuracy in Subject-Verb (SV) versus Verb-Subject (VS) word orders. Darker bars indicate SV order, while lighter bars indicate VS order. Individual black points represent the raw scores for every available model–language pairing in that condition. While Gender and Number maintain nearly identical high accuracy regardless of the subject's position, Person accuracy drops when the subject is not causally accessible (VS order).

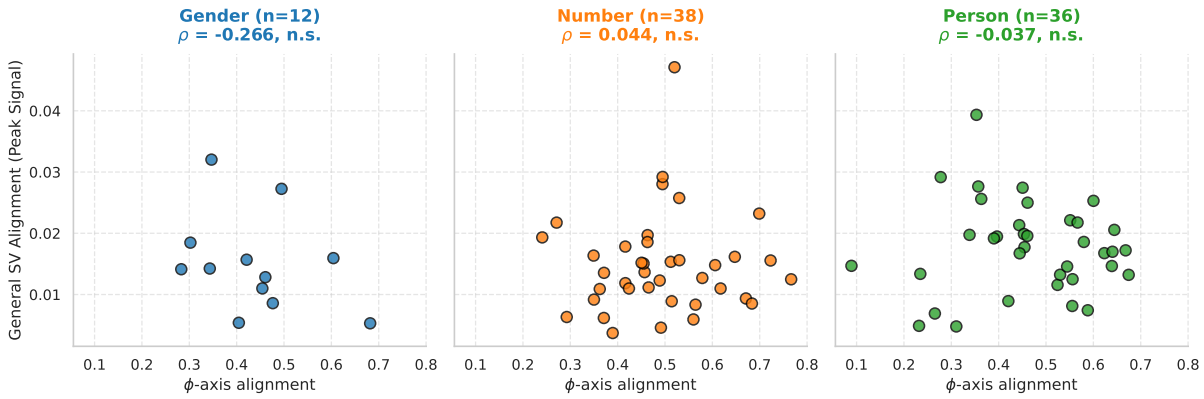


Figure 9: Scatter plot of general SV token alignment versus feature-specific ϕ-axis alignment. The lack of any significant correlation visually confirms that geometric feature alignment operates independently of generic token-level similarity.

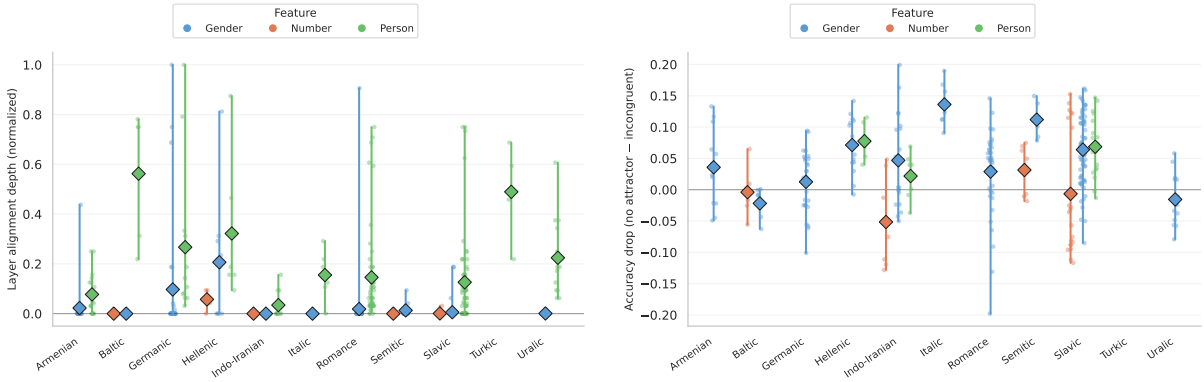
Model	Gender	Number	Person
Llama-3.2-1B	−0.014	0.034	0.039
Qwen3.5-2B	−0.002	0.035	0.092
Llama-3.2-3B	−0.001	0.024	0.050
Qwen3.5-4B	−0.010	0.038	0.066
Llama-3.1-8B	−0.011	0.035	0.064
Qwen3.5-9B	−0.018	0.039	0.079
Qwen3.5-27B	−0.017	0.047	0.077
Mean Drop	**−0.011**	**0.036**	**0.067**

Table 4: Absolute drop in verb-side accuracy caused by an incongruent attractor (No Attractor − Incongruent). Person agreement suffers substantial disruption across all scales, whereas Gender is unaffected.

