

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

Jieun Kim
University of Ulsan
kimje@ulsan.ac.kr

Suhyeon Park
UNIST
shheny@unist.ac.kr

Kunhui Lee
UNIST
krdkghszkfhs@unist.ac.kr

Seung-Hoon Na[†]
UNIST
nash@unist.ac.kr

Abstract

Scope ambiguity arises when a single surface string admits multiple logical interpretations depending on how operators such as negation and indefinites interact. Prior work (Kamath et al., 2024) primarily evaluates scope sensitivity through output-level behavior, leaving open whether alternative interpretations are internally represented or merely reflected in shifted token probabilities. We introduce an *evidence-first, ambiguity-following* diagnostic that asks a more specific question: when the surface string is held constant and only contextual evidence biases interpretation, do transformer language models form *distinct internal representations corresponding to alternative logical readings*?

To investigate this, we present SCOPEX, extending the dataset of (Kamath et al., 2024) by systematically manipulating preceding context while keeping the ambiguous sentence identical. We further introduce passivized variants as syntactically less ambiguous controls. Using SCOPEX, we show that transformer LMs (i) exhibit systematic scope-sensitive behavior conditioned on context, and (ii) form separable hidden-state representations for alternative scope readings of the same surface string. This divergence emerges in later layers and causally influences downstream predictions.

1 Introduction

Scope ambiguity is a pervasive phenomenon in natural language, whereby a sentence admits multiple interpretations depending on how scopal items—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 7, consider the sentence “*I didn’t feed a dog*” with two interpretations:

*Code, datasets, and experimental materials are publicly available at <https://github.com/Jieun-NLP/ScopeX-Scope-Ambiguity.git>.

[†]Corresponding author.

Evidence-first diagnosis of scope in LLMs

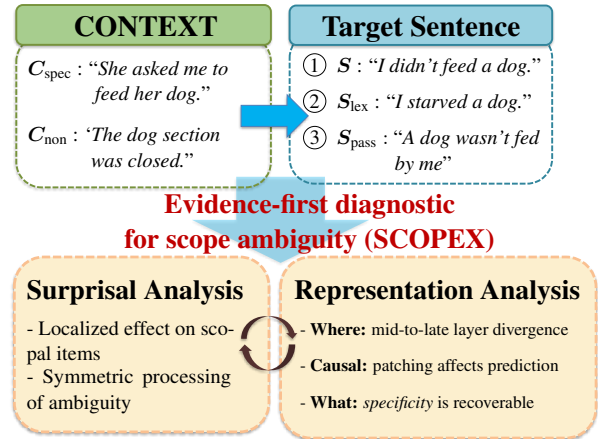


Figure 1: a high-level overview of our evidence-first diagnostic approach, summarizing the SCOPEX design and the roles of Experiments 1 and 2.

- *Specific*: “there exists a particular dog that was not fed.”
(inverse reading: $\exists x(DOG(x) \wedge \neg FEED(I, x))$)
- *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 ask **how modern transformer-based language models (LMs) resolve such ambiguity under contextual bias**.

Prior work (Kamath et al., 2024) adopts an *ambiguity-first, evidence-following diagnostic*: an ambiguous *preamble* is presented first, and sensitivity is measured via continuation probabilities. However, because autoregressive models condition on preceding tokens, this setup evaluates plausibility of continuations rather than directly examining how the internal representations of the ambiguous

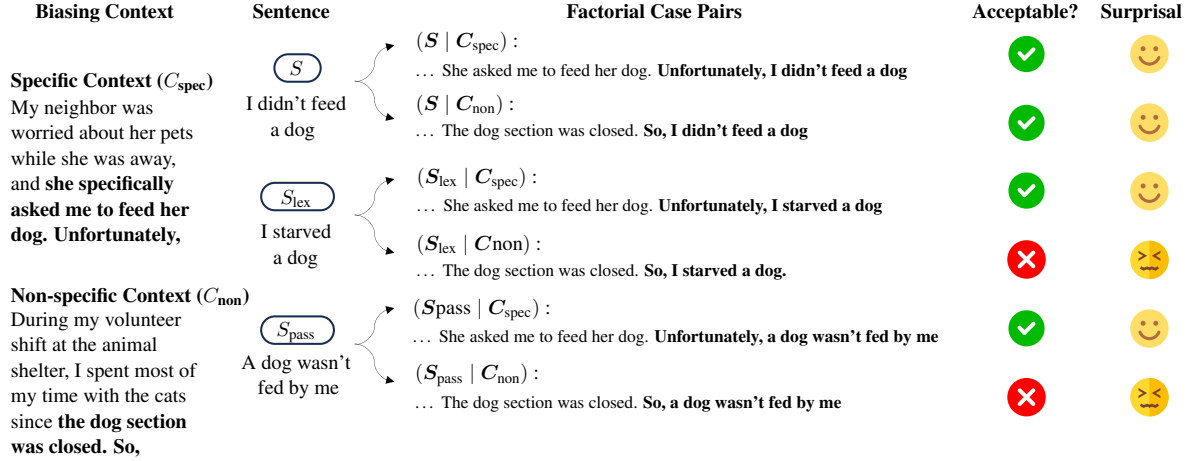


Figure 2: Overview of the SCOPEX dataset and Experiment 1. The figure illustrates the 2×3 factorial design, crossing contextual bias (specific vs. non-specific) with sentence type variations (ambiguous, lexical-semantically unambiguous, syntactically less ambiguous conditions).

sentence itself are shaped by context. We therefore adopt an **evidence-first, ambiguity-following diagnostic**. In this setup, contextual evidence precedes the scope-ambiguous sentence. As a result, the hidden states associated with the ambiguous sentence are adaptively determined by the preceding context, enabling *direct analysis of internal processing under contextual bias*. (see Table 10 in Appendix for a detailed comparison)

Operationalizing a theoretical question. In theoretical and psycholinguistics, a long-standing question concerns whether alternative logical interpretations are internally distinguished during comprehension (Altmann and Steedman, 1988; Tunstall, 1998). We adapt this question to transformer LMs in operational terms: when the surface string is held constant and only contextual evidence biases interpretation, do models exhibit systematic and causally relevant internal differentiation aligned with alternative logical readings?

To investigate this question, we construct SCOPEX (**SCOPE**-Expression Cross-factor Evaluation), extending Kamath et al. (2024) in two key ways. First, we introduce contextual evidence sentences preceding the ambiguous target. Second, we include syntactically less ambiguous passive variants alongside ambiguous and lexically disambiguated controls. This results in a 2×3 factorial design crossing contextual bias and sentence type (see Section 3.1 and Table 7 in Appendix).

Beyond trivial output adjustments. From an engineering perspective, one might argue that differ-

ences in output probabilities necessarily imply differences in hidden states at the final layer, rendering representational divergence unsurprising. Indeed, if scope resolution were implemented merely as a conditional language modeling adjustment, contextual variation could induce small vector shifts at the final layer, leading to logit changes without requiring structured internal differentiation.

However, our findings go beyond such minimal output-level effects. We observe (i) **systematic divergence emerging in mid-to-late layers**, (ii) **localization of representational differences to scopal items**, (iii) **causal influence of scope-conditioned representations on downstream prediction**, and (iv) **probing-based evidence** that the theoretically diagnostic property of *specificity* is internally represented. These convergent results indicate that alternative scope readings are tracked during forward computation rather than arising solely as final-layer probability adjustments.

2 Previous Research.

LMs and Ambiguity Recent work has examined whether large language models (LLMs) exhibit human-like 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 of-

ten 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 Overview. To support the evidence-first continuation diagnostic for scope ambiguity, we construct **SCOPEX** (SCOPE–Expression Cross-factor Evaluation), a controlled diagnostic dataset designed to probe whether LMs encode linguistically grounded scope distinctions in their representations. SCOPEX extends the dataset of Kamath et al. (2024) by introducing (i) explicit **contextual bias manipulation** and (ii) additional **sentence-level structural controls**. The dataset adopts a 2×3 **factorial design**, crossing two discourse contexts with three sentence types, yielding six systematically related conditions ($N = 139$ items per condition; 834 total) (See Table 7 and Table 8 in Appendix).

