

Testing Language Model Sensitivity to Object-Anaphoric Accessibility: Persian *ye(k)* and English indefinite *this*

Nikta Yekrang Safakar
Humboldt-Universität zu Berlin
Leibniz-Zentrum Allgemeine
Sprachwissenschaft (ZAS)
Berlin, Germany
niktaayekrang@gmail.com

Sebastian Padó
Natural Language Processing
University of Stuttgart, Germany
pado@ims.uni-stuttgart.de

Abstract

How accessible a discourse referent is for later mention depends on how it is introduced: indefinite marking can make an entity more accessible for anaphora. This dimension is computationally understudied. We ask whether language models are sensitive to this discourse function as encoded by Persian *ye(k)*. Linguistic theory for Persian predicts that *ye(k)*-marked objects support anaphoric uptake more readily than bare objects. Comparing a bare object against a minimally different *ye(k)*-marked object, we find that both encoder- and decoder-based models assign higher scores to object-anaphoric uptake material after *ye(k)* contexts. The effect is reliable at the object-anaphoric NP and, in six of seven decoders, exceeds both subject-span and neutral-continuation controls, but is weaker and more model-dependent over the whole continuation. We find a parallel directional effect for an English indefinite-*this* contrast, showing that a comparable downstream anaphoric-uptake pattern is observable with a different language-specific cue. As a secondary finding, we show that the scoring unit (only the anaphoric uptake site vs. the whole continuation) matters, and that the local site provides a clearer signal. We will release the Persian and English stimulus sets, item-level scores, and reproduction scripts upon publication.

1 Introduction

Discourse understanding requires tracking which expressions introduce referents that remain available for later mention. For language models, this raises a direct behavioral question: when a context contains a cue that should make a referent more accessible, does the model assign higher probability to later material that picks up that referent? Existing targeted evaluations have mostly examined morphosyntactic contrasts (Linzen et al., 2016; Marvin and Linzen, 2018; Warstadt et al., 2020). Work on discourse entities and anaphora is growing, but it

is still concentrated on English articles, sentential operators, definiteness, and dialogue-level continuations (Schuster and Linzen, 2022; Zhu and Frank, 2024; Zhu et al., 2025; Kim and Schuster, 2023; Kim and Misra, 2026).

We test the sensitivity of language models to discourse marking with Persian *ye(k)*. Persian nominal interpretation is shaped by the contrast between bare nominals, indefinite marking with *ye(k)*, and object marking with *rā* (Ghameshi, 2003; Jasbi, 2016; Karimi, 1990; Jasbi, 2020; Modarresi and Krifka, 2021). Our study isolates the contrast between bare objects and minimally different *ye(k)*-marked objects, without *rā* in the context sentence. The linguistic prediction: relative to a bare object, a *ye(k)*-marked object makes later object-anaphoric material more probable since it introduces the object as a more explicitly individuated discourse referent, which is predicted to be more accessible for later anaphora (Ariel, 1990). In Example (1), (1d) is more expected to follow after (1b) than after (1a) while there should be no difference for (1c) after (1a) versus (1b).

This gives the paper two linked questions:

Substantive question (RQ1). Are language models sensitive to discourse expectation? The same object-anaphoric continuation should receive a higher score after the *ye(k)* context than after the bare-object context.

Methodological question (RQ2). Where can we measure that effect? The full continuation contains the anaphoric expression, but is also influenced by plausibility, lexical choice, register, length, and model-specific fluency. We compare whole-continuation scores with local scores at the object-anaphoric uptake site.

To carry out this comparison, we create a 48-item minimal-pair dataset. Each item has a bare-object

context and a *ye(k)*-object context. The continuations are held fixed across conditions: one continuation follows the subject, and one picks up the object using either a demonstrative NP such as *in* OBJECT ‘this OBJECT’ or a same-NP form such as *hamān* OBJECT ‘the same OBJECT’. This design adds a Persian benchmark dataset to a literature in which discourse-referent evaluations have mostly used English cues. Cross-linguistic work shows that object marking and nominal form can affect discourse persistence in languages such as Romanian (Chiriacescu and von Heusinger, 2010) and Persian (Modarresi and Krifka, 2021). Persian *ye(k)* supplies a morphological indefiniteness cue whose hypothesized downstream consequence is increased anaphoric uptake.

As for LMs, we evaluate three encoder models—mBERT (Devlin et al., 2019), XLM-R (Conneau et al., 2020), and ParsBERT (Farahani et al., 2021)—and three autoregressive decoder families—Qwen2.5 (Qwen Team, 2024), mGPT (Shliazhko et al., 2024), and XGLM (Lin et al., 2022)—using both whole-continuation and local-span scoring.

The within-model item-level analyses provide evidence for the discourse-sensitivity claim (RQ1), with a clear measurement qualification. Whole-continuation scores are positive in several models but unstable across model families and scales. Local-span diagnostics are more consistent (RQ2): the full object-anaphoric NP provides the clearest and most reliable local evidence for the *ye(k)* effect, and subject-span and neutral-continuation controls support the object-anaphoric interpretation. As a functional comparison, an English *a*-versus-indefinite-*this* contrast shows the same direction across nine models (§6.6).

Our findings amount to four contributions. First, we present evidence that language models are sensitive to Persian *ye(k)* as an object-anaphoric accessibility cue. Second, we contribute the 48-item Persian minimal-pair benchmark dataset for testing discourse-sensitive object anaphora in a language that remains underrepresented in targeted discourse evaluation. Third, we show that the measured strength of this discourse effect depends on the scoring unit: whole-continuation scoring gives weak and model-sensitive evidence, while local uptake-site scoring gives a clearer signal. Fourth, as a functional comparison, we show that English indefinite *this* yields a downstream uptake effect in the same direction relative to *a*; this is consistent with a cross-linguistic parallel in model behavior

without implying semantic equivalence or a shared internal mechanism.

2 Background: Persian *ye(k)*

The core contrast pits a bare object against a minimally different *ye(k)*-marked object. We exclude *rā* from the context sentence because *rā* introduces additional object-marking and discourse-accessibility effects. This design isolates the narrower contribution of *ye(k)* to later object-anaphoric uptake. The rationale is that bare Persian objects can receive property-like or pseudo-incorporated interpretations and have reduced anaphoric potential, whereas *ye(k)*-marked objects introduce an ordinary indefinite individual more explicitly and are therefore predicted to support later uptake more readily (Ghameshi, 2003; Jasbi, 2016; Krifka and Modarresi, 2016; Modarresi and Krifka, 2021).

