(Where) do LLMs learn cross-linguistic pronoun biases?
Evidence from transformer layers

Miriam Schulz (mschulz@lst.uni-saarland.de)
Department of Language Science and Technology, Saarland University

A central question in psycholinguistics concerns cross-linguistic variation in pronoun resolution, which arises even in structurally similar constructions such as (1a-b):

(1) a. The pilot$_i$ saw the engineer$_j$ before he$_{i/j}$ went home.
b. Le pilote$_i$ a vu l'ingénieur$_j$ avant qu' il$_{i/j}$ rentre à la maison.

While English comprehenders tend to resolve ambiguous pronouns to the subject antecedent, French comprehenders show a preference for the object antecedent [5]. This contrast has been attributed to an interaction between general principles of discourse interpretation (e.g., subject/first-mention biases) and language-specific pragmatic inference based on the availability – as indexed by relative corpus frequency in each language – of alternative syntactic constructions that unambiguously express subject coreference [1, 9].

Recent work shows that Large Language Models (LLMs) exhibit coreference biases that often diverge from human interpretation biases [6, 7, 11, 12, though see also 3], with LLMs frequently showing a general object preference [7, 12]. To our knowledge, however, no work has compared cross-linguistic pronoun interpretation biases in humans vs. LLMs. We investigate whether encoder-only transformer models capture cross-linguistic differences in human pronoun resolution, and if so, at which representational layers such effects emerge. Since human cross-linguistic pronoun resolution biases have been linked to construction frequencies in corpora [1, 9] and LLMs are strongly influenced by corpus statistics [13], we predict that LLMs should capture these differences if they are able to leverage language-specific construction frequencies to modulate coreference biases. We further predict the strongest alignment with human judgments to emerge in mid-to-upper transformer layers, where reference-related information has been shown to peak [2, 10].

Using experimental materials from [9] consisting of 16 parallel ambiguous pronoun constructions in English and French (cf. (1)), we evaluate two encoder-only transformer models: English BERT [4] and French CamemBERT [8]. Encoder-only models were chosen because they provide bidirectionally contextualized representations of all sentence tokens, allowing direct comparison of pronouns and antecedents in the same sentence from a single forward pass. For each of the 12 transformer layers, we compute the cosine similarity between the mean-pooled representations of the pronoun and each candidate antecedent (subject/object). Item-level object bias is defined as the difference between pronoun–object and pronoun–subject similarity.

We observe that correlations between human and LLM object bias – with stronger object preferences for the French participants and LLM compared to English – emerge already in early layers (Layer 1: $r=.50$, adjusted $p<.001$), weaken or even reverse in intermediate layers, and peak in upper-middle layers (Layers 9-11: $r=.71$-$.75$, all adjusted $p<.001$; see Table 1, Fig. 1). In the final layer, however, correlations disappear ($r\approx0$), and both models exhibit a general object bias. These patterns are further supported by mixed effects models predicting item-level human object choice proportions from LLM object bias.

These findings show that transformer models are sensitive to cross-linguistic differences in pronoun resolution preferences, and that alignment with human biases peaks in upper-middle layers – linked to the emergence of reference related information [2, 10] – rather than in the final representations used for prediction. This suggests that while models learn language-specific pronoun biases, this may interfere with competing constraints at the output layer (see also [3]). Future work should test whether these findings generalize across multilingual models as well as different model architectures and additional languages.

Layer	1	2	3	4	5	6	7	8	9	10	11	12
ρ	**0.504**	**0.373**	0.062	**0.412**	-0.066	-0.122	**-0.434**	**0.318**	**0.715**	**0.752**	**0.731**	-0.049
p-value	**<.001**	**<.01**	.68	**<.01**	.68	.45	**<.001**	**.016**	**<.001**	**<.001**	**<.001**	.67

Table 1: Pearson's correlation between item-level human object choices and LLMs' object bias for each layer, as well as corresponding p-values. P-values were adjusted for multiple comparisons via Benjamini-Hochberg correction.

