Claude reinvestigates *herself*: A replication of different types of syntactic agreement recruit the same units in LLMs

Elena Polyakova*[1,2] Shivan Arora*[1,2] Suchir Salhan[1,2] Paula Buttery[1,2]

[1]Department of Computer Science and Technology, University of Cambridge [2]ALTA Institute

Motivation

- **Syntactic Interpretability**: How is grammar represented within language models?
- **Cross-linguistic and cross-phenomena patterns**: Are there similarities in grammatical representations across different syntactic structures and languages?
- Reproduced the key findings of Kryvosheieva et al. [2] through ablation, cross-overlap and cross-validation analyses on **BLiMP across seven LLMs, extending the analyses to RuBLiMP and SLING using Gemma-3 4B**.
- Our findings are consistent across different interpretability methods, suggesting that the **observed pattern of unit localisation generalises across methodological approaches**.
- Method-dependent results highlight the need for cross-method validation
- **Cross-Linguistic Extension** to Dutch, Russian and Turkish BLiMP-style datasets to assess cross-linguistic generalisability.

The Original Study

Using minimal pair benchmarks, Kryvosheieva et al. [2] investigated the localisation and overlap of syntax-responsive units in LLMs, identifying units by selecting the top-n% most responsive to each phenomenon. The paper has three main findings:

1. LLMs recruit localised units for specific phenomena.
2. Different instances of syntactic agreement recruit shared syntactic units.
3. The pattern replicates in Russian and Chinese; there is partial overlap cross-linguistically (typologically closer languages share more units for agreement).

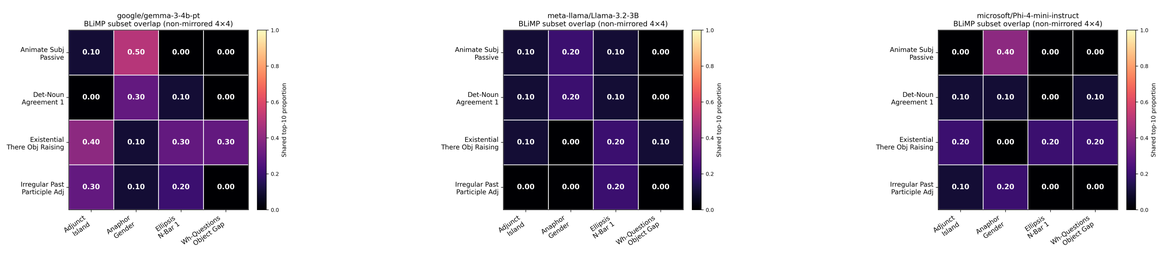
Extension Methods

- **Attention Head Analysis** identifies the top 10 attention heads per syntactic phenomena, then measures their overlap across syntactic phenomena to assess shared representations.
- **Differential Binary Masking** learns a binary mask over units, adapting the Hard-Concrete gate of Louizos et al. [3]
  - DBM's **discriminative objective** finds model units that **contain information about** a syntactic phenomena.
  - DBM's **causal objective** identifies units that are **responsible** for a phenomenon, by ablating selected units and checking whether the model's grammaticality judgements are worse.
- **Distributed Alignment Search** [1] locates a linear subspace of a model representation with a training objective defined by interchange interventions. identifies units that show a causal, not correlational relationship with grammaticality

Results

Attention-Head Results

- **Shared representations:** Related syntactic phenomena recruit overlapping attention heads across models.
- **Functional overlap:** Anaphor agreement, Principle A c-command, and determiner–noun agreement show recurring head sharing.
- **Model dependence:** The extent and organisation of overlap vary across architectures and languages.



Gemma-3 4B Llama-3.2-3B Phi-4-mini

Figure 1. Pairwise attention-head overlap across selected BLiMP phenomena for Gemma-3-4B, Llama-3.2-3B, and Phi-4 Mini. For each phenomenon, the ten heads with the largest grammatical–ungrammatical attention difference are selected. Each off-diagonal cell reports the proportion of these heads shared by a pair of phenomena; the diagonal is omitted.

Cross-Linguistic Dataset Results

- **Dutch:** Uneven localisation and weak functional grouping; parasitic gaps exhibit the clearest stability.
- **Turkish:** Strong cross-fold overlap (47–93%), with partial sharing between agreement phenomena.
- **Russian:** Cross-benchmark overlap exceeds chance (3.94% vs. 1%) but is lower than within-benchmark overlap (15.86–26.94%).

