LLM Interpretations of Null and Overt Pronouns in Portuguese

Anonymous ACL submission

Abstract

The concept of prominence has been used to decode pronoun resolution. Here, we would like to explore if LLM is able to mimic human prominence behaviour in pronoun resolution. We have focused on Portuguese; it allows speakers to drop subject pronouns, and human interpretations of null and overt pronouns vary. We used BERTimbau question-answer model to generate responses for stimuli used in an experiment by Fernandes et al. (2018). The results show that some aspects of prominence-based phenomena in pronoun resolution are not replicated by the LLM. However, examination of LLM confidence scores offers hope that the gap between human and LLM responses may be bridged by larger training corpora of language-specific data.

1 Introduction

Nowadays, large language models (LLM) are widely used as human assistants. Some studies deem that their outputs are human-like, i.e., similar to human responses (Hu et al. 2022). Yet, other studies (e.g. Leivada et al. 2024) show the opposite, with the LLMs studied incapable of generating truly human-like responses. These studies have focused primarily on text generation and question-answer models with the GPT family of LLMs.

The current study investigates pronoun resolution. In particular, we explore the contrast between the use of subject pronouns with greater form prominence (i.e. being actually pronounced, overt pronouns), in contrast with subject anaphors of less form prominence (covert, i.e. unrealized). We use Portuguese. It is a morphologically rich language in comparison to English and allows the use of both covert and overt subject anaphora.

Our human interpretation data comes from an experiment undertaken and published by Fernandes et al. (2018). We compare the human interpretation of European Portuguese pronouns with LLM interpretations. For the model, we used a Portuguese question-answer model based on BERTimbau available in HuggingFace (huggingface.co) to explore potential differences in the interpretation of these pronouns.

1.1 Covert vs. overt pronouns in Portuguese

Pronoun resolution is the process by which language users determine what referent is the intended denotation of anaphors in their linguistic input. Usually, this referent is one that has already been mentioned by another referential phrase in the same discourse. Morphologically rich languages like Spanish, Italian, and Portuguese, have an agreement system that aligns the person and number of subject referents with verbal inflection. Leaving subject pronouns covert thus leads to less ambiguity than would occur in the case of languages like English. See (1) for example.

(1) a. *Eu vou à escola todos os dias* - I go to school everyday (overt pronoun)

b. *Vou à escola todos os dias* - I go to school everyday (covert pronoun)

The lower risk of ambiguity offers a potential functional explanation for why these languages allow pro-drop.

Pronouns are, with occasional exceptions, non-initial elements in a *reference chain* of expressions which refer to the same referent. The immediately previous expression in the chain is called the *antecedent* of the pronoun. Interpreting a pronoun, therefore, is usually the same task as identifying its antecedent in the discourse. The

choice of interpretation has been found to depend on syntactic, semantic and pragmatic factors of the potential antecedents (Carminati 2002).

We can aggregate these factors using two notions of prominence. *Code prominence* expresses how attention-attracting a particular construction is (Ellison 2024). For example, a pronoun in English bearing phonetic word stress has higher code prominence than one without it. *Discourse prominence* describes how readily a particular referent comes to mind at a given point in a discourse (von Heusinger & Schumacher 2019). For example, if a referent has recently been the subject of a sentence, then other things being equal, it has higher discourse prominence than a referent that was the object (reference). The interpretation of a pronoun has been shown to depend on its code prominence, and the discourse prominence of potential referents. A general tendency has been found, namely that the lower the code prominence of the pronoun, the higher the discourse prominence of the referent it refers to. This has been explored in English (see Kameyama 1997), and in German (e.g. Tomaszewicz-Özakın & Schumacher 2022). English examples appear in (2) below.

(2) a. The chef phoned the supplier. She wanted the delivery early.

b. The chef phoned the supplier. **She** wanted the delivery early.

Note that when the pronoun has greater form prominence (2b) – read with strong stress on the pronoun, the anaphor is more likely to be interpreted as the low-discourse prominence referent, namely *the supplier*. The reverse is true in (2a) with a less form prominent anaphor.

