Investigating Syntactic Biases in Multilingual Transformers with RC Attachment Ambiguities in Italian and English

Michael Kamerath and Aniello De Santo
Dept. of Linguistics
University of Utah
{michael.kamerath, aniello.desanto}@utah.edu

Abstract

This paper investigates whether monolingual and multilingual LLMs show human-like preferences when presented with examples of relative clause attachment ambiguities in Italian and English. We also test whether these preferences can be modulated by lexical factors (the type of verb/noun in the matrix clause) which have been shown to be tied to subtle constraints on syntactic and semantic relations. Our results overall showcase how LLM behavior varies inconsistently across models and languages, and highlight the importance of leveraging subtle syntactic contrasts in exploring these models' ability to correctly align with human-like preferences.

1 Introduction

A recent but already classical line of work has focused on evaluating neural models' predictions on fine-grained syntactic phenomena/constructions, in order to probe whether the models have learned knowledge about the specific structural characteristics of (a) language (Linzen et al., 2016; Marvin and Linzen, 2018; Gauthier et al., 2020; Warstadt et al., 2019, 2020b,a; Sartran et al., 2022; Newman et al., 2021a; Jumelet et al., 2024; Arora et al., 2024). This type of comparison leverages psycholinguistic design to gain insight into the opaque (learned or architectural) biases of LLM (Linzen and Baroni, 2021; Futrell et al., 2019; Ettinger, 2020).

While a majority of past work has focused on evaluating LLMs' syntactic knowledge in terms of their ability to distinguish grammatical and ungrammatical constructions, an important component of human sentence comprehension is ambiguity resolution (Altmann, 1998; Gibson and Pearlmutter, 1998). In particular, it is worth investigating how neural models handle multiple *simultaneously correct interpretations* for a single sentence in the absence of disambiguating cues/context (Davis and van Schijndel, 2020; Bhattacharya et al., 2022; Liu et al., 2023; Zhou et al., 2024).

Consider the case of the relative clause (RC) "*that was running*" following the complex noun phrase "*son of the doctor*", as in (1):

(1) I saw the son of the doctor that was running.

There are two possible interpretations of this sentence: in one interpretation the RC modifies *the doctor* (low attachment; LA), and in the other it modifies *the son* (high attachment; HA). Famously, human preferences for HA or LA vary both individually and cross-linguistically, and are affected by a variety of syntactic and semantic factors (Cuetos and Mitchell, 1988; De Vincenzi and Job, 1993; Mitchell et al., 1990; Miyamoto, 1998; Maia et al., 2007; Abdelghany and Fodor, 1999). Moreover, some of these factors (e.g., the type of verb used in the matrix clause of the sentence) seem to be tied to subtle syntactic differences in each language (Cinque, 1992; Grillo et al., 2015; Grillo and Costa, 2014, a.o.).

RC attachment ambiguities thus present an interesting way of probing LLMs' syntactic knowledge and behavior. In fact, investigating LLMs' performance over ambiguous sentences cross-linguistically might provide crucial insights into the kind of linguistic biases available to these models through their training data, and the properties of the models tied to architectural choices (Davis and van Schijndel, 2020; Li et al., 2024). As differences in the frequency of HA vs. LA structures have been argued to account for the cross-linguistic variation of RC preferences at least to some degree, it seems reasonable that LLMs would be able to replicate (some of) these patterns. However, RC attachment seems to be understudied in the LLM syntactic evaluation literature (Davis and van Schijndel, 2020; Issa and Atouf, 2024; Lee et al., 2025; Scheinberg et al., 2025).

Here, we aim to add to this scarce literature, and evaluate a variety of LLMs to determine their disambiguation strategies for RCs in Italian and English. We compare Italian to English since the two lan-

Proceedings of the Society for Computation in Linguistics (SCiL) 2026, pages 191–218

guages have some shared structural properties (e.g., SVO, post-nominal RCs), but differ in RC interpretation: modulo other variables, English speakers generally exhibit a LA RC preference while Italian speakers a HA one (Frazier, 1983; Cuetos and Mitchell, 1988; De Vincenzi and Job, 1993). Additionally, Italian and English speakers respond differently to other variables affecting RC attachment, which have been argued to be also captured by some multilingual LLMs (e.g., type of matrix verb; Grillo et al., 2015; Grillo and Costa, 2014; Henot-Mortier, 2023). Therefore, building on the psycholinguistics and LLM literature on RC attachment, here we ask:

1. whether monolingual and multilingual LLMs tested on Italian and English show any type of attachment preference when presented with ambiguous RCs;

2. whether these preferences conform to those of Italian/English speakers;

3. whether these preferences show sensitivity to fine-grained structural information modulated by properties of the matrix clause.

2 Related Work

The cross-linguistic variability of attachment preferences for ambiguous RCs has been focus of many psycholinguistics debates, due to its direct relevance to questions about the mechanisms guiding human sentence processing (Frazier, 1983; Cuetos and Mitchell, 1988; De Vincenzi and Job, 1993; Gibson and Pearlmutter, 1998; Grillo et al., 2015; Hemforth et al., 2015; Lee and De Santo, 2024, a.o.).

Famously, when presented with a globally ambiguous sentence like in (1), and in the absence of a disambiguating context, English speakers tend to prefer a LA interpretation: an interpretation in which the RC gives us information about (*modifies)* the second noun of the preceding complex noun phrase (Frazier, 1983; Cuetos and Mitchell, 1988). This LA preference is well attested in other languages including Mandarin Chinese and Arabic (Shen, 2006; Abdelghany and Fodor, 1999; Ehrlich, 1999). In turn, a preference for the RC modifying the first noun — a HA interpretation — has been found in languages like Italian, Spanish, and Dutch (Cuetos and Mitchell, 1988; De Vincenzi and Job, 1993; Brysbaert, 1996; Frenck-Mestre and Pynte, 2000; Mitchell et al., 2000). Beyond these broader preferences at the language level, multiple factors have been shown to affect RC preferences across languages — for instance, referentiality of the modified nouns, lexical and structural frequency, and semantic or pragmatic plausibility (De Vincenzi and Job, 1993; MacDonald et al., 1994; Gilboy et al., 1995; Ferreira, 2003; Fernández, 2003; Swets et al., 2008; Acuna-Farina et al., 2009).

Recently, it has been argued that one important predictor of attachment disambiguation in Italian RCs is whether the verb in the main clause is non-perceptual (*marry, know, cook, etc*) or perceptual (*observe, hear, smell, etc*). When other semantic and syntactic aspects are controlled for, RCs of sentences containing non-perceptual verbs lead to a LA preference while perceptual verbs lead to a HA preference (Grillo and Costa, 2014; Lee and De Santo, 2024). More generally, reviewing past literature on RC attachment preferences in so-called HA languages, Grillo and Costa (2014) have related this verb-type sensitivity to a subtle structural ambiguity at the complementiser, beyond the classic LA RC vs. HA RC choice. Some languages allow for a construction known as a Pseudo-Relative Clause (PRs), which is string-identical to RCs but different at the semantic, syntactic, and prosodic levels (Cinque, 1992; Grillo and Costa, 2014; Aguilar and Grillo, 2021, a.o.). In particular, instead of providing information about the entity (noun) that is modified (2a), PRs denote direct perception of events (2b) and are thus only compatible with some specific subclasses of verbs (e.g., *photograph, record*) in the matrix clause (perception verbs, introducing events).

(2) Gianni vide il figlio che correva.

a. Gianni vide [$_{DP}$il [$_{NP}$figlio [$_{RC}$che correva]]].
Gianni saw [the [son [that ran]]].

b. Gianni vide [$_{SC}$[$_{DP}$il figlio] [che correva]].
Gianni saw [[the son] [running]].

While RCs modify nouns, PRs are analyzed as complements of perceptual verbs and are only compatible with what looks like a HA interpretation, leading to an apparent HA preference with verbs that license them (Cinque, 1992, a.o.). This hypothesis has found experimental support in a variety of languages including Italian (Grillo and Costa, 2014; Lee and De Santo, 2024) and Spanish (Aguilar and Grillo, 2021). RC attachment thus offers ways to explore the sensitivity of LLMs to important structural and semantic features within and, crucially, across languages.

Starting with Linzen et al. (2016), there has been a fruitful line of research using psycholinguistic tasks to explore neural models' knowledge of different lexical, structural, and semantic linguistic properties (Marvin and Linzen, 2018; Gauthier et al., 2020, 2022; Warstadt et al., 2019, 2020b,a; Sartran et al., 2022; Newman et al., 2021b; Jumelet et al., 2021; Arora, 2022; Gulordava et al., 2018; Sinclair et al., 2022; Goldberg, 2019; Wilson et al., 2023) — and to evaluate whether model behavior resembles the performance of humans tested on similar tasks/constructions (**?**Futrell et al., 2019; Ettinger, 2020).

While some work probing LLMs' ability to deal with (different types of) ambiguity exists (Van Schijndel and Linzen, 2018; Bhattacharya et al., 2022; Liu et al., 2023; Zhou et al., 2024; Li et al., 2024), little attention has been paid to the phenomenon of RC attachment in absence of a disambiguating context. In this sense, Davis and van Schijndel (2020) analyzed the ability of LSTMs to learn RC attachment preferences in English and Spanish. They showed that LSTMs preferred English-like attachment (LA) in both English and Spanish. More recently, Issa and Atouf (2024) tested RC attachment in Arabic with a variety of transformer models (Vaswani et al., 2017), using a zero-shot prompting method. They showed significant variability across model architectures, with some models' behavior being in line with the attachment preferences reported for Arabic speakers, while others showing no preference at all. Similarly, Lee et al. (2025) evaluated whether world knowledge biases help RC attachment disambiguation in six typologically diverse languages. Using prompting methods that simulated a force-choice decision task, they showed how three models (Claude 3.5 Sonnet, GPT-4o, and Llama 3.1) exhibited a strong attachment preference across languages (see also Scheinberg et al., 2025). Furthermore, going back to our discussion of linguistic factors that modulate RC preferences, Henot-Mortier (2023) has shown that monolingual and multilingual transformer architectures exhibit some sensitivity to PR-availability in French. However, this work evaluated PR-related properties only in contexts outside of ambiguous RC, and it thus unclear whether they would modulate an LLM's choice of attachment.

In sum, the complex interaction between RC attachment and other syntactic/semantic factors opens an exciting set of possibilities for the cross-linguistic evaluation of LLMs' linguistic performance. In what follows, building on the results of Davis and van Schijndel (2020) and Henot-Mortier (2023), we focus on evaluating a set of monolingual and multilingual models on the patterns of RC-attachment and PR-sensitivity reported in the psycholinguistic literature for Italian and English (Grillo and Costa, 2014; Grillo et al., 2015; Lee and De Santo, 2024).

Model Name	Lang.	Reference
GePpeTto	IT	De Mattei et al. (2020)
AlBERTo	IT	Polignano et al. (2019)
UmBERTo	IT	Parisi et al. (2020)
bert-base-multilingual-cased	M	Devlin et al. (2019a)
xlm-mlm-17-1280	M	CONNEAU and Lample (2019)
XLM-RoBERTa-large	M	Conneau et al. (2020)
GPT-2	ENG	Radford et al. (2019)
bert-base-uncased	ENG	Devlin et al. (2019b)

Table 1: Monolingual (IT: Italian; ENG: English) and Multilingual (M) models tested in this paper.

3 Italian Experiment

As mentioned, past literature suggests that in languages that allow for PRs (e.g., Italian) if the matrix verb is perceptual a PR parse takes precedence, resulting in a HA interpretation. Otherwise, a preference for an LA intepretation is observed. Grillo and Costa (2014) tested this prediction by evaluating Italian participants' behavior when exposed to globally ambiguous sentences containing a complex noun phrase followed by an RC. Sentences varied over the type of verb used in the matrix clause (perceptual/stative). As predicted, participants showed a HA preference only with perceptual verbs, and exhibited a LA preference with stative verbs.

Here, we exploit this design to explore whether the type of matrix verb in Italian sentences affects LLM attachment preferences. In the past, a common evaluation technique has been to check whether a model assigns a higher probability to a grammatical sentence compared to an ungrammatical one (Linzen et al., 2016; Gulordava et al., 2018). However, here we are interested in probing an LLM's preference in choosing one grammatical interpretation over another equivalently grammatical one, in the absence of other disambiguating factors (e.g., context). To do so, instead of using the globally ambiguous sentences of Grillo and Costa (2014), we follow Davis and van Schijndel (2020) and adopt sentences that are temporarily ambiguous.

