Do Language Models Understand Implicit Logical Meaning? A Case Study of Scope Ambiguity Resolution in Context

Anonymous ACL submission

Abstract

Scope ambiguity arises when a sentence admits multiple interpretations depending on how logical operators (e.g., negation and indefinites) interact. In contrast to prior work (Kamath et al., 2024), which adopts an *ambiguity-first, evidence-following* diagnostic and focuses primarily on surface-level behavioral sensitivity without directly examining internal mechanisms, we introduce an *evidence-first, ambiguity-following* diagnostic that allows the hidden representations of a scope-ambiguous sentence to be adaptively shaped by preceding contextual evidence. To support this analysis, we present SCOPEx, an extension of the dataset of (Kamath et al., 2024), constructed by semi-automatically generating preamble context sentences corresponding to different scope readings and by introducing passivized sentences as syntactically less ambiguous controls. Using SCOPEx, we find evidence that transformer language models (LMs) (i) exhibit systematic sensitivity to scope ambiguity as a function of preceding contextual evidence, and (ii) encode scope interpretations in their internal representations in ways that support reliable discrimination between inverse and surface scope readings.

1 Introduction

Scope ambiguity is a pervasive phenomenon in natural language, whereby a sentence admits multiple interpretations depending on how scope-bearing items(SBIs)—such as indefinites (*a/an*), negation (*not*), quantifiers (*some, every*), numerals, and others—interact (Barwise and Cooper, 1981; May, 1985). As illustrated in Figure 2 and Table 6, consider the sentence *"I didn't feed a dog"* with two interpretations:

- *Specific*: "there exists a particular dog that was not fed."
  (inverse reading: $\exists x(DOG(x) \wedge \neg FEED(I, x))$)

Evidence-first diagnosis of scope in LLMs

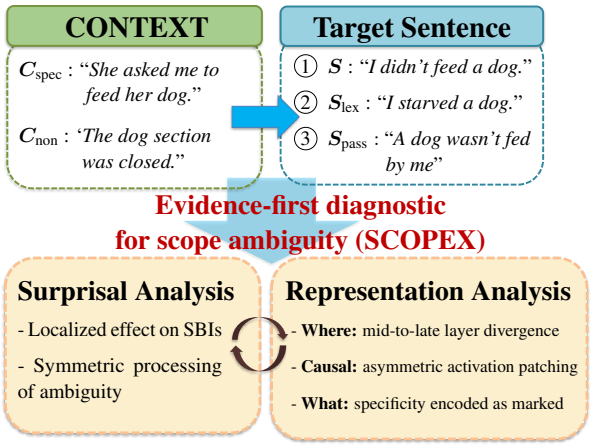

Figure 1: a high-level overview of our evidence-first diagnostic approach, summarizing the SCOPEx design and the roles of Experiments 1 and 2.

- *Non-specific*: ''no dogs were fed."
  (surface reading: $\neg\exists x(DOG(x) \wedge FEED(I, x))$)

Scope ambiguity is resolved when additional context is provided. In this example, an **inverse** reading is favored in a context such as *"she asked me to feed her dog,"* whereas a **surface** reading is favored in a context such as *"the dog section was closed."* In this paper, we examine **how modern transformer-based LMs resolve scope ambiguity in context**.

Regarding this question, (Kamath et al., 2024) explore an *ambiguity-first, evidence-following diagnostic*. In this setup, a scope-ambiguous or scope-unambiguous *preamble* sentence is presented first, and the LM's sensitivity is assessed by comparing the probabilities assigned to alternative follow-up *continuation* sentences. However, given the *unidirectional* attention architecture of most autoregressive language models, this *ambiguity-first* diagnostic setup conditions on the original scope-ambiguous preamble sentence and evaluates only the plausibility of subsequent continuations. As

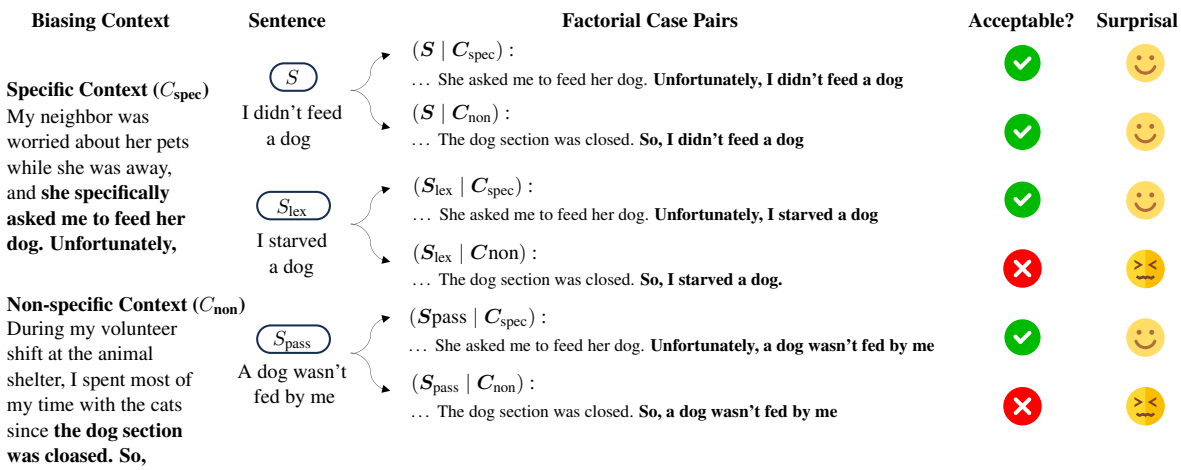

Figure 2: Overview of the SCOPEx dataset and Experiment 1. The figure illustrates the 2×3 factorial design, crossing contextual bias (specific vs. non-specific) with sentence type variations (ambiguous, lexical-semantically unambiguous, syntactically less ambiguous conditions).

a result, the sentence-level probabilities and internal hidden states associated with the target scope-ambiguous sentence are already fixed at the time ambiguity sensitivity is assessed, which limits the ability of this setup to directly analyze how scope ambiguity is resolved *internally* within language models.

To investigate scope ambiguity from both behavioral and internal perspectives, we adopt an **evidence-first, ambiguity-following diagnostic**, which we refer to as *evidence-first continuation*. In this setup, a preamble evidence sentence is presented first, followed by a scope ambiguous continuation sentence. Because the hidden states and sentence-level probabilities associated with the scope-ambiguous sentence are *contextually and adaptively determined as a function of the preceding evidence*, this evidence-first continuation setup allows us to more directly examine how LMs interpret and resolve scope ambiguity under preceding evidential context (See Table 8 in Appendix).

To support the evidence-first continuation setting, we construct a novel dataset, termed SCOPEx (**SCOP**e–**E**xpression Cross-factor Evaluation), by extending the dataset of (Kamath et al., 2024) in two important ways: First, we introduce *evidence sentences* that precede the target sentence. Second, in addition to the original scope-ambiguous and lexically unambiguous sentence types, we introduce a third sentence type that is less scope-ambiguous due to syntactic structure, namely *passivized sentences*, such as *"A dog wasn't fed by me."* (See Section 3.1). Building on this SCOPEx, we conduct a series of analyses to examine how LMs resolve scope ambiguity, substantially extending the analysis of (Kamath et al., 2024). We provide convergent evidence from surprisal analysis, representational similarity measures (centered kernel alignment(CKA) and Procrustes), activation patching, and probing.

- **Surprisal-based analysis.** We measure how model behavior varies as a function of contextual bias toward different readings. We find that LMs exhibit gradient sensitivity to scope-based semantic *congruence*. Sentences receive significantly lower surprisal in contexts that license their intended reading (Wilcoxon $p < .01$), with SBIs showing disproportionately larger effects than non-SBIs.

- **Representational similarity analysis.** We examine which layers contribute most to scope ambiguity resolution. We identify mid layers (L8–L16) and late layers (L30–L32) in LLaMA3-8B as critical for scope resolution.

- **Activation patching for steering reading type.** We perform activation patching to test whether and how model behavior on scope-ambiguous sentences can be steered by intervening on contextual representations, replacing them with representations associated with alternative readings. We find that interpretations corresponding to inverse-scope readings are more resistant to change, whereas surface-scope readings exhibit greater representational heterogeneity and thus constitute

the unmarked default.

- **Probing analysis for scopal specificity.** We train probing classifiers to distinguish between specific and non-specific interpretations of an NP. The results show that LMs encode the semantic property of *scopal specificity*: probing classifiers reliably distinguish inverse scope (specific) from surface scope (non-specific) SBIs, with performance varying systematically across sentence types that provide different degrees of disambiguation (syntactic > lexical > contextual).

Taken together, our findings indicate that LMs construct internally structured representations that differentiate scope readings, in addition to exhibiting behavioral sensitivity to ambiguity.

2 Previous Research.

LMs and Ambiguity Recent work has examined whether large language models (LLMs) exhibit similar sensitivity to ambiguity. Liu et al. (2023) find that even state-of-the-art models struggle to recognize or enumerate multiple readings across diverse ambiguity types. Stengel-Eskin et al. (2024) likewise show that pretrained models fail to capture the distribution of possible interpretations unless ambiguity is explicitly presented in the input. In contrast, Kamath et al. (2024) report that LLMs' forced-choice preferences over scope readings often align with human judgments, though the underlying mechanisms remain unclear. These mixed results suggest that LLMs may partially approximate human scope preferences but do not reliably manage ambiguity.

Geography of linguistic knowledge within LMs Prior work has investigated where linguistic information is localized within transformer-based LMs. Early analyses showed that transformers exhibit layer-wise specialization reminiscent of a classical NLP pipeline, with surface-level features encoded in lower layers and more abstract representations emerging in higher layers (Peters et al., 2018; Tenney et al., 2019; Liu et al., 2019; Jawahar et al., 2019).