C.2 Layer alignment depth and Grammaticality Baselines

Table 10 and Figure 10 present the dense typological metrics. The comprehensive data in Table 10 show that for Gender and Number, verb-side peak probe accuracy consistently exceeds subject-side peak accuracy (Verb. > Subj.) across almost all language families. As visualized in Panel A of Figure 10, this locally encoded processing is paired with shallow layer alignment depth. Person, conversely, requires large layer alignment depth in nearly all languages.

Finally, Figure 7 disaggregates the verb-side grammaticality probe across all evaluated languages. Despite the models' high behavioral accuracy (Table 1), the linear separability of abstract grammaticality at the single target verb token re-



(a) Layer alignment depth by language family.

(b) Incongruent attractor drop by language family.

Figure 10: Representational signatures by language family (seven models). **(a)** Mean layer alignment depth—the normalized first layer at which verb-side probe accuracy reaches within 0.02 of the subject-side peak. **(b)** Mean drop in verb-side probe accuracy under incongruent vs. no-attractor conditions. Diamonds aggregate over models and languages within each family; points show individual language–model cells. Person shows greater alignment depth and larger attractor disruption than Gender or Number across families (including Uralic and Slavic), whereas Gender and Number remain near-zero in depth and comparatively stable under intervention.

Predictor	**Coef.**	**Std. Err.**	z	$P > \|z\|$	**[0.025**	**0.975]**
Intercept	0.036	0.015	2.459	0.014	0.007	0.065
Feature [Number]	−0.028	0.016	−1.740	0.082	−0.059	0.004
Feature [Person]	−0.051	0.013	−3.851	< 0.001	−0.077	−0.025
ϕ-axis alignment	−0.083	0.024	−3.399	0.001	−0.131	−0.035
ϕ-axis × Feature [Number]	0.061	0.029	2.078	0.038	0.003	0.118
ϕ-axis × Feature [Person]	0.133	0.024	5.540	< 0.001	0.086	0.181
Group Var (Language)	0.002	0.011				
Model Var (Random Effect)	< 0.001					

Table 5: Linear Mixed-Effects Model (LMM) predicting the directionality effect ($\Delta_{\text{SV-VS}}$) as a function of ϕ-axis alignment, linguistic feature, and their interaction ($n = 427$). Language is treated as a clustering group and Model as a random effect to adjust standard errors. Gender is the reference category.

Diagnostic Split	Gender	Number	Person
General Encoding (All)	12	38	36
Word Order: SV	12	37	36
Word Order: VS	8	33	21
No Attractor	10	37	35
Congruent Attractor	6	27	2
Incongruent Attractor	8	32	6

Table 6: Effective language coverage for each diagnostic split. The numbers represent the count of unique languages evaluated in each specific analysis condition. Note that while broad encoding applies to nearly all languages, structural diagnostics (VS order, Attractors) are typologically bounded.

mains near chance (0.50) for Gender and Number across most typologies. This consistency validates our core methodological shift: resolving fine-grained agreement mechanisms requires probing for specific feature values, rather than relying on a generic signal of structural correctness.

Model	**Gender**	**Number**	**Person**
Llama-3.2-1B	0.002	0.006	0.020
Qwen3.5-2B	0.011	−0.007	0.018
Llama-3.2-3B	0.007	0.014	0.025
Qwen3.5-4B	0.005	−0.007	0.017
Llama-3.1-8B	0.002	0.003	0.032
Qwen3.5-9B	0.003	−0.006	0.021
Qwen3.5-27B	−0.003	0.000	0.022
Mean	**0.004**	**0.000**	**0.022**

Table 7: Detailed directionality effect ($\Delta_{\text{SV-VS}}$) across seven models. Values represent the difference in verb-side probing accuracy between subject–verb (SV) and verb–subject (VS) orders. Positive values indicate a representational benefit from direct causal access to the subject.