Notation. We use the notation $(S | C)$ to denote that a target sentence S is processed following a preceding context C . Contexts are written as C_{spec} and C_{non} , which bias interpretation toward inverse-scope (specific) and surface-scope (non-specific) readings, respectively. Target sentence variants are denoted as S (ambiguous), S_{lex} (lexically disambiguated), and S_{pass} (passivized).

Scope Operators and Lexical Coverage. We began with ambiguous sentences adapted from Kamath et al. (2024), but modified certain operators to ensure clearer scope interactions for probing. Example sentences and operator substitutions were constructed by a linguist. Importantly, the roles of scopal items are structurally balanced. Sentence-initial scopal items (SI1) predominantly host sen-

Operator Type	Count	%	SI1	SI2
Existential (a, some, several)	131	47.1	12	119
Frequency (always, often)	51	18.3	51	0
Negation (not, never)	41	14.7	41	0
Universal (every, all)	21	7.6	21	0
Proportional (most)	14	5.0	13	1
Numeral (one–twelve)	20	7.2	1	19

Table 1: Scope operator distribution in SCOPEX (278 scopal items per condition).

tential operators (e.g., negation), whereas the second scopal items (SI2) are primarily nominal operators (e.g., existential determiners). To support the specificity probing analysis, SI2 was systematically instantiated with existential-type operators, ensuring controlled variation in nominal scope interpretation. The resulting operator proportions are summarized in Table 1 (Also, see table 9 in Appendix).

Sentence Types and Structural Controls. Target sentences appear in three conditions:

- **Ambiguous (S):** canonical placement of two scopal items, permitting both surface and inverse scope readings.
- **Lexically disambiguated (S_{lex}):** one scopal item is replaced by a non-scopal item (e.g., *didn’t feed* $\rightarrow$ *starved*), removing ambiguity while preserving propositional content.
- **Passivized (S_{pass}):** syntactic reordering (e.g., object-to-subject promotion) introduces structural bias without lexical substitution.

This design isolates pragmatic, lexical-semantic, and syntactic contributions to scope interpretation.

Contextual Bias Generation. Preceding contexts (C_{spec} , C_{non}) were generated using GPT-4o via a controlled few-shot prompting procedure with a fixed prompt template conditioned on example pairs provided by a linguist (Figure 5). All generated contexts were manually verified and edited by the linguist to ensure that each pair clearly supported the intended wide- vs. narrow-scope interpretations and satisfied formatting constraints required for surprisal measurement. Further details of dataset construction and manual validation are provided in Appendix A.

3.2 Models

We evaluate **three transformer-based autoregressive LMs**: LLaMA-3 (Grattafiori et al.,

2024) (8B, Base), Mistral (Jiang et al., 2023) (7B, v0.3), and Qwen2.5 (Qwen et al., 2025) (7B). Due to space limitations, we present main results using **LLaMA-3-8B**. Full results for all models appear in the Appendix.

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 :

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

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:

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

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:

$$\Delta\mathcal{G} = \mathcal{G}(S | C) - \mathcal{G}(S). \quad (3)$$

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

Mean Surprisal (bits)			
Case	In-situ	Isolated	Diff.
$(S C_{\text{spec}})$	4.748	6.784	−2.036
$(S C_{\text{non}})$	4.812	6.785	−1.973
$(S_{\text{lex}} C_{\text{spec}})$	6.496	7.524	−1.028
$(S_{\text{lex}} C_{\text{non}})$	6.712	7.534	−0.822
$(S_{\text{pass}} C_{\text{spec}})$	5.603	6.933	−1.330
$(S_{\text{pass}} C_{\text{non}})$	5.817	6.944	−1.127
Pairwise Statistical Comparisons			
Pair	Mean $_{\Delta}$	p	d_z
$(S C_{\text{spec}}) - (S C_{\text{non}})$	−0.062	.277	−0.10
$(S_{\text{lex}} C_{\text{spec}}) - (S_{\text{lex}} C_{\text{non}})$	−0.204	.001**	−0.29
$(S_{\text{pass}} C_{\text{spec}}) - (S_{\text{pass}} C_{\text{non}})$	−0.203	.001**	−0.28

Table 2: Mean surprisal values and pairwise comparisons. Diff. = In-situ − Isolated (contextual facilitation). Mean $_{\Delta}$ = difference in contextual effects between cases. d_z = paired-samples Cohen’s d , computed from itemwise differences. 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
$(S C_{\text{spec}}), (S C_{\text{non}})$	0.843	0.790	−0.053	0.94×
$(S_{\text{lex}} C_{\text{spec}}), (S_{\text{lex}} C_{\text{non}})$	1.162	1.395	+0.233	1.20×
$(S_{\text{pass}} C_{\text{spec}}), (S_{\text{pass}} C_{\text{non}})$	1.020	1.399	+0.379	1.37×

Table 3: Sensitivity of scopal items. Mean $|\Delta|$ (bits) for scopal items vs. non-scopal items. Ratio = Scope/Non-scope effect magnitude.

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

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

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. These analyses quantify both the global processing cost associated with contextual disambiguation and the local contribution of scopal items to interpretive divergence.

4.2 Results and Discussion

Context-driven interpretive preference Table 2 reports mean surprisal values for in-situ and isolated conditions of LLaMA-3-8B. 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 scopal items Table 3 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, scopal items exhibit substantially larger mean $|\Delta_j|$ than non-scopal items ($1.20\times$ and $1.37\times$, 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.

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. Results for other models, Mistral-7B-v0.3 and Qwen2.5-7B, are reported in Appendix B, showing partial qualitative convergence, particularly in the localization of contextual effects to scopal items, but with model-dependent differences in sentence-level effect magnitude and statistical reliability.

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

Experiment 2 examines how contextual evidence shapes the internal representations of scope-ambiguous sentences. We investigate this 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 representations?

5.1 Experiment 2A: Layerwise Representation Similarity

In this subsection, we ask whether internal sentence representations are organized primarily by surface realization or by contextual interpretation. We compare a surface-identical but interpretation-different pair with two surface-varying but interpretation-matched pairs, asking which source of variation induces greater representational divergence across layers.

5.1.1 Method

To quantify representational divergence across layers, we compare sentence-level representations using two complementary similarity measures. Let $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{n \times d}$ 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 two measures $\text{sim}(\mathbf{X}, \mathbf{Y})$: **cosine similarity** and **CKA**(Kornblith et al., 2019). For comparison, we additionally construct a random baseline by pairing each $(S \mid C_{\text{spec}})$ sentence with a randomly sampled $(S \mid C_{\text{spec}})$ sentence from a different item and computing the same similarity measures. Implementation details are provided in Appendix C.2.

5.1.2 Results and Discussion

Figure 3 reports representational similarity across all three models. Sentence representations exhibit the characteristic high–low–high trajectory observed in previous studies (Tenney et al., 2019; Liu et al., 2019; Peters et al., 2018; Jawahar et al., 2019). The random-sentence baseline (gray) remains well separated from all target pairs throughout, confirming that the analysis captures fine-grained representational differences among semantically related sentences.

In the early layers, the surface-identical pair, $(S \mid C_{\text{spec}})$ versus $(S \mid C_{\text{non}})$ (red), is consistently more similar than the two surface-varying pairs, indicating that sentence representations are initially organized primarily by lexical identity and shallow syntactic structure. Beginning in the middle layers, however, this ordering reverses consistently across models: the similarity of the surface-identical pair

Layer-wise Representation Similarity (Sentence Means)

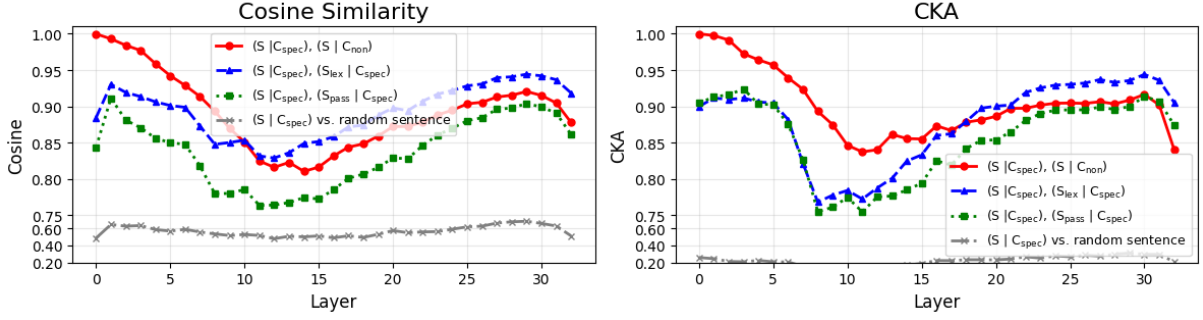


Figure 3: Layer-wise comparison of (1) Cosine Similarity and (2) CKA.