3 Materials and Methods

3.1 Dataset Construction

SCOPEx Dataset To support the evidence-first continuation diagnostic for scope ambiguity, we construct **SCOPEx**, a controlled benchmark designed to test whether LMs encode linguistic factors known to influence scope interpretation. SCOPEx extends the scope-ambiguity dataset of Kamath et al. (2024) by introducing **preceding contextual manipulation** and a broader set of **sentence-level controls**. Specifically, the dataset adopts a 2×3 **factorial design** that crosses two discourse contexts with three sentence types, yielding six systematically related conditions ($N = 139$ items per condition; 834 total) (Figure 2, Table 6).

Notation and Sentence Types We use the notation $(S \mid C)$ to indicate that a target sentence S is processed following a preceding context C. Contexts are denoted as C_{spec} and C_{non}, which bias interpretation toward inverse-scope (specific) and surface-scope (non-specific) readings, respectively. Target sentences appear in three conditions: an **ambiguous** condition (S), in which both scope-bearing items (SBIs) occur in their canonical positions; a **lexically unambiguous** condition (S_{lex}), where one SBI is replaced by a non-SBI (e.g., *n't feed* → *starved*), eliminating ambiguity while preserving propositional content; and a **passivized** condition (S_{pass}), in which the second SBI (e.g., *a dog*) is promoted to subject position, introducing a syntactic cue that systematically biases interpretation without lexical substitution.

Contextual Manipulation and Interpretational Logic Within each sentence type, pairing with C_{spec} or C_{non} yields either a **congruent** condition—where contextual and sentence-level cues jointly support the same interpretation—or an **incongruent** condition—where they conflict. This design cleanly isolates **pragmatic** (contextual), **lexical-semantic**, and **syntactic** contributions to scope interpretation, enabling precise tests of how LMs integrate these sources of information under controlled variation. All contextual and passivized variants are generated using the OpenAI GPT-4o model (OpenAI, 2024) and are subsequently manually verified for semantic consistency.

3.2 Models

We evaluate **three transformer-based autoregressive LMs**: LLaMA-3 (Dubey et al., 2024) (8B, Base), Mistral (Jiang et al., 2023) (7B, v0.3), and Qwen2.5 (Qwen Team: Yang et al., 2024) (7B). Due to space limitations, we present main results using **LLaMA-3-8B**. Full results for all models appear in the Appendix as ablation studies.

4 Experiment 1. Surprisal-Based Disambiguation

A central question in ambiguity research is whether LMs internally distinguish between alternative interpretations of structurally ambiguous sentences. To address this, we adopt *surprisal* as an incremental processing measure. (Shannon, 1948; Hale, 2001; Levy, 2008; Smith and Levy, 2013).

4.1 Method

We perform two complementary surprisal-based analyses to assess processing costs at the sentence level and at the level of local spans.

Suppose that $C = c_1 \cdots c_m$ is a given preamble context and $S = s_1 \cdots s_n$ is a target sentence, where c_i denotes the i-th context token and s_j denotes the j-th target token. We define the *token-level surprisal* of the j-th token in the target sentence S as the negative log-likelihood of that token given the preceding context C:

$$\mathcal{G}_j(S \mid C) = -\log p\Big(s_j \mid C, s_1^{j-1}\Big), \qquad (1)$$

where s_1^{j-1} denotes the subsequence consisting of the first $j-1$ tokens of the target sentence S.

The *in-situ surprisal* of the target sentence S given the preceding context C is then defined as the average token-level surprisal over the sentence:

$$\mathcal{G}(S \mid C) = \frac{1}{n}\sum_{j=1}^{n} \mathcal{G}_j(S \mid C). \qquad (2)$$

Isolated surprisal for the target sentence S is defined as the non-conditional surprisal obtained by presenting S without any preceding context and averaging the negative log-likelihood over all tokens in the sentence. We denote this quantity as $\mathcal{G}(S)$.

Given a preceding context C, we define the *contextual facilitation effect* as the difference between the in-situ and isolated surprisal values:

$$\Delta\mathcal{G} = \mathcal{G}(S \mid C) - \mathcal{G}(S). \qquad (3)$$

Negative values of $\Delta\mathcal{G}$ indicate reduced processing difficulty when contextual information is available.

Token-level scope sensitivity To localize where interpretive differences arise, we compare token-level surprisal between paired evidence contexts C_{spec} and C_{non}, corresponding to inverse- and surface-scope readings, within the same item. For each token position j, we compute

$$\Delta_j = \mathcal{G}_j(S \mid C_{\text{spec}}) - \mathcal{G}_j(S \mid C_{\text{non}}) \qquad (4)$$

and use its absolute magnitude $|\Delta_j|$ as a measure of interpretive sensitivity.

Tokens are categorized using a predefined scope lexicon (e.g., quantifiers, numerals, negation) into *scopal* and *non-scopal* items using SPACY. We then compute the mean of $|\Delta_j|$ separately for each group. Larger values indicate greater sensitivity to scope interpretation at the corresponding token position.

Taken together, these analyses quantify both the global processing cost associated with contextual disambiguation and the local contribution of SBIs to interpretive divergence.

Mean Surprisal (bits)			
Case	**In-situ**	**Isolated**	**Diff.**
$(S \mid C_{\text{spec}})$	4.748	6.784	−2.035
$(S \mid C_{\text{non}})$	4.812	6.785	−1.973
$(S_{\text{lex}} \mid C_{\text{spec}})$	6.499	7.524	−1.026
$(S_{\text{lex}} \mid C_{\text{non}})$	**6.712**	**7.534**	−0.822
$(S_{\text{pass}} \mid C_{\text{spec}})$	5.603	6.933	−1.330
$(S_{\text{pass}} \mid C_{\text{non}})$	5.817	6.944	−1.127
Pairwise Statistical Comparisons			
Pair	**Mean$_\Delta$**	p	d_z
$(S \mid C_{\text{spec}})$–$(S \mid C_{\text{non}})$	−0.062	.256	−0.10
$(S_{\text{lex}} \mid C_{\text{spec}})$–$(S_{\text{lex}} \mid C_{\text{non}})$	−0.204	.001**	−0.29
$(S_{\text{pass}} \mid C_{\text{spec}})$–$(S_{\text{pass}} \mid C_{\text{non}})$	−0.203	.001**	−0.28

Table 1: Mean surprisal values and pairwise comparisons. Diff. = In-situ − Isolated (contextual facilitation). Mean$_\Delta$ = difference in contextual effects between cases. Paired Wilcoxon signed-rank tests. green = favorable (largest effects/significance), red = highest difficulty, orange = largest pair difference. $^{**}p < .01$.

Pair	Non-scope	Scope	Diff.	Ratio
$(S \mid C_{\text{spec}})$–$(S \mid C_{\text{non}})$	0.843	0.790	−0.053	0.94×
$(S_{\text{lex}} \mid C_{\text{spec}})$–$(S_{\text{lex}} \mid C_{\text{non}})$	1.162	**1.395**	**+0.233**	1.20×
$(S_{\text{pass}} \mid C_{\text{spec}})$–$(S_{\text{pass}} \mid C_{\text{non}})$	1.020	**1.399**	**+0.379**	1.37×

Table 2: Sensitivity of SBIs. Mean $|\Delta|$ (bits) for SBIs vs. non-SBIs. Ratio = Scope/Non-scope effect magnitude.

4.2 Results and Discussion

Context-driven interpretive preference Table 1 reports mean surprisal values for in-situ and isolated conditions of LLaMa3-8B. The ambiguous pair ($(S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$) exhibits minimal contextual asymmetry, consistent with both

readings remaining viable. By contrast, the unambiguous ($(S_{\text{lex}} \mid C_{\text{spec}})$, $(S_{\text{lex}} \mid C_{\text{non}})$) and the passive ($(S_{\text{pass}} \mid C_{\text{spec}})$, $(S_{\text{pass}} \mid C_{\text{non}})$) pair show robust surprisal differences, with significantly higher surprisal in the mismatching-context conditions (Wilcoxon signed-rank tests, $p < .01$).

Contextual facilitation Contextual facilitation—defined as the reduction in surprisal when a biasing context is provided—is largest in Case $(S \mid C_{\text{spec}})(-2.04$ bits), where contextual bias resolves a genuinely ambiguous structure. The other two pairs (unambiguous and passive) show smaller facilitation effects, consistent with their lexical or syntactic cues already constraining interpretation.

Sensitivity of SBIs Table 2 shows that in the unambiguous ($(S_{\text{lex}} \mid C_{\text{spec}})$, $(S_{\text{lex}} \mid C_{\text{non}})$) and passive ($(S_{\text{pass}} \mid C_{\text{spec}})$, $(S_{\text{pass}} \mid C_{\text{non}})$) pairs, SBIs exhibit substantially larger mean $|\Delta_j|$ than non-SBIs (1.20× and 1.37×, respectively). In contrast, the ambiguous ($(S \mid C_{\text{spec}})$, $(S \mid C_{\text{non}})$) pair shows no such selective modulation, indicating that both readings are equally supported. Results for other models, Mistral-7B-v0.3 and Qwen2.5-7B, are reported in Appendix A, demonstrating convergent patterns with LLaMa3 but with attenuated magnitudes.

Together, these diagnostics recapitulate the structural distinctions encoded in our stimulus design. The observed pattern aligns with our linguistic hypothesis: for sentences whose scope readings remain legitimately underspecified, either interpretation can be pragmatically supported once an appropriate context is provided.

5 Experiment 2. Layerwise representations: where and what is differentiated?

Experiment 2 examines how contextual evidence shapes the internal representations of scope-ambiguous sentences, tracing a progression from geometric divergence, through causal readout, to the linear accessibility of a theoretically grounded semantic distinction. We address a single functional question from three complementary perspectives: (i) Do alternative scope interpretations diverge in representation space beyond surface-level variation? (ii) Are these representational differences behaviorally consequential? (iii) Which semantic distinction is encoded in these layers?

5.1 Experiment 2A: Layerwise Representation Similarity

In this subsection, we ask **where in the model alternative scope interpretations begin to diverge**. Specifically, we examine whether such divergence emerges in internal representations, at which layers it arises, and how its magnitude compares to that observed for sentence pairs that differ in surface form but converge on similar interpretations.