Specifically, we leverage a modification of the Grillo and Costa (2014)'s stimuli presented by Lee and De Santo (2024). We follow a 2×2 design, in which quartets of sentences vary across two

dimensions: Verb Type and Attachment Type. As in Grillo and Costa (2014), sentences include a main verb which is either perceptual (*heard*) or stative (*worked with*) and a complex noun phrase (*the grandma of the girls*) followed by an RC. Items are disambiguated towards HA or LA based on singular/plural agreement between one of the nouns in the matrix clause (*grandma/girls*), and the embedded verb. This is possible since Italian differentiates singular/plural morphology explicitly on the main verb. We use Lee and De Santo (2024)'s items, which include 24 sentence sets for a total of ninety-six sentences. Each set contains 4 sentences varying across the two dimensions mentioned above (Appendix A). In line with the models tested for French by Henot-Mortier (2023), we test three Italian-only models, and three multilingual models (GePpeTto; AlBERTo; bert-base-multilingual-cased; xlm-mlm-17-1280; XLM-roBERTa-large; see Table 1).

3.1 Surprisal Analysis

Following Davis and van Schijndel (2020), we evaluate LLMs using information-theoretic surprisal (Hale, 2001; Levy, 2008), which is usually defined as the inverse log probability assigned to a word in a sentence given its preceding context. Fixed verb-type, our stimuli include sentence pairs that are string identical until the singular/plural features on the disambiguating verb. Thus, for each item in the dataset, we compute surprisal at the embedded verb using the `minicons` library (Misra, 2022).[1] For the xlm-mlm–17-1280 model, three of the sets were excluded due to `minicons` not returning a surprisal value for the disambiguating verb.

For each LLM, we fit a linear mixed-effect (LME) model using Surprisal at the embedded verb as the dependent variable, and Verb Type and Attachment Type as fixed effects. We also include a random intercept for set, in order to account for lexical variation across sentence quartets.[2]. All analyses were performed using the lme4 package (version 1.1.35.5; Bates et al., 2015) in R (R version 4.4.1; R Core Team, 2021).

Statistical analyses show no significant attachment or verb type effects, nor their interaction, for any of the LLMs tested (see Figure 1).[3] These results can be interpreted as the absence of an attachment preference (in line with Italian speakers or not), and a lack of sensitivity to verb-type properties, in both the Italian-only and the multilingual models.

4 English Experiments

Since PR availability co-varies in Italian with semantic properties of the matrix verb (i.e., implicit causality), Grillo et al. (2015) acknowledge that a pragmatic explanation for their results could also be viable (Gilboy et al., 1995; Rohde et al., 2011). They conducted an English study probing similar variables modulating RC attachment as those manipulated in the Italian studies discussed above.

While not a PR-language, English allows for structures that are interpretatively similar to PRs — eventive small clauses (SC), which describe dynamic events as in: *Star heard the monster scream*. These constructions are licensed by perceptual verbs, and have many of the functional and interpretative properties of PRs in Italian. However, English SCs are not string equivalent to RC clauses with an explicit complementiser. Therefore, a PR-based account of attachment preferences predicts that PR-related verb-type effects should not arise with RC sentences in English. On the other hand, a pragmatic account would predict a perceptual/stative effects on attachment in English as observed by Grillo and Costa (2014) in Italian.

In an offline questionnaire, Grillo et al. (2015) showed that English participants consistently prefer LA, although they observed a small HA boost in the SC-licensing/Perceptual verb condition. They argue that these results are incompatible with a pragmatic account of the Italian findings, strengthening the PR-availability argument for their Italian results. This design offers us a way to further probe factors affecting RC attachment strategies in LLMs with a direct cross-linguistic comparison of the manipulated variables. Additionally, as implicit causality has been explored in the LLM literature to somewhat conflicting results, this experimental set up might lead to broader insights into LLMs' sensitivity to semantic/pragmatic variables (Kankowski et al., 2025; Ke-

[1] Specifically, for each of the LLMs tested, verbs were tokenized into stems and suffixes (e.g. *gridava/vano* is tokenized as [*grida*] and [*va/vano*]). Each LLM was verified to ensure tokenization occurred in this form. All LLMs tested were consistent with this tokenization approach with the exception of the *xlm-mlm-17-1280* model, which tokenized some verbs to include part of the stem in the token. For example, *saltava* was tokenized as [*sal*] and [*tava*] rather than [*salta*] and [*va*]. This linguistically opaque tokenization occurred for 4 out of the 48 Italian stimuli of the form *verb+va*.

[2] `Surprisal ~ Verb Type + Attachment Type + Verb Type*Attachment Type + (1|set)`

[3] For the sake of space, when discussing analyses we only report significance. The full output of each LME model can be found in Appendix B. Scripts and data for all the experiments in this paper can be found at https://shorturl.at/yYgIf

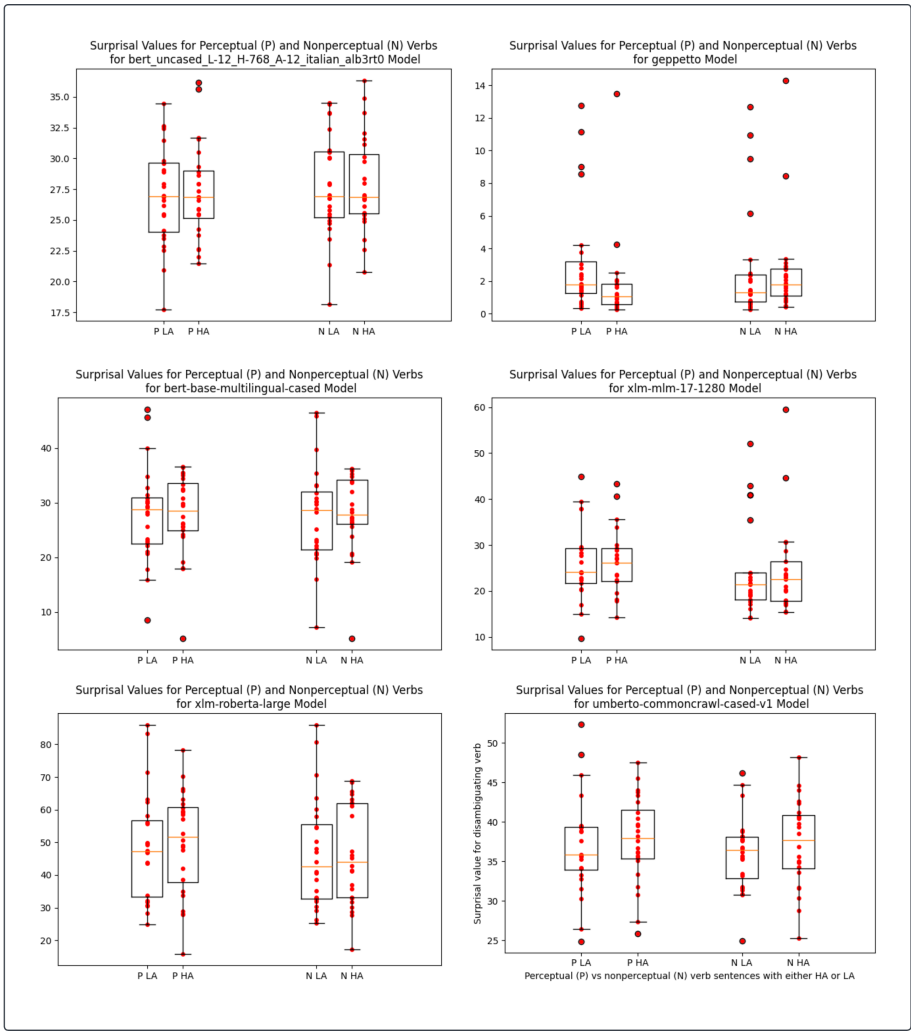

Figure 1: Surprisal values by condition, for each one of the models tested in the Italian Experiment.

mentchedjhieva et al., 2021). In addition to the verb-type manipulation of the Italian experiments, Grillo et al. (2015) also modulate the type of nominal used as the first noun in the complex noun — either licensing a SC or not (*heard* vs. *scream*). This noun-type manipulation implies testing RCs following a complex noun-phrase in the subject position of the main sentence, compared to the object modifying RCs used when manipulating verb type (Example 3).

(3) a. Kelly heard the grandma of the girl that was screaming.

b. The sounds of the grandma of the girl that was screaming is annoying.

We adopt this design, but we will again depart from the baseline psycholinguistic study by using disambiguated RCs over globally ambiguous ones. We thus split Grillo et al. (2015)'s experiment into two: Experiment 1 will test the effect of verb-type in English, while Experiment 2 will text the effects of noun-type/RC position. In line with the Italian experiment, we test two monolingual English models, and the three multilingual models evaluated in Section 1.

4.1 Experiment 1: Verb-Type Effects

First, we investigate Verb Type effects in English, using stimuli adapted from the first experiment in Grillo et al. (2015). These include sets of four lexically matched items holding all properties of a sentence constant except for the matrix verb, which is either a RC-only verb or a SC-licensing verb (Appendix A). Grillo et al. (2015) report that human participants tested on these stimuli showcase a general preference for LA, but a slight HA boost in the SC-licensing condition.

Similarly to the Italian experiment, all items were modified to disambiguate LA/HA based on singular/plural agreement on the embedded verb. Because of the properties of English, this disambiguation happens over an auxiliary verb

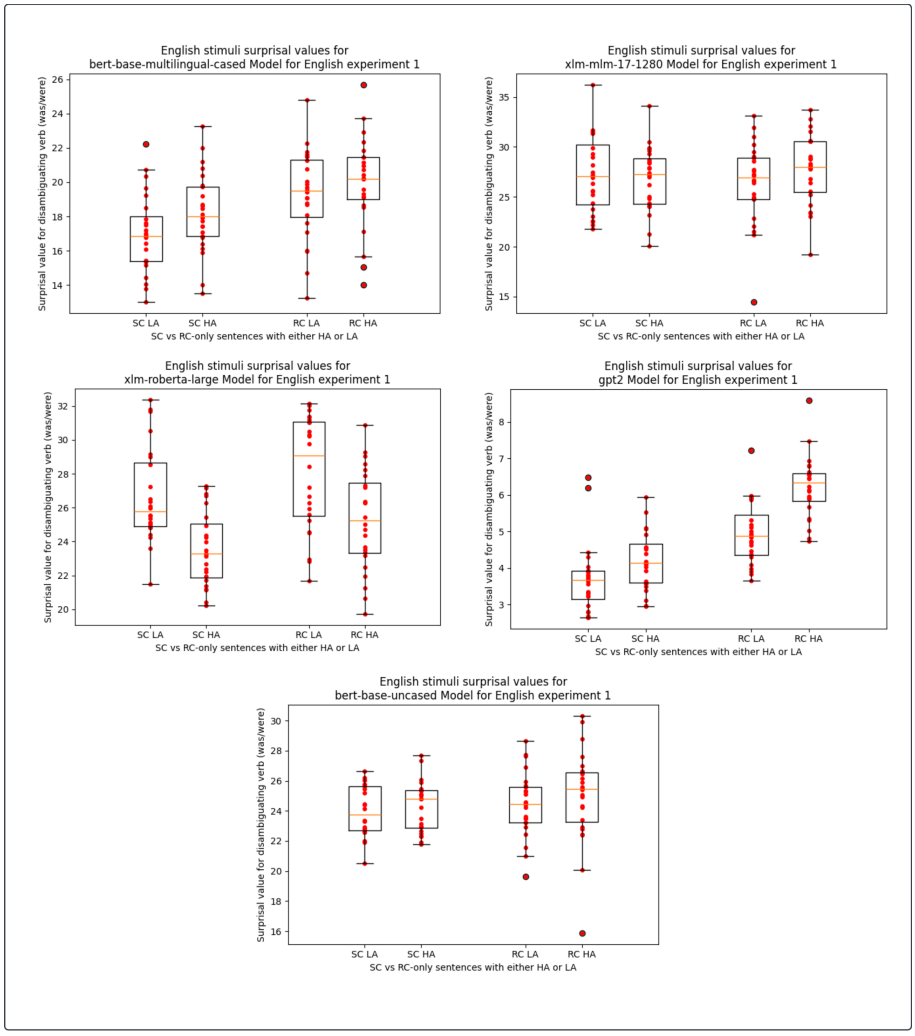

Figure 2: Surprisal values by condition, for each one of the models tested for Verb-type Effects in English (Exp. 1).