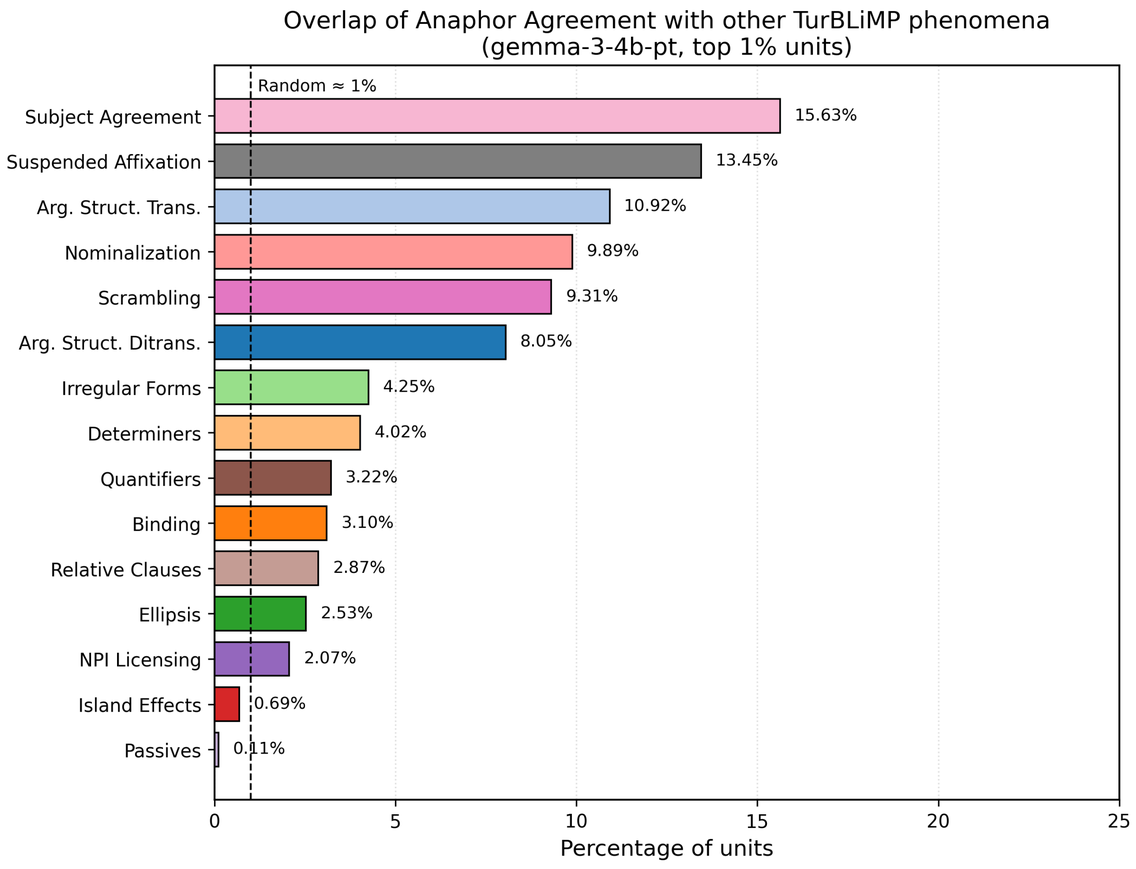


Figure 2. Cross-phenomenon overlap between Anaphor Agreement and the remaining TurBLiMP phenomena for Gemma-3 4B. Bars report the percentage of top-1% localised units shared with Anaphor Agreement. The dashed line indicates the analytical random expectation of 1%.

Findings

- Findings support the existence of **reusable syntactic and morphosyntactic representations** across models and languages.
- Attention-head overlap reveals shared components across phenomena, although this sharing is broader and less agreement-specific than in the residual-unit analysis of Kryvosheieva et al. [2].
- The DBM and DAS results further corroborate the presence of shared internal representations, while the neuron-level analyses provide cross-linguistic and cross-model evidence for morphosyntactic localisation.

References

[1] Geiger, A., Wu, Z., Potts, C., Icard, T., and Goodman, N. D. (2024). Finding alignments between interpretable causal variables and distributed neural representations.

[2] Kryvosheieva, D., de Varda, A. G., Fedorenko, E., and Tuckute, G. (2026). Different types of syntactic agreement recruit the same units within large language models. In Liakata, M., Moreira, V. P., Zhang, J., and Jurgens, D., editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 209–227, San Diego, California, United States. Association for Computational Linguistics.

[3] Louizos, C., Welling, M., and Kingma, D. P. (2018). Learning sparse neural networks through l_0 regularization.

BlackboxNLP Workshop | Empirical Methods in Natural Language Processing 2026 | Budapest, Hungary

*Corresponding Authors: ep757, sa2200@cam.ac.uk