5.1.1 Method

To quantify how internal representations diverge across layers, we compare sentence-level representations using three complementary similarity measures. Let $\boldsymbol{X}, \boldsymbol{Y} \in \mathbb{R}^{n \times d}$ denote matrices of sentence representations extracted from a given layer, where n is the number of tokens and d is the hidden dimension. We consider three measures $\mathsf{sim}(\boldsymbol{X}, \boldsymbol{Y})$: **cosine similarity**, **CKA**(Kornblith et al., 2019), and **Procrustes similarity** (Schönemann, 1966). Details are presented in Appendix B.2. These metrics allow us to assess representational divergence from complementary perspectives: local vector alignment (cosine), global second-order geometry (CKA), and explicit geometric correspondence under optimal alignment (Procrustes).

5.1.2 Results and Discussion

Figure 3 reports the representational similarity result. Across all conditions, the model exhibits a characteristic **high–low–high** trajectory, consistent with a progression from lexical encoding, through semantic abstraction, to output-oriented representations (Tenney et al., 2019; Liu et al., 2019; Peters et al., 2018; Jawahar et al., 2019). Crucially, however, the shape of this trajectory differs systematically across pairs, revealing how the model differentiates surface-level variation from interpretive contrasts.

As is well established in Tenney et al. (2019); Liu et al. (2019) that early layers primarily encode lexical information and shallow syntactic structure, the surface-varying pairs, $(((S \mid C_{\text{spec}}), (S_{\text{lex}} \mid C_{\text{spec}}))$ and $((S \mid C_{\text{spec}}), (S_{\text{pass}} \mid C_{\text{spec}})))$, exhibit substantially lower similarity than the surface-identical, $((S \mid C_{\text{spec}}), (S \mid C_{\text{non}}))$.

A critical distinction arises from the middle layers onward. From roughly L10 through the final layers (L30–L32), the three pairs exhibit markedly different recovery dynamics. In CKA, while the surface-identical pair $(((S \mid C_{\text{spec}}), (S \mid C_{\text{non}}))$

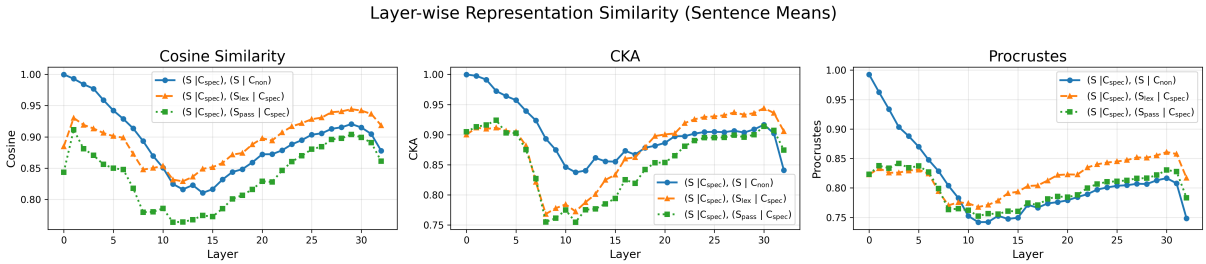

Figure 3: Layer-wise comparison of (1) Cosine Similarity, (2) CKA, and (3) Procrustes in LLaMa3-8B.

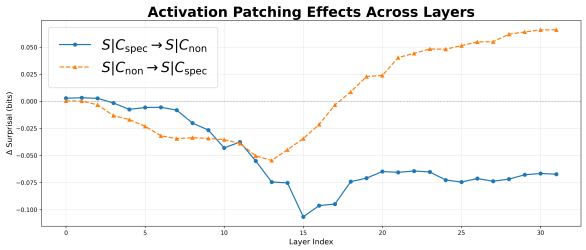

Figure 4: Layer-wise hidden-state patching effects for the $(S \mid C_{\textbf{spec}})$ **and** $(S \mid C_{\textbf{non}})$ pair in LLaMA3-8B.

shows only a **very gradual increase** in similarity, the two surface-varying pairs display a **much steeper rise**, reflecting rapid semantic convergence once event-level meaning has been abstracted. This contrast—most clearly visible in CKA, though less pronounced in cosine similarity—suggests that lexical and syntactic differences are progressively neutralized in later layers, whereas interpretive differences remain partially separated.

Notably, from approximately L10 onward, the surface-identical pair $(((S \mid C_{\text{spec}}), (S \mid C_{\text{non}})))$ exhibits consistently lower similarity than the surface-varying pairs in Procrustes, demonstrating **geometrical separation from surface-level variation**; in CKA and cosine similarity it occupies an intermediate position through most layers, diverging most clearly only in the final layers (L30+). For cross-model comparison, analogous trends are reported in Figure 5 and Figure 6 in Appendix B.1.

5.2 Experiment 2B: Activation Patching

Experiment 2B examines **whether representational differences identified in Experiment 2A become behaviorally consequential for model predictions**. Using activation patching as a coarse-grained causal intervention, we test whether context-conditioned hidden states at a given layer affect sentence-level surprisal when transferred across scope interpretations, without isolating specific neurons, heads, or circuits as in (Vig et al., 2020; Meng et al., 2022). This approach therefore provides a conservative causal probe of when layerwise representations are functionally utilized by the model.

5.2.1 Method

Intervention setup Given a source example E_{src} and a target example E_{tgt}, we replace the hidden states corresponding to the target sentence span at layer ℓ with those from the source example, while keeping the preceding context fixed. The sentence span is defined as a contiguous token interval $[s, e]$.

Patching operation Patching is implemented as a position-wise replacement of all token-level hidden states within the sentence span at the output of transformer block ℓ (i.e., the return value of `model.model.layers[`ℓ`]`).

Evaluation protocol We perform symmetric patching in both directions (e.g., $(S \mid C_{\text{spec}}) \rightarrow (S \mid C_{\text{non}})$ and $(S \mid C_{\text{non}}) \rightarrow (S \mid C_{\text{spec}})$). Let $\mathcal{G}^*_\phi(S \mid C)$ denote the sentence-level surprisal after applying an activation patching intervention ϕ (e.g., ϕ corresponds to patching from $(S \mid C_{\text{spec}}) \rightarrow (S \mid C_{\text{non}})$ or $(S \mid C_{\text{non}}) \rightarrow (S \mid C_{\text{spec}})$). We quantify the effect of the intervention by the resulting change in sentence-level surprisal:

$$\Delta\mathcal{G}^{(\ell)} = \mathcal{G}^*_\phi(S \mid C) - \mathcal{G}(S \mid C). \qquad (5)$$

Effects are interpreted comparatively across layers rather than in absolute terms.

Scope of inference It is designed to probe when context-conditioned representations become behaviorally consequential for model predictions, rather than to isolate the precise mechanisms by which they are constructed.

5.2.2 Results and Discussion

Directional Asymmetry Figure 4 reports surprisal differences induced by swapping hidden states across contexts. First, up to approximately

Layers 12, patching produces only mild and symmetric effects, suggesting that while interpretive differences may already be present in earlier layers, they are not yet expressed in a form that directly constrains the model's final predictions, as subsequent computation can still reinterpret or override these representations. Beyond this mid-layer region, however, the two patching directions diverge sharply. Injecting $(S \mid C_{\text{non}})$ hidden states into $(S \mid C_{\text{spec}})$ contexts yields increasingly **positive** surprisal differences, indicating a growing contextual mismatch, while injecting $(S \mid C_{\text{spec}})$ into $(S \mid C_{\text{non}})$ yields only modest and relatively stable **negative** shifts. Viewed through the lens of representational readout rather than representational formation, this asymmetry indicates that context-conditioned representations become behaviorally consequential only once they reach layers whose outputs are directly readable by the final language modeling head.

This directional asymmetry suggests that the two scope readings are not geometrically symmetric in the model's representational space. $(S \mid C_{\text{non}})$ behaves as a less marked and more general representational pathway whose activations tolerate substitution with minimal disruption, whereas $(S \mid C_{\text{spec}})$ appears to require a more constrained configuration that is less easily overridden. This finding aligns with the **context-free** preference for $(S \mid C_{\text{non}})$ reported in Kamath et al. (2024), and reveals a mechanistic correlate of this bias: $(S \mid C_{\text{non}})$-based representations exert weaker constraints on downstream computation and therefore dominate as the default representational trajectory.

5.3 Experiment 2C: Probing for Scopal Specificity

Experiment 2C addresses the question of **what semantic distinction is encoded in the layerwise representations identified in Experiments 2A and 2B**. In formal semantics, *specificity* distinguishes noun phrases that refer to a particular entity from those that do not (Abusch, 1993; Von Heusinger, 2019). Crucially, **specificity is tightly linked to scope**: a SBI taking wide scope over other operators is interpreted as specific, while one taking narrow scope remains non-specific. We now ask where within the model this theoretically grounded semantic distinction is made explicit and accessible. This motivates our use of both scalar-mixing edge probing (to infer global layer importance through learned mixing weights) and layer-wise logistic regression (to assess linear separability at each individual layer).

Case	Condition Type	Accuracy
$(S \mid C_{\text{spec}})$	Spec Ctx + Ambiguous	0.676
$(S \mid C_{\text{non}})$	Non-spec Ctx + Ambiguous	0.583
$(S_{\text{lex}} \mid C_{\text{spec}})$	Spec Ctx + Lexical Specific	**0.791**
$(S_{\text{lex}} \mid C_{\text{non}})$	Non-spec Ctx + Lexical Specific	**0.439**
$(S_{\text{pass}} \mid C_{\text{spec}})$	Spec Ctx + Passive Specific	**0.878**
$(S_{\text{pass}} \mid C_{\text{non}})$	Non-spec Ctx + Passive Specific	**0.353**

Table 3: Single-case edge probing accuracy in LLaMA-3-8B. Higher accuracy indicates stronger linear separability of specificity.

5.3.1 Method

Dataset construction. We construct a dedicated SPECIFICITY dataset for probing. Using GPT-4o, we generate 2,000 context–sentence pairs (1,000 per class) that induce *specific* vs. *non-specific* interpretations independently of scope manipulation (see Table 13 in Appendix C).