Covert subject pronouns have a very low level of code prominence. The only phonological expression of the referent is subject-agreement inflection on the verb, in languages where this happens. In contrast, overt pronouns offer an attentional anchor for the reference, and so have higher code prominence. Thus, by the generalization mentioned above, we expect that covert pronouns would be more likely to have referents with higher discourse prominence, while overt pronouns would be more likely to realise referents with lower discourse prominence. This is the case with Italian. Carminati (2002) argues that covert pronouns are often interpreted as having antecedents that were subjects, while overt pronouns more often non-subject, and so, non-topical antecedents. These accounts of pronominal usage are based on observations of human linguistic patterns. How well have LLM captured the subtleties of this behaviour? In what follows, we explore how well one LLM has internalized these clues to correctly resolve pronouns in Portuguese.

1.2 Fernandes et al. (2018)

Fernandes et al. (2018) report on a study of how speakers interpret pronouns in Portuguese pronouns, looking at both Brazilian and European Portuguese. Pro-drop is more common in European Portuguese than Brazilian (Barbosa et al., 2005). The data from their experiment provides an excellent window into pronoun resolution in this language. For the sake of simplicity, we only consider the European Portuguese component of their experiment. Their study presented speakers with context clauses containing both a subject and an object. These were followed by clauses that contained conjugated verbs in a predicate joined with or without an overt subject pronoun. The overt pronoun was always compatible in gender and number with both the previous subject or object, and thus grammatically, either could be selected as its antecedent. An example stimulus from the experiment is shown in (3).

(3) *A florista sossegou a peixeira no mercado quando Ø/ela divulgou os resultados do exame.*

The Florist calmed down the fishmonger at the market when ø/she disclosed the results of the exam.

In total their study included 32 sentences and 24 participants. They found significant differences in the interpretation of overt and covert anaphors. The distribution is shown in (Figure 1).

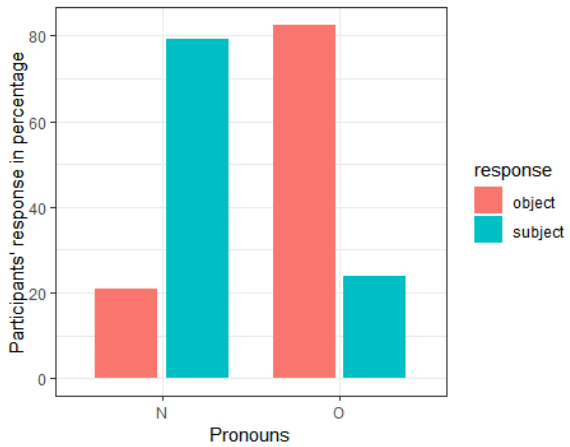

Figure 1: Results from the Fernandes et al. (2018) experimental study. N refers to covert and O to overt pronouns.

This difference was significant. A permutation test swapping per-stimulus overt and covert anaphor versions found few permutations resulting in more different distributions than the actual data. This shows that the difference in interpretations between the overt and covert conditions is unlikely to have arisen by chance (p=0.00233 99% c.i. 0.0019651-0.00275217, 10^5 permutations).

2 Comparing human and large language model antecedent choice

To compare LLM and human behaviour in Portuguese, we need an LLM trained for that language. We use the model BERTimbau data-set for Brazilian Portuguese subsequently trained on the SQUAD v1.1 by Pierre Guillou on SQUAD v1.1 question-answer dataset from European Portuguese (https://huggingface.co/pierreguillou/bert-large-cased-squad-v1.1-portuguese). We used them together with the "question-answering" pipeline from transformers.

We elicited pronoun interpretations from the LLM as follows. A stimulus item from the Fernandes et al. (2018) experiment was given to the LLM either with or without the overt pronoun. A question was then posed asking whether it was the subject or the object of the first clause (identified by the noun phrases used in those clauses) who was the subject of the predicate in the second clause. An example of overt and covert alternative stimuli are shown in (4), along with the common interpretation question used to determine the interpretation.

(4 .) Overt sentence : *A florista sossegou a peixeira no mercado quando ela divulgou os resultados do exame.*

Covert sentence : *A florista sossegou a peixeira no mercado quando divulgou os resultados do exame.* (see 3 for translation)

Question : *Que divulgou os resultados do exame, a florista ou a peixeira.* (Who disclosed the results of the exam, the florist or the fishermonger?)