The prediction is comparative: the same object-anaphoric continuation should receive a higher model score after the *ye(k)* context than after the bare-object context. Example (1) shows one item. The paper presents Persian examples in transliteration to improve readability; inputs to models were all provided in Persian script.

- (1) a. Bare-object context
Vahid keyk poxt.
Vahid cake baked
‘Vahid baked cake.’
- b. *ye(k)*-object context
Vahid yek keyk poxt.
Vahid one/a cake baked
‘Vahid baked a cake.’
- c. Subject-continuing continuation
Vahid payām-aš rā check kard.
Vahid message-POSS.3SG OM check did
‘Vahid checked his message.’
- d. Object-anaphoric continuation
in keyk bu-ye xub-i dāšt.
this cake smell-EZ good-INDEF had
‘This cake smelled good.’

Hypotheses. The contrast holds the continuation fixed and varies only the context (1a vs. 1b). We state the prediction at the level of referential accessibility and link it to model scores via surprisal theory (Levy, 2008): greater accessibility is expected to correspond to lower surprisal, i.e. higher length-normalized log-probability of the uptake material.

H1 (sensitivity). A *ye(k)*-marked object is a more accessible antecedent than a bare object, so object-anaphoric uptake material receives a higher score after the *ye(k)* context than after the bare context. Concretely, the object-anaphoric continuation (1d) should score higher after the *ye(k)* context (1b) than after the bare context (1a).

H2 (localization). Under H1, the clearest evidence should appear at the object-anaphoric uptake site: the object-NP span should show a more consistent *ye(k)* effect than whole-continuation scoring and should exceed the subject-span control. That is, the (1b)-vs-(1a) difference should be larger on the object span *in keyk* in (1d) than on the subject span *Vahid* in the subject-continuing (1c).

Continuation types Each item has a subject-continuing and an object-anaphoric continuation. Object-anaphoric continuations use either a demonstrative NP with *in* OBJECT ‘this OBJECT’ or a same-NP continuation with *hamān* OBJECT ‘the same OBJECT’. We keep these formats separate because they differ in register and anaphoric force.

The relevant spans are underlined in Example (1): *Vahid* in the subject-continuing continuation, *in keyk* in the object-anaphoric continuation. Within the object NP, the NP-initial anaphoric element is the demonstrative *in*. We call the NP-initial element *in* or *hamān* the anaphoric marker span. The object-anaphoric NP occupies the subject position in its clause. Because it is identical across the bare and *ye(k)* contexts, any effect of that syntactic position is held constant across conditions and cannot drive the bare-versus-*ye(k)* contrast.

3 Related Work

Discourse-referent sensitivity in language models. Targeted evaluations increasingly test whether models are sensitive to discourse referents and continuation preferences, including referential information, negated indefinites, discourse-entity recognition, anaphora accessibility, entity-state tracking, at-issue sensitivity, and coherence prediction (Sorodoc et al., 2020; Schuster and Linzen, 2022; Zhu and Frank, 2024; Zhu et al., 2025; Kim and Schuster, 2023; Kim and Misra, 2026; Beyer et al., 2021). Most of this work uses English articles, operators, or dialogue-level continuations. The gap we address is that these evaluations test discourse-referent sensitivity predominantly through English articles, operators, or dialogue-level continuations, so it is unknown

whether models are sensitive to accessibility cues carried by dedicated nominal morphology on the object. We extend this paradigm to a Persian nominal cue.

Persian and English do not lexicalize the same nominal categories, so the English comparison is functional rather than form-equivalent. The Persian core contrast compares a bare object with a *ye(k)*-marked object, manipulating whether the object is introduced as a more explicitly individuated indefinite discourse referent; Persian *ye(k)* is an ordinary indefinite cue, carrying an at-issue existence implication similar to English *a(n)*, and is not itself the language’s specificity marker (Jasbi, 2016; von Heusinger and Sadeghpour, 2020). English does not license bare singular count objects (**Vahid baked cake*), so the bare-versus-*ye(k)* contrast has no direct English counterpart, and an *a/the* contrast would manipulate definiteness and uniqueness rather than the Persian object-domain contrast. English does, however, have a marked indefinite construction, indefinite *this* (*Emma found this notebook*), which has been shown to increase later referential uptake relative to *a*. Indefinite *this* is a marked [+specific] indefinite (Ionin, 2006), and its use has been shown to increase later referential uptake relative to *a(n)* (Gernsbacher and Shroyer, 1989). English indefinite *this* can therefore serve as a functional comparison point for increased downstream uptake, without implying semantic equivalence to Persian *ye(k)*. What makes *this* versus *a* the relevant contrast is that both are indefinite: the pair holds definiteness constant and varies only how strongly the referent is individuated and flagged as noteworthy. A related object-domain diagnostic is Korean differential object marking (Shin and Trott, 2024), which targets an exhaustivity implicature; here the target is later anaphoric uptake.

Localized scoring. The theoretically predicted effect may be local even when the continuation is longer. Psycholinguistic and targeted-LM work therefore often evaluates probability or surprisal at critical regions rather than over an entire sentence (Levy, 2008; Smith and Levy, 2013; Futrell et al., 2020; Gauthier et al., 2020; Hu et al., 2020; Wilcox et al., 2021, 2023; Hanna and Mueller, 2025). Because tokenization, model family, and construction type can affect surprisal conclusions (Oh and Schuler, 2025; von der Malsburg and Padó, 2026), we compare full-continuation scores with targeted scores at the object-anaphoric marker, the

	Baseline context	Target context	Object-anaphoric continuation	Type
1	<i>Vahid keyk poxt.</i> ‘Vahid baked cake.’	<i>Vahid yek keyk poxt.</i> ‘Vahid baked a cake.’	<i>in keyk bu-ye xubi dāšt.</i> ‘this cake smelled good.’	dem.
2	<i>Ārād livān xarid.</i> ‘Ārād bought glass.’	<i>Ārād yek livān xarid.</i> ‘Ārād bought a glass.’	<i>hamān livān kenār-e sink</i> <i>xošk šod.</i> ‘the same glass dried by the sink.’	same-NP
3	<i>Bahār kulepošti xarid.</i> ‘Bahār bought backpack.’	<i>Bahār yek kulepošti xarid.</i> ‘Bahār bought a backpack.’	<i>in kulepošti jibhā-ye ziyādi</i> <i>dāšt.</i> ‘this backpack had many pockets.’	dem.
<i>English comparison set (§6.6): “a” vs. indefinite “this”</i>				
4	Emma found a notebook in the library.	Emma found this notebook in the library.	Later, the librarian put <u>it</u> on the desk.	pron.