Algorithm 2 ϕ-Axis Alignment Computation

Require: Subject hidden states $\mathbf{H}_{\text{subj}} \in \mathbb{R}^{N \times L \times d}$
Require: Verb hidden states $\mathbf{H}_{\text{verb}} \in \mathbb{R}^{N \times L \times d}$
Require: ϕ-feature labels $\mathbf{y} \in \{c_1, \ldots, c_k\}^N$
Ensure: ϕ-axis alignment profile over layers $\ell \in \{1, \ldots, L\}$

for $\ell = 1, \ldots, L$ **do**
  $\mathbf{X}^{(\ell)}_{\text{subj}} \leftarrow \mathbf{H}_{\text{subj}}[:, \ell, :]$
  $\mathbf{X}^{(\ell)}_{\text{verb}} \leftarrow \mathbf{H}_{\text{verb}}[:, \ell, :]$
  Initialize Sims $\leftarrow [\,]$
  Pairs $\leftarrow \{(c_i, c_j) \mid c_i, c_j \in \text{unique}(\mathbf{y}), i < j\}$
  for $(c_A, c_B) \in$ Pairs **do**
    ▷ Compute centroids for each class
    $\mathbf{c}_{\text{subj},A} \leftarrow \text{mean}(\mathbf{X}^{(\ell)}_{\text{subj}} \text{ where } \mathbf{y} = c_A)$
    $\mathbf{c}_{\text{subj},B} \leftarrow \text{mean}(\mathbf{X}^{(\ell)}_{\text{subj}} \text{ where } \mathbf{y} = c_B)$
    $\mathbf{c}_{\text{verb},A} \leftarrow \text{mean}(\mathbf{X}^{(\ell)}_{\text{verb}} \text{ where } \mathbf{y} = c_A)$
    $\mathbf{c}_{\text{verb},B} \leftarrow \text{mean}(\mathbf{X}^{(\ell)}_{\text{verb}} \text{ where } \mathbf{y} = c_B)$
    ▷ Compute feature-discrimination axes
    $\mathbf{d}^{(\ell)}_{\text{subj}} \leftarrow \mathbf{c}_{\text{subj},A} - \mathbf{c}_{\text{subj},B}$
    $\mathbf{d}^{(\ell)}_{\text{verb}} \leftarrow \mathbf{c}_{\text{verb},A} - \mathbf{c}_{\text{verb},B}$
    ▷ Compute cosine similarity
    sim $\leftarrow$ cosine_sim($\mathbf{d}^{(\ell)}_{\text{subj}}$, $\mathbf{d}^{(\ell)}_{\text{verb}}$)
    Sims.append(sim)
  end for
  $\text{Align}_\ell \leftarrow$ mean(Sims) ▷ Average across all pairs
end for
return $\{\text{Align}_\ell\}_{\ell=1}^{L}$

D Ruling Out a Positional Confound in VS-Order Analysis

The directionality diagnostic (Section 4.2) compares verb-side probe accuracy in subject–verb (SV) versus verb–subject (VS) order, attributing a positive $\Delta_{\text{SV}-\text{VS}}$ to the representational benefit of causal access to the preceding subject. A potential alternative explanation is *positional*: in verb-initial constructions the target verb occupies the first sequential position, whereas in SV constructions it is a later token. If sentence-initial tokens receive systematically different representations than later-occurring tokens—independent of syntactic function—a verb-side accuracy drop in VS order could reflect a positional artifact rather than evidence for subject-referential encoding.

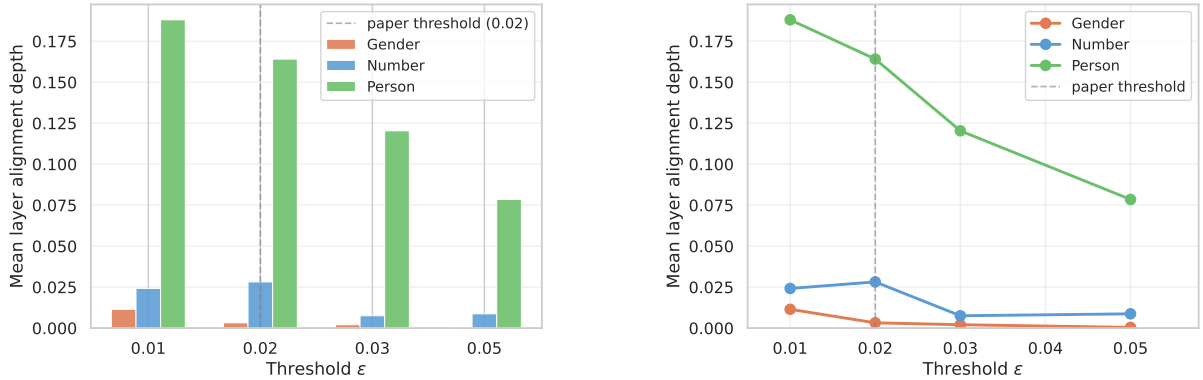
We address this concern using a within-construction control already present in our data. All three ϕ-features (Gender, Number, and Person) are evaluated on overlapping sets of VS-order sentences drawn from the same languages and models. The target verb token and its sequential position are therefore *held constant* across features within each VS-order sentence. If verb-initial position were the source of the accuracy drop, the effect should be equally present for all three features.