(*was/were*) instead of directly on the embedded verb. Experimental stimuli included twenty-four sets, for a total of ninety-six sentences. We fit an identical linear mixed-effect model model as for the Italian Experiment, using Surprisal at the embedded verb as the dependent variable, and Verb Type and Attachment Type as fixed effects.[4]

Compared to the Italian Experiments, results here are more mixed (Figure 2). For the multilingual bert-base model, we found a significant Verb Type effect, consistent with surprisal values being generally lower in the SC-licensing verb condition that in the RC-only condition. These differences are independent of Attachment Type, although with SC verbs we observe a (non-significant) trend in favor of the LA condition — which is line with the known LA preference in English, but somewhat in contrast with what Grillo et al. (2015) found with human participants. No significant effects were found with the xlm model, but there were marginal effects of Attachment Type and of the Verb Type/Attachment Type interaction. The xlm model's results do trend towards lower surprisal for LA in the RC-only condition (Figure 2). For the RoBERTa model we found significant Very Type and Attachment Type effects, but no interaction effects. Again, surprisal values in the SC condition are lower independently of Attachment Type (Figure 2). Additionally, surprisal values for HA items are significantly lower than those of LA items (thus indicating a HA preference). Considering now the monolingual models, with GPT-2 we found both Verb-Type and Attachment-Type effects. These results are related to higher surprisal values in the RC-Only and the HA conditions. We additionally found a significant interaction effect. Descriptively this is the closest the pattern of results in human experiments. It is worth observing the magnitude of these effects: human participants

[4] `Surprisal ~ Verb Type + Attachment Type + Verb Type*Attachment Type + (1|set)`

still exhibit a certain rate of HA choices in both conditions, compared to the heavy LA tendency of GPT-2. Finally we found no significant effects for bert-base-uncased across conditions.

4.2 English Experiment 2: Noun-Type Effects

In a second experiment, we leverage the stimuli in the nominal condition of Grillo et al. (2015)'s first experiment. This condition compares nominals that license SC (i.e., event-takin nouns compatible with the description of an event, like "*picture*" or "*sound*") to nominals that are only compatible with RCs (e.g., "*car*" , "*comb*"). As mentioned above, the nominal condition is also designed so that the complex noun phrase (and thus the following RC) occupies the subject position of the matrix clause (as in Table 4). On these stimuli, English participants show a LA preference, but no noun-type effect (Grillo et al., 2015).

For our LLM tests, we again modify all items to disambiguate LA/HA based on singular/plural agreement on the embedded verb, resulting in a 2×2 design. Results from linear-mixed effect models[5] for each LLM are mixed, but generally in line with those in the first English experiment (Figure 3). For the multilingual bert model, we find a strong effect of Attachment Type, no effect of Noun Type, and no interaction. These are compatible with bert strongly preferring LA items independently of the noun manipulation. For the xlm model, we found a significant clause type effect, but no effect of attachment, nor an interaction. This is compatible with a tendency for slightly higher surprisal values with SC-licensing nominals independently of Attachment Type. For the RoBERTa model we again found a strong Attachment Type effect, with no interaction with Noun Type. This is the result of a strong preference for HA items across Noun-Type conditions (Figure 3), in line with Roberta's behavior in Experiment 1. As for the English-only models, with GPT-2 we found a significant effect of Attachment type. Similarly, bert-base-uncased presents a significant sensitivity to attachment type and no other effect. It is also interesting to observe that although monolingual and multilingual bert behave similarly in their trends, their patterns are not human like.

[5] `Surprisal` $\sim$ `Noun Type + Attachment Type + Noun Type*Attachment Type + (1|set)`

5 Discussion and Further Work

This paper contributes to work evaluating the sensitivity of LLMs to a variety of linguistic properties, by exploring whether/how the preference for High or Low attachment of RCs is affected by syntactic/semantic properties of the matrix clause. Our results overall showcase how LLM behavior does not align with human data and, more importantly, varies inconsistently across models and languages.

Specifically, we measured the difference in surprisal of locally ambiguous sentences at the point of disambiguation (the embedded verb) to determine whether a number of (monolingual and multilingual) LLMs align with human-like RC attachment preferences in Italian and English. Furthermore, we tested whether these preferences can be modulated by lexical factors in the matrix clause (Verb Type or Noun Type), which have been argued to be related to subtle differences between RCs and other constructions.

For Italian, statistical results over raw surprisal values indicate that none of the models we tested exhibits any attachment preference at all, whether in line with the human results or not. Furthermore, we do observe high item-level variability, which is also not human-like. Even though we control for item-level lexical effects in our statistical models, because of this stark item-based variability we do note interesting (non statistically significant) tendencies in some of the models that beg for deeper inquiry in future work. For instance, modulo some high surprisal LA items, the GePpeTto model shows a general qualitative preference towards LA, in particular with perceptual verbs.

English results across two experiments were more mixed. While some models showcase some type of attachment preference, and at times verb and noun type effects on these preferences, these were not exactly in line with human data. For instance, while the multilingual bert-base model does show a slight preference for LA items, the RoBERTa model shows a strong bias towards HA items, in contrast with the reported LA preference for English. The mirrored behavior of BERT and RoBERTa across the two English experiments is also of note, and opens question for future comparisons — as does the fact that surprisal values across models were slightly higher in the RC-only condition.

Notably, our statistical results are also somewhat in contrast with what previous work found for Spanish and Arabic (Davis and van Schijndel, 2020; Issa and Atouf, 2024), which found a general LA

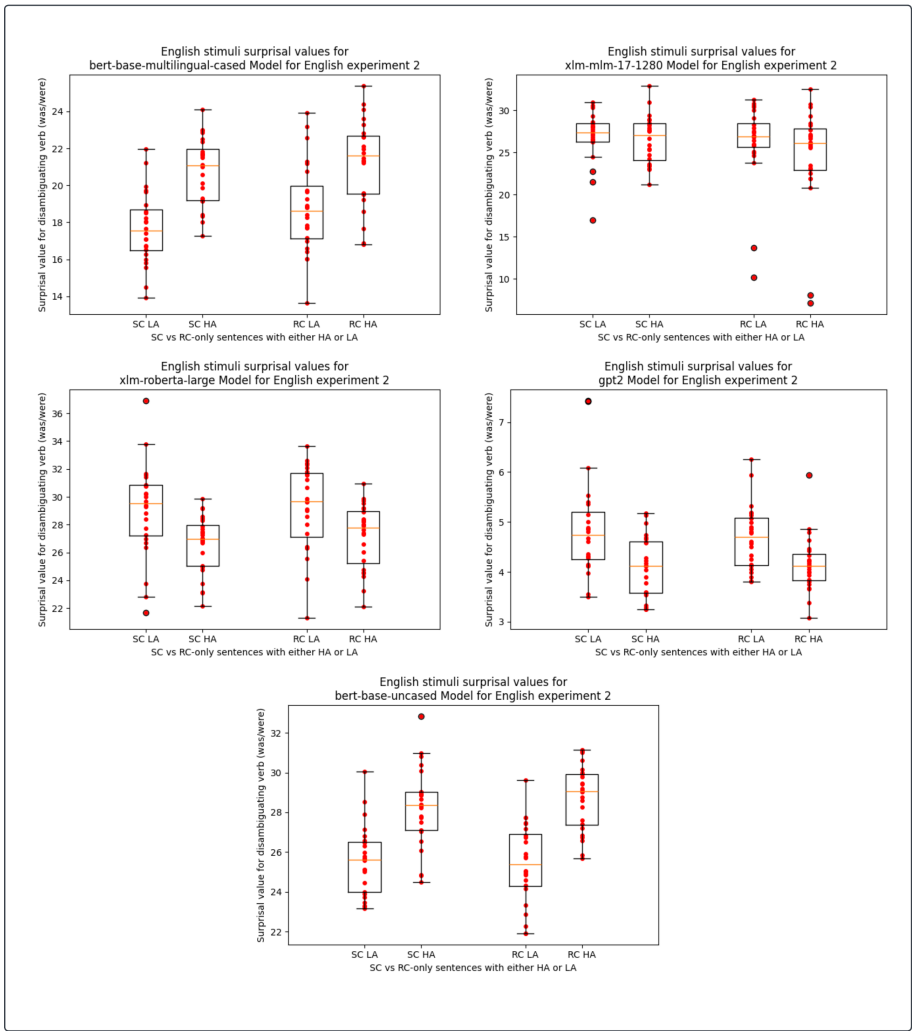

Figure 3: Surprisal values by condition, for each one of the models tested for Noun-Type effects in English (Exp. 2).

preference even in HA languages. However, Davis and van Schijndel (2020) tested models with an LSTM architecture, while Issa and Atouf (2024) used prompting methods as opposed to the surprisal measurements used here. Future work should then probe differences between architectures and tasks/measures more in depth. With respect to our comparison with some of the recent literature on LLM attachment preferences, it is worth considering the effects of task. Some studies on LLMs and RCs have adopted prompting (Brown et al., 2020) as a method of studying LLMs's sensitivity to RC properties in ambiguous attachment contexts (Issa and Atouf, 2024; Scheinberg et al., 2025; Lee et al., 2025). Importantly, prompting tasks have been criticized with respect of their ability to provide insights into the linguistic representations and decision processes of LLMs (Hu and Levy, 2023; Hu and Frank, 2024). Crucially, prompting tasks rely on specific meta-linguistic information that might confound the evaluation of model sensitivity to linguistic properties. For instance, Lee et al. (2025) set-up an evaluation pipe-line dependent on a model's ability to first identify an RC and DPs related to it. Generally, this tension between past work and our results suggest the necessity of testing similar constructions while keeping architectural (and task) details constant.

Overall, our work suggests a primary role for structural ambiguity in the study of LLMs' capabilities cross-linguistically, problematizing general, broad-stroke claims about "human-like" performance of LLMs, and strengthening the argument in favor of psycholinguistically motivated, cross-linguistics benchmarks. We argue for the importance of studying LLM performance over ambiguous, grammatically well-formed linguistic stimuli, as a fine-grained lens into LLM knowledge and processes.

Limitations

In this paper we argued for the importance of leveraging well-designed psycholinguistic stimuli in investigating LLM's behavior. Therefore, we relied on experimental items available from two psycholinguistic studies of interest in conducting our evaluations. However, this meant that the number of items used in the paper is relatively low compared to the size of test sets in the LLM literature (but comparable to other work on LLM/RC attachment). This choice might also have made our results particularly sensitive to item-level (set) variability, which deserves further investigation. Additionally, a limitation of comparing Italian to English is that in Italian surprisal is measured at the disambiguating verb, which varies lexically across sets, but in English the disambiguating continuation is always measured on the *was/were* contrast consistently. While we controlled for set in our statistical models to partially address item-level confounds, future work on RC attachment and noun/verb type effects should explore this difference, and be extended to multiple languages with and without pseudo-relative constructions.

Acknowledgments

We thank members of the Utah Computational Linguistics Group, the audiences at the Utah Undergraduate Linguistics Conference 2025 and at the University of Utah's first "SPEAKING TO MACHINES: Cross- Disciplinary dialogues in Linguistics, AI, and Computer Science" workshop, and SCiL's anonymous reviewers for helpful feedback on this project.

References

Hala Abdelghany and Janet Dean Fodor. 1999. Low attachment of relative clauses in arabic. *Poster presented at AmlaP (Architectures and mechanisms of language Processing), edinburgh, uk*.

Carlos Acuna-Farina, Isabel Fraga, Javier García-Orza, and Ana Piñeiro. 2009. Animacy in the adjunction of spanish rcs to complex nps. *European Journal of Cognitive Psychology*, 21(8):1137–1165.

Miriam Aguilar and Nino Grillo. 2021. Spanish is not different: On the universality of minimal structure and locality principles. *Glossa: a journal of general linguistics*, 6.

Gerry TM Altmann. 1998. Ambiguity in sentence processing. *Trends in cognitive sciences*, 2(4):146–152.

Aryaman Arora. 2022. Universal Dependencies for Punjabi. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5705–5711, Marseille, France. European Language Resources Association.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. 2024. CausalGym: Benchmarking causal interpretability methods on linguistic tasks. pages 14638–14663.

Douglas Bates, Martin Maechler, Ben Bolker, Steven Walker, Rune Haubo Bojesen Christensen, Henrik Singmann, Bin Dai, Gabor Grothendieck, Peter Green, and M Ben Bolker. 2015. Package 'lme4'. *convergence*, 12(1):2.

