

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

Elena Polyakova*   Shivan Arora*  

Suchir Salhan   Paula Buttery  

 Department of Computer Science & Technology, University of Cambridge, U.K.

 ALTA Institute, Cambridge, U.K.

Abstract

The representation of grammatical knowledge in individual LLM neurons and attention heads remains poorly understood. We replicate the feature-based approach of Kryvosheieva et al. (2026) that uses functional localisation to illustrate that distinct syntactic phenomena can recruit the same units. We reproduce and extend the paper to new minimal pair datasets (Dutch, Turkish, and Russian) and additional interpretability techniques: Attention-based analysis, Differential Binary Masking (DBM), and Distributed Alignment Search (DAS). We find reproducible and robust evidence for syntactic localisation in LLMs across several methods and languages. However, localisation strength depends on interpretability techniques (particularly for DBM) and varies across languages, highlighting the value of validating mechanistic claims using several techniques.¹

1 Introduction

Large Language Models (LLMs) are capable of distinguishing grammatical and ungrammatical sentences, yet the representation of grammatical knowledge within model architecture is not well understood (Wilcox et al., 2018; Gulordava et al., 2018; An et al., 2019; Warstadt et al., 2019; Gauthier et al., 2020; Linzen et al., 2016). Syntactic Minimal Pairs, including BLiMP (Warstadt et al., 2020), evaluate whether LMs show a probabilistic preference for grammatical sentences over minimally ungrammatical sentences. These datasets test on phenomena such as anaphor agreement (e.g. “Claude reinvestigates *herself*”). These have been developed for several languages, including RuBLiMP for Russian (Taktasheva et al., 2024) and SLING for Chinese (Song et al., 2022). Kryvosheieva et al. (2026) investigate whether

different syntactic phenomena recruit shared or distinct units using *feature-based* mechanistic interpretability methods, which identify human-interpretable properties encoded in model representations. (Chi et al., 2020; Papadimitriou et al., 2021; Brinkmann et al., 2025). They first identify the top- k units that distinguish between language and non-language stimuli; subsequent ablation of these units was found to decrease language performance more than ablating the same number of arbitrarily selected units. They primarily investigate English, as well as examining overlap in units used across different syntactic phenomena in Russian and Chinese.

The Original Study Using minimal pair benchmarks, Kryvosheieva et al. (2026) 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: (i) LLMs recruit localised units for specific phenomena (C_1); (ii) different instances of syntactic agreement recruit shared syntactic units (C_2); (iii) the pattern replicates in Russian and Chinese; there is partial overlap cross-linguistically (typologically closer languages share more units for agreement) (C_3).

Our Study We reproduce the key findings of Kryvosheieva et al. (2026) by conducting the original ablation, cross-overlap, and cross-validation analyses on BLiMP across seven LLMs, and on RuBLiMP and SLING using Gemma-3 4B. Complete replication results are reported in Appendix A. Although our numerical results are not identical to those of the original study, the differences are small and the same overall patterns are recovered, supporting the correctness and robustness of the implementation across repeated runs. We use three other mechanistic interpretability approaches: two different unit localisation methods that replace the original top- $n\%$ sampling in the ablation experi-

*Corresponding authors: {ep757, sa2200}@cam.ac.uk

¹Code and data available via <https://github.com/shivster2000/syntax-interpretability>

ment, and an attention-head analysis pipeline based on Clark et al. (2019).

Findings We illustrate that the findings of Kryvosheieva et al. (2026) hold across different methods of unit localisation, as well as at the level of attention-heads. Crucially, different interpretability methods yield different results and different localized units, highlighting the need for more cross-method studies. We corroborate their findings by additionally performing cross-validation and cross-overlap experiments on minimal pair-based datasets in other languages: Tur-BLiMP (Başar et al., 2025), BLiMP-NL (Suijkerbuijk et al., 2025), and CLAMS-Ru (Mueller et al., 2020).

2 Methodology: Background, Replication and Extension of Kryvosheieva et al. (2026)

2.1 Alternative Localisation Methods

Kryvosheieva et al. (2026) localise condition-specific units via last-token pooling: LLM unit activations are extracted for each condition, their absolute values computed, and a Welch’s t-test applied to identify the top- k most discriminative units. Kryvosheieva et al. (2026) introduce a two-fold cross-validation method to test whether identified LLM units are a generalisable syntactic feature, rather than a coincidence of localisation, by splitting minimal pairs into two halves and measuring the overlap of the top- n most syntactically responsive units that are activated for a specific minimal pair phenomena. They also analyse cross-overlap, where they carry out localisation independently for each minimal pair paradigm, and measure the overlaps for the unit sets identified for each paradigm.

However their original top- $n\%$ selection procedure may overlook more distributed syntactic representations and provide only a partial picture of the underlying computation. We therefore test whether the central representational claim—that syntactic phenomena recruit shared internal components—remains stable under three alternative interpretability methods: **Attention head analysis**, **Differential Binary Masking (DBM)** and **Distributed Alignment Search (DAS)**. Convergence across methods would provide evidence that the original result reflects genuine syntactic organisation in language models, rather than an artefact of a particular localisation procedure.

Attention Heads Attention heads provide a useful level of analysis for syntax because they encode token-to-token relations relevant to dependencies such as agreement, binding, and modification (Clark et al., 2019; Zheng et al., 2024). We identify heads most sensitive to grammatical contrasts in minimal pairs and test whether related phenomena recruit overlapping head sets across models, since individual heads and layers are not directly comparable between models.

Differential Binary Masking DBM learns a binary mask over units, adapting the Hard-Concrete gate of Louizos et al. (2018)—originally proposed for network pruning—to localise units involved in specific linguistic phenomena via a sparsity penalty weighted by a coefficient λ . We consider two variants, differing only in the training signal used to shape the mask. DBM’s **discriminative objective** finds model units that contain information about a syntactic phenomena, separating grammatical/ungrammatical sentences by training a small gate to select useful units and a linear probe to classify sentences, without changing the original model. 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 (De Cao et al., 2020, 2022). The discriminative objective operates on a frozen model, whereas the causal objective forms a real intervention via forward hooks on the live model, every training step. Whilst the discriminative objective identifies units that correlate with grammaticality, the causal objective identifies units that are causally implicated, by interventions on the model. A full description of the methodology can be found in Appendix H.2.