The data directly falsify this prediction. As shown in Figure 2 and Table 7, Gender and Number exhibit $\Delta_{\text{SV}-\text{VS}} \approx 0$ (means: 0.004 and 0.000 respectively), whereas Person shows a consistent positive Δ (mean: 0.022). This dissociation holds across all seven models and all applicable languages. Because the only factor that differs between features within the same VS sentence is the feature type itself—not the verb's sequential position—the Person-specific accuracy drop cannot be attributed to a positional confound. Gender and Number serve as matched within-study baselines, confirming that the Person deficit in VS order reflects a feature-specific mechanism that depends on the causal presence of a prior structural subject.

E Sensitivity Analysis: Layer Alignment Depth Threshold

The definition of layer alignment depth (Section 4.4) requires setting a tolerance ϵ, such that $\ell_{\text{align}} = \min\{\ell : a_{\text{verb}}(\ell) \geq a^*_{\text{subj}} - \epsilon\}$. The main text uses $\epsilon = 0.02$. To verify that our findings are not artefacts of this specific choice, we recompute mean layer alignment depth across all model–language conditions for $\epsilon \in \{0.01, 0.02, 0.03, 0.05\}$.

As shown in Table 8 and Figure 11, the qualitative ordering Gender $<$ Number $<$ Person holds across every threshold value tested. As ϵ increases (tolerance becomes more lenient), all three features shift toward shallower depths, but the gap between Person and Gender/Number is preserved throughout. Notably, Number remains consistently above Gender at every threshold, confirming its intermediate position. These results demonstrate that the locally encoded versus subject-referential distinction is robust to the precise threshold choice and is not an artefact of the $\epsilon = 0.02$ setting.



(a) Layer alignment depth per threshold

(b) Trend across thresholds

Figure 11: Threshold sensitivity of layer alignment depth. The Gender–Number–Person ordering is stable across $\epsilon \in \{0.01, 0.02, 0.03, 0.05\}$; the dashed line marks $\epsilon = 0.02$ used in the paper.

Threshold ϵ	Gender	Number	Person
0.01	0.011	0.024	0.188
0.02*	0.003	0.028	0.164
0.03	0.002	0.008	0.120
0.05	0.000	0.009	0.078

Table 8: Mean layer alignment depth across all model–language conditions ($n = 602$) at four threshold values. * denotes the value used in the main text. The Gender $<$ Number $<$ Person ordering holds at every threshold. Values are averaged over conditions that reached the threshold; stricter thresholds exclude more NaN conditions, which can produce minor non-monotonicity in Number.

F Typological Constraints and Methodological Rigor

A core contribution of our work is adapting structural properties of agreement—such as directionality and intervention—into representational diagnostics. However, applying these diagnostics across 39 languages requires careful attention to typological availability. Not all languages permit the syntactic configurations required for every diagnostic test in the dataset.

Table 6 details the effective language coverage for each diagnostic analysis presented in the main text. Values indicate the number of unique languages that contain a sufficient number of minimal pairs (at least 15 per feature class) to train and evaluate linear probes in each condition.

While broad cross-position feature decoding (comparing subject and verb hidden states) can be evaluated across almost all sampled languages (12 for gender, 38 for number, 36 for person), targeted mechanistic tests are constrained. For instance, testing the effect of word order (SV vs. VS directionality) for Person agreement is restricted to a specific subset of 21 languages in our dataset, predominantly from the Romance, Slavic, Germanic, and classical families. Similarly, testing the interference of incongruent attractors is highly restricted by dataset availability for Gender and Person.

The multilingual nature of our study highlights a critical methodological point for interpretability research: diagnostic availability is inherently bound by linguistic typology. Rather than viewing this as a mere dataset limitation, we argue it must be an explicit parameter in multilingual evaluation. Claims about universal neural mechanisms cannot rest on syntactic constructions that only a fraction of the world's languages support.

We explicitly structure our claims in the main text around this availability. We use broad diagnostics to establish general patterns of feature representation across the multilingual spectrum, while treating directionality and attractor tests as targeted, opportunistic checks valid only within the specific typological subsets that support them. By separating broad encoding evidence from targeted, coverage-restricted mechanistic tests, our analysis avoids overclaiming universal rules from skewed subsets.