We restricted the model to providing a single response. Thus, we eliminate the other possible response that appeared in human response. For instance, if the LLM provides *'a florista'* as a response for sentence (4), *'a peixeira'* response from human participants is eliminated. Hence, each sentence only has one response, and it is the same response for both human and LLM. We, then, compared the proportion percentage obtained from human response with the score obtained from the question-answer model. This score reflects how confidence a model in giving its response.

The LLM tends to interpet the anaphor as coreferential with the subject of the first (European covert: n =25, overt: n = 21) rather than with its object (European covert: n =5, overt: n = 7), regardless of whether an overt pronoun was supplied in the second clause or not.

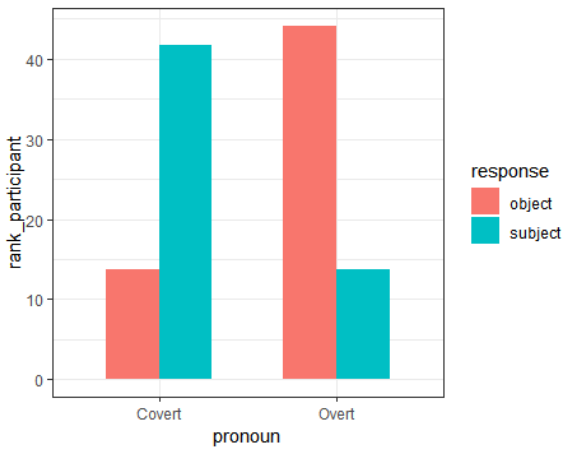

Figure 2: Human responses for covert and overt pronouns in the ranking format

Further, to have an equivalent comparison we converted the proportion response of the

participants and the score from the question-answering model into a ranking. Visually, by observing Figure 2, for the human response, and Figure 3, for the LLM response, we can see that they behaved differently. The LLM responses for overt pronouns do not match speaker behaviour in the case of the overt anaphors. Rather than reversing the interpretation, as seen in human results, instead we see almost identical interpretations of covert and overt anaphors.

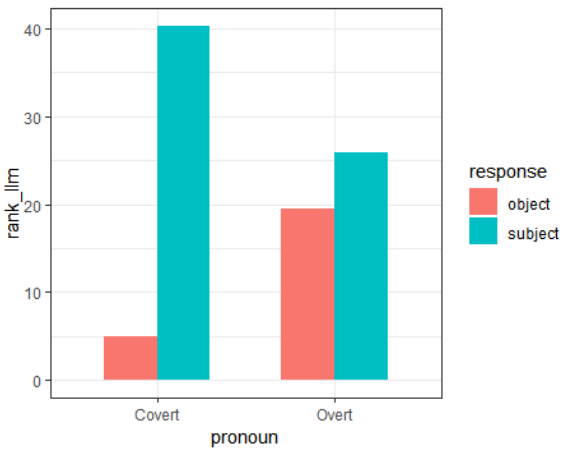

Figure 3: Human responses for covert and overt pronouns in the ranking format

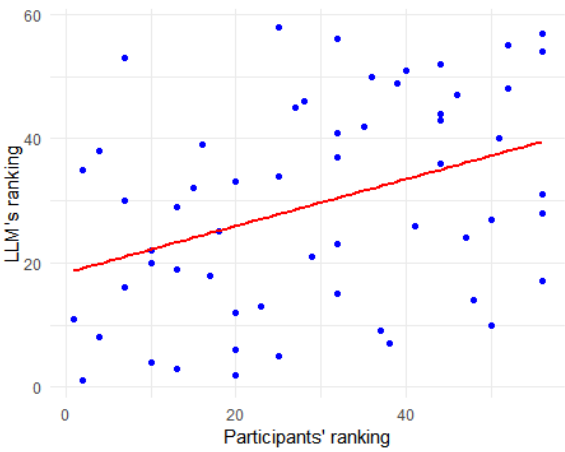

Figure 4: Correlation plot between the human participants' ranking and LLM rankingon. The interpretation preferences seen of experimental participants reflects the influence of the discourse prominence of potential referents, and the form prominence of pronoun.