Distributed Alignment Search DAS (Geiger et al., 2024) locates a linear subspace of a model representation with a training objective defined by interchange interventions (Huang et al., 2024). In the original formulation, the subspace is parameterised as an $n \times n$ orthogonal matrix $Q = [u_1 \dots u_n]$ rotating the representation into a new coordinate system, with the first k rotated dimensions designated as the subspace of interest. As this is computationally expensive, we use the implementation of Huang et al. (2024), which learns only the $k \ll n$ orthogonal vectors forming the feature F_A . This differs from DBM and the original methodology in that it does not assume the rele-

vant computation is aligned with an individual unit, but rather a linear subspace; it can find a feature distributed across many units. Notably, to allow this method to be compared to the other unit-level approaches, we convert the learned subspace to a per-unit mask. This method also identifies units that show a causal, not correlational relationship with grammaticality. Our implementation is explicated in detail in Appendix H.3.

DAS learns a subspace of units, whereas DBM assumes individual units are causally linked. In intervention, DBM on a causal objective zeroes identified units, whereas DAS’s interchange intervention swaps in the activation values of a different sentence on the aligned subspace, allowing DAS to find things DBM structurally cannot (distributed features), and DBM to find things DAS might dilute (sparse, localized features). Based on the differences in methodology, we would expect each method to identify different results and different localised units, with DBM causal and DAS being aligned with each other.

Experimental Setup We apply DBM and DAS to the top-1% ablation, 2-fold cross-validation, and cross-overlap experiments of Kryvosheieva et al. (2026), comparing the sensitivity and specificity of the localisation methods. We also analyse attention head overlap to determine whether different syntactic phenomena recruit shared sets of heads. These experiments are conducted on 12 paradigms from BLiMP and subsequently on a RuBLiMP subset, examining variation across phenomena and models (see Appendix B for a full description of the selected phenomena and models). This restricted selection enables comparison across models, methods, and languages while remaining computationally feasible; Mistral and Falcon are excluded due to resource and implementation constraints.

2.2 Benchmark Extensions

We further extend the unit-localisation analysis of Kryvosheieva et al. (2026) to three additional benchmarks: BLiMP-NL (Suijkerbuijk et al., 2025), TurBLiMP (Başar et al., 2025), and CLAMS (Mueller et al., 2020). Using the released codebase, Gemma-3-4B model and original experimental framework, we test whether syntax-specific localisation generalises across additional languages and benchmark designs.

Cross-linguistic Evaluation BLiMP-NL contains 8,400 Dutch minimal pairs across 84

paradigms and 22 syntactic phenomena, while TurBLiMP covers 16 Turkish phenomena with 1,000 pairs each. Dutch and Turkish provide typologically distinct extensions: Dutch is a Germanic Indo-European language, whereas Turkish belongs to the Turkic family (Dryer and Haspelmath, 2013). As BLiMP-NL does not support a systematic analysis of agreement as a functional category, we instead examine whether A’-dependencies exhibit a comparable shared-component structure.

Generalisation to Other Linguistic Metrics

We additionally consider CLAMS, the Cross-Linguistic Assessment of Models on Syntax, to extend the original study’s cross-dataset generalisation analysis to a cross-linguistic setting. CLAMS contains subject–verb agreement paradigms across multiple languages; we focus on Russian. As CLAMS-Russian does not include a directly equivalent subject–predicate agreement category, we use its subject–verb agreement paradigms—including long verb-phrase coordination and agreement across object relative clauses—as the closest available syntactic counterpart to RuBLiMP. We apply the same unit-localisation procedure to both datasets and compare the resulting unit sets, testing whether localisation patterns generalise across benchmarks targeting the same broad agreement phenomenon or remain dataset-specific.

3 Results

We present the results of our analyses of localisation patterns across different methods, using BLiMP and RuBLiMP phenomena, and different minimal pair datasets.

3.1 Mechanistic Interpretability Results

We report our findings across various interpretability methods, including attention heads, DBM, and DAS, examining how syntactic localisation varies across different methodological implementations.

3.1.1 Attention Heads

Across models, recurring overlap patterns emerge among the top-10 attention heads. Figure 1 presents representative matrices for three models, with the remaining results reported in Appendix C. The most consistent overlap occurs between anaphor gender agreement, and Principle A c-command, while anaphor gender agreement also overlaps with determiner–noun agreement, particularly in DeepSeek and Gemma. Similar sharing is

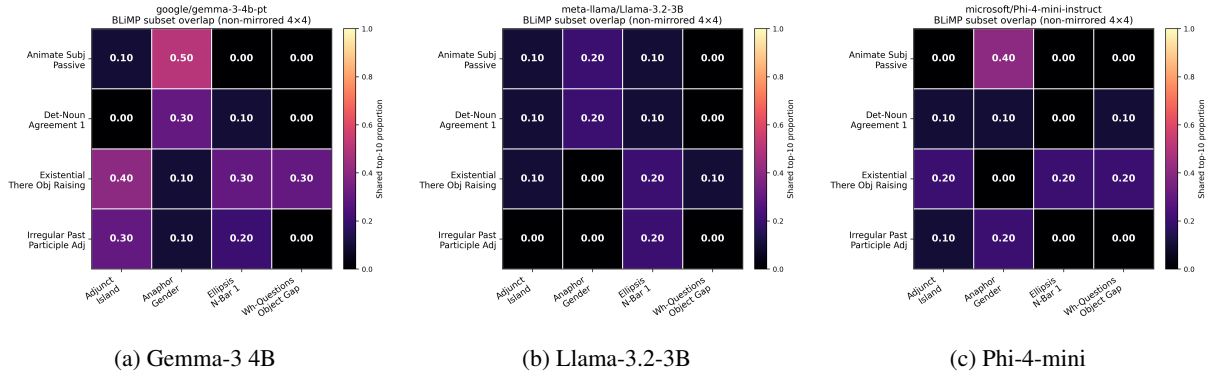


Figure 1: Pairwise attention-head overlap across selected BLiMP phenomena for Gemma-3-4B, Llama-3.2-3B, and Phi-4 Mini. For each phenomenon, the ten heads with the largest grammatical–ungrammatical attention difference are selected. Each off-diagonal cell reports the proportion of these heads shared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

observed among other agreement- and dependency-related phenomena. The recurrence of these relationships across models suggests that **attention heads are not exclusively tied to individual constructions, but may be reused for related syntactic computations**. This provides complementary support for Kryvosheieva et al. (2026)’s claim that distinct phenomena recruit shared internal components.