Table 1: Example items. Rows 1–3 are Persian and romanized: the two contexts differ only in *yek*, and the object-anaphoric continuation is identical across contexts. Row 4 is an item from the English comparison set (§6.6), where the contexts differ only in *a* versus indefinite *this*. In both arms, the corresponding prediction is that the underlined object-anaphoric span scores higher after the target context. ‘Type’ is the anaphor form: demonstrative *in* ‘this’, same-NP *hamān* ‘the same’, or English pronoun.

full object-anaphoric NP, and control spans.

4 Data and Models

4.1 Benchmark: Stimuli and validation

The stimulus set contains 48 manually constructed Persian items, each with a bare-object context and a minimally different *ye(k)*-object context, yielding 96 context rows. Context pairs differ only by insertion of *yek*. Each item has one subject-continuing continuation and one object-anaphoric continuation, both held fixed across context conditions. The object-anaphoric continuations are balanced across two formats: 24 demonstrative-NP items with *in* OBJECT ‘this OBJECT’ and 24 same-NP items with *hamān* OBJECT ‘the same OBJECT’. Items were designed as simple transitive events with concrete objects and human subjects. This resulted in 48 unique subjects, 48 unique objects, and 12 predicates. Each item instantiates the RQ1 contrast: the two context rows differ only in the presence of *yek*, while the object-anaphoric continuation is held fixed. The score for its material is predicted to rise under the *ye(k)* context. Table 1 shows three Persian items and one English item.

Items were checked by one native-speaker Persian linguist for naturalness and for the intended contrast between property-like bare-object contexts and referentially stronger *ye(k)*-object contexts. Structural validation verified one bare and one *ye(k)* context per item, exact paired continua-

tion identity, target-span presence, absence of subject names from object-anaphoric continuations, absence of exact continuation duplicates, object-domain labels, and successful target-span matching after tokenization. Because some semantic domain skew remains, we report leave-one-predicate and leave-one-domain robustness checks.¹

4.2 Models

We evaluate encoders, mBERT (bert-base-multilingual-cased), XLM-R (xlm-roberta-base), and ParsBERT (HooshvareLab/bert-fa-base-uncased), and decoders with Persian coverage: Qwen2.5 0.5B/1.5B/3B, mGPT-Persian 1.3B, and XGLM 1.7B/2.9B/4.5B. This covers a range of LMs across architectures, training data, tokenizers, vocabulary, language coverage, and scale.

5 Scoring

Table 2 summarizes the contrast space. The main linguistic question is whether the *ye(k)* context increases the score of object-anaphoric uptake material relative to the bare-object context (RQ1).

The localization question (RQ2) contrasts the two scoring schemes: Whole-continuation scoring asks whether *ye(k)* changes the relative score of an entire object-anaphoric continuation. Local span

¹The stimulus items, item-level scores, and reproduction scripts will be released in a repository upon publication.

Measure	Contrast	Target comparison	Hypothesis role
Whole continuation	$M^{\text{yek}} - M^{\text{bare}}$	Object continuation versus subject continuation	Broad H1 diagnostic
Object NP span	Δ^{objNP}	Object-anaphoric NP after $ye(k)$ vs. bare context	Main local H1 test
Anaphoric marker span	Δ^{marker}	<i>in/hamān</i> after $ye(k)$ vs. bare context	Secondary local H1 diagnostic
Subject span	Δ^{subj}	Subject mention after $ye(k)$ vs. bare context	Control for general context shift
Object-minus-subject	$\Delta^{\text{objNP}} - \Delta^{\text{subj}}$	Object uptake effect beyond subject-span effect	H2 object-uptake test
Neutral baseline, reported for decoders	$\Delta^{\text{objNP}} - \Delta^{\text{neutral}}$	Object uptake effect beyond generic continuation shift	Fluency control

Table 2: Measure and contrast space. H1 predicts a positive $ye(k)$ -over-bare effect for object-anaphoric uptake material. H2 predicts that this effect is localized at the object-anaphoric uptake site and exceeds the subject-span control. The neutral-continuation comparison provides a complementary decoder fluency control.

scoring asks whether $ye(k)$ changes the score of the specific object-uptake anaphoric material.

Detailed example. To make the measures concrete, take item (1). The two contexts are the bare *Vahid keyk poxt* (1a) and the $ye(k)$ variant *Vahid yek keyk poxt* (1b). Both are paired with the same subject-continuing continuation (1c), whose scored subject span is *Vahid*, and the same object-anaphoric continuation (1d), whose scored object-NP span is *in keyk*. The whole-continuation margin M^c (Eq. 3) is the score of the object continuation (1d) minus the score of the subject continuation (1c) within a single context c ; the whole-continuation effect $\Delta^{\text{whole}} = M^{\text{yek}} - M^{\text{bare}}$ asks whether adding *yek* raises that margin (H1, RQ1). The local object-NP effect Δ^{objNP} compares the score of the span *in keyk* after context (1b) versus context (1a); the subject-span control Δ^{subj} does the same for *Vahid* in (1c). H2 predicts $\Delta^{\text{objNP}} > \Delta^{\text{subj}}$: the $ye(k)$ effect should concentrate on the object-anaphoric span, not on unrelated continuation material. These are exactly the quantities in Table 4; for XLM-R, for instance, $\Delta^{\text{objNP}} = 0.67$ and $\Delta^{\text{subj}} = -0.01$, a +0.67 object-minus-subject contrast.