Edge probing with scalar mixing Our primary analysis adopts a scalar-mixing edge probing framework (Tenney et al., 2019). We stack layerwise hidden representations of the target noun phrase span into a tensor $\boldsymbol{X} \in \mathbb{R}^{N \times L \times d}$ and train a scalar-mixing MLP that learns a softmax-weighted combination of layers(details in Appendix E).

Span selection and evaluation. Probes are applied to hidden representations of the target span, defined as the second SBI in surface order (e.g., *a dog*) in our SCOPEX dataset, extracted using span-mean pooling. Decision thresholds are selected per model by maximizing F1 on a held-out validation split. To assess probe selectivity, we follow Hewitt and Liang (2019), and compare performance on SCOPEX against matched control dataset.

Auxiliary probe. As an auxiliary analysis, we additionally train independent layerwise logistic regression probes (Alain and Bengio, 2018) to verify that the observed layerwise trends are not specific to the scalar-mixing architecture; full details and results are reported in the Appendix F.

.

5.3.2 Results and Discussion

Asymmetry between *specificity* and *non-specificity* Table 3 reveals a systematic asymmetry between the two readings. Specific interpretations yield reliably high accuracy ($(S \mid C_{\text{spec}})$:

Pair	PR-AUC	ROC-AUC	Acc.	F1
Baseline (Random)	0.4785	0.4684	0.5	0.6213
$(S \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.6549	0.6773	0.6295	0.6460
$(S_{\text{lex}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.7347	0.7521	0.6871	0.7166
$(S_{\text{pass}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	**0.7999**	**0.8276**	**0.7302**	**0.7649**
$(S \mid C_{\text{spec}})$ vs. $(S_{\text{lex}} \mid C_{\text{non}})$	**0.5519**	**0.5931**	**0.5576**	**0.6045**

Table 4: Pairwise edge probing performance in LLaMA-3-8B. Each row trains a binary classifier on span representations for the two cases in the pair.

Category	Mean (%)	vs. Expected
Layer groups		
Early (L0–L10)	2.862	−5.6%
Middle (L11–L21)	3.003	−0.9%
Late (L22–L32)	**3.292**	**+8.7%**
Expected (uniform)	3.030	—
Top-5 layers		
1. Layer 32 (Late)	**3.300**	**+8.9%**
2. Layer 25 (Late)	3.298	+8.8%
3. Layer 30 (Late)	3.294	+8.7%
4. Layer 31 (Late)	3.294	+8.7%
5. Layer 27 (Late)	3.291	+8.6%
Range	0.443%	—
Non-uniform?	✓ Yes (>1.0% threshold)	

Table 5: Scalar mixing weights over LLaMA-3-8B layers learned by the edge probe.

0.68; $(S_{\text{lex}} \mid C_{\text{spec}})$: **0.79**; $(S_{\text{pass}} \mid C_{\text{spec}})$: **0.88**), indicating that the model forms a coherent and linearly separable representation whenever lexical or syntactic cues unambiguously support *specificity*. By contrast, the non-specific cases show uniformly low accuracy, but for distinct linguistic reasons. For the $(S \mid C_{\text{non}})$ case (0.58), the reduced accuracy does not reflect noise but rather the fact that ***non-specificity* does not constitute a single semantic category**, as argued in Degano and Aloni (2025). The probe's difficulty in isolating a single *non-specific* class is therefore not a model limitation but an expected consequence of the theoretical semantics. In contrast, the $(S_{\text{lex}} \mid C_{\text{non}})$ case (0.44) and the $(S_{\text{pass}} \mid C_{\text{non}})$ case (0.35) are low primarily due to **cue conflict**: while context biases a non-specific interpretation, the sentence-internal cues (lexical or syntactic) strongly bias *specificity*. The resulting representational tension prevents the model from stabilizing a consistent encoding.

Graded hierarchy: syntactic > lexical > contextual Table 4 evaluates the **discriminability** between readings. The pairwise probes reveal a **graded hierarchy**: when cues align (e.g. the pair of $(S_{\text{pass}} \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$), separability is high (PR-AUC: **0.88**), but purely pragmatic contrasts (the pair of $(S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$) achieve only moderate separability (0.67). Crucially, when **contextual cues oppose sentence-internal cues** (the pair of $((S \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}}))$, performance barely exceeds baseline (**0.55**), confirming that *specificity* encoding becomes robust only when multiple linguistic signals converge.

Layerwise distribution Scalar-mixing weights (Table 5) further indicate that while *specificity* information is distributed across layers, it becomes increasingly accessible toward the top of the network, with late layers receiving +8–9% greater weight than expected under uniformity. Layerwise probing yields a consistent pattern (see Appendix F).

Together, Experiments 2B and 2C provide converging evidence for Degano and Aloni (2025)'s claim that *specificity* is encoded as a marked semantic category, whereas *non-specificity* corresponds to a heterogeneous, non-categorical representational space.

6 Conclusion

This work investigated whether transformer language models encode implicit logical meaning, focusing on scope ambiguity as a diagnostic domain requiring compositional integration of lexical, structural, and contextual cues. Across three complementary experiments, we found convergent evidence that LLaMA3-8B (and to a lesser extent other models) constructs internally structured representations that distinguish alternative scope readings.

The convergence of behavioral, representational, and causal diagnostics suggests that scope ambiguity provides a promising window into how models internalize abstract semantic operations beyond surface-level distribution.

Overall, this work demonstrates that transformer LMs develop structured internal representations that reflect the logical organization posited in formal semantics. Clarifying how these representations arise, generalize, and support reasoning remains a central challenge for future research.

7 Limitations

Despite these encouraging findings, several limitations remain. (1) **Weak purely contextual disambiguation.** Models show relatively shallow sensitivity to context-only contrasts (the pair of $((S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}}))$, whereas human comprehenders resolve such cases efficiently. This gap highlights continuing limitations in discourse-level integration. (2) **Causal underdetermination.** Activation patching identifies layers carrying interpretation-relevant information, but does not by itself specify the mechanisms that compute scope or *specificity*. Stronger causal modeling (e.g., mediation analysis, neuron-level circuit discovery) is needed. (3) **Model variability.** Qwen2.5 and Mistral exhibited weaker and less stable effects, suggesting that architectural differences (e.g., grouped-query attention) may influence the encoding of fine-grained semantic distinctions, though further controlled studies are required.

References

Dorit Abusch. 1993. The scope of indefinites. *Natural Language Semantics*, 2(2):83–135.

Guillaume Alain and Yoshua Bengio. 2018. Understanding intermediate layers using linear classifier probes. *Preprint*, arXiv:1610.01644.

John Barwise and Robin Cooper. 1981. Generalized quantifiers and natural language. *Linguistics and Philosophy*, 4(2):159–219.

Marco Degano and Maria Aloni. 2025. How to be (non-)specific? *Unknown Journal (Accepted)*. Accepted: 23 June 2025.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

John Hale. 2001. A probabilistic Earley parser as a psycholinguistic model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*.

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, Hong Kong, China. Association for Computational Linguistics.

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, Florence, Italy. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, and 1 others. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics*, 12:738–754.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. 2019. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 3519–3529. PMLR.

Roger Levy. 2008. Expectation-based syntactic comprehension. *Cognition*, 106(3):1126–1177.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic knowledge and transferability of contextual 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 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

Nelson F. Liu, Rima Taslakian, Xiang Lisa Chen, and Najoung Kim. 2023. We are afraid language models aren't modelling ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*.

Robert May. 1985. *Logical Form: Its Structure and Derivation*. MIT Press.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

OpenAI. 2024. Gpt-4o. `https://openai.com/index/gpt-4o/`. Multimodal large language model accessed via OpenAI API.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. 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 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

An Qwen Team: Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and 1 others. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Peter H. Schönemann. 1966. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423.

Nathaniel J. Smith and Roger Levy. 2013. The effect of word predictability on reading time is logarithmic. *Cognition*, 128(3):302–319.

Elias Stengel-Eskin, Yonatan Belinkov, and Benjamin Van Durme. 2024. Zero and few-shot semantic parsing with ambiguous inputs. In *Findings of the Association for Computational Linguistics*.

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, Florence, Italy. Association for Computational Linguistics.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. In *Advances in Neural Information Processing Systems*, volume 33, pages 12388–12401. Curran Associates, Inc.

Klaus Von Heusinger. 2019. Indefiniteness and specificity. In *The Oxford Handbook of Reference*. Oxford University Press.

Case	Context	Sentence Type	Target Sentence	Cong.
$(S \mid C_{\textbf{spec}})$	*My neighbor was worried about her pets while she was away, and she specifically asked me to feed her dog. Unfortunately,*	**Ambiguous**	*I did**n't** feed **a** dog* $\exists x(DOG(x) \wedge \neg FEED(I, x))$	✓
$(S \mid C_{\textbf{non}})$	*During my volunteer shift at the animal shelter, I spent most of my time with the cats since the dog section was closed. So,*	**Ambiguous**	*I did**n't** feed **a dog*** $\neg\exists x(DOG(x) \wedge FEED(I, x))$	✓
$(S_{\textbf{lex}} \mid C_{\textbf{spec}})$	**(My neighbor...)** Same as $(S \mid C_{\text{spec}})$	**Lexical Unambiguous**	*I starved **a dog*** $\exists x(DOG(x) \wedge STARVE(I, x))$	✓
$(S_{\textbf{lex}} \mid C_{\textbf{non}})$	**(During my volunteer shift...)** Same as $(S \mid C_{\text{non}})$	**Lexical Unambiguous**	*I starved **a dog*** $\exists x(DOG(x) \wedge STARVE(I, x))$	✗
$(S_{\textbf{pass}} \mid C_{\textbf{spec}})$	**(My neighbor...)** Same as $(S \mid C_{\text{spec}})$	**Passivized**	***A dog** was**n't** fed by me* $\exists x(DOG(x) \wedge \neg FEED(I, x))$	✓
$(S_{\textbf{pass}} \mid C_{\textbf{non}})$	**(During my volunteer shift...)** Same as $(S \mid C_{\text{non}})$	**Passivized**	***A dog** was**n't** fed by me* $\exists x(DOG(x) \wedge \neg FEED(I, x))$	✗

Table 6: Complete dataset design (2×3 factorial, $N = 139$ items per cell; 834 total). Congruency (✓/✗) indicates whether contextual cues support the same interpretation as the lexical or syntactic cues of the target sentence.