G Structural Diagnostics: Disaggregating SV vs. VS Word Order

In the main text (Section 5.2), we introduced the directionality effect ($\Delta_{\text{SV-VS}}$) to quantify the representational benefit of causal subject access. To provide full transparency, we disaggregate this metric into the absolute verb-side probe accuracies for both Subject-Verb (SV) and Verb-Subject (VS) or-

Feature	Spearman ρ	p-value	Sig.
Gender (n=12)	-0.266	0.404	n.s.
Number (n=38)	0.044	0.792	n.s.
Person (n=36)	-0.037	0.832	n.s.

Table 9: Correlation between general SV alignment and feature-specific ϕ-axis alignment. All correlations are non-significant (n.s.).

ders.

As visualized in Figure 8, all three ϕ-features exhibit high decoding accuracy (> 0.85) when the subject precedes the verb (SV Order). The critical divergence emerges in the VS order, where the verb representation must be computed before the nominal source is observed.

For Gender and Number, the verb-side accuracy remains nearly identical across both structural conditions ($\Delta \approx 0$). This visual symmetry confirms our assertion of *locally encoded processing*: because the local verbal morphology is sufficiently unambiguous for these features, the causal visibility of the structural subject provides no additional representational gain.

Conversely, Person agreement demonstrates a distinct asymmetry. When the subject is relegated to the post-verbal position (VS Order), verb-side accuracy drops substantially across the evaluated model–language pairs. This drop isolates the exact *subject-referential* nature of Person processing: its successful representation at the verb token heavily depends on dynamic, forward-looking information transfer from a prior structural goal.

H Cross-Linguistic Consistency: Analysis by Language Family

A persistent concern in multilingual NLP evaluation is typological imbalance: because comprehensive datasets like MultiBLiMP heavily sample from Indo-European languages, aggregate means can mask idiosyncratic, family-specific behaviors. To verify that our proposed representational distinction (locally encoded vs. subject-referential) is not a mere artifact of averaging across skewed language distributions, we disaggregate our two primary physical signatures—Layer Alignment Depth and Attractor Drop—by language family.

As visualized in Figure 10, the feature-specific divergence remains remarkably consistent across distinct language families.

Temporal Consistency. In Panel A, Gender and Number show near-zero layer alignment depth across Slavic, Romance, and Semitic families: verb-side ϕ reaches subject-peak decodability from shallow layers. Person requires substantially greater depth in nearly every family. Notably, even in highly inflected Uralic (e.g., Finnish) and Slavic (e.g., Russian) languages, Person exhibits a markedly larger layer alignment depth than Gender or Number.

Intervention Consistency. Panel B mirrors this consistency under structural stress. The disruption caused by incongruent attractors is highly pronounced for Person across all applicable families. Crucially, the invariance of Gender and Number holds typologically: despite the immense morphological variation across these families, incongruent nominal interveners consistently fail to disrupt their decoding at the verb.

This disaggregated analysis provides strong evidence that our identified representational profiles are stable cross-family patterns employed by LLMs, rather than spurious statistical artifacts of Indo-European dominance.

I Disentangling ϕ-Axis Alignment from General Token Similarity

A potential confounding hypothesis in our geometric analysis (Section 5.4) is that high ϕ-axis alignment might merely reflect a general representational convergence between the subject and verb tokens in deeper layers. If the representations of these two tokens simply blend together contextually, their feature-discrimination axes would appear artificially parallel, masking true feature-specific geometric mapping.

To conclusively rule out this confound, we computed a baseline Subject-Verb (SV) alignment: the probe-independent cosine similarity between the full, unmodified hidden states of the subject and target verb tokens. If ϕ-axis alignment is merely a byproduct of tokens blending together, it should correlate strongly with this general SV alignment.

As visualized in Figure 9 and quantified in Table 9, there is no significant correlation between general SV alignment and our feature-specific ϕ-axis alignment across any of the three linguistic features. The points are randomly dispersed, confirming that high ϕ-axis alignment is not an artifact of generic token-level proximity. Instead, it rigorously isolates the specific geometric axis the model

uses to discriminate the ϕ-feature, validating our interpretation of locally encoded processing versus geometric re-encoding.

J Full Language-Level Results

Table 10 reports subject-side peak probe accuracy (Subj.), verb-side peak probe accuracy (Verb.), and layer alignment depth (Align.), averaged across seven models. Missing values (-) indicate that the respective feature is not syntactically marked on the verb in that specific language according to our typological extraction constraints.