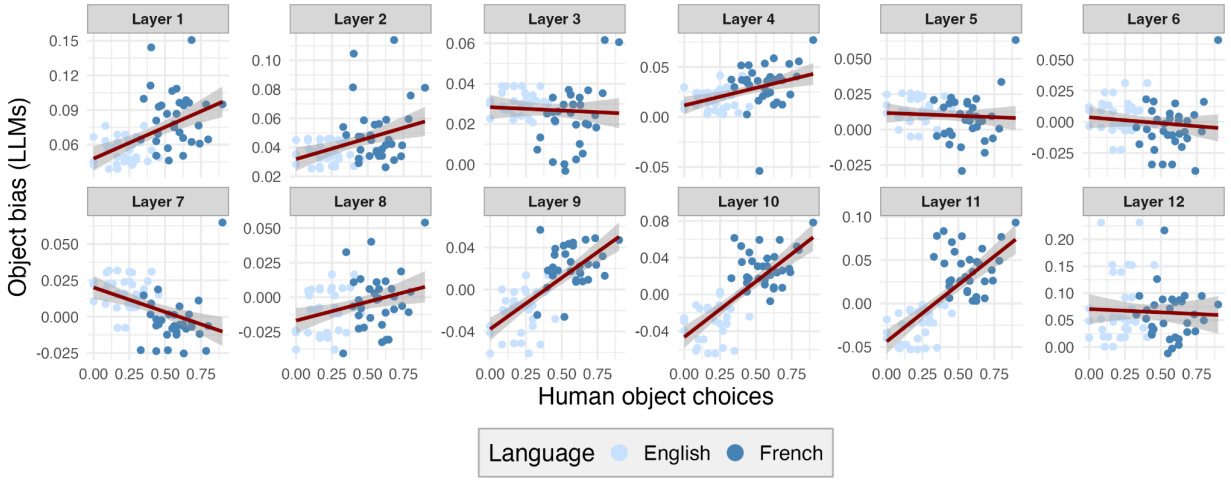

Figure 1: Proportion of human object choices from two experiments in [9] vs. LLMs' object bias for each item across each of the 12 layers of the encoder transformer models.

References

[1] Baumann, P., Konieczny, L., & Hemforth, B. (2014). *Psycholinguistic approaches to meaning and understanding across languages*.

[2] Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). *Proceedings of the 2019 ACL workshop BlackboxNLP: analyzing and interpreting neural networks for NLP*.

[3] Davis, F., & van Schijndel, M. (2020). *Proceedings of the 24th Conference on Computational Natural Language Learning*.

[4] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies.*

[5] Hemforth, B., Konieczny, L., Scheepers, C., Colonna, S., Schimke, S., Baumann, P., & Pynte, J. (2010). *Proceedings of the Annual Meeting of the Cognitive Science Society*.

[6] Kankowski, F., Solstad, T., Zarriess, S., & Bott, O. (2025). *arXiv preprint arXiv:2501.12980*.

[7] Kementchedjhieva, Y., Anderson, M., & Søgaard, A. (2021). *Findings of the Association for Computational Linguistics.*

[8] Martin, L., Muller, B., Suarez, P. O., Dupont, Y., Romary, L., de La Clergerie, É. V., ... & Sagot, B. (2020). *Proceedings of the 58th annual meeting of the association for computational linguistics*.

[9] Schulz, M., Burnett, H., & Hemforth, B. (2021). *Glossa: a journal of general linguistics*.

[10] Tenney, I., Das, D., & Pavlick, E. (2019). *Proceedings of the 57th annual meeting of the association for computational linguistics*.

[11] Upadhye, S., Bergen, L., & Kehler, A. (2020). *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[12] Zarrieß, S., Groener, H., Solstad, T., & Bott, O. (2022). *Proceedings of the 18th Conference on Natural Language Processing.*

[13] Zhong, Y., Todd, S., Xu, N., & Brehm, L. (2025). *Research Methods in Applied Linguistics*.