Table 7: Complete SCOPEx examples for four items (IDs 8, 13, 27, 112), each showing all six cases in the 2×3 factorial design.

ID 8 — SBI1: "ocationally", SBI2: "two dentists"		
$(S \mid C_{\textbf{spec}})$	Whenever I have a dental issue, I usually check with my regular dentist, Dr. Smith, and sometimes for a second opinion, I consult my sister, who is also a dentist.	I **occasionally** consult **two dentists**.
$(S \mid C_{\textbf{non}})$	Living in a small town, there aren't many dental professionals around. So when I need advice, if there happen to be that many available,	I **occasionally** consult **two dentists**.
$(S_{\textbf{lex}} \mid C_{\textbf{spec}})$	(Whenever I have a dental issue,...) Same as C_{spec}.	I consulted **two dentists**.
$(S_{\textbf{lex}} \mid C_{\textbf{non}})$	(Living in a small town,...) Same as C_{non}.	I consulted **two dentists**.
$(S_{\textbf{pass}} \mid C_{\textbf{spec}})$	(Whenever I have a dental issue,...) Same as C_{spec}.	**Two dentists** are **occasionally** consulted.
$(S_{\textbf{pass}} \mid C_{\textbf{non}})$	(Living in a small town,...) Same as C_{non}.	**Two dentists** are**occasionally** consulted.
ID 13 — SBI1: "n't", SBI2: "an error"		
$(S \mid C_{\textbf{spec}})$	My colleague pointed out that there was an issue in the Python script I submitted yesterday. Surprisingly,	I did**n't** notice **an error** in my code
$(S \mid C_{\textbf{non}})$	I ran my program several times, and everything seemed to be working perfectly without any hiccups.	I did**n't** notice **an error** in my code
$(S_{\textbf{lex}} \mid C_{\textbf{spec}})$	(My colleague...) Same as C_{spec}.	I overlooked **an error** in my code
$(S_{\textbf{lex}} \mid C_{\textbf{non}})$	(I ran my program...) Same as C_{non}.	I overlooked **an error** in my code
$(S_{\textbf{pass}} \mid C_{\textbf{spec}})$	(My colleague...) Same as C_{spec}.	**An error** in my code was**n't** noticed.
$(S_{\textbf{pass}} \mid C_{\textbf{non}})$	(I ran my program...) Same as C_{non}.	**An error** in my code was**n't** noticed.
ID 27 — SBI1: "all", SBI2: "a few boxes of pizza"		
$(S \mid C_{\textbf{spec}})$	At the end of the conference, the organizers ordered several boxes of pizza for the urban planners attending the event.	**All** urban planners eat **a few boxes of pizza**.
$(S \mid C_{\textbf{non}})$	When discussing the late-night work culture in their field, the students joked about how common it was to order pizza.	**All** urban planners eat **a few boxes of pizza**.
$(S_{\textbf{lex}} \mid C_{\textbf{spec}})$	(At the end of the conference...) Same as C_{spec}.	**All** urban planners eat just these few boxes of pizza.
$(S_{\textbf{lex}} \mid C_{\textbf{non}})$	(When discussing the late-night work culture...) Same as C_{non}.	**All** urban planners eat just these few boxes of pizza.
$(S_{\textbf{pass}} \mid C_{\textbf{spec}})$	(At the end of the conference...) Same as C_{spec}.	**a few boxes of pizza** are eaten by **all** urban planners.
$(S_{\textbf{pass}} \mid C_{\textbf{non}})$	(When discussing the late-night work culture...) Same as C_{non}.	**a few boxes of pizza** are eaten by **all** urban planners.
ID 112 — SBI1: "frequently", SBI2: "a few books"		
$(S \mid C_{\textbf{spec}})$	I have a favorite set of novels that I borrow from the library every few months.	I **frequently** read **a few books** from the library.
$(S \mid C_{\textbf{non}})$	Whenever I visit the library, I aim to explore different genres and authors.	I **frequently** read **a few books** from the library.
$(S_{\textbf{lex}} \mid C_{\textbf{spec}})$	(I have a favorite set...) Same as C_{spec}.	I **frequently** read these few books from the library
$(S_{\textbf{lex}} \mid C_{\textbf{non}})$	(Whenever I visit the library...) Same as C_{non}.	I **frequently** read these few books from the library
$(S_{\textbf{pass}} \mid C_{\textbf{spec}})$	(I have a favorite set...) Same as C_{spec}.	**A few books** from the library are **frequently** read by me.
$(S_{\textbf{pass}} \mid C_{\textbf{non}})$	(Whenever I visit a library...) Same as C_{non}.	**A few books** from the library are **frequently** read by me.

A Model Variability in Experiment 1: Mistral-7B-ver0.3 and Qwen2.5-7B

While LLaMA-3-8B exhibited the predicted context-sensitive surprisal pattern, the two other 7B-scale models—Qwen2.5-7B and Mistral-7B—showed quantitative variations while preserving core qualitative patterns. We applied the same statistical significance threshold ($p < .01$) for comparability across models.

A.1 Mistral-7B: Convergent Patterns with Attenuated Magnitudes

Mistral-7B replicates LLaMA-3-8B's core findings despite weaker effect sizes (Table 9). Like LLaMA-3-8B, Mistral-7B shows significant scope-driven modulation in the unambiguous pair (the pair of $(S_{\text{lex}} \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}})$) (Mean$_\Delta$ = -0.279 bits, $p < .001$, $r = +0.349$)) and localizes surprisal effects to SBIs (1.24× ratio; Table 10), mirroring LLaMA-3-8B's 1.20× ratio.

The passive condition also exhibits scope-element sensitivity (1.05× ratio), consistent with LLaMA-3-8B (1.37× ratio), though Mistral-7B shows weaker overall facilitation. The ambiguous condition shows no significant difference ($p = .549$), similar to LLaMA-3-8B ($p = .256$).

These results demonstrate that scope-sensitivity mechanisms generalize across architectures: both models concentrate contextual effects at SBIs in conditions requiring inverse scope (the pair of $(S_{\text{lex}} \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}})$ and the pair of $(S_{\text{pass}} \mid C_{\text{spec}})$ and $(S_{\text{pass}} \mid C_{\text{non}})$), while showing symmetry in genuinely ambiguous cases (the pair of $(S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$).

Table 8: Comparison between SCOPEX and Kamath et al. (2024)

Aspect	**ScopeX (This work)**	**Kamath et al. (2024)**
Dataset design	• 2×3 factorial design • Ambiguous / lexically controlled / syntactically biased sentences • Preceding context disambiguating inverse vs. surface scope	• 2×2 design • Ambiguous vs. unambiguous sentences • Continuations favoring inverse vs. surface scope
Presentation Design to models	• *evidence-first, ambiguity-following*: Context + Sentence	• *ambiguity-first, evidence-following*: Sentence + Continuation
Experimental methodology	• Surprisal-based analysis • Layerwise representation similarity (Cosine Similarity/ CKA / Procrustes) • Causal intervention (activation patching) • Probing for scopal specificity	• Human–LM alignment analysis • Probability comparison of continuations
Research question	• When disambiguating context precedes the sentence, are scope-sensitive interpretations reflected in the internal representations of language models? • If so, are these representations localized to particular layers or distributed across models? • Do these encodings causally affect the model behavior?	• Can LLMs detect and respond to scope ambiguity? • Do LLMs show human-like default preferences?

Mean Surprisal (bits)			
Case	**In-situ**	**Isolated**	**Diff.**
$(S \mid C_{\text{spec}})$	4.230	6.336	-2.106
$(S \mid C_{\text{non}})$	4.166	6.303	-2.137
$(S_{\text{lex}} \mid C_{\text{spec}})$	4.843	7.100	-2.257
$(S_{\text{lex}} \mid C_{\text{non}})$	5.050	7.028	-1.978
$(S_{\text{pass}} \mid C_{\text{spec}})$	4.975	6.650	-1.676
$(S_{\text{pass}} \mid C_{\text{non}})$	5.023	6.675	-1.652
Pairwise Statistical Comparisons			
Pair	**Mean$_\Delta$**	p	r
$(S \mid C_{\text{spec}})$–$(S \mid C_{\text{non}})$	$+0.031$	.549	$+0.059$
$(S_{\text{lex}} \mid C_{\text{spec}})$–$(S_{\text{lex}} \mid C_{\text{non}})$	-0.279	$< .001^{***}$	$+0.349$
$(S_{\text{pass}} \mid C_{\text{spec}})$–$(S_{\text{pass}} \mid C_{\text{non}})$	-0.024	.861	$+0.017$

Table 9: Mean surprisal values and pairwise comparisons for Mistral-7B. Diff. = In-situ − Isolated (contextual facilitation). Mean$_\Delta$ = difference in contextual effects between cases. Paired Wilcoxon signed-rank tests. Colored values indicate column-wise extrema: green = favorable (strongest facilitation or significance), red = highest difficulty (weakest facilitation), orange = largest pairwise difference. $^{***}p < .001$.

Pair	**Non-scope**	**Scope**	**Diff.**	**Ratio**
$(S \mid C_{\text{spec}})$–$(S \mid C_{\text{non}})$	1.592	1.477	-0.115	0.93×
$(S_{\text{lex}} \mid C_{\text{spec}})$–$(S_{\text{lex}} \mid C_{\text{non}})$	1.754	**2.174**	**+0.420**	**1.24×**
$(S_{\text{pass}} \mid C_{\text{spec}})$–$(S_{\text{pass}} \mid C_{\text{non}})$	1.511	1.593	$+0.082$	1.05×

Table 10: Sensitivity of SBIs for Mistral-7B. Mean $|\Delta|$ (bits) for SBIs vs. non-SBIs. Ratio = Scope/Non-scope effect magnitude

A.2 Qwen2.5-7B: Scope-Element Sensitivity Without Statistical Significance

Qwen2.5-7B demonstrates the same SBI sensitivity pattern observed in both LLaMA-3-8B and Mistral-7B (Table 12), with SBIs showing enhanced contextual effects in unambiguous (1.21× ratio) and passive (1.19× ratio) conditions—nearly identical to LLaMA-3-8B's 1.20× and Mistral-7B's 1.24×. The ambiguous condition exhibits symmetric processing (1.00× ratio), mirroring the other models.