The RuBLiMP results show substantial cross-phenomenon attention-head overlap, although the extent of sharing varies by model. Gemma shows strong overlap between verb–accusative object constructions and adposition government, while Llama exhibits particularly dense sharing, including between verb conjugation and duration aspect. These patterns suggest **reuse of attention heads across related morphosyntactic phenomena involving case, agreement, tense and aspect in Gemma and Llama**. Phi, however, produces a much sparser overlap structure, indicating that **localisation of TAM is model-specific**. Overall, the attention-head analysis supports the existence of shared syntactic components across languages, while showing that their organisation and degree of reuse vary according to both the language and the model.

3.1.2 Alternative Unit Localisation Methods: DBM and DAS

Implementing DBM and DAS localisation methods showed similar findings to the original experiments of Kryvosheieva et al. (2026). Tables 4 and 6 in Appendix D illustrate a decrease in performance for different localisation methods across a range of syntactic phenomena when the top 1% of units are ablated, suggesting a level of functional locali-

sation that supports the claim of the initial experiments. Ablations with DBM’s causal objective has larger reductions in grammaticality accuracy than units identified by DBM’s discriminative objective, suggesting that causal localisation more effectively identifies relevant functional units. We attribute this finding to the fact that the DBM with a causal objective is trained directly on the effect of ablating gated units—these tangibly affect performance, whereas the discriminative objective learns units that are predictive of differences in performance. Our results suggest that predictive units are not always causally involved in model behaviour.

Cross-Phenomena Overlap We find variance in overlap magnitude with different localisation methods across different datasets, with DAS having higher cross-phenomena overlap than the original method of Kryvosheieva et al. (2026). Figure 27 & Figure 25 in Appendix D illustrate similar findings in the cross-overlap experiments to the original paper. However, there is strong variation in the size of overlap when different unit localisation methods are employed; DAS shows higher overlap across phenomena, e.g. 40.58% overlap for subject-predicate agreement, compared to 13.79% overlap with the original method. These findings hold across datasets too; the same pattern is observed with the RuBLiMP data, highlighting that different mechanistic interpretability methods can diverge significantly depending on the methods used to identify the subset of neurons implicated with a phenomenon. This underscores the need to compare across methods, rather than relying entirely on one method.

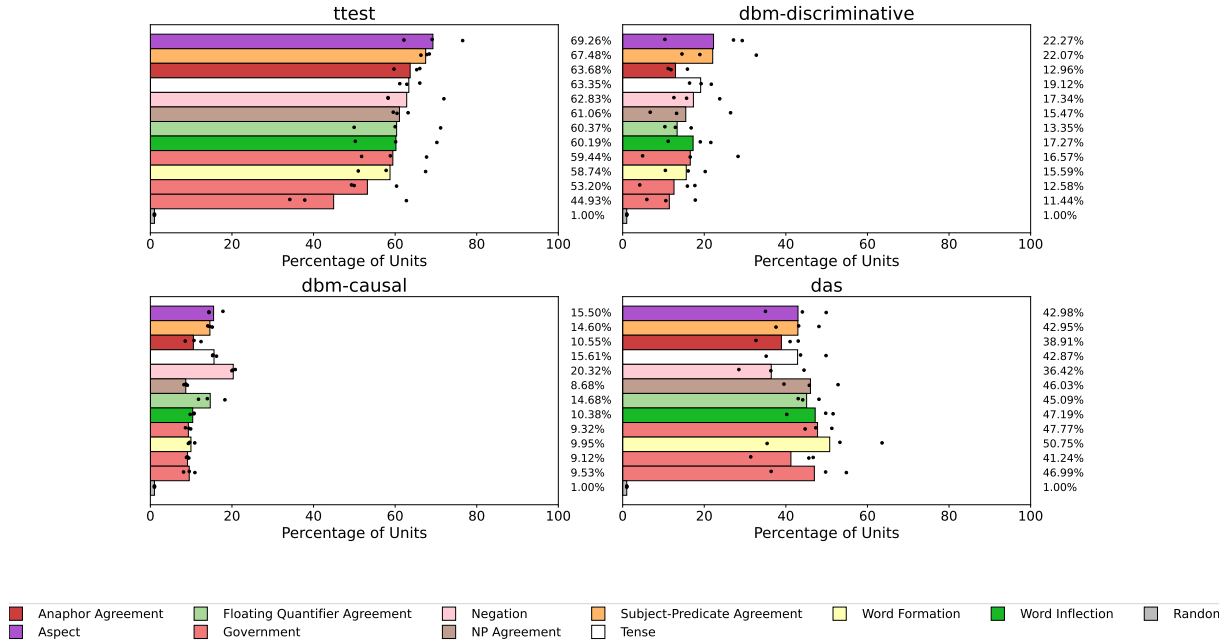


Figure 2: Effects of cross-validation on a subset of RuBLiMP, aggregated across Gemma-3-4B, Llama-3.2-3B, and Phi-4 Mini. The ttest method reproduces the original methodology of Kryvosheieva et al. (2026), while the remaining methods show the results of the alternative functional localisation approaches used in this work.