Scoring. Let i index items and m index models. Each item has a bare-object context x_i^{bare} , a $ye(k)$ context x_i^{yek} , a subject-continuing continuation y_i^{subj} , and an object-anaphoric continuation y_i^{obj} . We write $S_m(x, z)$ for the mean log score that model m assigns to target string or target span z in context x . Higher values are better be-

cause log probabilities closer to zero mean that the model assigned higher probability to the target. For decoders, when the target is a continuation y , $S_m(x, y)$ is its length-normalized log probability given context x :

$$S_m^{\text{dec}}(x, y) = \frac{1}{T_y} \sum_{t=1}^{T_y} \log p_m(y_t \mid x, y_{<t}). \quad (1)$$

For decoder local-span scoring, we apply the same left-to-right rule only to the target tokens at their original positions in the continuation, retaining the context and all preceding continuation tokens.

For encoders, we use pseudo-log-likelihood scoring (Salazar et al., 2020). The primary encoder result uses PLL-word-l2r (Kauf and Ivanova, 2023). Let target span z contain W_z words, and let word w be tokenized into K_w subtokens $s_{w,1}, \dots, s_{w,K_w}$. For each target subtoken $s_{w,k}$, PLL-word-l2r masks that subtoken and all future subtokens of the same word. The score is the mean log probability over scored target subtokens:

$$S_m^{\text{pll-wl2r}}(x, z) = \frac{1}{N_z} \sum_{w=1}^{W_z} \sum_{k=1}^{K_w} \ell_{m,w,k}, \quad (2)$$

$$\ell_{m,w,k} = \log p_m(s_{w,k} \mid r_{w,k}),$$

$$r_{w,k} = \text{mask}_{w,k:K_w}(C(x, z)).$$

where $N_z = \sum_{w=1}^{W_z} K_w$, and $C(x, z)$ is the scored sequence containing the context and the relevant continuation material. For targeted span scoring, z is a selected span inside the continuation. For

whole-continuation encoder scoring, the score is the length-normalized PLL over the full context–continuation sequence, whereas the decoder whole-continuation score (Eq. 1) is length-normalized over the continuation tokens only.

5.1 Whole-continuation effect

For each context condition $c \in \{\text{bare}, \text{yek}\}$, the object-continuation margin is:

$$M_{i,m}^c = S_m(x_i^c, y_i^{\text{obj}}) - S_m(x_i^c, y_i^{\text{subj}}). \quad (3)$$

The whole-continuation $ye(k)$ effect is the change in this margin from the bare context to $ye(k)$:

$$\Delta_{i,m}^{\text{whole}} = M_{i,m}^{\text{yek}} - M_{i,m}^{\text{bare}}. \quad (4)$$

A positive value means that $ye(k)$ increased the model’s relative score for an object-anaphoric over a subject-continuing continuation.

5.2 Local spans and controls

The local diagnostics target the anaphoric marker z_i^{marker} , the full object-anaphoric NP z_i^{objNP} , and the subject control span z_i^{subj} . Their effects are:

$$\begin{aligned} \Delta_{i,m}^{\text{marker}} &= S_m(x_i^{\text{yek}}, z_i^{\text{marker}}) - S_m(x_i^{\text{bare}}, z_i^{\text{marker}}), \\ \Delta_{i,m}^{\text{objNP}} &= S_m(x_i^{\text{yek}}, z_i^{\text{objNP}}) - S_m(x_i^{\text{bare}}, z_i^{\text{objNP}}), \\ \Delta_{i,m}^{\text{subj}} &= S_m(x_i^{\text{yek}}, z_i^{\text{subj}}) - S_m(x_i^{\text{bare}}, z_i^{\text{subj}}). \end{aligned}$$

The main object-uptake tests subtract the subject control from the object-span effects:

$$I_{i,m}^{\text{marker-subj}} = \Delta_{i,m}^{\text{marker}} - \Delta_{i,m}^{\text{subj}}, \quad (5)$$

$$I_{i,m}^{\text{objNP-subj}} = \Delta_{i,m}^{\text{objNP}} - \Delta_{i,m}^{\text{subj}}. \quad (6)$$

For decoder models, we additionally score a fixed neutral continuation, romanized as *kami ba’d, hame chiz ārām bud* ‘a little later, everything was calm’. It provides a decoder-specific anchor for general fluency: a $ye(k)$ context could make following strings more or less fluent independently of anaphoric uptake. Encoder PLL scores are bidirectional compatibility scores, so a fixed neutral string has no comparable interpretation; the primary encoder control is the subject-span subtraction.

All deltas are computed at the item level. For each model and metric, we report the mean item-level delta and number of positive (hypothesis-conforming) items. We also report 95% item-bootstrap confidence intervals over item-level resamples with fixed random seeds. These intervals quantify uncertainty over the sampled items

	Model	Score	95% CI	Pos.
Enc.	mBERT	0.22	[0.12, 0.31]	36
Enc.	ParsBERT	0.13	[-0.00, 0.26]	31
Enc.	XLM-R	0.05	[-0.05, 0.15]	25
Dec.	Qwen2.5-0.5B	0.11	[0.08, 0.15]	39
Dec.	Qwen2.5-1.5B	0.09	[0.05, 0.14]	34
Dec.	Qwen2.5-3B	0.08	[0.04, 0.13]	35
Dec.	mGPT-Persian	0.05	[-0.00, 0.09]	30
Dec.	XGLM-1.7B	-0.44	[-0.76, -0.19]	14
Dec.	XGLM-2.9B	-0.08	[-0.17, 0.01]	20
Dec.	XGLM-4.5B	0.16	[0.09, 0.24]	35

Table 3: Whole-continuation $ye(k)$ effects. Positive Δ^{whole} scores indicate that the $ye(k)$ context increased the object-continuation margin relative to the bare-object context, as predicted by H1. Encoders are scored as PLL-word-l2r (Eq. 2), and decoders as mean token log probability (Eq. 1). Scores are thus not directly comparable, but the sign, confidence interval, and item direction can be compared descriptively. Pos.: Positive out of 48. CI bounds are rounded to two decimals; -0.00 denotes a bound marginally below zero.

within each model and metric, rather than uncertainty over a broader population of models. The reported measures follow a predefined hierarchy reflecting their different inferential roles: the object-NP contrast constitutes the primary local test of H1, while whole-continuation and marker-span measures provide complementary diagnostics of the same underlying prediction. Subject-span and neutral-continuation contrasts serve primarily as controls rather than independent confirmatory tests. Statistical uncertainty is estimated over items separately within each model. Because the tested models are few and heterogeneous in architecture, training data, tokenization, language coverage, scale, and scoring method, we do not make population-level inferences over models; consistency across the tested models is therefore interpreted descriptively.