However, unlike LLaMA-3-8B and Mistral-7B, none of Qwen2.5-7B's pairwise comparisons reach statistical significance under the $p < .01$ threshold (Table 11): the unambiguous pair (the pair of $(S_{\text{lex}} \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}})$) shows $p = .680$, and the passive pair (the pair of $(S_{\text{pass}} \mid C_{\text{spec}})$ and $(S_{\text{pass}} \mid C_{\text{non}})$) shows $p = .592$.

Despite this lack of statistical reliability, the qualitative pattern remains consistent: SBIs concentrate larger surprisal differences in conditions requiring scope computation (the pair of $(S_{\text{lex}} \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}})$ and the pair of $(S_{\text{pass}} \mid C_{\text{spec}})$ and $(S_{\text{pass}} \mid C_{\text{non}})$) while showing uniform processing in genuinely ambiguous cases (the pair of $(S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$). This suggests that Qwen2.5-7B encodes scope-sensitive representations with reduced magnitude.

The convergence in scope-element ratios (1.20×–1.24× across all three models for the pair of $(S_{\text{lex}} \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}})$ indicates that **the core localization mechanism generalizes across architectures, though statistical detectability varies with model-specific factors**.

Mean Surprisal (bits)			
Case	**In-situ**	**Isolated**	**Diff.**
$(S \mid C_{\text{spec}})$	4.651	6.575	−1.924
$(S \mid C_{\text{non}})$	4.481	6.570	−2.089
$(S_{\text{lex}} \mid C_{\text{spec}})$	5.413	7.450	−2.037
$(S_{\text{lex}} \mid C_{\text{non}})$	5.502	7.469	−1.967
$(S_{\text{pass}} \mid C_{\text{spec}})$	5.324	6.800	−1.476
$(S_{\text{pass}} \mid C_{\text{non}})$	4.891	6.758	−1.868
Pairwise Statistical Comparisons			
Pair	**Mean$_\Delta$**	p	r
$(S \mid C_{\text{spec}})$–$(S \mid C_{\text{non}})$	+0.165	.072	+0.176
$(S_{\text{lex}} \mid C_{\text{spec}})$–$(S_{\text{lex}} \mid C_{\text{non}})$	−0.070	.680	+0.040
$(S_{\text{pass}} \mid C_{\text{spec}})$–$(S_{\text{pass}} \mid C_{\text{non}})$	−0.087	.592	+0.052

Table 11: Mean surprisal values and pairwise comparisons for Qwen2.5-7B. Diff. = In-situ − Isolated (contextual facilitation). Mean$_\Delta$ = difference in contextual effects between cases. Paired Wilcoxon signed-rank tests. Colored values indicate column-wise extrema: green = favorable (strongest facilitation), red = highest difficulty (weakest facilitation), orange = largest pairwise difference.

Pair	**Non-scope**	**Scope**	**Diff.**	**Ratio**
$(S \mid C_{\text{spec}})$–$(S \mid C_{\text{non}})$	1.848	1.843	−0.006	1.00×
$(S_{\text{lex}} \mid C_{\text{spec}})$–$(S_{\text{lex}} \mid C_{\text{non}})$	2.051	**2.482**	**+0.430**	**1.21×**
$(S_{\text{pass}} \mid C_{\text{spec}})$–$(S_{\text{pass}} \mid C_{\text{non}})$	1.770	2.114	+0.344	1.19×

Table 12: Sensitivity of SBIs for Qwen2.5-7B. Mean $|\Delta|$ (bits) for SBIs vs. non-SBIs. Ratio = Scope/Non-scope effect magnitude

B Model Variability in Experiment 2A: Mistral-7B-ver0.3 and Qwen2.5-7B

B.1 Cross-Model Comparison of Representation Similarity with Fast Procrustes

Figures 5–6 present layerwise representation similarity for Mistral-7B and Qwen2.5-7B. For Procrustes analysis, we employ a fast approximation (see below), which we validated against the full Procrustes computation on LLaMA-3-8B in Figure 3 and found to yield qualitatively identical layerwise trends in Figure 7.

Across models, all exhibit the characteristic **high–low–high** trajectory also observed in the main text (Section 5.1), reflecting a common progression from lexical encoding through intermediate abstraction to output-oriented representations. Crucially, however, the models differ in how *interpretive contrasts* are geometrically separated from surface-level variation across layers.

Mistral-7B. Mistral-7B exhibits clear representational separability beginning in the mid-to-late layers. Across both cosine similarity and Procrustes alignment, all sentence pairs show a marked divergence starting around layers L8–L9 and persisting through the final layer. Linear CKA reveals a similar overall trajectory, although the separation for the pair of $(S \mid C_{\text{spec}})$ and $(S_{\text{pass}} \mid C_{\text{non}})$ is more pronounced. Notably, the pair of $(S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$ reaches a comparable degree of separability to the pair of $(S \mid C_{\text{spec}})$ and $(S_{\text{pass}} \mid C_{\text{non}})$ from approximately layer L20 onward. Aside from this difference in relative magnitude, the layerwise trends are internally consistent across metrics within Mistral-7B and closely mirror the patterns observed for Llama3.

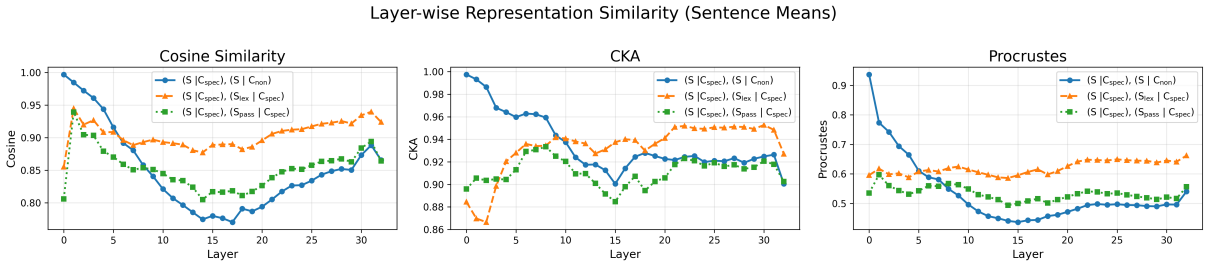

Figure 5: Representaion Similarity of Mistral-7B Model

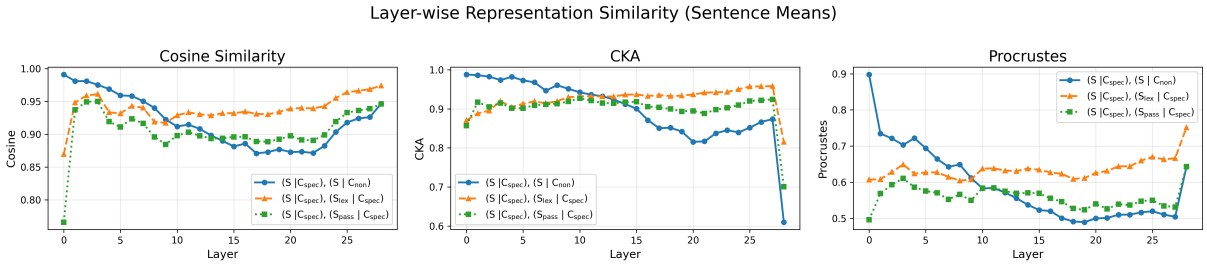

Figure 6: Representation Similarity of Qwen2.5-7B

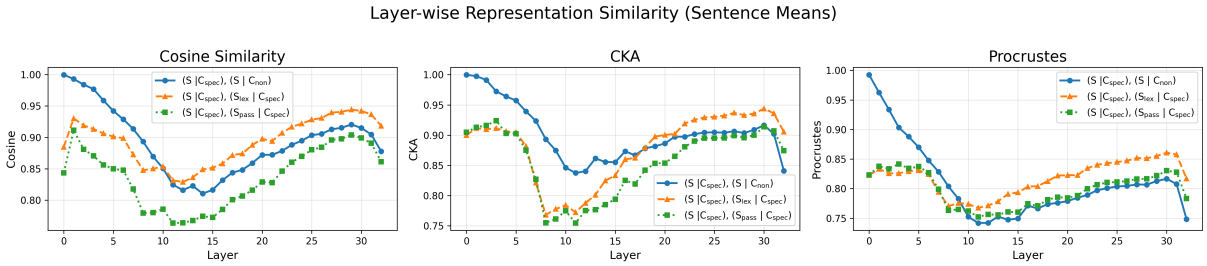

Figure 7: Representaion Similarity of LLaMa3-8B Model with Fast Procrustes

Qwen2.5-7B. Qwen2.5-7B largely follows the representational trajectory observed for Llama3 and Mistral. The pair of $(S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$ remains more convergent than other pairs through approximately layers L13–L14, after which it diverges steadily toward the final layer. A distinctive characteristic of Qwen2.5, however, is the pronounced non-monotonic behavior at the extremes of the network. Specifically, we observe a sharp increase in the representational similarity between layers L0 and L1 in Cosine Similarity, followed by a comparably sharp drop at the final layers (L27–L28) in CKA. These early and late discontinuities are substantially more abrupt than in the other models, suggesting heightened sensitivity at the embedding interface and at the output-oriented layers.