Table 10: Comprehensive per-language structural agreement profiles (averaged across 7 models). For each evaluated feature (Gender, Number, Person), we report Subject-side accuracy, Verb-side accuracy, and Layer align. depth. Missing values (-) indicate the feature is not syntactically marked on the verb in that language. A dash in Align. indicates that verb-side accuracy never reached within 0.02 of the subject-side peak.

Family	**Language**	**Gender**			**Number**			**Person**		
		Subj.	**Verb.**	**Align.**	**Subj.**	**Verb.**	**Align.**	**Subj.**	**Verb.**	**Align.**
Armenian	Cl.Arm.	-	-	-	0.75	0.84	0.07	0.84	0.84	0.13
	E.Arm.	-	-	-	0.65	0.87	0.00	0.85	0.89	0.09
	W.Arm.	-	-	-	0.66	0.86	0.00	0.88	0.92	0.02
Baltic	Latvian	0.68	0.83	0.00	0.74	0.83	0.00	0.94	0.93	0.56
Germanic	Dutch	-	-	-	0.80	0.93	0.00	0.91	0.91	0.12
	German	-	-	-	0.83	0.94	0.00	0.92	0.90	0.45
	Gothic	-	-	-	0.82	0.83	0.25	0.97	0.85	-
	Icelandic	-	-	-	0.72	0.83	0.00	0.93	0.90	0.33
	LowGerman	-	-	-	0.91	0.89	0.37	0.89	0.85	-
Hellenic	Anc.Greek	0.82	0.79	0.06	0.84	0.86	0.32	0.86	0.86	0.32
	Mod.Greek	-	-	-	0.86	0.85	0.00	0.97	0.94	-
Indo-Iranian	Hindi	0.74	0.92	0.00	0.68	0.87	0.00	-	-	-
	Persian	-	-	-	0.73	0.91	0.00	0.85	0.90	0.00
	Sanskrit	0.67	0.86	0.00	0.66	0.79	0.00	0.86	0.84	0.11
Italic	Latin	-	-	-	0.76	0.91	0.00	0.91	0.93	0.16
Romance	Catalan	-	-	-	0.80	0.85	0.00	0.94	0.96	0.10
	French	-	-	-	0.84	0.89	0.00	0.94	0.97	0.06
	Italian	-	-	-	0.82	0.85	0.13	0.88	0.89	0.10
	OldFrench	-	-	-	0.78	0.88	0.00	0.89	0.89	0.27
	Portuguese	-	-	-	0.88	0.91	0.00	0.82	0.84	0.02
	Romanian	-	-	-	0.75	0.91	0.00	0.94	0.94	0.22
	Spanish	-	-	-	0.83	0.89	0.00	0.88	0.88	0.24
Semitic	Arabic	0.79	0.90	0.00	0.96	0.97	0.03	-	-	-
	Hebrew	0.66	0.82	0.00	0.69	0.85	0.00	-	-	-
Slavic	Belarusian	0.72	0.85	0.00	0.73	0.87	0.00	0.88	0.87	0.16
	Bulgarian	-	-	-	0.79	0.89	0.00	0.93	0.96	0.08
	Czech	-	-	-	0.75	0.92	0.00	0.91	0.94	0.01
	OCS	-	-	-	0.68	0.72	0.04	0.93	0.83	-
	OldRussian	0.68	0.80	0.00	0.76	0.82	0.03	0.90	0.89	0.30
	Polish	0.66	0.87	0.00	0.76	0.91	0.00	0.95	0.95	0.08
	Russian	0.67	0.88	0.00	0.73	0.93	0.00	0.90	0.92	0.10
	S-Cr.	-	-	-	0.74	0.91	0.00	0.95	0.92	0.57
	Slovak	0.77	0.89	0.00	0.75	0.89	0.00	0.98	0.98	0.08
	Slovenian	-	-	-	0.76	0.88	0.00	0.91	0.97	0.00
	Ukrainian	0.61	0.88	0.00	0.75	0.90	0.00	0.92	0.92	0.10
Turkic	Turkish	-	-	-	-	-	-	0.93	0.91	0.49
Uralic	Estonian	-	-	-	0.72	0.83	0.00	0.89	0.90	0.20
	Finnish	-	-	-	0.77	0.90	0.00	0.95	0.95	0.25
	N.Sami	-	-	-	0.71	0.81	0.00	0.97	0.91	-