Sunit Bhattacharya, Vilém Zouhar, and Ondrej Bojar. 2022. Sentence ambiguity, grammaticality and complexity probes. In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 40–50, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Marc Brysbaert. 1996. Modifier attachment in sentence parsing: Evidence from dutch. *The Quarterly Journal of Experimental Psychology: Section A*, 49(3):664–695.

Guglielmo Cinque. 1992. *The pseudo-relative and ACC-ing constructions after verbs of perception*. Università degli studi di Venezia.

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, Online. Association for Computational Linguistics.

Alexis CONNEAU and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Fernando Cuetos and Don C. Mitchell. 1988. Cross-linguistic differences in parsing: Restrictions on the use of the late closure strategy in spanish. *Cognition*, 30(1):73–105.

Forrest Davis and Marten van Schijndel. 2020. Recurrent neural network language models always learn

English-like relative clause attachment. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1979–1990, Online. Association for Computational Linguistics.

Lorenzo De Mattei, Michele Cafagna, Felice Dell'Orletta, Malvina Nissim, and Marco Guerini. 2020. Geppetto carves italian into a language model. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 97–104.

Marica De Vincenzi and Remo Job. 1993. Some observations on the universality of the late-closure strategy. *Journal of Psycholinguistic Research*, 22:189–206.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Karen Ehrlich. 1999. Low attachment of relative clauses: New data from swedish, norwegian and romanian. In *the 12th Annual CUNY Conference on Human Sentence Processing. New York, NY., 1999*.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Eva M Fernández. 2003. Bilingual sentence processing.

Fernanda Ferreira. 2003. The misinterpretation of noncanonical sentences. *Cognitive Psychology*, 47(2):164–203.

Lyn Frazier. 1983. Processing sentence structure. *Eye movements in reading: Perceptual and language processes*, pages 215–236.

Cheryl Frenck-Mestre and Joël Pynte. 2000. Chapter 21 - 'romancing' syntactic ambiguity: Why the french and the italians don't see eye to eye. In Alan Kennedy, Ralph Radach, Dieter Heller, and Joël Pynte, editors, *Reading as a Perceptual Process*, pages 549–564. North-Holland, Oxford.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Elodie Gauthier, Papa Séga Wade, Thierry Moudenc, Patrice Collen, Emilie De Neef, Oumar Ba, Ndeye Khoyane Cama, Ahmadou Bamba Kebe, Ndeye Aissatou Gningue, and Thomas Mendo'O Aristide. 2022. Preuve de concept d'un bot vocal dialoguant en wolof (proof-of-concept of a voicebot speaking Wolof). In *Actes de la 29e Conférence sur le Traitement Automatique des Langues Naturelles. Volume 1 : conférence principale*, pages 403–412, Avignon, France. ATALA.

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.

Edward Gibson and Neal J Pearlmutter. 1998. Constraints on sentence comprehension. *Trends in cognitive sciences*, 2(7):262–268.

Elizabeth Gilboy, Josep-MMaria Sopena, Charles Cliftrn, and Lyn Frazier. 1995. Argument structure and association preferences in spanish and english complex nps. *Cognition*, 54(2):131–167.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Nino Grillo and João Costa. 2014. A novel argument for the universality of parsing principles. *Cognition*, 133(1):156–187.

Nino Grillo, João Costa, Bruno Fernandes, and Andrea Santi. 2015. Highs and lows in english attachment. *Cognition*, 144:116–122.

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

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*.

Barbara Hemforth, Susana Fernandez, Charles Clifton, Lyn Frazier, Lars Konieczny, and Michael Walter. 2015. Relative clause attachment in german, english, spanish and french: Effects of position and length. *Lingua*, 166:43–64.

Adele Henot-Mortier. 2023. Do language models discriminate between relatives and pseudorelatives? In *Proceedings of the 2023 CLASP Conference on Learning with Small Data (LSD)*, pages 55–61, Gothenburg, Sweden. Association for Computational Linguistics.

Jennifer Hu and Michael Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *First Conference on Language Modeling*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Elsayed Issa and Noureddine Atouf. 2024. Context-biased vs. structure-biased disambiguation of relative clauses in large language models. *Procedia Computer Science*, 244:425–431. 6th International Conference on AI in Computational Linguistics.

Jaap Jumelet, Lisa Bylinina, Willem Zuidema, and Jakub Szymanik. 2024. Black big boxes: Do language models hide a theory of adjective order? *arXiv preprint arXiv:2407.02136*.

Jaap Jumelet, Milica Denic, Jakub Szymanik, Dieuwke Hupkes, and Shane Steinert-Threlkeld. 2021. Language models use monotonicity to assess NPI licensing. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4958–4969, Online. Association for Computational Linguistics.

Florian Kankowski, Torgrim Solstad, Sina Zarriess, and Oliver Bott. 2025. Implicit causality-biases in humans and llms as a tool for benchmarking llm discourse capabilities. *arXiv preprint arXiv:2501.12980*.

Yova Kementchedjhieva, Mark Anderson, and Anders Søgaard. 2021. John praised mary because _he_? implicit causality bias and its interaction with explicit cues in lms. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4859–4871.

So Young Lee and Aniello De Santo. 2024. Online evidence for pseudo-relative effects on italian rc attachment resolution. *Language, Cognition and Neuroscience*, 39(9):1212–1229.

So Young Lee, Russell Scheinberg, Amber Shore, and Ameeta Agrawal. 2025. Who relies more on world knowledge and bias for syntactic ambiguity resolution: Humans or LLMs? 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 3484–3498, Albuquerque, New Mexico. Association for Computational Linguistics.

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

Andrew Li, Xianle Feng, Siddhant Narang, Austin Peng, Tianle Cai, Raj Sanjay Shah, and Sashank Varma. 2024. Incremental comprehension of garden-path sentences by large language models: Semantic interpretation, syntactic re-analysis, and attention. *Preprint*, arXiv:2405.16042.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7(1):195–212.

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.

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.

Maryellen C MacDonald, Neal J Pearlmutter, and Mark S Seidenberg. 1994. The lexical nature of syntactic ambiguity resolution. *Psychological review*, 101(4):676.

Marcus Maia, Eva M Fernández, Armanda Costa, and Maria do Carmo Lourenço-Gomes. 2007. Early and late preferences in relative clause attachment in portuguese and spanish. *Journal of Portuguese Linguistics*, 6(1).

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, Brussels, Belgium. Association for Computational Linguistics.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

Don C. Mitchell, Marc Brysbaert, Stefan Grondelaers, and Piet Swanepoel. 2000. Chapter 18 - modifier attachment in dutchdutch: Testing aspects of construal theory. In Alan Kennedy, Ralph Radach, Dieter Heller, and Joël Pynte, editors, *Reading as a Perceptual Process*, pages 493–516. North-Holland, Oxford.

Don C Mitchell, Fernando Cuetos, and Daniel Zagar. 1990. Reading in different languages: Is there a universal mechanism for parsing sentences? In *Comprehension processes in reading*, pages 285–302. Routledge.

Edson Tadashi Miyamoto. 1998. *Relative clause processing in Brazilian Portuguese and Japanese*. Ph.D. thesis, Massachusetts Institute of Technology.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021a. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Benjamin Newman, Kai-Siang Ang, Julia Gong, and John Hewitt. 2021b. Refining targeted syntactic evaluation of language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3710–3723, Online. Association for Computational Linguistics.

Loreto Parisi, Simone Francia, and Paolo Magnani. 2020. Umberto: an italian language model trained with whole word masking. https://github.com/musixmatchresearch/umberto.

Marco Polignano, Pierpaolo Basile, Marco De Gemmis, Giovanni Semeraro, Valerio Basile, and 1 others. 2019. Alberto: Italian bert language understanding model for nlp challenging tasks based on tweets. In *CEUR workshop proceedings*, volume 2481, pages 1–6. CEUR.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

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

H. Rohde, R. Levy, and A. Kehler. 2011. Anticipating explanations in relative clause processing. *Cognition*, 118(3):339–358.

Laurent Sartran, Samuel Barrett, Adhiguna Kuncoro, Miloš Stanojević, Phil Blunsom, and Chris Dyer. 2022. Transformer grammars: Augmenting transformer language models with syntactic inductive biases at scale. *Transactions of the Association for Computational Linguistics*, 10:1423–1439.

Russell Scheinberg, So Young Lee, and Ameeta Agrawal. 2025. Missing the cues: Llms' insensitivity to semantic biases in relative clause attachment. *Proceedings of the Linguistic Society of America*, 10(1):5902–5902.

Xingjia Shen. 2006. *Late assignment of syntax theory: Evidence from Chinese and English*. University of Exeter (United Kingdom).

Arabella Sinclair, Jaap Jumelet, Willem Zuidema, and Raquel Fernández. 2022. Structural persistence in language models: Priming as a window into abstract language representations. *Transactions of the Association for Computational Linguistics*, 10:1031–1050.

Benjamin Swets, Timothy Desmet, Charles Clifton, and Fernanda Ferreira. 2008. Underspecification of syntactic ambiguities: Evidence from self-paced reading. *Memory & Cognition*, 36:201–216.

Marten Van Schijndel and Tal Linzen. 2018. Modeling garden path effects without explicit hierarchical syntax. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, volume 40.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural network acceptability judgments. *Transactions of the Association for Computational Linguistics*, 7:625–641.

Alex Warstadt, Yian Zhang, Xiaocheng Li, Haokun Liu, and Samuel R. Bowman. 2020b. Learning which features matter: RoBERTa acquires a preference for linguistic generalizations (eventually). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 217–235, Online. Association for Computational Linguistics.

Michael A Wilson, Zhenghao Zhou, and Robert Frank. 2023. Subject-verb agreement with seq2seq transformers: Bigger is better, but still not best. *Proceedings of the Society for Computation in Linguistics*, 6(1):278–288.

Lingling Zhou, Suzan Verberne, and Gijs Wijnholds. 2024. Tree transformer's disambiguation ability of prepositional phrase attachment and garden path effects. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12291–12301, Bangkok, Thailand. Association for Computational Linguistics.

A Templates for Experimental Stimuli

Table 2: Italian stimuli by condition, adapted from Lee and De Santo (2024).

	Verb Type	Attachment	Target					
a.	Perceptual (P)	LA	*Gianni vide il figlio dei*	*medici*	*che*	*correvano*	*la*	*maratona*
			Gianni saw the son-SG of the	*doctors-PL*	*who*	*were running-PL*	*the*	*marathon*
b.	Perceptual (P)	HA	*Gianni vide il figlio dei*	*medici*	*che*	*correva*	*la*	*maratona*
			Gianni saw the son-SG of the	*doctors-PL*	*who*	*was running-SG*	*the*	*marathon*
c.	Non-Perceptual (N)	LA	*Gianni amò il figlio dei*	*medici*	*che*	*correvano*	*la*	*maratona*
			Gianni loved the son-SG of the	*doctors-PL*	*who*	*were running-PL*	*the*	*marathon*
d.	Non-Perceptual (N)	HA	*Gianni amò il figlio dei*	*medici*	*che*	*correva*	*la*	*maratona*
			Gianni loved the son-SG of the	*doctors-PL*	*who*	*was running-SG*	*the*	*marathon*

Table 3: Stimuli by condition for English Exp. 1, adapted from Grillo et al. (2015).

	Verb Type	Attachment		Target	
a.	RC-only	HA	Jim saw the son of the doctors that	was	having dinner.
b.	RC-only	LA	Jim saw the son of the doctors that	were	having dinner.
c.	SC	HA	Jim shares the house with the son of the doctors that	was	having dinner.
d.	SC	LA	Jim shares the house with the son of the doctors that	were	having dinner.

Table 4: Stimuli by condition for English Exp. 2, adapted from Grillo et al. (2015).

	Noun Type	Attachment		Target	
a.	RC-only	HA	The picture of the son of the doctors that	was	having dinner is old.
b.	RC-only	LA	The picture of the son of the doctors that	were	having dinner is old.
c.	SC	HA	The car of the son of the doctors that	was	having dinner is old.
d.	SC	LA	The car of the son of the doctors that	were	having dinner is old.