Overall, these representational patterns bear a tentative resemblance to the behavioral trends observed in Experiment 1. In particular, LLaMA-3's sustained mid-to-late divergence coincides with stronger scope sensitivity, while Mistral's more compressed geometry co-occurs with attenuated effects. For Qwen2.5, the instability observed at the final layers may be consistent with its lack of statistical significance, although this interpretation remains speculative. Importantly, we do not claim a direct causal link between representational geometry and behavioral outcomes, but note this alignment as a suggestive pattern warranting further investigation.

B.2 Similarity Metrics for Hidden Representations

Let $X, Y \in \mathbb{R}^{n \times d}$ denote matrices of sentence representations extracted from a given layer, where n is the number of items and d the hidden dimension.

- **Cosine similarity** measures the average angular alignment between corresponding representations:

$$\text{CosSim}(X, Y) = \frac{1}{n} \sum_{i=1}^{n} \frac{X_i \cdot Y_i}{\|X_i\| \, \|Y_i\|}. \quad (6)$$

  This metric is sensitive to surface-level differences and serves as a baseline measure of representational overlap.

- **Centered Kernel Alignment (CKA)** provides an orthogonally invariant measure of representational similarity that compares second-order statistics rather than individual vectors (Kornblith et al., 2019). Let $K = XX^\top$ and $L = YY^\top$ be Gram matrices, and let $H = I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top$ be the centering matrix. Linear CKA is defined as

$$\text{CKA}(X, Y) = \frac{\text{tr}(HKH\,HLH)}{\sqrt{\text{tr}((HKH)^2)\,\text{tr}((HLH)^2)}}. \quad (7)$$

  CKA captures similarity in representational geometry independent of orthogonal transformations.

- **Procrustes similarity** measures the geometric alignment between two representation spaces by optimally rotating one space to match the other, while abstracting away from absolute coordinate orientation (Schönemann, 1966). Formally, Procrustes similarity is defined as

$$\text{Procrustes}(X, Y) = 1 - \frac{\|XR^* - Y\|_F}{\sqrt{2}}, \quad (8)$$

  where R^* is the optimal orthogonal transformation minimizing the Frobenius distance between the aligned representations.

Fast vs. Full Procrustes Analysis. Due to computational constraints, for Mistral-7B and Qwen2.5-7B we used a computationally efficient variant. We verified on LLaMA-3-8B that the fast Procrustes variant yields qualitatively identical layer-wise similarity patterns to the full version (Figure X in Appendix), confirming that our conclusions are robust to this computational approximation. The fast version of Procrustes is computed as follows:
After centering and normalizing as before, instead of computing the full SVD of $\tilde{X}^\top \tilde{Y} \in \mathbb{R}^{d \times d}$, we exploit the property that for unit-norm matrices,

$$\|\tilde{X}R^* - \tilde{Y}\|_F^2 = 2 - 2\,\text{tr}(R^* \tilde{Y}^\top \tilde{X}) = 2 - 2 \sum_{i=1}^{\min(n,d)} \sigma_i, \quad (9)$$

where σ_i are the singular values of the smaller matrix $\tilde{Y}\tilde{X}^\top \in \mathbb{R}^{n \times n}$. When $n \ll d$ (e.g., $n \approx 100$ sentences vs. $d = 4096$ dimensions), this reduces computational cost from $O(d^3)$ to $O(n^2 d)$.

C SPECIFICITY Dataset for Probing Classifier Training

ID	full_text	target_NP	specific label
"0_0_2"	"I grew up in that town before moving to the city. There is a small town where I spent my childhood summers."	"a small town"	**SPECIFIC**
"0_1_1"	"Many people enjoy the charm and close-knit community of a small town. If you live in a small town, you might know everyone."	"a small town"	**NON-SPECIFIC**
"1_0_18""	"At the jewelry store downtown, there was an expensive watch in the display case. The thief's attention was immediately caught by an expensive watch."	"an expensive watch"	**SPECIFIC**
"1_1_19"	"Many families pass down valuable items, like an expensive watch, through generations. He might inherit an expensive watch from his grandfather."	"an expensive watch"	**NON-SPECIFIC**
"7_0_37""	"While I was gardening, I noticed two stray cats exploring the flower beds. I saw two cats that wandered into our garden yesterday and have made themselves at home."	"two cats"	**SPECIFIC**
"7_1_38"	"She has always wanted to have some pets, especially two cats, in her new place. She hopes to adopt two cats when she moves to her new apartment."	"two cats"	**NON-SPECIFIC**
"100_0_2"	"I met some people at the event last night. I found it really fun to talk to two friends I met."	"two friends"	**SPECIFIC**
"100_1_1"	"It would be nice to have two friends who share my love for adventure and exploration. I wish I had two friends to travel with."	"two friends"	**NON-SPECIFIC**
"102_0_1"	"During the camping trip last weekend, we looked up at the sky. I spotted seven stars that were unusually bright."	"seven stars"	**SPECIFIC**
"102_1_2"	"On a clear night, you can often see seven stars shining brightly in the sky. If you count seven stars, you might find a constellation."	"seven stars"	**NON-SPECIFIC**
"124_0_5"	"I ordered seven roses for my anniversary, and they arrived just in time. The room was filled with fragrance by seven roses that were delivered this morning."	"seven roses"	**SPECIFIC**
"124_1_1"	"For special occasions, many people often choose to give a bouquet of roses as a thoughtful gesture. He plans to buy seven roses for their anniversary."	"seven roses"	**NON-SPECIFIC**

Table 13: Representative examples from the synthetic specificity probing dataset (one per label). The table shows the item identifier, full input text, the target NP span, and the gold specificity label.

D Details of Edge Probing in Experiment 2C

We distinguish between two input distributions. The probing classifier is trained on a labeled specificity dataset SPECIFICITY, here denoted as $\mathcal{D}_{\text{spec}}$, while inference is performed on the original scope-ambiguity dataset SCOPEx, here denoted as $\mathcal{D}_{\text{scope}}$ which is unlabeled with respect to *specificity*. Each training instance from SPECIFICITY is a tuple $(x_i^{\text{spec}}, a_i, b_i, y_i) \sim \mathcal{D}$

Edge-probing module (inputs/outpus and neurel form). Let a transformer encoder M with layers $\ell \in \{0, \ldots, L\}$ map an input sequence $x = (x_1, \ldots, x_T)$ to hidden states $H^{(\ell)} \in \mathbb{R}^{T \times d}$. The probing module is shared across two input distributions: (i) a labeled specificity dataset $\mathcal{D}_{\text{spec}}$ used for training, an (ii) the scope-ambiguity dataset $\mathcal{D}_{\text{scope}}$ used only at inference. Each training instance $i \sim \mathcal{D}_{\text{spec}}$ consists of $(x_i^{\text{spec}}, a_i, b_i, y_i)$, where $[a_i, b_i]$ is the (inclusive) token span of the target expression an $y_i \in \{0, 1\}$ is the specificity label. At inference time, inputs $(x_j^{\text{scope}}, a_j, b_j) \sim \mathcal{D}_{\text{scope}}$ are unlabeled with respect to specificity.

Span representation (Span-Mean). For a given input sequence x with target span $[a, b]$ and for each layer ℓ, we compute a span representation by mean pooling over the span tokens:

$$\mathbf{h}^{(\ell)} = \frac{1}{b - a + 1} \sum_{t=a}^{b} H_t^{(\ell)}(x) \in \mathbb{R}^d. \tag{10}$$

Stacking representations across layers yields

$$\mathbf{H} = \left[\mathbf{h}^{(0)}, \ldots, \mathbf{h}^{(L)}\right] \in \mathbb{R}^{(L+1) \times d}. \tag{11}$$

Learnable scalar mixing. We learn layer-importance parameters $\boldsymbol{\alpha} \in \mathbb{R}^{L+1}$ and obtain normalized mixing weights $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{\alpha})$, where $\sum_\ell \pi_\ell = 1$. The mixed span representation is

$$\mathbf{s} = \sum_{\ell=0}^{L} \pi_\ell \mathbf{h}^{(\ell)} \in \mathbb{R}^d. \tag{12}$$

Probe classifier (2-layer MLP). The probing head is a binary classifier parameterized by $W_1, \mathbf{b}_1, W_2, b_2$:

$$\begin{aligned} \mathbf{u} &= \mathrm{ReLU}(W_1 \mathbf{s} + \mathbf{b}_1), \\ \tilde{\mathbf{u}} &= \mathrm{Dropout}(\mathbf{u}), \\ z &= W_2 \tilde{\mathbf{u}} + b_2, \end{aligned} \tag{13}$$

Case	Condition Type	Accuracy
$(S \mid C_{\text{spec}})$	Spec Ctx + Ambiguous	0.741
$(S \mid C_{\text{non}})$	Non-spec Ctx + Ambiguous	0.540
$(S_{\text{lex}} \mid C_{\text{spec}})$	Spec Ctx + Lexical Specific	**0.907**
$(S_{\text{lex}} \mid C_{\text{non}})$	Non-spec Ctx + Lexical Specific	**0.180**
$(S_{\text{pass}} \mid C_{\text{spec}})$	Spec Ctx + Passive Specific	**0.928**
$(S_{\text{pass}} \mid C_{\text{non}})$	Non-spec Ctx + Passive Specific	**0.295**

Table 14: Single-case edge probing accuracy in Mistral-7B. Higher accuracy indicates stronger linear separability of specificity.

Pair	PR-AUC	ROC-AUC	Acc.	F1
Baseline (Random)	0.5012	0.5059	0.5108	0.6421
$(S \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.6677	0.6935	0.6403	0.6732
$(S_{\text{lex}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.8503	0.8479	0.7230	0.7660
$(S_{\text{pass}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	**0.8758**	**0.8778**	**0.7338**	**0.7771**
$(S \mid C_{\text{spec}})$ vs. $(S_{\text{lex}} \mid C_{\text{non}})$	**0.5519**	**0.5931**	**0.5576**	**0.6045**

Table 15: Pairwise edge probing performance in Mistral-7B. Each row trains a binary classifier on span representations for the two cases in the pair.

where z is the logit and

$$\hat{p} = \sigma(z) \tag{14}$$

is the predicted probability of specificity.