2-fold Cross-Validation 2-fold CV experiments reveal sensitivity of the technique on specific unit localisation techniques, but we broadly identify similar high-overlap trends as Kryvosheieva et al. (2026). We find DBM has substantially lower cross-phenomena overlap than the baseline or DAS methods; this difference further highlights the necessity of cross-method analyses.

Extent of Unit Localisation varies across Models

We also report significant variation across different models in the extent of unit localisation; across a variety of methods, models such as Llama showed a smaller decrease in performance post ablation; Tables 3 and 5 in Appendix D illustrate a *decrease in performance of 27.60%* for Gemma using DAS localisation, whereas for Llama the *change was only 7.46%*, when tested on a subset of BLiMP. Our extension of Kryvosheieva et al. (2026) using DAS and DBM, therefore, complicates a universal definition of syntactic localisation across all models, with Gemma representing more syntax-responsive units than Llama’s more diffuse representations.

3.2 Cross-Linguistic Evaluation

Cross-linguistic dataset extensions on Dutch, Turkish, and Russian shows that syntax-localised units are not uniformly distributed.

3.2.1 BLiMP-NL

Our results for Dutch generally show non-uniform syntactic localisation with substantial cross-category overlap and limited localisation. The BLiMP-NL cross-phenomenon results, compared with the previous RuBLiMP findings in Figure 3, provide uneven support for syntax-responsive localisation. Verb Second, Adverbial Modification, and Parasitic Gaps exhibit the highest within-category overlap, whereas Crossing Dependencies and Quantifiers show considerably lower values. For several categories, overlap with unrelated phenomena is comparable to, or greater than, within-category overlap, weakening evidence for phenomenon-specific localisation. The 2-fold analysis in Figure 28 in Appendix E shows similarly variable stability, with cross-fold overlap ranging from approximately 10% to 72%, against a 1% random expectation. Parasitic Gaps provides the clearest evidence of stable localisation, exhibiting both high within-category and cross-fold overlap.

The weaker pattern elsewhere may partly reflect the smaller size of BLiMP-NL, which contains approximately 8.4K minimal pairs compared with 67K in BLiMP and 45K in RuBLiMP, potentially making unit selection noisier. The A’-dependency analysis, found in Figure 30 in Appendix E, also provides limited evidence for a

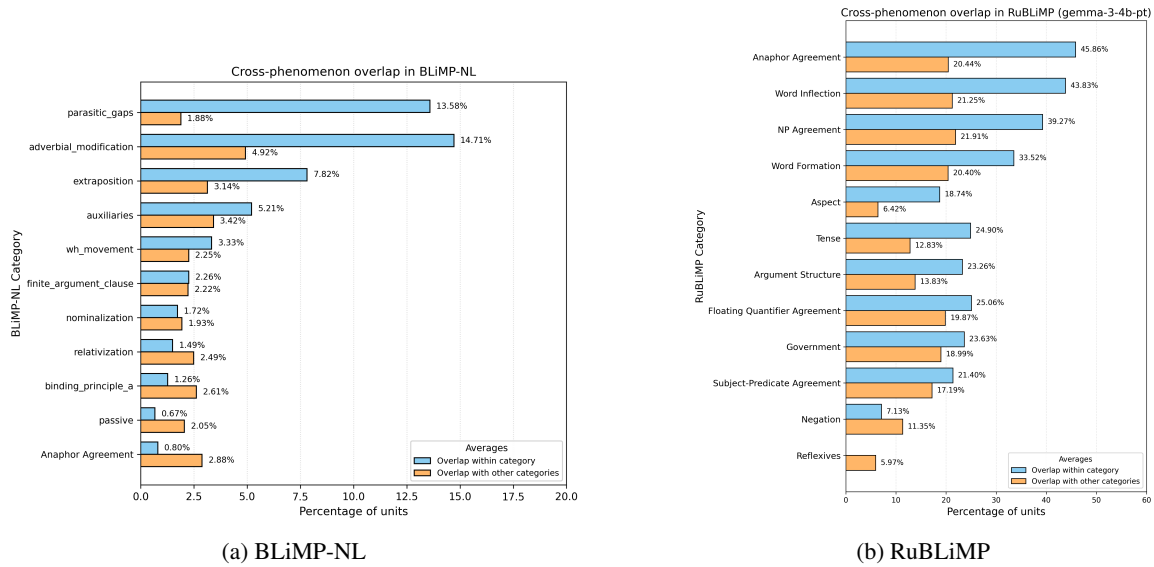


Figure 3: **Within- and cross-category unit overlap in BLiMP-NL and RuBLiMP.** Bars show the mean overlap among paradigms within each syntactic category and with paradigms outside that category for Dutch (left) and Russian (right), using the Gemma model. Categories are ordered by the difference between within- and cross-category overlap.

broader functional grouping. For A'-Movement Restrictions, Relativization, Wh-Movement, and Topicalization, within-category overlap is similar to overlap with other A' and non-A' phenomena. Parasitic Gaps is the main exception, with approximately 13.5% within-category overlap compared with 1.4–2.7% in the other conditions. The results therefore do not support A' dependencies as a uniformly represented functional category, although some constructions recruit stable or partially shared units.

3.2.2 TurBLiMP

The TurBLiMP results provide strong converging evidence for stable syntax-responsive localisation. As shown in Figure 31 in Appendix F, overlap between units localised from two independent folds ranges from approximately 47% to 93%, with every phenomenon substantially exceeding the 1% random baseline. This indicates that similar unit sets are recovered across different instances of the same construction, replicating the cross-validation findings of Kryvosheieva et al. (2026). Figure 4 presents the overlap between Anaphor Agreement and the remaining TurBLiMP phenomena. The highest cross-phenomenon overlap is observed with Subject Agreement, at approximately 15.6%, while the other comparisons range from 13.45% to 0.69%. A corresponding analysis centred on Subject Agreement is reported in Figure 32

in Appendix F. The strong overlap between the two agreement phenomena supports the claim that related agreement constructions recruit partially shared units. However, elevated overlap with Suspended Affixation and transitive Argument Structure indicates that this sharing is not exclusively agreement-specific. Agreement therefore forms a partially coherent functional category in TurBLiMP while also sharing resources with broader morphosyntactic processes. Overall, TurBLiMP provides stronger and more uniform evidence for stable syntactic localisation than BLiMP-NL, alongside selective sharing between related phenomena.