6 Results

6.1 Whole-continuation scoring gives weak and model-sensitive evidence for H1

Table 3 reports the item-level whole-continuation effect. The whole-continuation pattern provides only partial evidence for H1. Most models show a positive mean shift, but the result is not stable across families or scales. Among encoders, mBERT is clearly positive, ParsBERT is border-

Model	Target span	Score	95% CI	Pos.	-Sub
mBERT	object NP	0.65	[0.47, 0.84]	43	0.33
mBERT	anaph. marker	1.04	[0.63, 1.51]	37	0.73
mBERT	subject span	0.31	[0.16, 0.49]	31	—
ParsBERT	object NP	0.57	[0.43, 0.71]	42	0.65
ParsBERT	anaph. marker	0.93	[0.68, 1.18]	44	1.01
ParsBERT	subject span	-0.08	[-0.44, 0.24]	25	—
XLM-R	object NP	0.67	[0.54, 0.80]	46	0.67
XLM-R	anaph. marker	1.56	[1.32, 1.80]	46	1.57
XLM-R	subject span	-0.01	[-0.14, 0.13]	17	—

Table 4: Local span effects for encoders. Scores are mean Δ^{objNP} , Δ^{marker} , and Δ^{subj} , computed as the score after the $ye(k)$ context minus the score after the bare-object context. Pos.: Positive out of 48. Final column (-Sub) reports the object-minus-subject and marker-minus-subject uptake-site contrasts (H2); all six are positive. Positive object-NP and marker effects support H1.

line, and XLM-R is near flat. Among decoders, the Qwen2.5 models are positive, mGPT is weak, and XGLM is scale-sensitive, with XGLM-1.7B and XGLM-2.9B not supporting the predicted direction. Whole-continuation scoring therefore suggests that $ye(k)$ can affect continuation preferences, but it is noisy as a broad diagnostic.

6.2 Local object-anaphoric spans show clearer effects

Table 4 reports local span effects for the anaphoric marker, the full object-anaphoric NP, and the subject-span control. The object-span effects are clearer than the whole-continuation effects. The full object NP is positive across all three models, with 43/48 positive items in mBERT, 42/48 in ParsBERT, and 46/48 in XLM-R. The anaphoric marker is also positive for all three encoders, ranging from 0.93 in ParsBERT to 1.56 in XLM-R, but marker spans are short and token-type-distinct from subject-name spans, so we treat them as supportive rather than as the conservative diagnostic. Subject spans behave substantively differently: mBERT shows a smaller positive subject-span shift, while ParsBERT and XLM-R are near zero.

6.3 Object-minus-subject interactions

The rightmost column of Table 4 reports the main encoder uptake-site test: object and marker spans minus the subject control. All six interactions are positive. Among these, we treat the object-NP-minus-subject interactions as primary, since they compare a referential object span against the subject-span control. Marker-minus-subject inter-

actions are length- and token-type-confounded similar to the decoder marker spans, so we treat them as secondary evidence. The results support an object-anaphoric interpretation of the local effect.

6.4 Anaphor type and robustness

The anaphor-type breakdown shows model- and target-dependent variation. Demonstrative-NP continuations yield larger full-object-NP effects in mBERT and ParsBERT, while XLM-R shows a larger full-object-NP effect for same-NP continuations. Marker-span effects also vary by model and should not be treated as a clean independent anaphor-type result. We therefore interpret anaphor type as a diagnostic and robustness factor, not as a separate theoretical conclusion. The demonstrative form *in* OBJECT can realize object uptake in the controlled continuation, but it can also have deictic force outside this design. The same-NP form *hamān* OBJECT is more explicitly anaphoric, but it is also register-sensitive. Across the full item set, leave-one-predicate and leave-one-domain checks show that the mean object-minus-subject interaction remains positive for every exclusion in all three encoders. These checks indicate that the object-minus-subject effect is not driven by any single predicate or semantic domain. The observed differences between demonstrative-NP and same-NP continuations remain descriptive, however, because anaphor form covaries with lexical realization and register and was included as a robustness factor rather than as an independently controlled manipulation.

Model	ObjNP–Neutral	ObjNP–Subject
Qwen2.5-0.5B	0.46 [0.38, 0.55]	0.50 [0.39, 0.61]
Qwen2.5-1.5B	0.50 [0.40, 0.60]	0.37 [0.26, 0.49]
Qwen2.5-3B	0.49 [0.40, 0.58]	0.27 [0.16, 0.39]
mGPT-Persian	0.36 [0.27, 0.45]	0.39 [0.29, 0.50]
XGLM-1.7B	-0.02 [-0.18, 0.14]	-0.06 [-0.27, 0.15]
XGLM-2.9B	0.15 [0.03, 0.27]	0.33 [0.15, 0.52]
XGLM-4.5B	0.72 [0.57, 0.87]	0.86 [0.61, 1.12]

Table 5: Decoder local-span controls for the full object-anaphoric NP ($n=48$). ObjNP–Neutral subtracts the fixed neutral-continuation baseline from the full object-NP effect. ObjNP–Subject subtracts the subject-span effect from the full object-NP effect. Positive values indicate that the object-NP effect exceeds the corresponding general-fluency or subject-span control. Brackets show item-bootstrap 95% confidence intervals.

6.5 Decoder local-span controls

We now aim to confirm that the decoder effects are localized to object-anaphoric material. Our first test is the subject-span control, which tests whether the $ye(k)$ effect is stronger for object-anaphoric material than for subject-continuing material. The second control uses a fixed neutral continuation baseline (cf. Section 5.2) which estimates the changes in general continuation fluency after bare-object versus $ye(k)$ -object contexts. The critical comparison is whether the object-NP effect exceeds this generic-continuation shift.