B Summary of LME Models

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	27.94089	0.82105	31.02296	34.031	<2e-16
Verb Type	-0.63548	0.50688	69.00000	-1.254	0.214
Attachment Type	-0.19745	0.50688	69.00000	-0.390	0.698
Verb Type : Attachment Type	-1.34373	4.39996	69.00000	-0.305	0.761

(a) AlBERTo

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	2.5262	0.6478	80.3745	3.900	0.000199
Verb Type	-0.7534	0.8093	69.0000	-0.931	0.355153
Attachment Type	0.1703	0.8093	69.0000	0.210	0.833988
Verb Type : Attachment Type	1.2688	1.1446	69.0000	1.109	0.271492

(b) GePpeTto

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	28.1589	1.6636	32.7576	16.927	<2e-16
Verb Type	-0.4183	1.1124	69.0000	-0.376	0.708
Attachment Type	-0.5246	1.1124	69.0000	-0.472	0.639
Verb Type : Attachment Type	0.6105	1.5732	69.0000	0.388	0.699

(c) bert_base_multilingual_case

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	24.6305	2.0505	54.8460	12.012	<2e-16
Verb Type	1.7987	2.2630	60.0000	0.795	0.430
Attachment Type	0.3184	2.2630	60.0000	0.141	0.889
Verb Type : Attachment Type	-0.4746	3.2003	60.0000	-0.148	0.883

(d) xlm-mlm-17-1280

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	46.35441	3.29053	47.98074	14.087	<2e-16
Verb Type	3.62229	3.11124	69.00000	1.164	0.248
Attachment Type	0.08279	3.11124	69.00000	0.027	0.979
Verb Type : Attachment Type	-1.34373	4.39996	69.00000	-0.305	0.761

(e) XLM-roBERTa-large

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	2.5262	0.6478	80.3745	3.900	0.000199
Verb Type	-0.7534	0.8093	69.0000	-0.931	0.355153
Attachment Type	0.1703	0.8093	69.0000	0.210	0.833988
Verb Type : Attachment Type	1.2688	1.1446	69.0000	1.109	0.271492

(f) UmBERTo

Table 5: LMER Summary for all models in the Italian Experiment. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	19.9798	0.5074	55.6898	39.377	<2e-16
Verb Type	-1.7528	0.5243	69.0000	-3.343	0.00134**
Attachment Type	-0.7160	0.5243	69.0000	-1.366	0.17648
Verb Type : Attachment Type	-0.5337	0.7414	69.0000	-0.720	0.47407

(a) bert_base_multilingual_case

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	27.7412	0.7469	45.2554	37.144	<2e-16
Verb Type	-0.9368	0.6790	69.0000	-1.380	0.1721
Attachment Type	-1.3458	0.6790	69.0000	-1.982	0.0515
Verb Type : Attachment Type	1.8169	0.9602	69.0000	1.892	0.0627

(b) xlm-mlm-17-1280

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	25.2791	0.5831	53.7553	43.350	<2e-16
Verb Type	-1.7132	0.5907	69.0000	-2.900	0.00499**
Attachment Type	2.8251	0.5907	69.0000	4.783	9.46e-06***
Verb Type : Attachment Type	0.2826	0.8353	69.0000	0.338	0.73616

(c) xlm-roberta-large

Table 6: LMER Summary for all models in the English Experiment 1. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05.

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	21.2704	0.4423	47.9400	48.094	<2e-16
Noun Type	-0.5328	0.4179	69.0000	-1.275	0.207
Attachment Type	-2.4803	0.4179	69.0000	-5.935	1.06e-07***
Noun Type : Attachment Type	-0.5808	0.5911	69.0000	-0.983	0.329

(a) bert_base_multilingual_case

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	24.6088	0.9001	51.2856	27.339	<2e-16
Noun Type	1.9332	0.8871	69.0000	2.179	0.0327*
Attachment Type	1.5127	0.8871	69.0000	1.705	0.0926
Noun Type : Attachment Type	-1.1707	1.2545	69.0000	-0.933	0.3540

(b) xlm-mlm-17-1280

	Estimate	Std. Error	df	t-value	Pr(>|t|)
(Intercept)	27.1706	0.5632	44.8643	48.243	<2e-16
Noun Type	-0.5667	0.5089	69.0000	-1.114	0.269274
Attachment Type	2.0496	0.5089	69.0000	4.028	0.000143***
Noun Type : Attachment Type	0.3905	0.7197	69.0000	0.589182	0.589182

(c) xlm-roberta-large

Table 7: LMER Summary for all models in the English Experiment 2. Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05

C Additional Analyses

C.1 Preferences as Categorical Pairwise Comparisons

Following standard practices in psycholinguistics, for the paper's core analyses statistical robustness of the effects/contrasts has been determined by running linear-mixed effects models using the original distribution of surprisal values over items. This is also consistent with what is done with human data during (for instance) online tasks involving locally ambiguous sentences like the ones we used, and it is also how Davis and van Schijndel (2020) discuss significant effects for their results. However, a qualitative understanding of model's trend, in line with results from human participants from forced choices tasks targeting globally ambiguous sentences, can be achieved by coding a model's preference for HA/LA categorically for each item pair in a set (Davis and van Schijndel, 2020). That is, in each set items can be paired by keeping Verb Type/Noun Type consistent. Then, if surprisal for the LA disambiguated item was lower than the surprisal for the HA disambiguated item, attachment is coded as LA. See Example 4 for a summary of this coding approach across the experiments in this paper.

(4) Interpretation of Pairwise comparisons for each experiment

a. `Attachment Preference` ← `LOW` if Verb Surprisal(a) > Verb Surprisal(b)

b. `Attachment Preference` ← `HIGH` if Verb Surprisal(a) < Verb Surprisal(b)

c. `Attachment Preference` ← `LOW` if Verb Surprisal(c) > Verb Surprisal(d)

d. `Attachment Preference` ← `HIGH` if Verb Surprisal(c) < Verb Surprisal(d)

C.1.1 Italian Pairwise Comparisons

In terms of qualitative interpretation, when comparing sentence types a higher percentage of low surprisal values for LA items compared to their paired HA items would indicate a LA preference, and vice versa. Verb-type sensitivity would show these high/low surprisal values at the embedded verb also modulated by the properties of the matrix verb. Figure 4 shows a visualization of model preferences given this kind of coding schema in the Italian Experiment. In contrast with the quantitative analyses over raw surprisal scores, these qualitative contrasts do show some trends compatible with attachment preferences — although still not exactly aligned with human data. In particular, we observe a profile compatible with a LA preference for the Perceptual condition for GePpeTto and AlBERTo. The xlm and RoBerta models show an opposite trend, with an LA-leaning pattern in the Non-Perceptual condition. Finally, the Umberto and multilingual bert-base models show an HA profile across both Verb-Type conditions.

C.1.2 English Pairwise Comparisons: Exp. 1

For the first English experiment, we similarly observe differences between the quantitative comparisons reported in the main paper and the qualitative pairwise differences. Notably, it seems that the roberta model prefers HA items in almost every set — in contrast with the pattern of preferences usually reported for human English participants (see Figure 5). This is showcased in the model complete lack of HA choices in the RC-Only condition, and in a small increase in HA rates in the SC condition.

It is also interesting to observe that the magnitude of these effects is strikingly different: human participants still exhibit a certain rate of HA choices in both conditions, compared to the heavy LA tendency of GPT-2. Finally, we can observe a LA trend in the qualitative comparisons for bert-base-uncased across conditions, in contrast with the not significant statistical effects.

C.1.3 English Pairwise Comparisons: Exp. 2

For the second English Experiment, qualitative contrasts for multilingual bert showcase an almost absolute preference for LA attachment across all conditions (see Figure 6). Similarly, bert-base-uncased presents a significant sensitivity to attachment type and no other effect, again in line with qualitative contrasts revealing an almost absolute favoring of LA. It is also interesting to observe that although monolingual and multilingual bert behave similarly in their striking preference for a LA interpretation across conditions, bert-base-uncased shows a slight increase of HA choices in the SC-condition, while multilingual-bert presents the same increase in the RC-only condition. Notably, neither of these patterns is human-like. In contrast, qualitative comparisons for the xlm model a tendency towards HA, stronger in the RC-only condition. For roberta, qualitative contrasts also showcase a strong, almost absolute HA preference across all conditions, with no effects of Noun-type.

C.1.4 Pairwise Comparisons by Set

Finally, even though in the main analyses of the paper we accounted for lexical effects by including random intercept for set in our statistical models, here we plot pairwise comparisons by set in order to (preliminary) evaluate model sensitivity to lexical variation. In future work, this preliminary step could be used to evaluate potential causes for the observed variation (e.g. by exploring whether a model's expectation for RC or PRs are tied to the frequency of the main clause verbs and/or the target verbs), potentially connected to corpus analyses.

C.1.5 Pairwise Comparisons: Final Notes

Across all experiments, we found some differences between statistically significant results for raw surprisal analyses, and qualitative trends in forced-choice contrasts. Davis and van Schijndel (2020) also presented both kinds of evaluations, but in their case the two methodologies overlapped with respect to the overall trend of the results. While some of the trends highlighted in the forced choice contrasts might warrant further investigation on the effect of task type, it is worth stressing that in our results these trends seem to arise from really small numerical differences in surprisal values in the item-by-item comparisons — which thus do not result in significant differences at the global level. Therefore, we must be cautious in over-interpreting these qualitative effects as overstating/emphasizing small numerical differences in values.

C.2 Prompting

While past work has leveraged prompting, in the main paper we privileged surprisal-based analyses since prompting tasks rely on specific meta-linguistic information that might confound the evaluation of model sensitivity to linguistic properties. Another issue with prompting methods lies in their generalizability across architectures — for instance, the differences in training objectives for the various LLMs tested in this paper (all of which are not considered conversational models), makes it complicated to adopt prompting as a consistent evaluation strategy. Nonetheless, we attempted to implement a prompting-based evaluation pipeline similar to Lee et al. (2025). However, while the models tested by Lee et al. (2025) were able to correctly identify which parts of a sentence corresponded to a RC and a possible DP for the RC to attach to 86 percent of the time, in our preliminary tests the three Italian models were able to identify a DP correctly only in 5.2% of the cases (on average across the three models). Since correct identification of RCs and candidate DPs is a preliminary first step in Lee et al. (2025)'s approach to investigating ambiguity resolution, failure on this constitutes an immediate roadblock in the attempt to replicate a prompt-based approach to attachment biases within the models investigates here *and* for any future attempt to conduct cross-architectural comparisons of model performance.

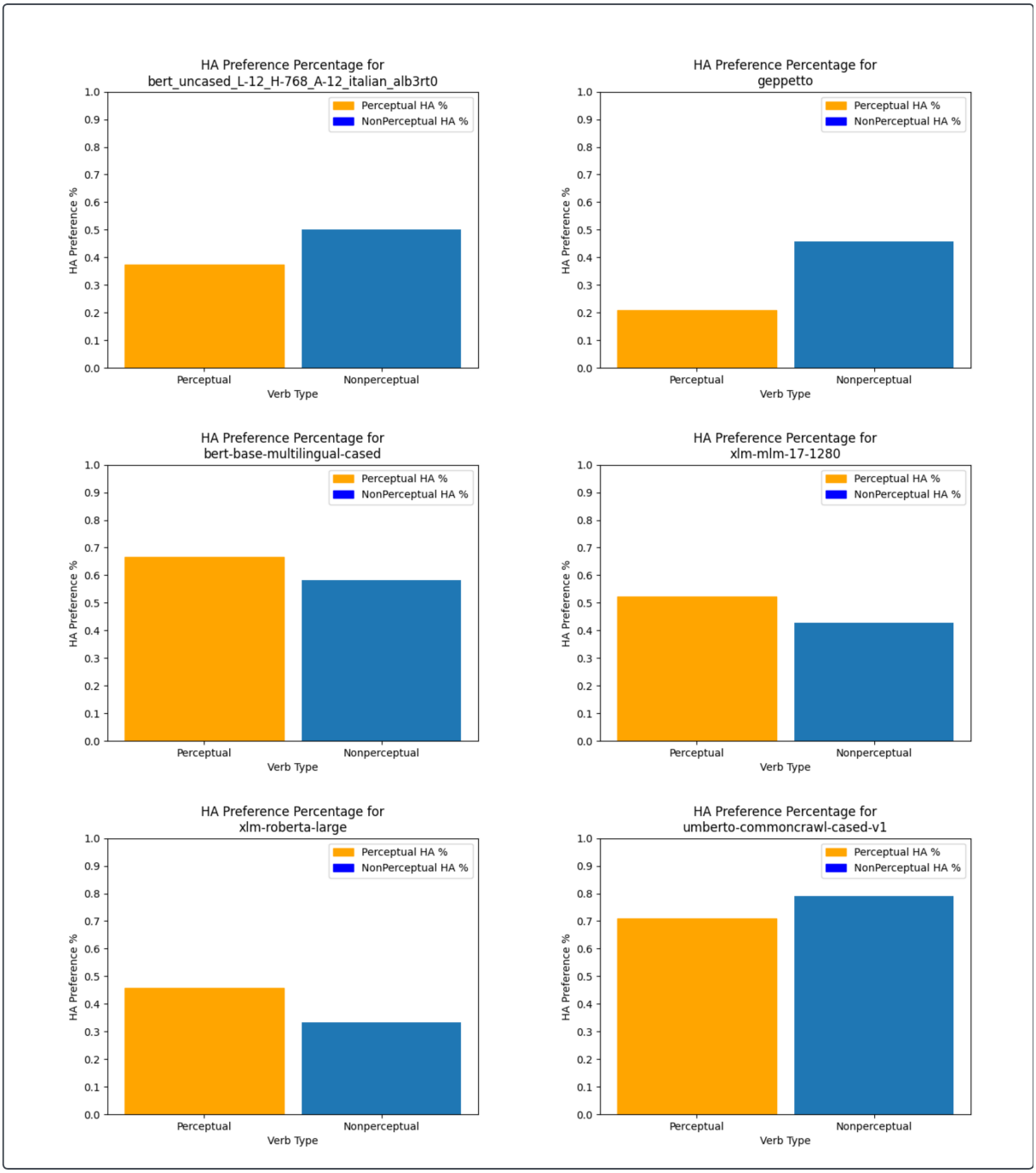

Figure 4: Proportion of HA vs. LA in the Italian Experiment, derived from categorical pairwise comparisons within sets.