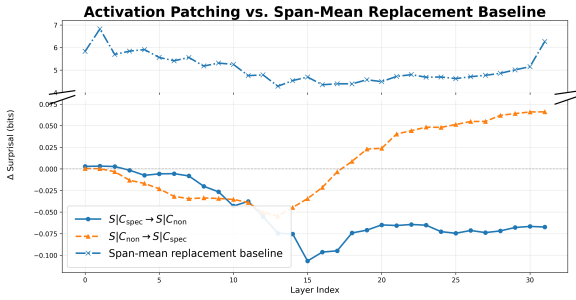


Figure 4: Layer-wise hidden-state patching effects for the $(S | C_{\text{spec}})$ and $(S | C_{\text{non}})$ pair in LLaMA3-8B.

falls below that of the lexical-variation pair (blue), despite differing only in preceding contextual interpretation. Thus, changing contextual interpretation while keeping the surface realization unchanged induces greater representational divergence than changing surface lexical or syntactic realization while preserving the intended interpretation. This pattern persists through the final layers, although the transition occurs somewhat later in LLaMA-3 than in Mistral and Qwen. The same qualitative ordering is observed across all three architectures, while model-specific behavior of the passive-voice pair (green) is discussed in Appendix C.1.

5.2 Experiment 2B: Activation Patching

Experiment 2B examines whether the 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 representations encoded in the post-MLP residual stream at a given layer causally influence sentence-level surprisal when transferred across scope interpretations. Unlike fine-grained causal tracing approaches that intervene on individ-

ual neurons, attention heads, or circuits (Vig et al., 2020; Meng et al., 2022), our intervention replaces the entire layer output and therefore provides a conservative test of when layerwise representations begin to contribute to downstream prediction.

5.2.1 Method

Intervention setup Given a source example E_{src} and a target example E_{tgt} , we replace the post-MLP residual-stream representations 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 $[s, e]$. As a baseline, we additionally perform span-mean replacement, replacing all token-level representations in the same sentence span with the span-wise mean representation, $\bar{h}^{(\ell)} = \frac{1}{e-s+1} \sum_{j=s}^e h_j^{(\ell)}$.

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

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

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

5.2.2 Results and Discussion

Figure 4 reports layerwise surprisal changes induced by activation patching together with the span-mean replacement baseline. Across layers, the baseline exhibits a qualitatively distinct profile, indicating that the observed effects cannot be reduced to non-specific perturbation of the residual

stream. Instead, transferring representations between the two contextual conditions systematically changes surprisal, providing causal evidence that the representations identified in Experiment 2A are functionally utilized to prediction. Patching effects remain modest through approximately Layer 12 and more pronounced in later layers, suggesting that their influence strengthens toward the model’s final predictions.

Notably, LLaMA-3 exhibits a directional asymmetry between the two patching directions. Injecting $(S \mid C_{\text{non}})$ representations into $(S \mid C_{\text{spec}})$ contexts yields increasingly positive surprisal changes, whereas the reverse intervention produces comparatively modest negative effects. As this asymmetry is not consistently observed across models, it appears to reflect a model-specific feature of LLaMA-3, motivating Experiment 2C, which investigates whether it is associated with differential encoding of semantic *specificity*. Cross-model comparisons with Mistral and Qwen are provided in Appendix D.

5.3 Experiment 2C: Probing for Scopal Specificity

Experiment 2C asks **what semantic distinction is recoverable from the layerwise representations examined in Experiments 2A and 2B**. In formal semantics, *specificity* distinguishes noun phrases interpreted as referring to particular entities from those receiving non-specific interpretations (Abusch, 1993; Von Heusinger, 2019). ***Specificity is closely related to scope***: wide-scope existential expressions generally receive specific interpretations, whereas narrow-scope expressions tend to receive non-specific interpretations. We therefore test whether the hidden representations distinguish these two interpretive classes. Scalar-mixing edge probing estimates their global layerwise distribution, while independent layerwise logistic-regression probes measure their linear recoverability at individual layers.

5.3.1 Method

Dataset construction. We construct a dedicated SPECIFICITY dataset consisting of 1,998 context–sentence items, comprising 997 specific and 1,001 non-specific instances. The dataset was constructed around specific–non-specific contrasts, with corresponding items sharing the same target NP but differing in whether the context–sentence combination licenses a specific or non-specific interpre-

Case	Condition Type	Accuracy
$(S \mid C_{\text{spec}})$	Spec Ctx + Ambiguous	0.734
$(S \mid C_{\text{non}})$	Non-spec Ctx + Ambiguous	0.554
$(S_{\text{lex}} \mid C_{\text{spec}})$	Spec Ctx + Lexical Specific	0.899
$(S_{\text{lex}} \mid C_{\text{non}})$	Non-spec Ctx + Lexical Specific	0.201
$(S_{\text{pass}} \mid C_{\text{spec}})$	Spec Ctx + Passive Specific	0.957
$(S_{\text{pass}} \mid C_{\text{non}})$	Non-spec Ctx + Passive Specific	0.273

Table 4: Single-case specificity-probe accuracy in LLaMA-3-8B. Accuracy indicates how reliably each condition is classified according to its expected *specific* or *non-specific* interpretation.

tation. Each item includes a discourse context, a sentence, a marked target NP span, and a binary *specificity* label. For example, *During the camping trip last weekend, we looked up at the sky. I spotted **seven stars** that were unusually bright* licenses a specific interpretation of *seven stars*, whereas *On a clear night, you can often see stars shining brightly in the sky. If you count **seven stars**, you might find a constellation* licenses a non-specific interpretation. A linguist manually validated a random 20% sample (400 items), assessing whether the target-span *specificity* label was correct and whether the context–sentence pair was discourse-coherent. Further examples are provided in Appendix E.

Edge probing with scalar mixing Our primary analysis adopts a scalar-mixing edge probing framework (Tenney et al., 2019). A probe is trained on the labeled SPECIFICITY dataset and then applied to hidden representations extracted from the original SCOPEX dataset, allowing us to measure whether the semantic property of *specificity* is recoverable from the scope representations analyzed in Experiments 2A and 2B. The probe operates on span-mean representations of the target span, defined as the second scopal item (e.g., *a dog*) in our SCOPEX dataset, stacking layerwise span representations into a tensor $\mathbf{X} \in \mathbb{R}^{N \times L \times d}$ and learning a softmax-weighted combination of layers. Decision thresholds are selected per model by maximizing F1 on a held-out 20% validation split of SPECIFICITY dataset and are then applied to SCOPEX without further optimization. To assess probe selectivity, we follow Hewitt and Liang (2019), and compare performance on SCOPEX against matched control dataset (details in Appendix F).

Auxiliary probe. As an auxiliary analysis, we additionally train independent layerwise logistic regression probes (Alain and Bengio, 2018) to verify

Pair	PR-AUC	ROC-AUC	Acc.	F1
Baseline (Random)	0.4785	0.4684	0.5	0.6213
$(S \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.6697	0.6952	0.6439	0.6733
$(S_{\text{lex}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.8461	0.8481	0.7266	0.7669
$(S_{\text{pass}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.8947	0.9008	0.7554	0.7964
$(S \mid C_{\text{spec}})$ vs. $(S_{\text{lex}} \mid C_{\text{non}})$	0.4300	0.3928	0.4676	0.5795

Table 5: Pairwise edge probing performance in LLaMA-3-8B.