Table 5 shows the two decoder controls. In six of seven decoder models, the full object NP exceeds the neutral baseline with a bootstrap confidence interval above zero. The same six models also show a positive full-object-NP-minus-subject-span contrast. XGLM-1.7B is the exception in both diagnostics. Thus, the decoder local-span results support the same main conclusion as the encoder span analysis: the $ye(k)$ effect is localized to the object-anaphoric uptake site.

6.6 English comparison: indefinite *this* and anaphoric uptake

A final concern is the question of whether the LMs are merely picking up the distribution of a specific surface cue – $ye(k)$ – or whether there are other instances of sensitivity to later referential uptake.

To test this, we build a 48-item English comparison set matched in structure to the Persian one. Each item has an *a* context and an indefinite-*this* context, a fixed pronominal object-anaphoric continuation, a demonstrative-NP variant, a negated context used only for the sanity check in Appendix B, and a neutral control. The English effect $\Delta^{\text{this-a}}$ is the local score of the pronoun after the

this context minus after the *a* context. Row 4 of Table 1 shows an English item.

This is not a direct semantic analogue of the Persian Δ^{objNP} : Persian $ye(k)$ is an ordinary indefinite cue, while English indefinite *this* is a marked indefinite associated with later uptake and noteworthiness. We use the pronominal anaphor (*it*) as the primary form (Gernsbacher and Shroyer, 1989) rather than demonstrative anaphors, which would be confounded with *this*→*that*. As discussed in Section 3, we do not consider a definiteness contrast (*a* versus *the*), which would introduce a separate set of predictions outside our present scope.

We evaluate three encoders for English: BERT-base-cased, mBERT, and XLM-R. The latter two are the multilingual encoders shared with the Persian arm. The English decoder set comprises Qwen2.5 0.5B/1.5B/3B and XGLM 1.7B/2.9B/4.5B. We do not include mGPT in the English analysis: the Persian arm uses the Persian fine-tune mGPT-Persian, and substituting the base mGPT checkpoint would mix checkpoints.

Our results are shown in Table 6. We find that $\Delta^{\text{this-a}}$ is positive with a 95% CI above zero in every model, encoders and decoders alike (40–44 of 48 items positive). Indefinite *this* is associated with higher pronominal object-uptake scores relative to *a*, consistent with previous psycholinguistic findings. This parallels the Persian result at the level of downstream uptake, but it does not imply that English indefinite *this* and Persian $ye(k)$ have the same semantics. The existence sanity check (affirmative versus negated indefinite context) is positive with a 95% CI above zero in all nine models and for both anaphor forms (Appendix B), consistent with Schuster and Linzen (2022).

Model	$\Delta^{\text{this-a}} (\text{this}-a)$		
	mean	95% CI	Pos.
BERT (en)	0.45	[0.31, 0.61]	44
mBERT	1.09	[0.78, 1.42]	42
XLM-R	0.52	[0.34, 0.72]	44
Qwen2.5-0.5B	0.54	[0.34, 0.74]	41
Qwen2.5-1.5B	0.58	[0.35, 0.80]	43
Qwen2.5-3B	0.69	[0.43, 0.99]	40
XGLM-1.7B	0.59	[0.46, 0.73]	44
XGLM-2.9B	0.55	[0.39, 0.72]	42
XGLM-4.5B	0.43	[0.29, 0.59]	42

Table 6: English local anaphor-span effects (pronoun continuation). The indefinite-*this* versus *a* contrast is positive with a 95% CI above zero in all nine models. This gives a functional point of comparison for an accessibility-raising indefinite contrast. Pos.: positive cases out of 48.

7 Discussion and Conclusion

In this study, we investigated whether language models treat Persian *ye(k)* as a discourse-accessibility cue. Mean local object-NP effects are positive across all tested encoders and decoders, although XGLM-1.7B does not survive the subject and neutral controls and the whole-continuation results remain model-sensitive. The local results therefore support the prediction from Persian linguistics that a *ye(k)*-marked object is associated with greater availability for later anaphoric uptake than a bare object.

The effect is clearer at the local uptake site than over the whole continuation. This is theoretically expected: the dependency is realized when the continuation picks up the object referent, while a full continuation also contains event plausibility, lexical frequency, register, fluency, object-domain effects, and model-specific preferences. The measured visibility of this discourse cue therefore depends on the scoring unit.

We have carried out a series of control analyses to validate the claim against possible confounders: the full object-NP span tests referential material beyond a single marker, the subject-span control isolates the object-anaphoric span from a general context shift, and the decoder neutral baseline tests whether the object-NP effect exceeds a generic continuation-fluency shift.

Two aspects of these results are worth discussing in more detail. First, we observe comparable di-

rectional patterns in underrepresented Persian and high-resource English, despite the different nominal cues used in the two languages. This supports a functional parallel in downstream uptake behavior, but does not distinguish a shared accessibility mechanism from separate cue-specific sensitivities, nor does it establish semantic equivalence between Persian *ye(k)* and English indefinite *this*. Our current study also cannot establish whether information associated with *ye(k)* is merely correlated with the behavioral effect or is causally recruited during object-referent integration. We conducted a pilot distance-based analysis of encoder representations, but it produced no consistent cross-model result (Appendix A). We leave an intervention-based investigation to future work.

Second, because we operationalize accessibility through model surprisal (Levy, 2008), the direction of the effect is compatible with a processing prediction: object-anaphoric uptake should be easier after a context that makes the object referent more available. Model surprisal is only a linking measure here, not a direct human-processing measure, and recent work cautions against reading it as one (Oh and Schuler, 2025; von der Malsburg and Padó, 2026). Whether human readers show the same direction and magnitude remains an empirical question. Consistent with this cautious reading, the comparable uptake effect for a Persian morphosyntactic cue and for English indefinite *this* motivates treating downstream anaphoric availability as the functional point of comparison, rather than identical specificity marking. Further work should test the experimental materials more rigorously against potential confounds. First, for the Persian arm, an item-matched neutral-control set applicable to both encoders and decoders could replace the single fixed neutral string and provide a uniform fluency baseline. Second, human reading-time or judgment data on the same items would calibrate the model scores against processing measures, especially for the demonstrative-NP versus same-NP forms.