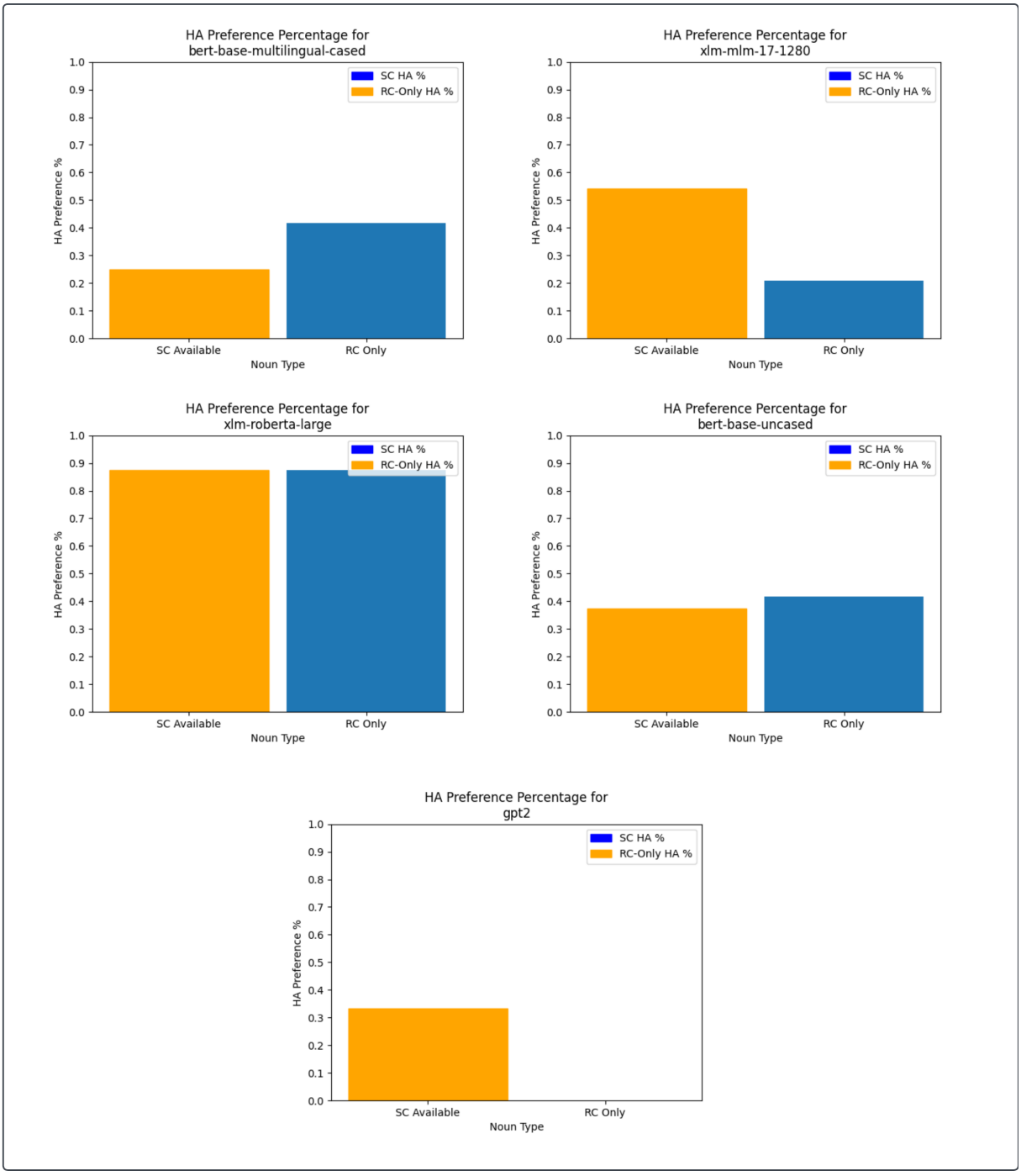

Figure 5: Proportion of HA vs. LA in the English Experiment 1, derived from categorical pairwise comparisons within sets.

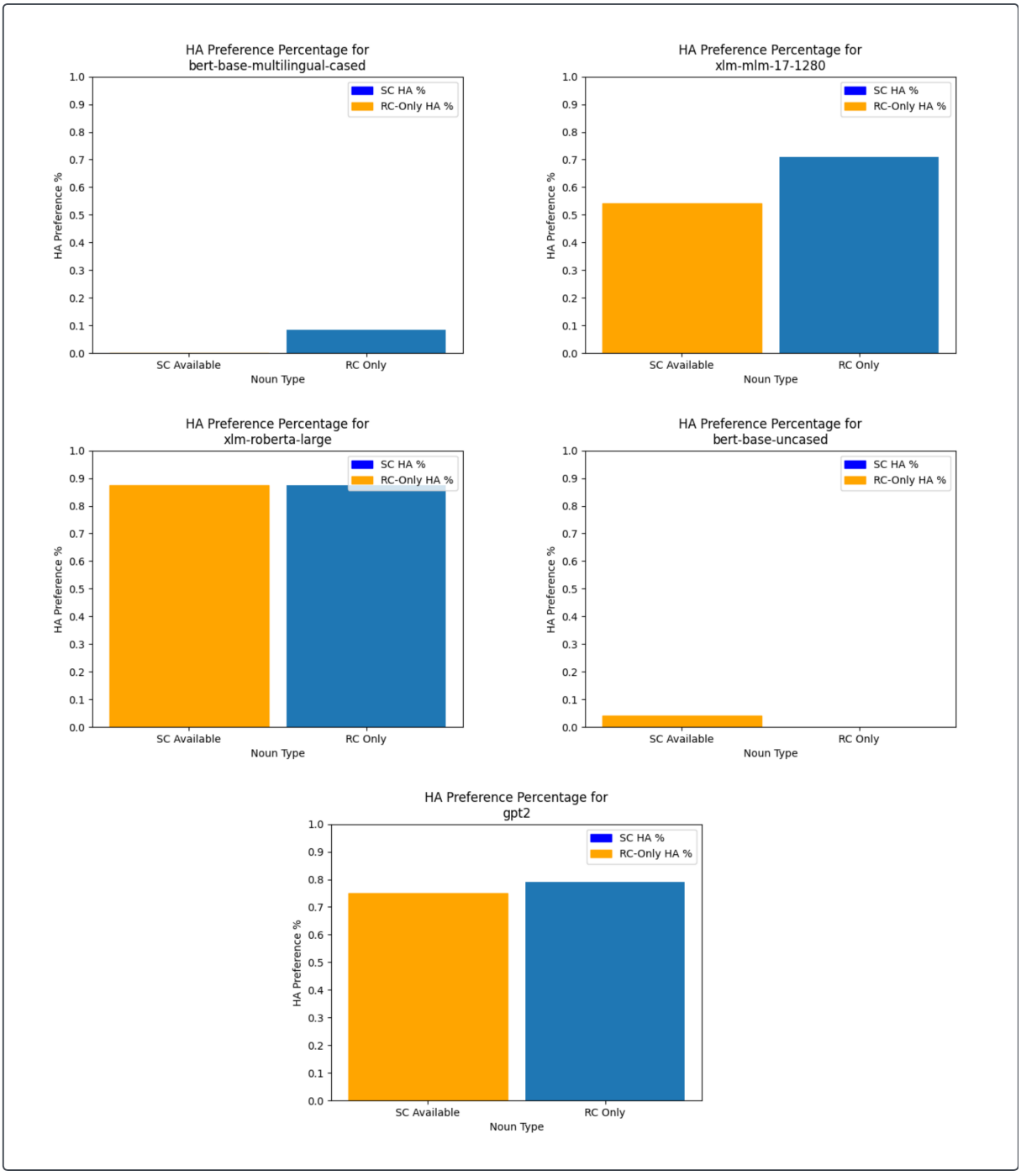

Figure 6: Proportion of HA vs. LA in the English Experiment 2, derived from categorical pairwise comparisons within sets.

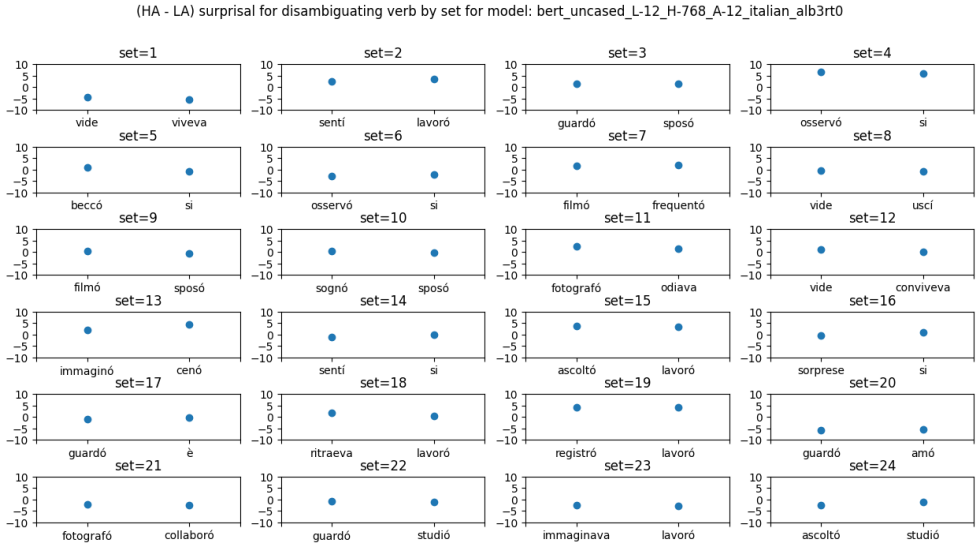

Figure 7: Surprisal Comparisons by Set and Verb Type (Perceptual vs. NonPerceptual) for alb3rt0.