Category	Mean (%)	vs. Expected
<i>Layer groups</i>		
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	—
<i>Top-5 layers</i>		
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 6: Scalar mixing weights over LLaMA-3-8B layers learned by the edge probe.

that the observed layerwise trends are not specific to the scalar-mixing architecture; full details and results are reported in the Appendix H.

5.3.2 Results and Discussion

Asymmetry between *specificity* and *non-specificity* In LLaMA-3, Table 4 reveals an asymmetry between the two readings. Specific interpretations yield reliably high accuracy ($(S \mid C_{\text{spec}})$: **0.73**; $(S_{\text{lex}} \mid C_{\text{spec}})$: **0.90**; $(S_{\text{pass}} \mid C_{\text{spec}})$: **0.96**), indicating that *specificity* is more reliably recoverable from the model’s representations when lexical or syntactic cues unambiguously support the specific interpretation. By contrast, the non-specific cases show lower accuracy, but for distinct linguistic reasons. For the $(S \mid C_{\text{non}})$ case (0.55), the probe’s difficulty in isolating a single *non-specific* class may partly reflect the semantic heterogeneity of non-specific interpretations, as expected from their theoretical semantics consistent with Degano and Aloni (2025), rather than representational noise alone. This interpretation, however, remains a hypothesis to be tested directly in the future work. In contrast, the $(S_{\text{lex}} \mid C_{\text{non}})$ case (0.20) and the

$(S_{\text{pass}} \mid C_{\text{non}})$ case (0.27) 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 appears to make it difficult for the model to stabilize a consistent encoding.

Graded hierarchy: syntactic > lexical > contextual Table 5 evaluates the **discriminability** between readings using pairwise comparisons of the frozen specificity-probe scores. These comparisons 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: **0.89**), 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 falls below the baseline (PR-AUC: **0.43**), consistent with the view that *specificity* encoding becomes robust when multiple linguistic signals converge. This relative hierarchy is also observed in Mistral-7B and Qwen2.5-7B, despite important model-specific differences in the single-condition asymmetry (Appendix G).

Layerwise distribution Scalar-mixing weights (Table 6) 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 H).

Experiments 2B and 2C in LLaMA-3 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 how language models handle scope ambiguity. The combination of behavioral measurements, representational analyses, and probing results suggests that scope ambiguity provides a useful diagnostic setting for studying how language models organize information relevant to semantic interpretation. In particular, probing reveals that *specificity*—a property closely tied to scope in linguistic theories of implicit logical structure—is recoverable from model representations. This indicates that models encode signals associ-

ated with distinct scope-induced structures, including semantic properties that are not overtly marked in the surface form.

Overall, our findings show that transformer LMs exhibit consistent and interpretation-sensitive internal differentiation aligned with alternative scope readings. Understanding how such representations arise, how they vary across model architectures, and how they contribute to downstream reasoning remains an important direction for future work.

Limitations

Despite these encouraging findings, several limitations remain. (1) **Weak purely contextual disambiguation.** Models show relatively modest 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.** The precise organization of scope-related representations varies across models. In particular, the asymmetric pattern observed in LLaMA-3-8B—where *specificity* emerges as a comparatively marked semantic category—is not consistently reproduced in Mistral-7B and Qwen2.5-7B. Interestingly, however, this cross-model variability is itself consistent across Experiments 2B and 2C: the LLaMA-3-specific asymmetry observed under activation patching is accompanied by a corresponding asymmetry in probing, whereas Mistral and Qwen do not exhibit the same pattern in either analysis. Thus, rather than treating cross-model differences simply as a weaker or less stable semantic encoding, the within-model convergence across two distinct diagnostics raises the possibility that model families organize the same semantic distinction differently. Detailed cross-model results are provided in Appendix G. Future work should investigate what architectural, training, or representational factors give rise to these alternative organizations.

Ethics Statement

This work uses manually constructed linguistic examples and pretrained language models. No human

participants, personal data, or sensitive information were involved in this study. Our experiments are intended as diagnostic analyses of language model behavior, and we caution against interpreting the results as evidence of human-like semantic or logical understanding. Since the authors are non-native English speakers, AI-assisted grammar-level editing assistance was used to improve writing clarity.

Acknowledgments

This work was supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Korea government (MSIT) (No. RS-2020-II201336, Artificial Intelligence Graduate School Support (UNIST)).

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.
- Gerry Altmann and Mark Steedman. 1988. [Interaction with context during human sentence processing](#). *Cognition*, 30(3):191–238.
- 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?](#) *Linguistics and Philosophy*, 48(5):955–1003.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- 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, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *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.
- Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. [We’re afraid language models aren’t modeling ambiguity](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.
- 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.
- 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. *Advances in Neural Information Processing Systems*, 36. ArXiv:2202.05262.
- 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.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Claude E. 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, Kyle Rawlins, and Benjamin Van Durme. 2024. [Zero and few-shot semantic parsing with ambiguous inputs](#). In *The Twelfth International Conference on Learning Representations*.
- 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.
- Susanne L. Tunstall. 1998. *The Interpretation of Quantifiers: Semantics and Processing*. Phd dissertation, University of Massachusetts Amherst.
- 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 Jeanette Gundel and Barbara Abbott, editors, *The Oxford Handbook of Reference*. Oxford University Press.

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

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

ID 8 — scopal item1: “ocationally”, scopal item2: “two dentists”		
($S \mid C_{\text{spec}}$)	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 .
($S \mid C_{\text{non}}$)	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 .
($S_{\text{lex}} \mid C_{\text{spec}}$)	(Whenever I have a dental issue,...) Same as C_{spec} .	I consulted two dentists .
($S_{\text{lex}} \mid C_{\text{non}}$)	(Living in a small town,...) Same as C_{non} .	I consulted two dentists .
($S_{\text{pass}} \mid C_{\text{spec}}$)	(Whenever I have a dental issue,...) Same as C_{spec} .	Two dentists are occasionally consulted.
($S_{\text{pass}} \mid C_{\text{non}}$)	(Living in a small town,...) Same as C_{non} .	Two dentists are occasionally consulted.
ID 13 — scopal item1: “n’t”, scopal item2: “an error”		
($S \mid C_{\text{spec}}$)	My colleague pointed out that there was an issue in the Python script I submitted yesterday. Surprisingly,	I didn’t notice an error in my code
($S \mid C_{\text{non}}$)	I ran my program several times, and everything seemed to be working perfectly without any hiccups.	I didn’t notice an error in my code
($S_{\text{lex}} \mid C_{\text{spec}}$)	(My colleague...) Same as C_{spec} .	I overlooked an error in my code
($S_{\text{lex}} \mid C_{\text{non}}$)	(I ran my program...) Same as C_{non} .	I overlooked an error in my code
($S_{\text{pass}} \mid C_{\text{spec}}$)	(My colleague...) Same as C_{spec} .	An error in my code wasn’t noticed.
($S_{\text{pass}} \mid C_{\text{non}}$)	(I ran my program...) Same as C_{non} .	An error in my code wasn’t noticed.
ID 27 — scopal item1: “all”, scopal item2: “a few boxes of pizza”		
($S \mid C_{\text{spec}}$)	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 .
($S \mid C_{\text{non}}$)	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 .
($S_{\text{lex}} \mid C_{\text{spec}}$)	(At the end of the conference...) Same as C_{spec} .	All urban planners eat just these few boxes of pizza.
($S_{\text{lex}} \mid C_{\text{non}}$)	(When discussing the late-night work culture...) Same as C_{non} .	All urban planners eat just these few boxes of pizza.
($S_{\text{pass}} \mid C_{\text{spec}}$)	(At the end of the conference...) Same as C_{spec} .	a few boxes of pizza are eaten by all urban planners.
($S_{\text{pass}} \mid C_{\text{non}}$)	(When discussing the late-night work culture...) Same as C_{non} .	a few boxes of pizza are eaten by all urban planners.
ID 112 — scopal item1: “frequently”, scopal item2: “a few books”		
($S \mid C_{\text{spec}}$)	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.
($S \mid C_{\text{non}}$)	Whenever I visit the library, I aim to explore different genres and authors.	I frequently read a few books from the library.
($S_{\text{lex}} \mid C_{\text{spec}}$)	(I have a favorite set...) Same as C_{spec} .	I frequently read these few books from the library
($S_{\text{lex}} \mid C_{\text{non}}$)	(Whenever I visit the library...) Same as C_{non} .	I frequently read these few books from the library
($S_{\text{pass}} \mid C_{\text{spec}}$)	(I have a favorite set...) Same as C_{spec} .	A few books from the library are frequently read by me.
($S_{\text{pass}} \mid C_{\text{non}}$)	(Whenever I visit a library...) Same as C_{non} .	A few books from the library are frequently read by me.