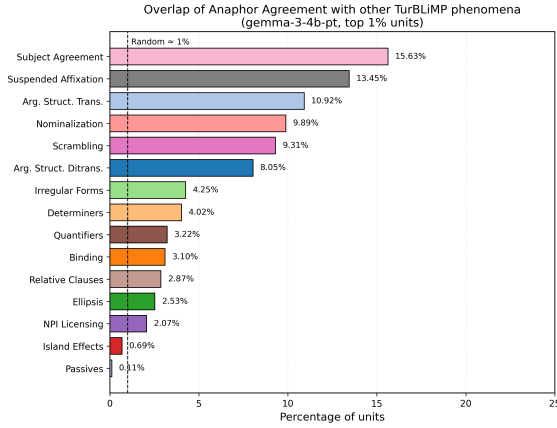


Figure 4: Cross-phenomenon overlap between Anaphor Agreement and the remaining TurBLiMP phenomena for Gemma-3 4B. Bars report the percentage of top-1% localised units shared with Anaphor Agreement. The dashed line indicates the analytical random expectation of 1%.

3.3 CLAMS vs RuBLiMP

Figure 5 compares units localised for subject-predicate agreement in RuBLiMP with those localised for subject-verb agreement in CLAMS-Russian. Across 77 cross-benchmark paradigm pairs, the mean overlap was 27.23 units, equivalent to 3.94% of the 691-unit target set. This was significantly greater than the random expectation of 6.91 units, or 1%, $t = 14.54$, $p < .001$.

Despite exceeding chance, cross-benchmark overlap remained substantially lower than overlap within either benchmark. RuBLiMP paradigm pairs shared an average of 186.16 units, or 26.94%, whereas CLAMS-Russian pairs shared 109.57 units, or 15.86%. Welch’s t -tests confirmed that cross-benchmark overlap was significantly lower than both within-RuBLiMP overlap, $t = -16.22$, $p < .001$, and within-CLAMS-Russian overlap, $t = -3.70$, $p = .0014$. This pattern is unlikely to result from unstable localisation in CLAMS-Russian, as the 2-fold overlap reported in Appendix G ranged from 68.85% to 91.84%, well above the 1% random expectation.

Together, these findings indicate that agreement localisation includes a shared cross-benchmark component, while remaining strongly sensitive to benchmark-specific constructions and materials. The relatively modest cross-benchmark overlap may also reflect the absence of an exact subject-predicate agreement counterpart in CLAMS-Russian, requiring subject-verb agreement to be used as the closest available compar-

ison.

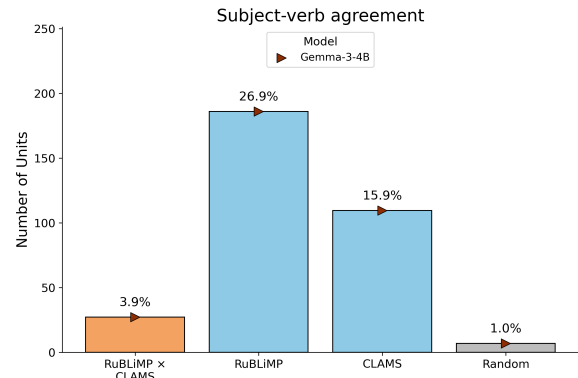


Figure 5: Agreement-unit overlap within and across RuBLiMP and CLAMS-Russian for Gemma-3 4B. Bars show the mean number of shared units across benchmarks, within RuBLiMP, within CLAMS-Russian, and under the analytical random baseline. Percentages are calculated relative to the 691-unit target set.

4 Conclusion

Overall, our 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. (2026). 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. There are nonetheless distinct differences in results and units localised through different mechanistic interpretability methods, highlighting the need for further cross-method analyses. Moreover, the extent and organisation of this localisation vary across models, suggesting that specialised syntactic information is encoded differently across architectures. The dataset extensions provide similarly qualified support: syntax-responsive units generalise across languages and benchmarks, but their stability and degree of sharing depend on the construction and dataset. These results strengthen the original claims by providing further evidence for syntactic localisation.

5 Limitations

Several limitations should be considered when interpreting the Dutch, Turkish, and methodological extensions. For BLiMP-NL, the relatively small amount of available data may have reduced localisation stability, as each paradigm contains only 100 minimal pairs. TurBLiMP, by contrast, is not segmented at the paradigm level, which restricted the analysis to broader phenomena and prevented a more fine-grained assessment of whether overlap was concentrated within closely related paradigms or extended across different syntactic categories. The methodological extensions were also limited by computational constraints, requiring the use of only 12 paradigms from BLiMP and RuBLiMP rather than the full datasets. This reduced the amount of evidence available for evaluating intra-category variation and the broader generality of the observed localisation patterns.

Acknowledgements

This work is supported by Cambridge University Press & Assessment. With thanks to Professor Andrew Caines, Dr Diana Galvan-Sosa and Lily Goulder.