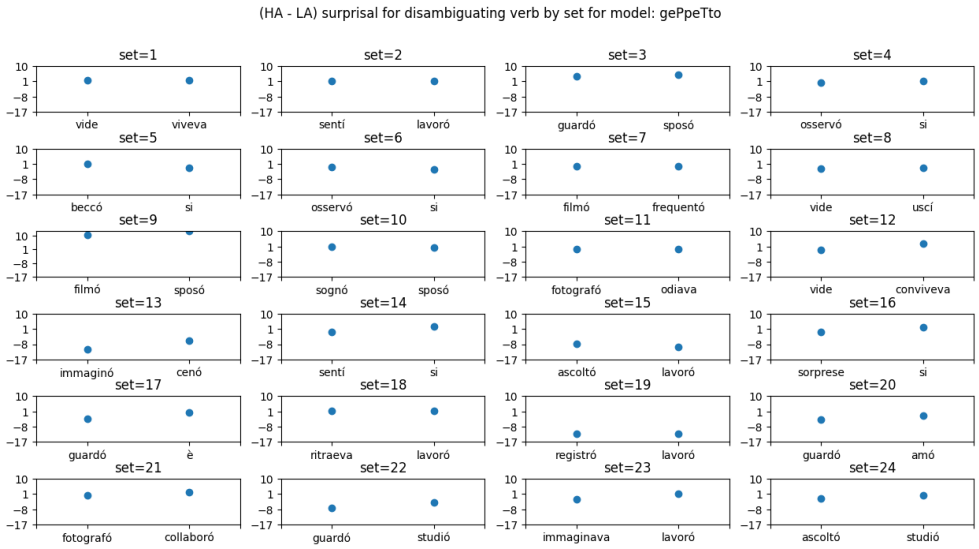

Figure 8: Surprisal Comparisons by Set and Verb Type (Perceptual vs. NonPerceptual) for gePpeTto.

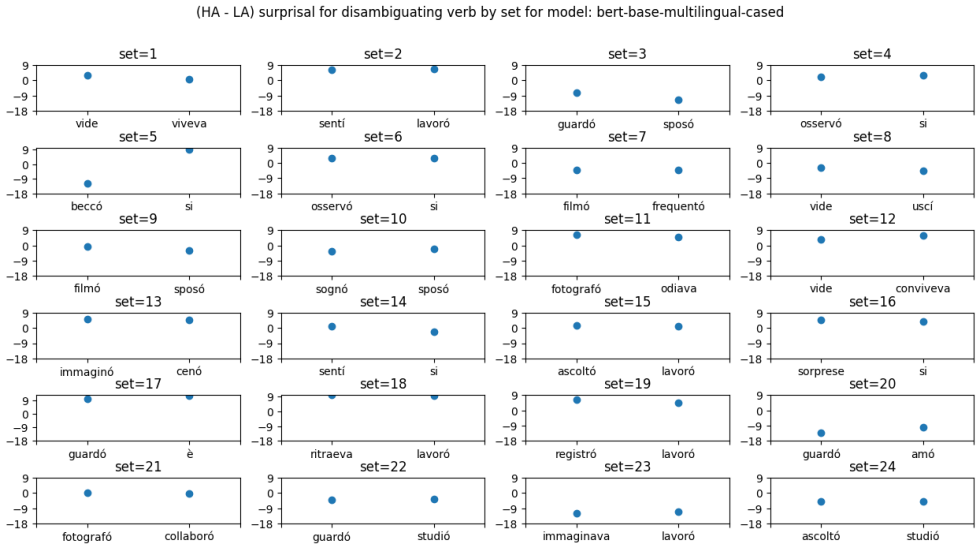

Figure 9: Surprisal Comparisons by Set and Verb Type (Perceptual vs. NonPerceptual) for Italian bert-base-multilingual-cased.

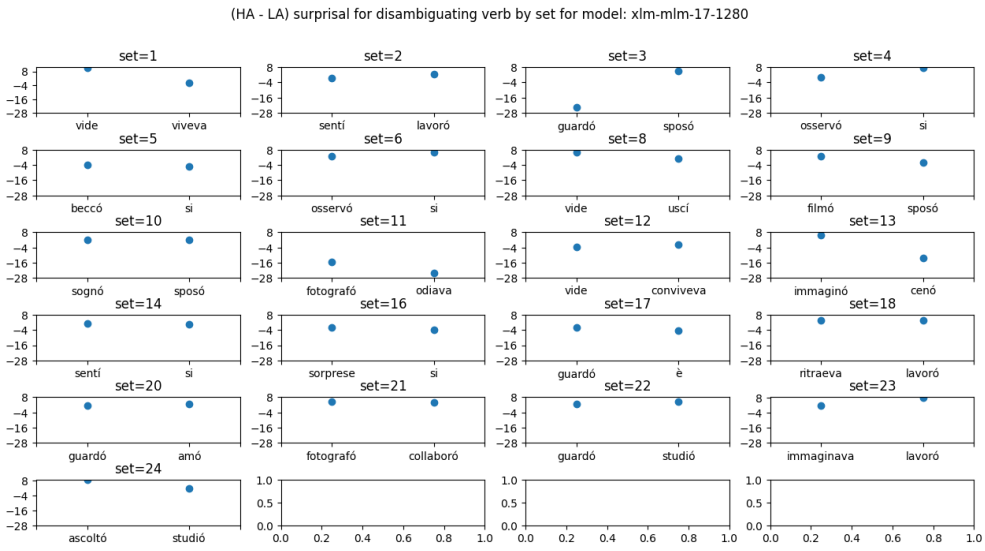

Figure 10: Surprisal Comparisons by Set and Verb Type (Perceptual vs. NonPerceptual) for xlm-mlm-17-1280

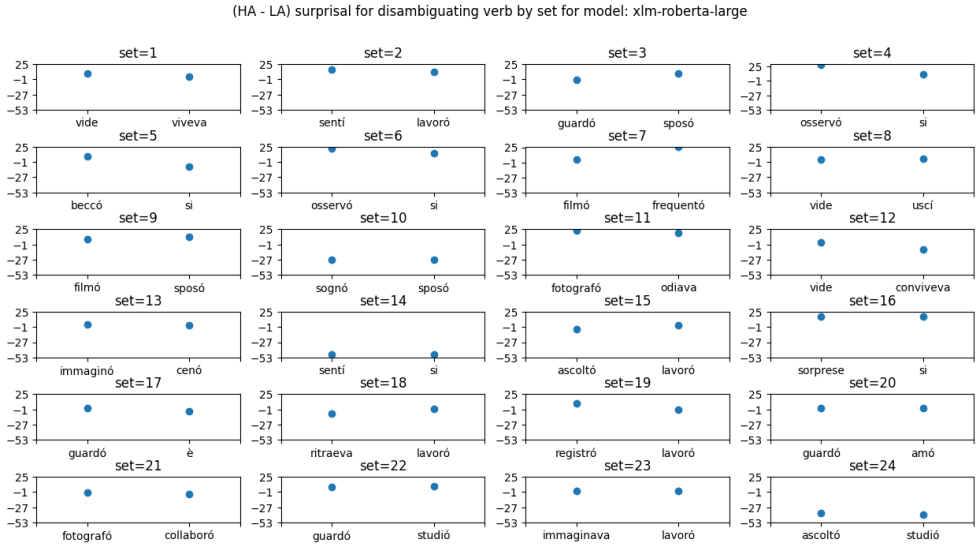

Figure 11: Surprisal Comparisons by Set and Verb Type (Perceptual vs. NonPerceptual) for Italian roberta.

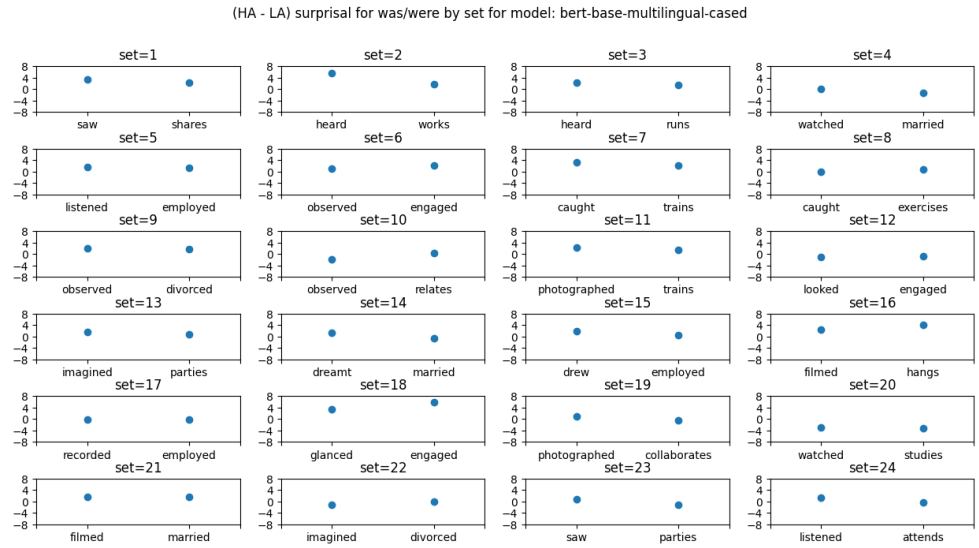

Figure 12: Surprisal Comparisons by Set and Verb Type (SC vs. RC Only) for English bert-base-multilingual-cased.

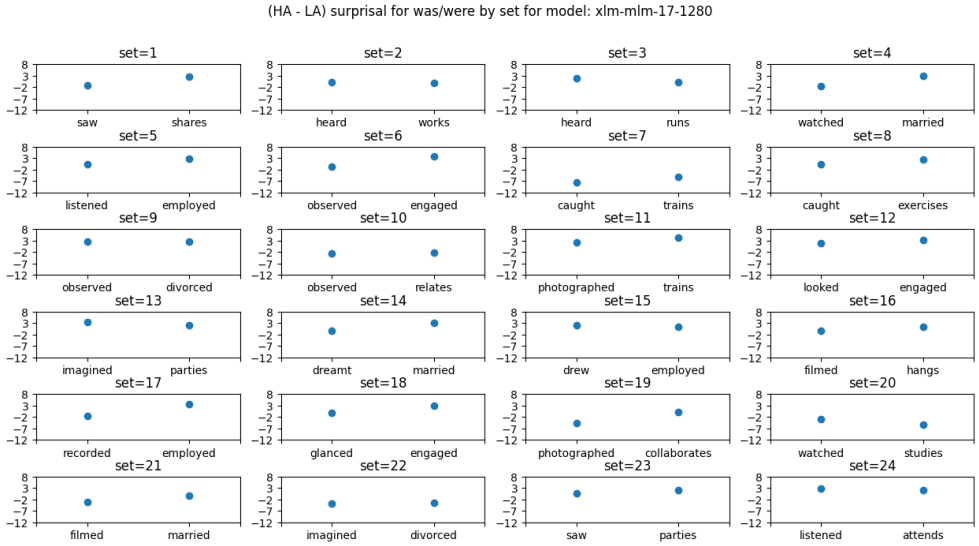

Figure 13: Surprisal Comparisons by Set and Verb Type (SC vs. RC Only) for English xlm-mlm-1280.

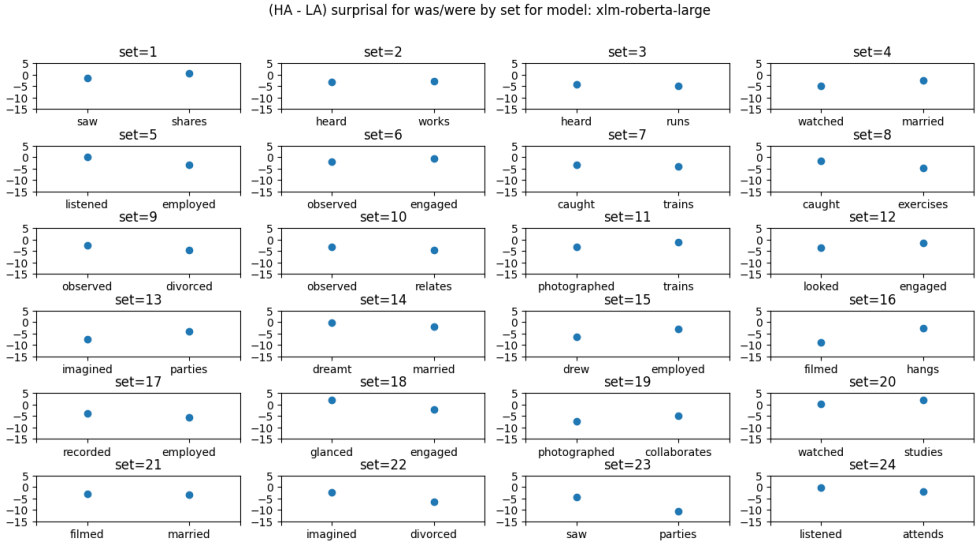

Figure 14: Surprisal Comparisons by Set and Verb Type (SC vs. RC Only) for English roberta.