A Dataset Construction Details

The SCOPEX dataset was constructed through a combination of LLM-assisted generation and manual linguistic validation.

Base Sentences and Operator Design. We began with ambiguous sentences adapted from the dataset of Kamath et al. (2024). However, their examples were not always suitable for the probing setup used in our experiments, which requires a controlled interaction between two scope-bearing items (scopal items). In particular, the second operator position (scopal item2) needed to consistently instantiate operators that participate in existential or numeral scope relations.

To achieve this, some universal quantifiers (e.g., *every*, *all*) appearing in the original dataset were replaced with existential determiners or numerals. These substitutions were performed by a linguist with over 20 years of experience in linguistic research, ensuring that the resulting sentences remained natural and semantically coherent while preserving the intended scope ambiguity.

Context Generation with GPT-4o. Preceding contexts were generated using GPT-4o with a fixed prompt template (Figure 5). The model was given example pairs consisting of an ambiguous sentence and two short contexts illustrating the specific (wide-scope) and non-specific (narrow-scope) readings. Conditioned on these examples, GPT-4o generated context pairs for new target sentences.

Manual Validation and Editing. All generated contexts were manually reviewed by the linguist. This step was necessary because a substantial portion of the raw generations did not reliably produce clear disambiguation between the two readings. In some cases, both contexts inadvertently supported the same interpretation, or the generated text included unintended continuations despite explicit instructions to provide only preceding context.

The validation process therefore involved verifying that each context pair unambiguously supported the intended scope reading, removing unintended continuations, and correcting formatting issues.

Additionally, because surprisal measurements were computed only over the target ambiguous sentence, the formatting of the context– sentence boundary had to be carefully controlled. For example, target sentences were lowercased when appearing as continuations within a sentence, and

capitalized when starting a new sentence after a period.

Through this combination of LLM-assisted generation and systematic manual verification, we obtained the final SCOPEX dataset consisting of 139 base items with six experimental conditions each (834 total sentence instances).

Operator Type	Count	%	Lexical Items	Description
Existential	131	47.1	a, an, some, a few, several, someone, something, somebody	Existential / indefinite determiners
Frequency	51	18.3	always, usually, often, frequently, sometimes, occasionally, generally, mostly	Frequency adverbs
NEG	41	14.7	not, n't	Sentential negation
Universal	21	7.6	every, each, all	Universal quantifiers
Proportional	14	5.0	most, most questions	Proportional quantifier
Num	20	7.2	one–twelve (digit and word form)	Cardinal numerals

Table 9: Distribution and taxonomy of scope operators in the dataset. Percentages are computed over 278 total scopal items (Case0 condition).

Prompt for Context Generation

System:

You are a writer who identifies the possible interpretations of an ambiguous sentence and creates natural and logically coherent contexts to disambiguate it.

User:

Here is an ambiguous sentence. It can be interpreted in two ways – one assuming a specific reference and another indicating a general, existential meaning or even non-existence.

Your task is to write two distinct contexts that naturally and coherently support each respective reading of the sentence.

Please make each context short and concise (1–2 sentences).
Provide only the preceding context (to the left).
Do not include any continuation.

Label the two contexts exactly as:

[Context 1: specific/wide-scope reading]

...

[Context 2: non-specific/narrow-scope reading]

...

Sentence:

I didn't feed a dog.

{Example Outputs}

Sentence: I didn't feed a dog.

[Context 1: specific/wide-scope reading]

My uncle asked me to take care of his dog while he was traveling.
However, I didn't feed a dog.

[Context 2: non-specific/narrow-scope reading]

When I visited the new animal shelter, there were no dogs there.
So, I didn't feed a dog.

Figure 5: Prompt template and example outputs used for generating preceding contexts.

B 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 preserv-

ing core qualitative patterns. We applied the same statistical significance threshold ($p < .01$) for comparability across models.

Table 10: 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	• <i>evidence-first, ambiguity-following: Context + Sentence</i>	• <i>ambiguity-first, evidence-following: Sentence + Continuation</i>
Experimental methodology	• Surprisal-based analysis • Layerwise representation similarity (Cosine Similarity/ CKA) • Causal intervention (activation patching) • Probing for scopal specificity	• Human–LM alignment analysis • Probability comparison of continuations
Research question	• <i>When disambiguating context precedes the sentence, are scope-sensitive interpretations reflected in the internal representations of language models?</i> • <i>If so, are these representations localized to particular layers or distributed across models?</i> • <i>Do these encodings causally affect the model behavior?</i>	• <i>Can LLMs detect and respond to scope ambiguity?</i> • <i>Do LLMs show human-like default preferences?</i>

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

Mistral-7B broadly replicates LLaMA-3-8B’s core pattern, although the magnitude and statistical reliability of the effects vary across conditions (Table 11). Most clearly, Mistral-7B shows significant scope-driven modulation in the lexically unambiguous pair (the pair of $(S_{\text{lex}} \mid C_{\text{spec}})$ and $(S_{\text{lex}} \mid C_{\text{non}})$) ($\text{Mean}_{\Delta} = -0.279$ bits, $p < .001$, $d_z = -0.30$) and localizes larger surprisal effects to scopal items ($1.24\times$ ratio; Table 12), closely paralleling LLaMA-3-8B’s $1.20\times$ ratio.

The passive condition also shows a modest concentration of contextual effects at scopal items ($1.05\times$ ratio), although its overall pairwise difference is small and not statistically significant ($\text{Mean}_{\Delta} = -0.024$ bits, $p = .860$, $d_z = -0.03$). The ambiguous condition likewise shows no reli-

able pairwise difference ($\text{Mean}_{\Delta} = +0.031$ bits, $p = .543$, $d_z = +0.03$), consistent with the relatively weak context-only contrast observed in LLaMA-3-8B.

Mistral-7B provides the clearest convergence with LLaMA-3-8B in the lexically disambiguated condition, where contextual effects are both statistically reliable and concentrated at scopal items. The weaker passive and ambiguous contrasts indicate that the strength of this sensitivity depends on the form of disambiguating information rather than generalizing uniformly across conditions.

Mean Surprisal (bits)			
Case	In-situ	Isolated	Diff.
$(S C_{\text{spec}})$	4.230	6.336	-2.106
$(S C_{\text{non}})$	4.166	6.304	-2.137
$(S_{\text{lex}} C_{\text{spec}})$	4.843	7.100	-2.256
$(S_{\text{lex}} C_{\text{non}})$	5.050	7.027	-1.977
$(S_{\text{pass}} C_{\text{spec}})$	4.975	6.650	-1.676
$(S_{\text{pass}} C_{\text{non}})$	5.023	6.675	-1.652
Pairwise Statistical Comparisons			
Pair	Mean $_{\Delta}$	p	d_z
$(S C_{\text{spec}}) - (S C_{\text{non}})$	+0.031	.543	+0.03
$(S_{\text{lex}} C_{\text{spec}}) - (S_{\text{lex}} C_{\text{non}})$	-0.279	< .001***	-0.30
$(S_{\text{pass}} C_{\text{spec}}) - (S_{\text{pass}} C_{\text{non}})$	-0.024	.860	-0.03

Table 11: Mean surprisal values and pairwise comparisons for Mistral-7B. Diff. = In-situ – Isolated (contextual facilitation). Mean $_{\Delta}$ = difference in contextual effects between cases. d_z = paired-samples Cohen’s d , computed from itemwise differences. 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
$(S C_{\text{spec}}), (S C_{\text{non}})$	1.592	1.477	-0.115 0.93×
$(S_{\text{lex}} C_{\text{spec}}), (S_{\text{lex}} C_{\text{non}})$	1.754	2.174	+0.420 1.24×
$(S_{\text{pass}} C_{\text{spec}}), (S_{\text{pass}} C_{\text{non}})$	1.511	1.593	+0.082 1.05×

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

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

Qwen2.5-7B shows evidence of scopal-item sensitivity in the localization analysis (Table 14). Scopal items exhibit enhanced contextual effects in the lexically unambiguous (1.21× ratio) and passive (1.19× ratio) conditions, whereas the ambiguous condition shows an approximately uniform distribution (1.00× ratio). This localization pattern is qualitatively similar to that observed in LLaMA-3-8B and Mistral-7B.