The mean confidence scores of the LLM for these answers are low (i.e., less than 0.50): covert pronoun interpreted as subject - 0.45 and as object 0.06; overt pronoun interpreted as subject - 0.29 and as object: 0.23. Higher confidence scores in the LLM response correspond to higher frequencies of those referents in the human responses. This was confirmed with a permutation test (p=0.0011 99% c.i. 0.00086-0.00140, 10^5 permutations), and a correlation with a significance measure (r = 0.38 and *p*=.003). The confidence scores are plotted against the response rates in Figure 3, with a line reflecting the correlation. These results can be interpreted as the LLM having internalized some human-like behaviour. Perhaps larger models, trained on more data might reflect increased human-like responses in their question-answering as well.

3 Discussion

We have seen that there is a sharp difference in the interpretation of particularly overt pronouns in European Portuguese by native speakers and an LLM trained on Portuguese data. This is despite the general success of the model in producing intelligible Portuguese.

It seems that the LLM is blind to the difference in discourse prominence in referents where the antecedent was in subject vs object position. If this problem is more widespread than this model of this language, then this it may be symptomatic of a wider problem in LLM behaviour – a blindness towards discourse prominence. It is unlikely that the problem is pervasive, as it would result in very visible problems in generated text. More likely, this discourse prominence is rendered invisible when there is sufficient interword constraint exerted by either collocation or semantic constraint.

A number of major world languages show a contrast in pronouns, e.g. English stressed and unstressed pronouns, German *er/sie/es* vs *der/die/das*. In these cases, we see the same pattern of contrasting interpretations found in Portuguese. The lack of sensitivity seen in the LLM to the overt/covert contrast in interpretation, could potentially affect LLM utility in the future. In follow up work, we propose to explore whether this potential blind spot in LLM language use is present for other languages, and still present in later, more advanced, language models.

Limitations

The work is limited and could be expanded in a number of ways. Firstly, while it addresses a wide issue of prominence-sensitivity in LLMs, it draws on data looking only at one phenomenon in one dialect of one language. Secondly, while the paper describes human experimental results, it does not include a computational model of the human processing that can account for this data, e.g. a Bayesian model. Thirdly, there are not so many updated European Portuguese question-answer model. The one that was used here was based on BERTimbau that was developed for the Brazilian variety but the SQUAD database that was used to train the model was mixed with European Portuguese.

Finally, given the speed with which AI and the development of LLMs is changing, the work is limited by focusing on the predictions of a small number of models.

Acknowledgments

We thank Eunice Fernandes for sharing the data from Fernandes et al. (2018).

References

Blakemore, D. (1990, August). Constraints on interpretation. In *Annual Meeting of the Berkeley Linguistics Society* (pp. 363-370).

Carminati, M. N. (2002). *The processing of Italian subject pronouns*. University of Massachusetts Amherst.

Fernandes, E. G., Luegi, P., Correa Soares, E., de La Fuente, I., & Hemforth, B. (2018). Adaptation in pronoun resolution: Evidence from Brazilian and European Portuguese. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *44*(12), 1986.

Hu, J., Floyd, S., Jouravlev, O., Fedorenko, E., & Gibson, E. (2022). A fine-grained comparison of pragmatic language understanding in humans and language models. *arXiv preprint arXiv:2212.06801*.

Kameyama, M. (1997). Stressed and unstressed pronouns: Complementary preferences. *arXiv preprint cmp-lg/9707008*.

Larson, R., & Luján, M. (1989). Emphatic pronouns. Ms., Stony Brook University, Stony Brook, NY, and University of Texas at Austin.

Leivada, E., Dentella, V., & Günther, F. (2024). Evaluating the Language Abilities of Large Language Models vs. Humans: Three Caveats. *Biolinguistics*, *18*, 1-12.

Tomaszewicz-Özakın, B., & Schumacher, P. B. (2022). Anaphoric pronouns and the computation of prominence profiles. *Journal of Psycholinguistic Research*, *51*(3), 627-653.

A Supplementary Material

Data and codes are available in this OSF anonymous repository(https://osf.io/qscpk/?view_only=ee74769332e14743b44038fa612bb297 .