Training objective. The probe parameters θ and the mixing parameters $\boldsymbol{\alpha}$ are jointly optimized by minimizing the binary cross-entropy loss over the specificity dataset:

$$\mathcal{L}_{\text{spec}} = \mathbb{E}_{(x,y)\sim\mathcal{D}_{\text{spec}}}\left[-y \log \hat{p} - (1-y) \log(1-\hat{p})\right]. \tag{15}$$

Threshold selection and inference. A decision threshold $\tau \in (0, 1)$ is selected on a development split of $\mathcal{D}_{\text{spec}}$ by maximizing the F1 score. At inference time, the trained probe is applied to inputs from $\mathcal{D}_{\text{scope}}$ without further optimization, and $\hat{p}$ is used as a continuous *specificity score* rather than a discrete classification decision. The learned mixing weights $\boldsymbol{\pi} = \mathrm{softmax}(\boldsymbol{\alpha})$ are reported as the layerwise importance distribution.

E Model Variability in Experiment 2C: Mistral-7B-ver0.3 and Qwen2.5-7B

Comparison to LLaMA-3 (Main). Across all three models, edge probing achieves near-ceiling overall performance (ROC-AUC/PR-AUC ≈ 1.0), and the scalar-mixing analysis consistently assigns larger weights to *late* layers, indicating that scopal specificity is most salient toward the top of the

Category	Mean (%)	vs. Expected
Layer groups		
Early (L0–L10)	2.948	−2.7%
Middle (L11–L21)	2.961	−2.3%
Late (L22–L32)	**3.169**	**+4.6%**
Expected (uniform)	3.030	—
Top-5 layers		
1. Layer 32 (Late)	**3.286**	**+8.4%**
2. Layer 31 (Late)	3.284	+8.3%
3. Layer 30 (Late)	3.282	+8.3%
4. Layer 29 (Late)	3.275	+8.1%
5. Layer 28 (Late)	3.256	+7.5%
Range	0.191%	—
Non-uniform?	✓ Yes (>1.0% threshold)	

Table 16: Scalar mixing weights over Mistral-7B layers learned by the edge probe. As in LLaMA-3, the probe assigns the largest weights to late layers, indicating that specificity-related information is most strongly encoded toward the top of the network.

Case	Condition Type	Accuracy
$(S \mid C_{\text{spec}})$	Spec Ctx + Ambiguous	0.489
$(S \mid C_{\text{non}})$	Non-spec Ctx + Ambiguous	0.770
$(S_{\text{lex}} \mid C_{\text{spec}})$	Spec Ctx + Lexical Specific	**0.784**
$(S_{\text{lex}} \mid C_{\text{non}})$	Non-spec Ct + Lexical Specific	**0.381**
$(S_{\text{pass}} \mid C_{\text{spec}})$	Spec Ctx + Passive Specific	**0.813**
$(S_{\text{pass}} \mid C_{\text{non}})$	Non-spec Ctx + Passive Specific	**0.547**

Table 17: Single-case edge probing accuracy in Qwen2.5-7B. Higher accuracy indicates stronger linear separability of specificity.

network (Tables 16 and 19; cf. main-text LLaMA results).

Key differences. Mistral-7B most closely matches LLaMA-3 in exhibiting a clear late-layer skew with a non-uniform mixing profile ("Non-uniform?" ✓), whereas Qwen2.5-7B shows a noticeably flatter distribution ("Non-uniform?" ×), suggesting a more distributed (less top-heavy) encoding of specificity across layers. In addition, both models reproduce the same qualitative pattern in pairwise separability: the pair of $(S_{\text{pass}} \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$ is the strongest contrast, while the pair of $(S \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}})$ is consistently the weakest (Tables 15 and 18).

Pair	PR-AUC	ROC-AUC	Acc.	F1
Baseline (Random)	0.4684	0.4386	0.4640	0.4016
$(S \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.7176	0.7205	0.6295	0.5690
$(S_{\text{lex}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.8850	0.8816	0.7770	0.7786
$(S_{\text{pass}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	**0.8861**	**0.8871**	**0.7914**	**0.7958**
$(S \mid C_{\text{spec}})$ vs. $(S_{\text{lex}} \mid C_{\text{non}})$	**0.4559**	**0.4237**	**0.4353**	**0.4642**

Table 18: Pairwise edge probing performance in Qwen2.5-7B. Each row trains a binary classifier on span representations for the two cases in the pair.

Category	Mean (%)	vs. Expected
Layer groups		
Early (L0–L9)	3.334	−3.3%
Middle (L10–L18)	3.443	−0.1%
Late (L19–L28)	**3.578**	**+3.8%**
Expected (uniform)	3.448	—
Top-5 layers		
1. Layer 23 (Late)	**3.627**	**+5.2%**
2. Layer 24 (Late)	3.625	+5.1%
3. Layer 22 (Late)	3.617	+4.9%
4. Layer 21 (Late)	3.612	+4.8%
5. Layer 20 (Late)	3.592	+4.2%
Range	0.301%	—
Non-uniform?	× No (<1.0% threshold)	

Table 19: Scalar mixing weights over Qwen2.5-7B layers learned by the edge probe. While late layers still receive higher weights on average, the overall distribution is comparatively flatter than in LLaMA-3 and Mistral, indicating a more diffuse encoding of specificity across layers.

F Layerwise Logistic Regression Probing as an auxiliary probing of Edge Probing in Experiment 2C

F.1 Details of Layerwise Logistic Regression Probing

As an auxiliary analysis, we perform layerwise logistic regression probing to assess how strongly specificity is linearly decodable at each layer. The probes are trained on the labeled specificity dataset $\mathcal{D}_{\text{spec}}$ and applied independently at each layer.

Inputs. For each input sequence Xx with target span $[a, b]$ and for each layer $\ell \in \mathcal{L}$, we use the same span-mean representation $\mathbf{h}^{(\ell)}$ defined in the edge-probing module above.

Outputs. For each layer ℓ, a logistic regression probe produces a probability $\hat{p}^{(\ell)}$ indicating the likelihood that the span is specific. A decision threshold $\tau^{(\ell)}$ is selected on a development split of $\mathcal{D}_{\text{spec}}$ by maximizing the F1 score. At inference

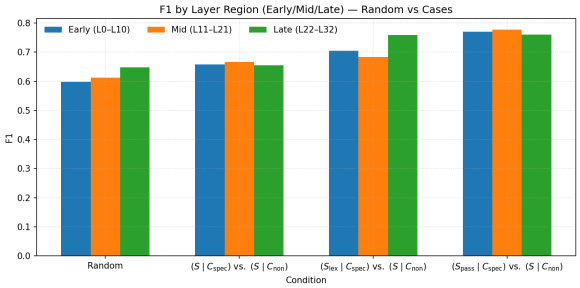

Figure 8: Layerwise Comparison Logistic Regression PR-AUC of LLaMA3-8B

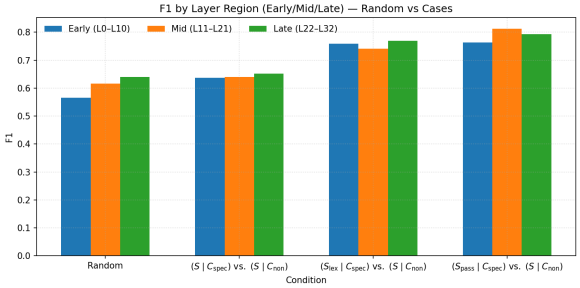

Figure 9: Layerwise Comparison Logistic Regression PR-AUC of Mistral-7B

time, the probes are applied to inputs from $\mathcal{D}_{\text{Scope}}$ without further optimization, and $\hat{p}^{(ell)}$ is used as a continuous specificity score.

F.2 Results of Layerwise Logistic Regression

Across LLAMA3, MISTRAL-7B, and QWEN2.5, the layerwise logistic-regression results closely mirror the trends observed in the Edge Probing analysis reported in the main section, Section 5.3. For the Random control, PR-AUC remains near chance level (approximately 0.5) across early, mid, and late layer regions for all models. This confirms the Edge Probing finding that, in the absence of meaningful interpretive structure, the probing classifiers fail to extract any reliable signal. By contrast, all designed sentence pairs achieve substantially higher PR-AUC values, indicating that the learned representations support effective linear separation when the contrast aligns with theoretically motivated distinctions.

Consistent with the Edge Probing results, the pragmatics-only contrast pair($(S \mid C_{\text{spec}})$ and $(S \mid C_{\text{non}})$) yields the weakest performance among the designed pairs, yet still clearly outperforms the Random baseline. This pattern suggests that specificity differences driven purely by contextual inference are encoded more diffusely and are therefore harder to recover than contrasts involving stronger sentence-internal cues, such as the pair of $(S \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{spec}})$ and the pair of $(S \mid C_{\text{spec}})$ and $(S_{\text{pass}} \mid C_{\text{spec}})$. The fact that this contrast remains distinguishable across all models reinforces the Edge Probing conclusion that pragmatic specificity is nevertheless systematically represented.

Finally, the layerwise distribution of PR-AUC closely aligns with the depth-wise trends observed in Edge Probing. Specificity-related information is not localized to a narrow set of layers but is distributed throughout the network. At the same time, performance consistently improves from early to mid layers and remains strong into late layers, with mid-layer representations often yielding the highest PR-AUC. This mid-to-late concentration echoes the Edge Probing finding that specificity becomes increasingly explicit and linearly accessible as representations transition from lexical encoding toward higher-level semantic abstraction.

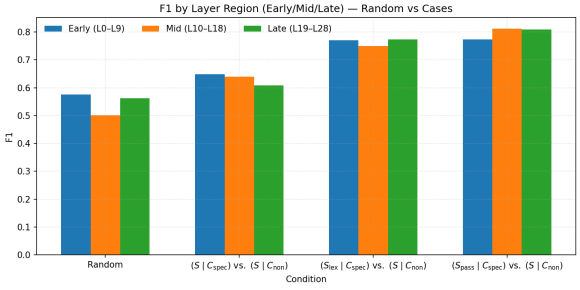

Figure 10: Layerwise Comparison Logistic Regression PR-AUC of Qwen2.5