In contrast, the aggregate pairwise surprisal comparisons are substantially weaker in Qwen2.5-7B (Table 13). None of the three contrasts reaches statistical significance under the $p < .01$ threshold. The ambiguous pair shows the largest numerical difference (Mean $_{\Delta}$ = +0.165 bits, p = .070, d_z = +0.16), whereas the lexically unambiguous pair (Mean $_{\Delta}$ = -0.070 bits, p = .674, d_z = -0.06) and the passive pair (Mean $_{\Delta}$ = -0.087

Mean Surprisal (bits)			
Case	In-situ	Isolated	Diff.
$(S C_{\text{spec}})$	4.650	6.574	-1.923
$(S C_{\text{non}})$	4.481	6.569	-2.089
$(S_{\text{lex}} C_{\text{spec}})$	5.412	7.449	-2.037
$(S_{\text{lex}} C_{\text{non}})$	5.502	7.469	-1.967
$(S_{\text{pass}} C_{\text{spec}})$	5.324	6.801	-1.477
$(S_{\text{pass}} C_{\text{non}})$	5.434	6.824	-1.390
Pairwise Statistical Comparisons			
Pair	Mean $_{\Delta}$	p	d_z
$(S C_{\text{spec}}) - (S C_{\text{non}})$	+0.165	.070	+0.16
$(S_{\text{lex}} C_{\text{spec}}) - (S_{\text{lex}} C_{\text{non}})$	-0.070	.674	-0.06
$(S_{\text{pass}} C_{\text{spec}}) - (S_{\text{pass}} C_{\text{non}})$	-0.087	.593	-0.10

Table 13: Mean surprisal values and pairwise comparisons for Qwen2.5-7B. Diff. = In-situ – Isolated (contextual facilitation). Mean $_{\Delta}$ = difference in contextual effects between cases. d_z = paired-samples Cohen’s d , computed from itemwise differences. 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
$(S C_{\text{spec}}), (S C_{\text{non}})$	1.848	1.843	-0.006 1.00×
$(S_{\text{lex}} C_{\text{spec}}), (S_{\text{lex}} C_{\text{non}})$	2.051	2.482	+0.430 1.21×
$(S_{\text{pass}} C_{\text{spec}}), (S_{\text{pass}} C_{\text{non}})$	1.770	2.114	+0.344 1.19×

Table 14: Sensitivity of scopal items for Qwen2.5-7B. Mean $|\Delta|$ (bits) for scopal items vs. non-scopal items. Ratio = Scope/Non-scope effect magnitude

bits, p = .593, d_z = -0.10) show only small pairwise effects.

Thus, Qwen2.5-7B presents a more differentiated picture than a simple attenuated replication of the other models. The localization analysis indicates that contextual effects are preferentially concentrated at scopal items in the lexically and syntactically disambiguated conditions, but this localization does not translate into reliable condition-level differences in overall contextual facilitation. This dissociation suggests that sensitivity to where scope-relevant effects occur can be present even when their aggregate magnitude is weak or variable across conditions.

Across models, the relatively similar scope-element ratios in the lexically disambiguated condition (approximately 1.20×–1.24×) provide the clearest cross-model convergence in Experiment 1. At the same time, the differences in pairwise statistical reliability caution against treating scope

sensitivity as quantitatively uniform across model architectures.

C Model Variability in Experiment 2A: Mistral-7B-v0.3 and Qwen2.5-7B

C.1 Cross-Model Comparison of Representation Similarity

Figures 6 and 7 present the corresponding representation similarity analyses for Mistral-7B-v0.3 and Qwen2.5-7B.

Both models exhibit the same qualitative pattern reported in the main text. Representations follow the characteristic high–low–high trajectory, while the random baseline remains well separated from all target pairs across layers. Most importantly, the relative ordering of the sentence pairs reverses from the middle layers onward: although the surface-identical pair $((S \mid C_{\text{spec}}), (S \mid C_{\text{non}}))$; (red) is initially the most similar, it becomes less similar than the lexical-variation pair (blue), indicating that contextual interpretation ultimately induces greater representational divergence than surface lexical variation.

Mistral-7B-v0.3. Mistral exhibits an earlier transition than LLaMA-3, with the ordering reversal emerging around Layers 6–9 under both cosine similarity and CKA. Following this transition, the representations stabilize quickly, maintaining a relatively constant separation through the remaining layers. The passive-voice pair (green) shows a trajectory broadly similar to the red pair and remains comparable to, or slightly above, it through much of the middle and late layers.

Qwen2.5-7B. Qwen2.5 also exhibits the same qualitative reversal, beginning approximately around Layers 9–12. Unlike LLaMA-3, the contextual-interpretation pair (red) remains the least similar of the three throughout most of the middle and late layers, indicating particularly persistent separation associated with context-dependent interpretation. Qwen additionally shows sharper changes at the network boundaries, including a rapid increase during the first layer in cosine similarity and an abrupt decrease in the final CKA layer, suggesting greater representational instability near the embedding and output interfaces.

Overall, despite differences in the exact transition layer and late-layer dynamics, all three architectures exhibit the same qualitative organization: sentence representations are initially dominated by

surface form but become increasingly differentiated by contextual interpretation in the middle and late layers

C.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 tokens and d is the hidden dimension.

- **Cosine similarity** measures the average angular alignment between corresponding token 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 captures local representational alignment.

- **Centered Kernel Alignment (CKA)** provides an orthogonally invariant measure of representational similarity by comparing second-order statistics rather than individual vectors (Kornblith et al., 2019).

Let

$$K = XX^\top, \quad L = YY^\top, \quad H = I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top$$

and let

$$K_c = HKH, \quad L_c = HLH.$$

Linear CKA is then defined as

$$\text{CKA}(X, Y) = \frac{\text{tr}(K_c L_c)}{\sqrt{\text{tr}(K_c^2) \text{tr}(L_c^2)}} \quad (7)$$

CKA captures similarity in representational geometry independent of orthogonal transformations.

D Supplementary Cross-model Analysis for Experiment 2B

The directional asymmetry reported for LLaMA-3 naturally raises the question of whether the same intervention pattern generalizes across other transformer language models. To address this question, we apply the identical activation- patching protocol to Mistral-7B-v0.3 and Qwen2.5-7B, allowing a direct comparison of the layerwise intervention effects.