References

- Aixiu An, Peng Qian, Ethan Wilcox, and Roger Levy. 2019. [Representation of constituents in neural language models: Coordination phrase as a case study](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2888–2899, Hong Kong, China. Association for Computational Linguistics.
- Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. [Turblimp: A turkish benchmark of linguistic minimal pairs](#). In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, page 16506–16521. Association for Computational Linguistics.
- Jannik Brinkmann, Chris Wendler, Christian Bartelt, and Aaron Mueller. 2025. [Large language models share representations of latent grammatical concepts across typologically diverse languages](#). 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 6131–6150, Albuquerque, New Mexico. Association for Computational Linguistics.
- Ethan A. Chi, John Hewitt, and Christopher D. Manning. 2020. [Finding universal grammatical relations in multilingual BERT](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5564–5577, Online. Association for Computational Linguistics.
- Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. [What does BERT look at? an analysis of BERT’s attention](#). In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.
- Nicola De Cao, Michael Sejr Schlichtkrull, Wilker Aziz, and Ivan Titov. 2020. [How do decisions emerge across layers in neural models? interpretation with differentiable masking](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3243–3255, Online. Association for Computational Linguistics.
- Nicola De Cao, Leon Schmid, Dieuwke Hupkes, and Ivan Titov. 2022. [Sparse interventions in language models with differentiable masking](#). In *Proceedings of the Fifth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pages 16–27, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.
- Google DeepMind. 2025. [Gemma 3 technical report](#). *arXiv preprint arXiv:2503.19786*.
- DeepSeek. 2024. [DeepSeek LLM: Scaling open-source language models with longtermism](#). *arXiv preprint arXiv:2401.02954*.
- Matthew S. Dryer and Martin Haspelmath, editors. 2013. [WALS Online](#). Max Planck Institute for Evolutionary Anthropology, Leipzig.
- Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 2020. [SyntaxGym: An online platform for targeted evaluation of language models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 70–76, Online. Association for Computational Linguistics.
- Atticus Geiger, Zhengxuan Wu, Christopher Potts, Thomas Icard, and Noah D. Goodman. 2024. [Finding alignments between interpretable causal variables and distributed neural representations](#).
- 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, New Orleans, Louisiana. Association for Computational Linguistics.

- Jing Huang, Zhengxuan Wu, Christopher Potts, Mor Geva, and Atticus Geiger. 2024. [RAVEL: Evaluating interpretability methods on disentangling language model representations](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8669–8687, Bangkok, Thailand. Association for Computational Linguistics.
- Daria Kryvosheieva, Andrea Gregor de Varda, Evelina Fedorenko, and Greta Tuckute. 2026. [Different types of syntactic agreement recruit the same units within large language models](#). In *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.
- 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.
- Christos Louizos, Max Welling, and Diederik P. Kingma. 2018. [Learning sparse neural networks through \$l_0\$ regularization](#).
- Meta. 2024. [Llama can now see and run on your device - welcome Llama 3.2](#).
- Microsoft. 2025. [Phi-4-mini technical report: Compact yet powerful multimodal language models via Mixture-of-LoRAs](#). *arXiv preprint arXiv:2503.01743*.
- Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. [Cross-linguistic syntactic evaluation of word prediction models](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.
- Isabel Papadimitriou, Ethan A. Chi, Richard Futrell, and Kyle Mahowald. 2021. [Deep subjecthood: Higher-order grammatical features in multilingual BERT](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2522–2532, Online. Association for Computational Linguistics.
- Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.
- Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. 2022. [Sling: Sino linguistic evaluation of large language models](#).
- Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. [BLiMP-NL: A corpus of Dutch minimal pairs and acceptability judgments for language model evaluation](#). *Computational Linguistics*, 51(4):1267–1301.
- Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. [RuBLiMP: Russian benchmark of linguistic minimal pairs](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9268–9299, Miami, Florida, USA. Association for Computational Linguistics.
- Alex Warstadt, Yu Cao, Ioana Grosu, Wei Peng, Hagen Blix, Yining Nie, Anna Alsop, Shikha Bordia, Haokun Liu, Alicia Parrish, Sheng-Fu Wang, Jason Phang, Anhad Mohananey, Phu Mon Htut, Paloma Jeretic, and Samuel R. Bowman. 2019. [Investigating BERT’s knowledge of language: Five analysis methods with NPIs](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2877–2887, Hong Kong, China. Association for Computational Linguistics.
- Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. [BLiMP: The benchmark of linguistic minimal pairs for English](#). *Transactions of the Association for Computational Linguistics*, 8:377–392.
- Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. [What do RNN language models learn about filler–gap dependencies?](#) In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.
- Zifan Zheng, Yezhaohui Wang, Yuxin Huang, Shichao Song, Mingchuan Yang, Bo Tang, Feiyu Xiong, and Zhiyu Li. 2024. [Attention heads of large language models: A survey](#).

A Reproduced results

A.1 English

We reproduce the original BLiMP experiments to verify the findings of the original study. The figures below illustrate the comparison between the original and replicated 1% ablation, 2-fold cross-validation, cross-overlap, and agreement-overlap results. Across these experiments, we find the same patterns as the original results, corroborating the claims of syntactic unit localisation of the original paper for English.