Limitations

The Persian stimulus set is small and controlled. It contains 48 items with everyday concrete objects and a restricted set of transitive predicates. This restriction supports tightly controlled minimal contrasts and item-level analysis, similar to well-controlled psycholinguistic stimulus sets, but limits generalization to broader Persian discourse.

Larger and more varied stimulus sets are needed to test generalization across registers, abstract nouns, predicate classes, and naturalistic discourse contexts.

The primary English *this*-versus-*a* analysis is restricted to the pronominal anaphor. Demonstrative-NP continuations are included in the existence sanity check, but not treated as another primary test of the *this*-versus-*a* contrast. Broader comparisons involving specificity, definiteness, demonstrative re-access, and human validation remain future work.

The Persian naturalness check was performed by one native-speaker linguist involved in the study. The present study, therefore, does not include independent naturalness ratings or inter-annotator agreement for the stimulus set.

The evidence comes from transformer scoring behavior under controlled manipulations and has not yet been calibrated against independent human judgments or processing measures. Human continuation judgments or reading-time data on the same items would provide that independent behavioral comparison, especially for demonstrative-NP and same-NP continuations. The model comparison is descriptive because the models differ in architecture, training data, tokenizer, vocabulary, language coverage, and scale. For the same reason, raw log-score magnitudes should be compared within a given model and scoring method, rather than treated as directly comparable effect sizes across model families. Encoder pseudo-log-likelihood is also an approximation rather than a true sentence probability, which is why scoring choice is treated as part of the object of analysis. The Persian neutral-continuation baseline is coarse and reported only for decoders in this version. It controls for a generic left-to-right fluency shift, but not for item-specific neutral-continuation effects. An item-matched Persian neutral-control set for both encoders and decoders would provide a stronger test of whether the object-NP effect exceeds general context fluency.

References

- Mira Ariel. 1990. *Accessing Noun-Phrase Antecedents*. Routledge, London.
- Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. *Is incoherence surprising? targeted evaluation of coherence prediction from language models*. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computa-*

tional Linguistics: Human Language Technologies, pages 4164–4173, Online. Association for Computational Linguistics.

- Sofiana Chiriacescu and Klaus von Heusinger. 2010. *Discourse prominence and pe-marking in Romanian*. *International Review of Pragmatics*, 2(2):298–332.
- Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. *Unsupervised cross-lingual representation learning at scale*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. *BERT: Pre-training of deep bidirectional transformers for language understanding*. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186.
- Mehrdad Farahani, Mohammad Gharachorloo, Marzieh Farahani, and Mohammad Manthouri. 2021. *Pars-BERT: Transformer-based model for persian language understanding*. *Neural Processing Letters*, 53:3831–3847.
- Richard Futrell, Edward Gibson, and Roger P. Levy. 2020. *Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing*. *Cognitive Science*, 44(3):e12814.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. *SyntaxGym: An online platform for targeted evaluation of language models*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.
- Morton Ann Gernsbacher and Suzanne Shroyer. 1989. *The cataphoric use of the indefinite this in spoken narratives*. *Memory & Cognition*, 17(5):536–540.
- Jila Ghomeshi. 2003. *Plural marking, indefiniteness and the noun phrase*. *Studia Linguistica*, 57(2):47–74.
- Michael Hanna and Aaron Mueller. 2025. *Incremental sentence processing mechanisms in autoregressive transformer language models*. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3181–3203, Albuquerque, New Mexico. Association for Computational Linguistics.
- Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P. Levy. 2020. *A systematic assessment of syntactic generalization in neural language models*. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744.

- Tania Ionin. 2006. [This is definitely specific: Specificity and definiteness in article systems](#). *Natural Language Semantics*, 14(2):175–234.
- Masoud Jasbi. 2016. [Three types of indefinites in persian: Simple, complex, and antidefinite](#). In *Proceedings of Semantics and Linguistic Theory 26*, pages 244–263.
- Masoud Jasbi. 2020. [The meaning of the persian object marker rā: What it is not, and what it \(probably\) is](#). In Richard K. Larson, Sedigheh Moradi, and Vida Samiiian, editors, *Advances in Iranian Linguistics*, pages 119–136. John Benjamins.
- Simin Karimi. 1990. Obliqueness, specificity, and discourse functions: *rā* in persian. *Linguistic Analysis*, 20(3–4):139–191.
- Carina Kauf and Anna A. Ivanova. 2023. [A better way to do masked language model scoring](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 925–935.
- Najoung Kim and Sebastian Schuster. 2023. [Entity tracking in language models](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3835–3855, Toronto, Canada. Association for Computational Linguistics.
- Sanghee J. Kim and Kanishka Misra. 2026. [Hey, wait a minute: on at-issue sensitivity in language models](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 214–224, Rabat, Morocco. Association for Computational Linguistics.
- Manfred Krifka and Fereshteh Modarresi. 2016. [Number neutrality and anaphoric update of pseudo-incorporated nominals in Persian \(and weak definites in English\)](#). In *Proceedings of Semantics and Linguistic Theory 26*, pages 874–891.
- Roger Levy. 2008. [Expectation-based syntactic comprehension](#). *Cognition*, 106(3):1126–1177.
- Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O’Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona Diab, and 2 others. 2022. [Few-shot learning with multilingual generative language models](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. [Assessing the ability of LSTMs to learn syntax-sensitive dependencies](#). *Transactions of the Association for Computational Linguistics*, 4:521–535.
- Rebecca Marvin and Tal Linzen. 2018. [Targeted syntactic evaluation of language models](#). In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202.
- Fereshteh Modarresi and Manfred Krifka. 2021. [Pseudo incorporation and anaphoricity: Evidence from persian](#). *Glossa: A Journal of General Linguistics*, 6(1):124.
- Byung-Doh Oh and William Schuler. 2025. [The impact of token granularity on the predictive power of language model surprisal](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4150–4162, Vienna, Austria. Association for Computational Linguistics.
- Qwen Team. 2024. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. [Masked language model scoring](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712.
- Sebastian Schuster and Tal Linzen. 2022. [When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics.
- Hagyeong Shin and Sean Trott. 2024. Do language models capture implied discourse meanings? an investigation with exhaustivity implicatures of Korean morphology. In *Proceedings of the Society for Computation in Linguistics 2024*, pages 150–161, Irvine, CA. Association for Computational Linguistics.
- Oleh Shliakhko, Alena Fenogenova, Maria Tikhonova, Anastasia Kozlova, Vladislav Mikhailov, and Tatiana Shavrina. 2024. [mGPT: Few-shot learners go multilingual](#). *Transactions of the Association for Computational Linguistics*, 12:58–79.
- Nathaniel J. Smith and Roger Levy. 2013. [The effect of word predictability on reading time is logarithmic](#). *Cognition*, 128(3):302–319.
- Ionut-Teodor Sorodoc, Kristina Gulordava, and Gemma Boleda. 2020. [Probing for referential information in language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4177–4189, Online. Association for Computational Linguistics.
- Titus von der Malsburg and Sebastian Padó. 2026. [Diverging transformer predictions for human sentence processing: A comprehensive analysis of agreement attraction effects](#). In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing*, Budapest, Hungary. Association for Computational Linguistics. Accepted for publication.