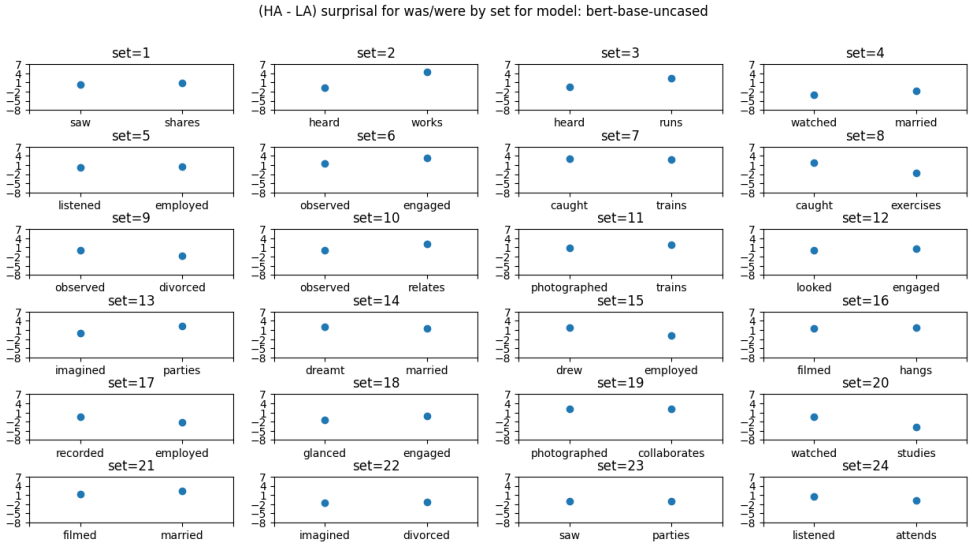

Figure 15: Surprisal Comparisons by Set and Verb Type (SC vs. RC Only) for English bert-base-uncased.

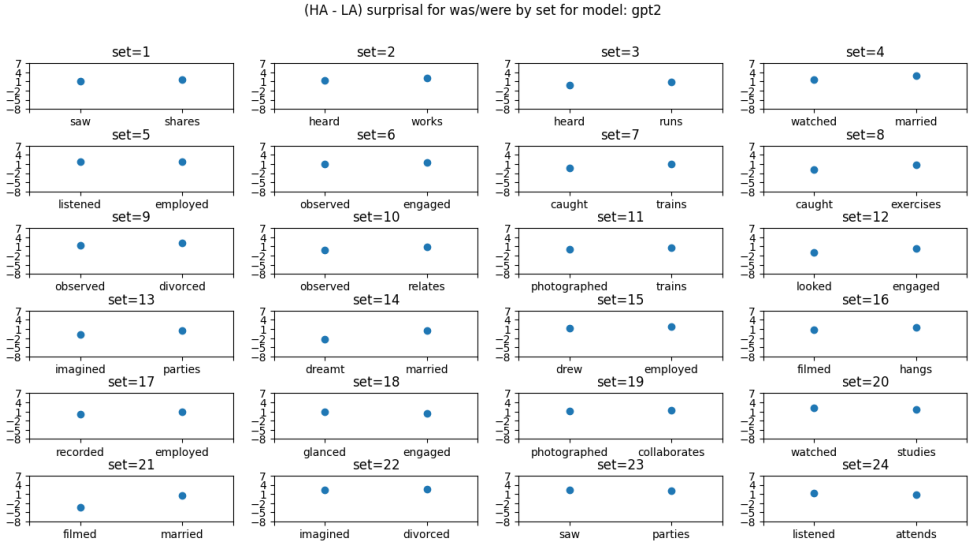

Figure 16: Surprisal Comparisons by Set and Verb Type (SC vs. RC Only) for English gpt2.

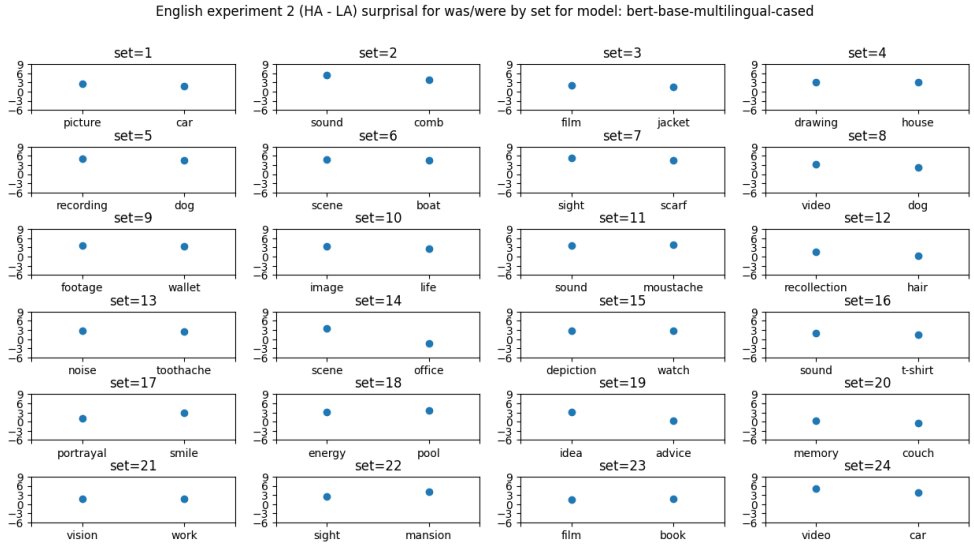

Figure 17: Surprisal Comparisons by Set and Noun Type for English bert-base-multilingual-cased.

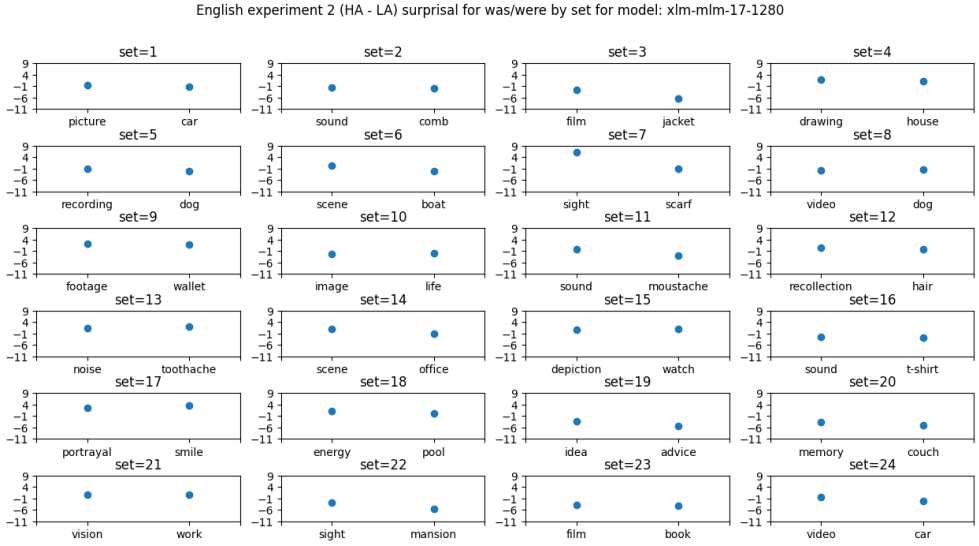

Figure 18: Surprisal Comparisons by Set and Noun Type for English xlm-mlm-17-1280.

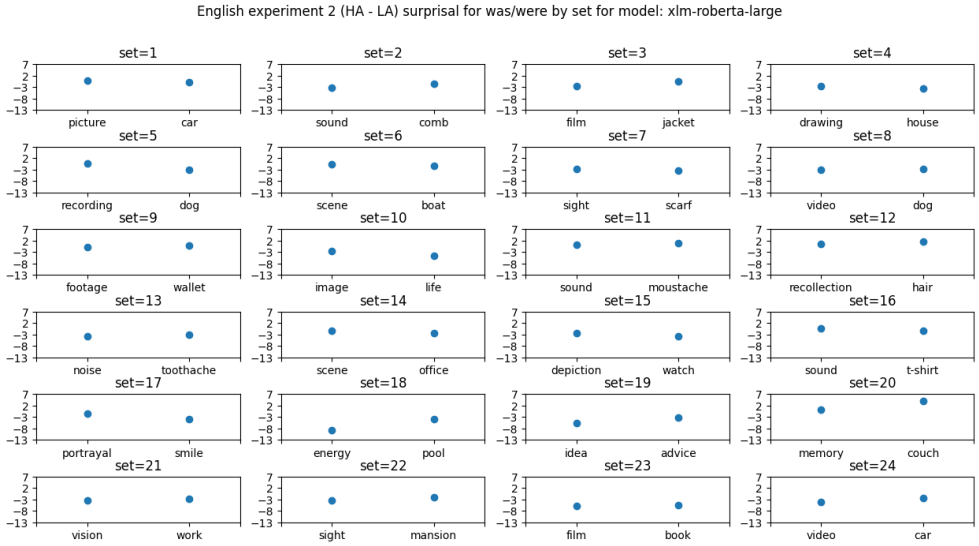

Figure 19: Surprisal Comparisons by Set and Noun for English roberta.

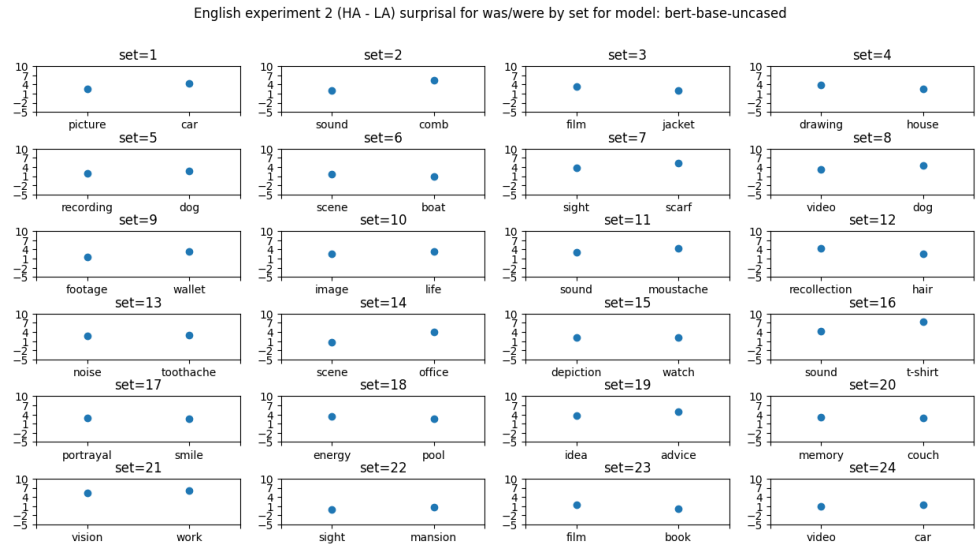

Figure 20: Surprisal Comparisons by Set and Noun Type for English bert-base-uncased.

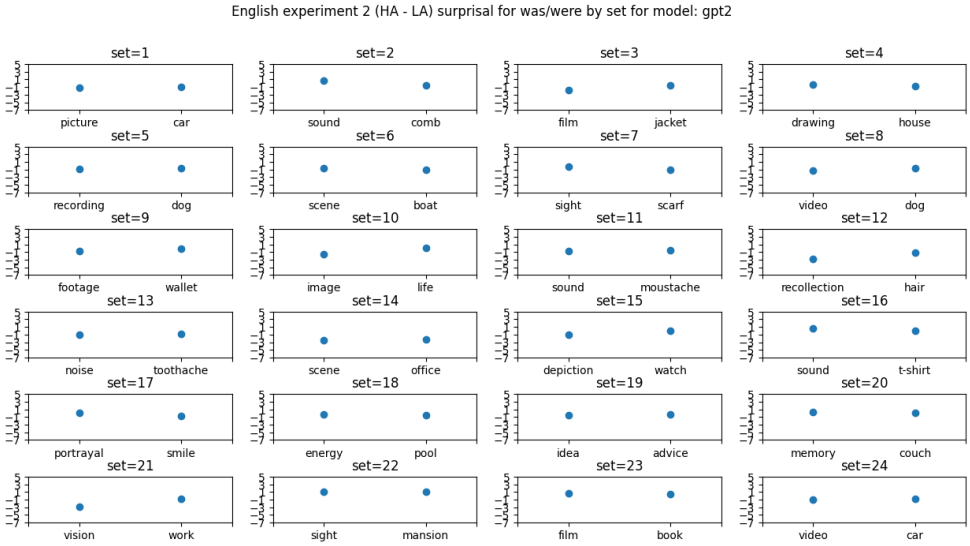

Figure 21: Surprisal Comparisons by Set and Noun Type (for English gpt2.