Layer-wise Representation Similarity (Sentence Means)

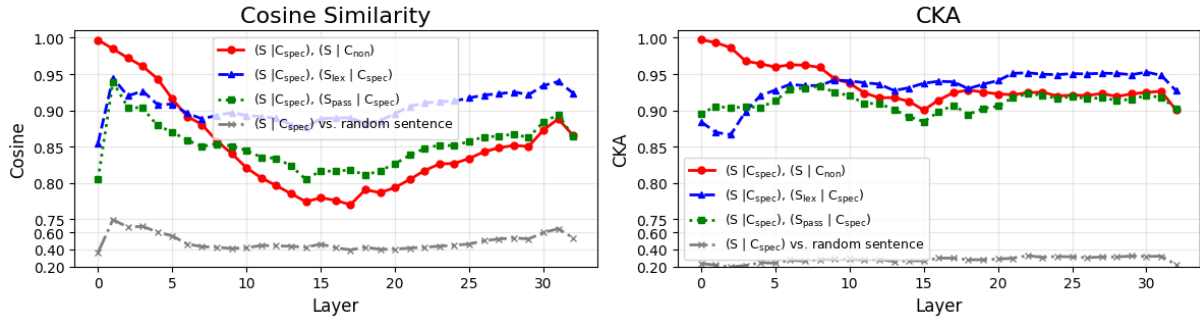


Figure 6: Representation Similarity of Mistral-7B Model

Layer-wise Representation Similarity (Sentence Means)

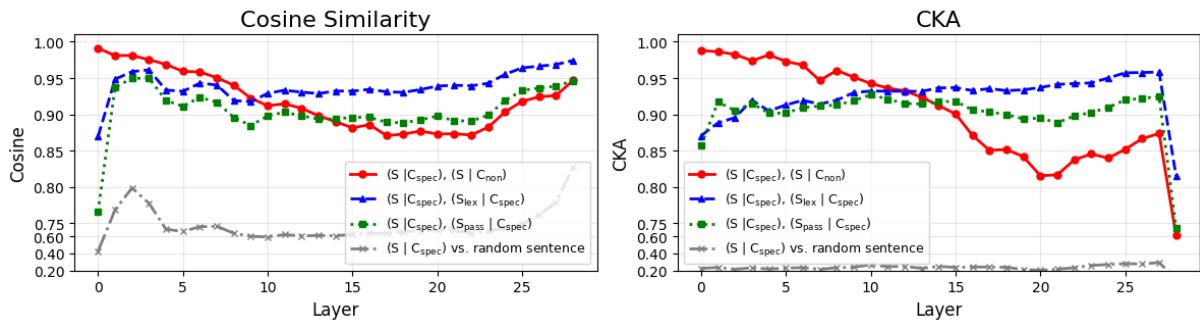


Figure 7: Representation Similarity of Qwen2.5-7B

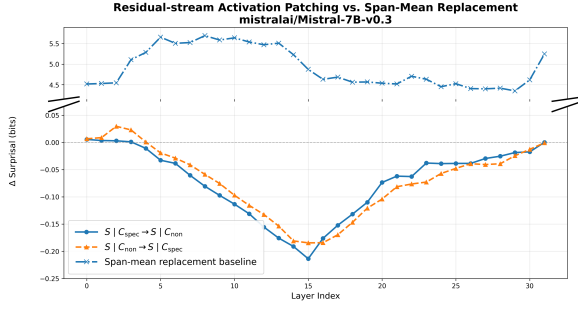


Figure 8: Layer-wise activation-patching effects for Mistral-7B-v0.3.

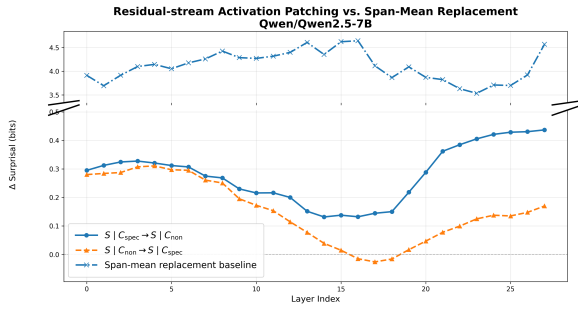


Figure 9: Layer-wise activation-patching effects for Qwen2.5-7B.

Figures 8 and 9 show that both models exhibit intervention profiles that remain clearly separated from the span-mean replacement baseline, indicating that the transferred representations produce structured effects beyond non-specific perturbation. However, unlike LLaMA-3, neither model reproduces the pronounced directional asymmetry observed in the main text. Mistral shows largely symmetric intervention effects across the two directions, whereas Qwen displays a different intervention profile without reproducing the LLaMA-3 pattern. Thus, while activation patching consistently reveals the functional importance of scope-conditioned representations, the representational geometry underlying these effects appears to differ across models.

The present experiments do not determine why these representational organizations differ. The observed variation may reflect differences in architectural design, training data, optimization dynamics, or alternative internal strategies for implementing comparable linguistic computations. Distinguishing among these possibilities requires finer-grained causal analyses, such as component-level or feature-level interventions, which can identify the computational pathways responsible for these representational differences.

Accordingly, we interpret the directional asymmetry reported in the main text as a model-specific observation rather than a universal property of scope representations. More broadly, the cross-model comparison suggests that while behaviorally consequential scope-conditioned representations emerge robustly across contemporary transformer language models, the internal organization of those representations remains model dependent.

E Supplementary Details of the SPECIFICITY Dataset

The SPECIFICITY dataset contains 1,998 context-sentence items: 997 specific and 1,001 non-specific instances. Each item is represented as a tuple (x_i, a_i, b_i, y_i) , where x_i is the full context-sentence sequence, $[a_i, b_i]$ identifies the target NP span, and $y_i \in \{0, 1\}$ is its specificity label. The items were generated using GPT-4o. Linguists manually reviewed a random 20% sample (400 items), evaluating the correctness of the target-span specificity labels and the discourse coherence of the context-sentence pairs.

ID	Full input text	Target NP	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.	<i>a small town</i>	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.	<i>a small town</i>	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.	<i>an expensive watch</i>	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.	<i>an expensive watch</i>	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.	<i>two cats</i>	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.	<i>two cats</i>	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.	<i>two friends</i>	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.	<i>two friends</i>	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.	<i>seven stars</i>	SPECIFIC
102_1_2	On a clear night, you can often see stars shining brightly in the sky. If you count seven stars, you might find a constellation.	<i>seven stars</i>	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.	<i>seven roses</i>	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.	<i>seven roses</i>	NON-SPECIFIC

Table 15: Representative items from the synthetic SPECIFICITY dataset. Each item contains a full context–sentence input, a marked target NP, and its gold specificity label.

F Details of Edge Probing in Experiment 2C

This appendix specifies the scalar-mixing edge-probing architecture used in Experiment 2C. 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/outputs and neural form). Let a transformer encoder M with layers $\ell \in \{0, \dots, L\}$ map an input sequence $x = (x_1, \dots, x_T)$ 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, and (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 and $y_i \in \{0, 1\}$ 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:

$$\mathbf{h}^{(\ell)} = \frac{1}{b - a + 1} \sum_{t=a}^b H_t^{(\ell)}(x) \in \mathbb{R}^d. \quad (8)$$

Stacking representations across layers yields

$$\mathbf{H} = [\mathbf{h}^{(0)}, \dots, \mathbf{h}^{(L)}] \in \mathbb{R}^{(L+1) \times d}. \quad (9)$$

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

$$\mathbf{s} = \sum_{\ell=0}^L \pi_{\ell} \mathbf{h}^{(\ell)} \in \mathbb{R}^d. \quad (10)$$

Probe classifier (2-layer MLP). The probing head is a binary classifier parameterized by $W_1, \mathbf{b}_1, W_2, \mathbf{b}_2$:

$$\begin{aligned} \mathbf{u} &= \text{ReLU}(W_1 \mathbf{s} + \mathbf{b}_1), \\ \tilde{\mathbf{u}} &= \text{Dropout}(\mathbf{u}), \\ z &= W_2 \tilde{\mathbf{u}} + \mathbf{b}_2, \end{aligned} \quad (11)$$

Case	Condition Type	Accuracy
$(S \mid C_{\text{spec}})$	Spec Ctx + Ambiguous	0.5180
$(S \mid C_{\text{non}})$	Non-spec Ctx + Ambiguous	0.7698
$(S_{\text{lex}} \mid C_{\text{spec}})$	Spec Ctx + Lexical Specific	0.7914
$(S_{\text{lex}} \mid C_{\text{non}})$	Non-spec Ctx + Lexical Specific	0.3741
$(S_{\text{pass}} \mid C_{\text{spec}})$	Spec Ctx + Passive Specific	0.8561
$(S_{\text{pass}} \mid C_{\text{non}})$	Non-spec Ctx + Passive Specific	0.4604

Table 16: Single-case edge probing accuracy in Mistral-7B. Accuracy indicates how reliably the probe recovers the expected *specificity* label for each condition.