Klaus von Heusinger and Roya Sadeghpour. 2020. [The specificity marker -e with indefinite noun phrases in modern colloquial Persian](#). In Kata Balogh, Anja Latrouite, and Robert D. Van Valin, Jr., editors, *Nominal Anchoring: Specificity, Definiteness and Article Systems across Languages*, Topics at the Grammar-Discourse Interface 5, pages 115–147. Language Science Press, Berlin.

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

Ethan Gotlieb Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P. Levy. 2023. [Testing the predictions of surprisal theory in 11 languages](#). *Transactions of the Association for Computational Linguistics*, 11:1451–1470.

Ethan Gotlieb Wilcox, Pranali Vani, and Roger P. Levy. 2021. [A targeted assessment of incremental processing in neural language models and humans](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 939–952.

Xiaomeng Zhu and Robert Frank. 2024. [LIEDER: Linguistically-informed evaluation for discourse entity recognition](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13835–13850, Bangkok, Thailand. Association for Computational Linguistics.

Xiaomeng Zhu, Zhenghao Zhou, Simon Charlow, and Robert Frank. 2025. [Meaning beyond truth conditions: Evaluating discourse level understanding via anaphora accessibility](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8824–8842, Vienna, Austria. Association for Computational Linguistics.

A Layer-wise Representation Diagnostic

This supplementary diagnostic is not part of the evidence for H1 or H2: under the pre-specified include/drop rule described below, it did not meet the criteria for inclusion as supporting evidence. We also tested whether the behavioral object-NP $ye(k)$ effect leaves a consistent trace in encoder hidden states. For mBERT, ParsBERT, and XLM-R, we extracted per-layer hidden states and computed a distance-based object-minus-subject separation diagnostic. For each item and layer, the diagnostic measured the $ye(k)$ -versus-bare representational distance at the object-anaphoric NP minus the corresponding $ye(k)$ -versus-bare distance

for the subject-control span. Positive values therefore indicate greater representational separation at the object-anaphoric uptake site than at the subject-control site; they do not indicate a higher model score, lower surprisal, or a directional movement toward an anaphoric representation.

The analysis was evaluated under a pre-specified include/drop rule. We included the layer-wise diagnostic as supporting evidence only if all three conditions were met: (i) the non-early object-minus-subject distance contrast had a confidence interval above zero in at least two of the three encoders; (ii) the peak object-minus-subject distance contrast was not in layers 0–2 for at least two of the three encoders; and (iii) the best non-early object-minus-subject distance contrast was positive in all three encoders.

Criteria (i) and (iii) failed while (ii) held: only one of three encoders had a non-early peak confidence interval above zero (i), and only one had a positive best non-early contrast (iii), though all three had peaks outside layers 0–2 (ii). The per-model peaks were: mBERT layer 10, mean = -0.000574, 95% CI [-0.003882, 0.002294]; ParsBERT layer 12, mean = 0.002557, 95% CI [0.001332, 0.003813]; and XLM-R layer 12, mean = -0.000113, 95% CI [-0.000220, -0.000010]. Item-level representation-behavior correlations were weak: Pearson correlations ranged from 0.122 to 0.286, and Spearman correlations ranged from 0.124 to 0.213.

We therefore do not treat the layer-wise representation diagnostic as evidence for a consistent cross-model hidden-state signature. The behavioral result reported in the main paper is an input-output-level scoring effect. Locating its internal basis, and testing whether $ye(k)$ -related information is causally recruited during object-referent integration, requires intervention-based analyses such as activation patching.

The per-layer summaries and the include/drop decision file will be included in that repository.

B Existence Sanity Check

As a validation check on the English item set, we compare the affirmative indefinite context against a negated counterpart (*Emma did not find a notebook in the library*). The measure is whole-continuation: for each context c we take the score of the anaphoric continuation minus the score of the item-matched neutral continuation, and report

Model	pronoun		dem. NP	
	mean	95% CI	mean	95% CI
BERT (en)	0.12	[0.08, 0.16]	0.14	[0.10, 0.18]
mBERT	0.07	[0.03, 0.12]	0.11	[0.07, 0.14]
XLM-R	0.16	[0.13, 0.19]	0.10	[0.07, 0.14]
Qwen2.5-0.5B	0.31	[0.22, 0.41]	0.17	[0.09, 0.25]
Qwen2.5-1.5B	0.22	[0.12, 0.32]	0.21	[0.11, 0.32]
Qwen2.5-3B	0.39	[0.27, 0.52]	0.37	[0.24, 0.50]
XGLM-1.7B	0.09	[0.03, 0.14]	0.10	[0.05, 0.15]
XGLM-2.9B	0.10	[0.06, 0.15]	0.10	[0.05, 0.14]
XGLM-4.5B	0.11	[0.07, 0.16]	0.06	[0.02, 0.10]

Table 7: Existence sanity check on the English item set: affirmative minus negated indefinite context, whole-continuation measure with the item-matched neutral continuation subtracted. All 18 contrasts are positive with a bootstrap 95% CI above zero.

the affirmative-minus-negated difference. A referent introduced under negation should not support later anaphoric uptake, so positive values are expected. Table 7 confirms this in all nine models for both the pronominal and the demonstrative-NP anaphor.