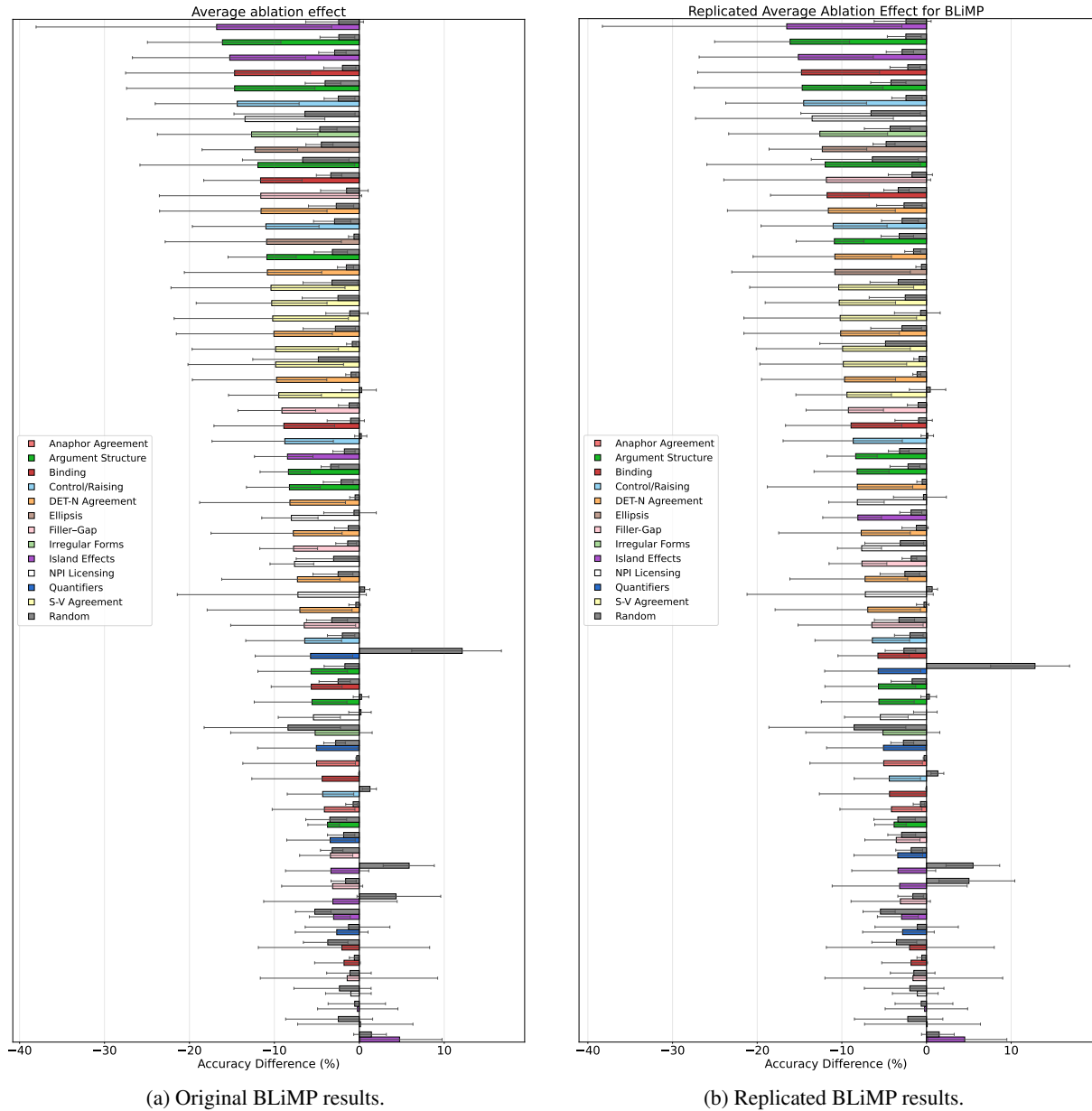


Figure 6: **Ablation effects of syntax-responsive units on BLiMP performance.** For each of the 67 BLiMP phenomena, coloured bars show the mean change in accuracy after ablating the top 1% of localised units relative to the unablated model, averaged across seven models. Phenomena are ordered from the largest to the smallest performance decrease. Grey bars show the corresponding effect of ablating a randomly selected 1% of units, averaged across seven models and four random seeds. Error bars indicate 95% confidence intervals across models. Results show the original BLiMP (left) and replicated BLiMP (right), on all 7 models.