Pair	PR-AUC	ROC-AUC	Acc.	F1
Baseline (Random)	0.5012	0.5059	0.5108	0.6421
$(S \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.6830	0.7100	0.6439	0.5926
$(S_{\text{lex}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.8620	0.8565	0.7806	0.7829
$(S_{\text{pass}} \mid C_{\text{spec}})$ vs. $(S \mid C_{\text{non}})$	0.8897	0.8876	0.8129	0.8207
$(S \mid C_{\text{spec}})$ vs. $(S_{\text{lex}} \mid C_{\text{non}})$	0.4292	0.4070	0.4460	0.4832

Table 17: Pairwise edge probing performance of the frozen specificity probe in Mistral-7B. Each row evaluates the probe, trained on SPECIFICITY, on the corresponding pair of SCOPEX conditions.

where z is the logit and

$$\hat{p} = \sigma(z) \quad (12)$$

is the predicted probability of *specificity*.

Training objective. The probe parameters θ and the mixing parameters α are jointly optimized by minimizing the binary cross-entropy loss over the *specificity* dataset:

$$\mathcal{L}_{\text{spec}} = \mathbb{E}_{(x,y) \sim \mathcal{D}_{\text{spec}}} [-y \log \hat{p} - (1-y) \log(1-\hat{p})]. \quad (13)$$

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 $\pi = \text{softmax}(\alpha)$ are reported as the layerwise importance distribution.

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

Model-specific asymmetry in specificity. The single-condition results reveal an important cross-model difference in the organization of *specificity* and *non-specificity*. In LLaMA-3, the main-text results show an asymmetric pattern in which

Category	Mean (%)	vs. Expected
<i>Layer groups</i>		
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	—
<i>Top-5 layers</i>		
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 18: 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
$(S C_{\text{spec}})$	Spec Ctx + Ambiguous	0.4460
$(S C_{\text{non}})$	Non-spec Ctx + Ambiguous	0.7986
$(S_{\text{lex}} C_{\text{spec}})$	Spec Ctx + Lexical Specific	0.7410
$(S_{\text{lex}} C_{\text{non}})$	Non-spec Ctx + Lexical Specific	0.4388
$(S_{\text{pass}} C_{\text{spec}})$	Spec Ctx + Passive Specific	0.8705
$(S_{\text{pass}} C_{\text{non}})$	Non-spec Ctx + Passive Specific	0.5180

Table 19: Single-case edge probing accuracy in Qwen2.5-7B. Accuracy indicates how reliably each condition is classified according to its expected *specific* or *non-specific* interpretation.

specific interpretations are more reliably recovered than non-specific ones, consistent with recent formal-semantic accounts of *specificity* as a comparatively marked semantic category (Degano and Aloni, 2025). This asymmetry is not reproduced in Mistral-7B or Qwen2.5-7B. In particular, for the purely contextual contrast, both models show higher accuracy for $(S | C_{\text{non}})$ than for $(S | C_{\text{spec}})$ (Tables 16 and 19). The markedness-like asymmetry observed in LLaMA-3 should therefore be understood as a model-specific representational pattern rather than a uniform property across the three model families examined here.

Consistency across Experiments 2B and 2C. Interestingly, this cross-model variability is consistent within models across two distinct analyses. The LLaMA-3-specific asymmetry in Experiment 2C accompanies the directional asymmetry observed under activation patching in Exper-

Pair	PR-AUC	ROC-AUC	Acc.	F1
Baseline (Random)	0.4684	0.4386	0.4640	0.4016
$(S C_{\text{spec}})$ vs. $(S C_{\text{non}})$	0.7127	0.7255	0.6223	0.5415
$(S_{\text{lex}} C_{\text{spec}})$ vs. $(S C_{\text{non}})$	0.8778	0.8787	0.7698	0.7630
$(S_{\text{pass}} C_{\text{spec}})$ vs. $(S C_{\text{non}})$	0.9055	0.9078	0.8345	0.8403
$(S C_{\text{spec}})$ vs. $(S_{\text{lex}} C_{\text{non}})$	0.4571	0.4325	0.4424	0.4444

Table 20: Pairwise edge probing performance of the frozen specificity probe in Qwen2.5-7B. Each row evaluates the probe, trained on SPECIFICITY, on the corresponding pair of SCOPEX conditions.

Category	Mean (%)	vs. Expected
<i>Layer groups</i>		
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	—
<i>Top-5 layers</i>		
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 21: 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.

iment 2B, whereas Mistral-7B and Qwen2.5-7B do not consistently exhibit the corresponding LLaMA-3-like asymmetry in either analysis. Thus, despite differences across model families, the intervention and probing results show notable within-model convergence.

Shared structure and layerwise variation. Despite this model-specific asymmetry, the pairwise analyses reveal a stable qualitative hierarchy across models. For both Mistral-7B and Qwen2.5-7B, the contrast between $(S_{\text{pass}} | C_{\text{spec}})$ and $(S | C_{\text{non}})$ yields the strongest separability, whereas $(S | C_{\text{spec}})$ versus $(S_{\text{lex}} | C_{\text{non}})$ yields the weakest (Tables 17 and 20). The scalar-mixing profiles further reveal differences in how this information is distributed across layers: Mistral-7B, like LLaMA-3, shows a relatively stronger late-layer contribution, whereas Qwen2.5-7B exhibits a flatter layerwise distribution (Tables 18 and 21). These results sug-

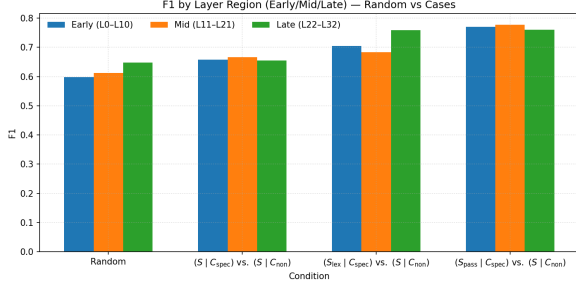


Figure 10: Layerwise Comparison Logistic Regression PR-AUC of LLaMA-3-8B

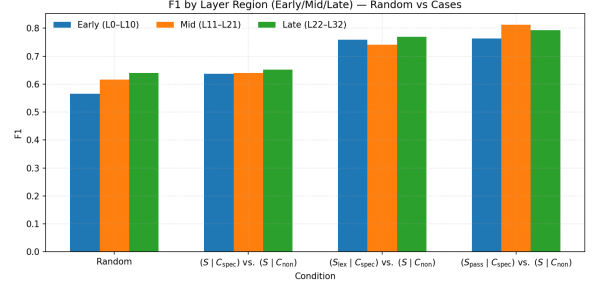


Figure 11: Layerwise Comparison Logistic Regression PR-AUC of Mistral-7B

gest that the models share sensitivity to the relative strength of linguistic cues while differing in how scope-related semantic information is organized across representations.

H Layerwise Logistic Regression Probing as an Auxiliary Probing of Edge Probing in Experiment 2C

H.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 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.

H.2 Results of Layerwise Logistic Regression

Across LLAMA-3, 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,

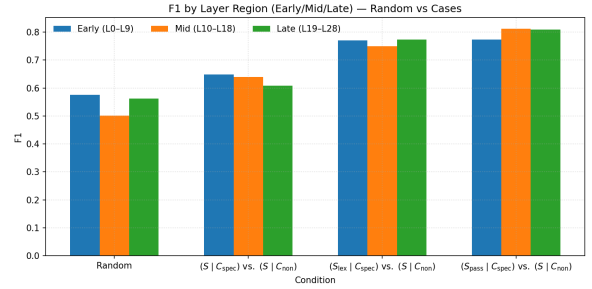


Figure 12: Layerwise Comparison Logistic Regression PR-AUC of Qwen2.5

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 $((S \mid C_{\text{spec}}) \text{ and } (S \mid C_{\text{non}}))$ 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 high-

est 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.