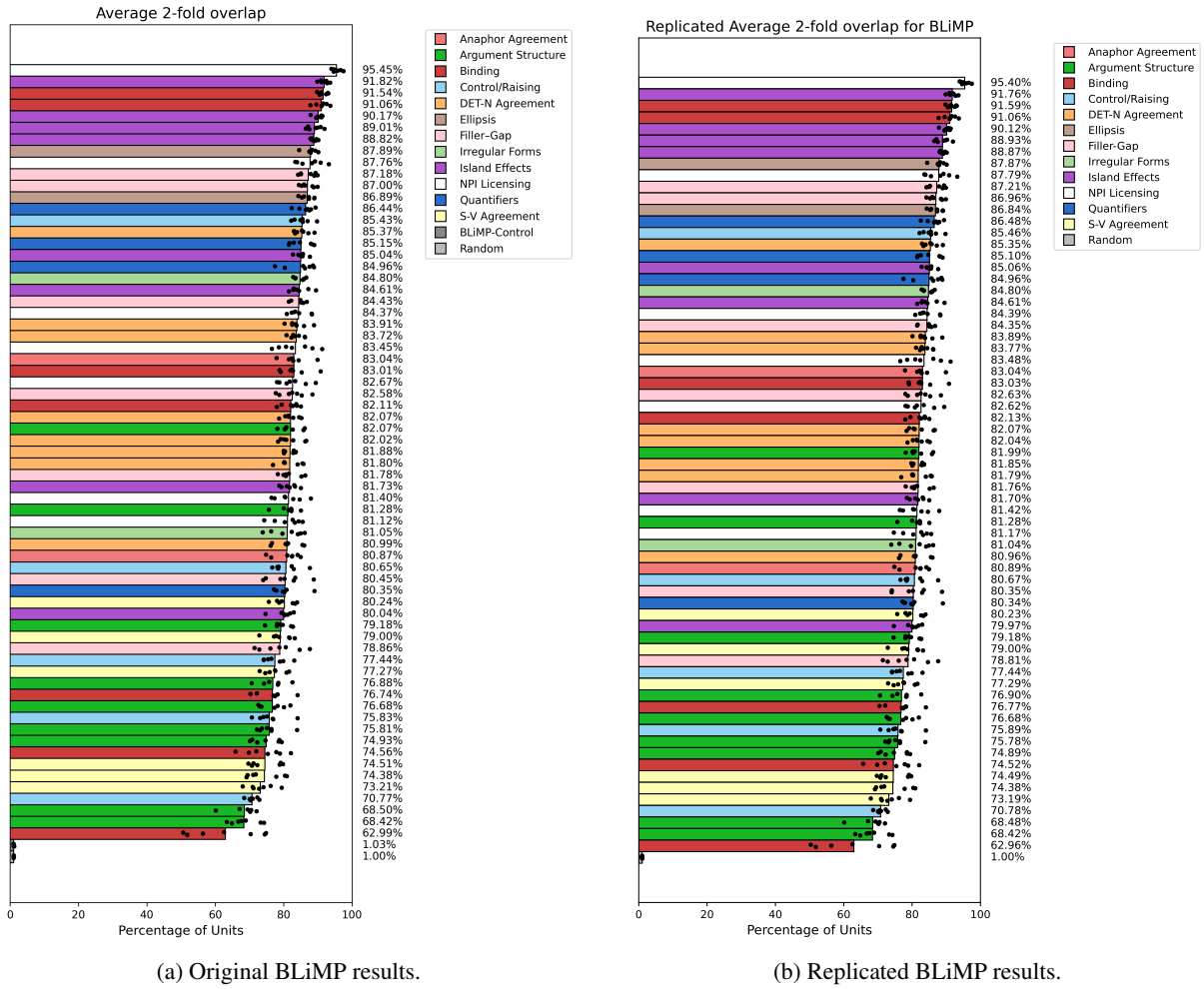


Figure 7: **Comparison of the original and replicated two-fold localisation results for BLiMP.** Bars show the mean percentage overlap between the top-1% unit sets identified from two independent halves of each paradigm. In the replication, black markers show the results for the individual evaluated models. Bar colours denote the corresponding BLiMP syntactic categories, and the grey bar marks the 1% random expectation.

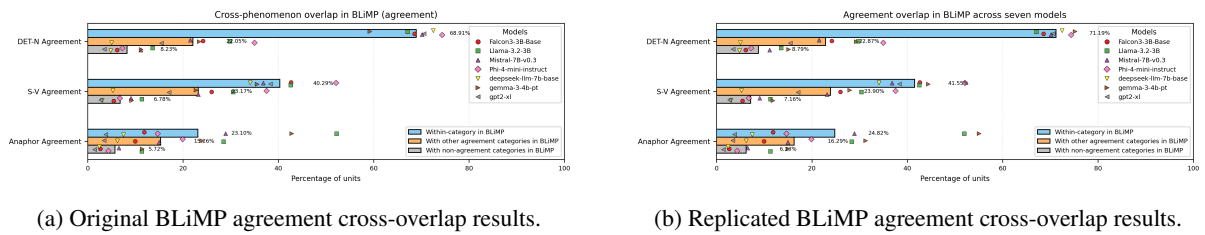
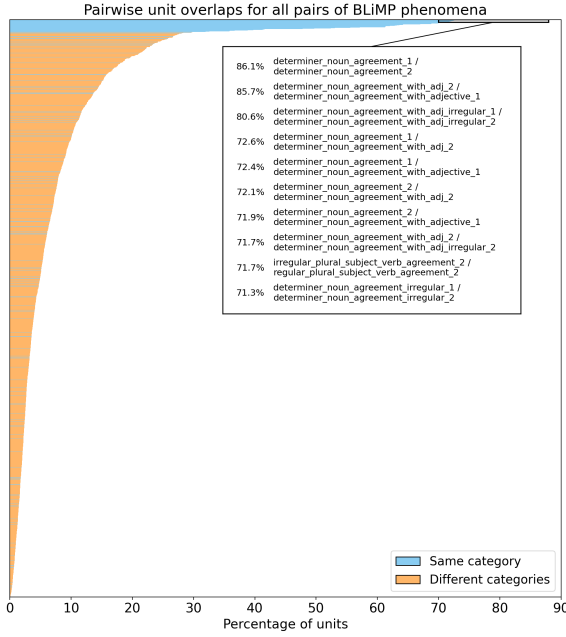
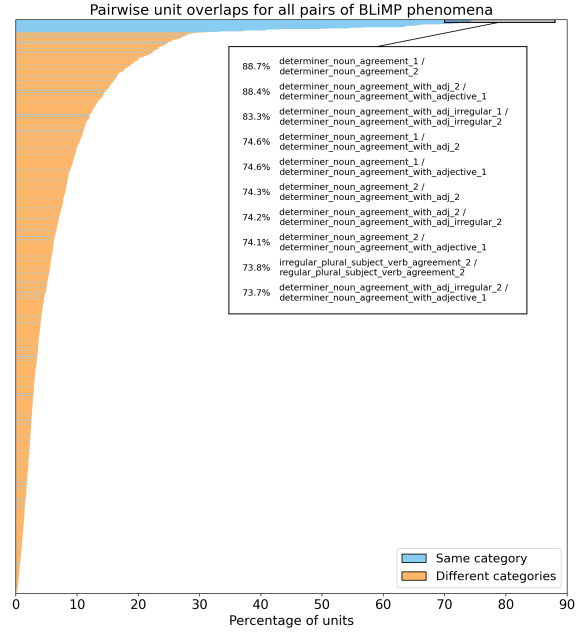


Figure 8: **Comparison of the original and replicated agreement-specific unit overlap results for BLiMP.** For determiner–noun, subject–verb, and anaphor agreement, bars compare overlap within the corresponding agreement category, with other agreement categories, and with non-agreement categories. Markers show the results for individual models.

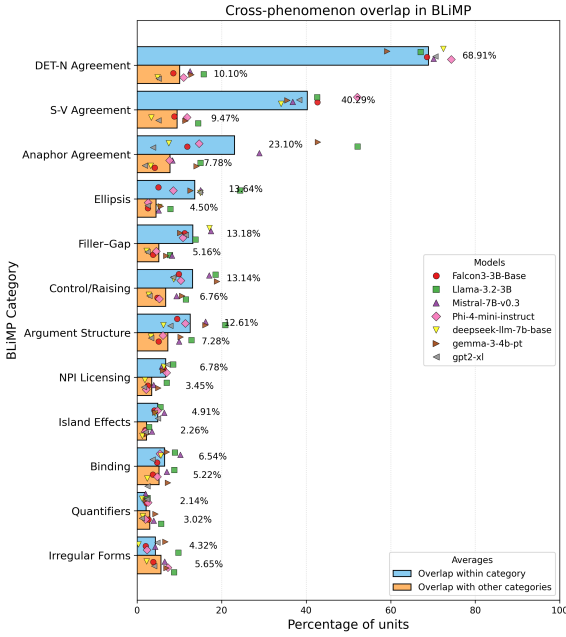


(a) Original BLiMP results.

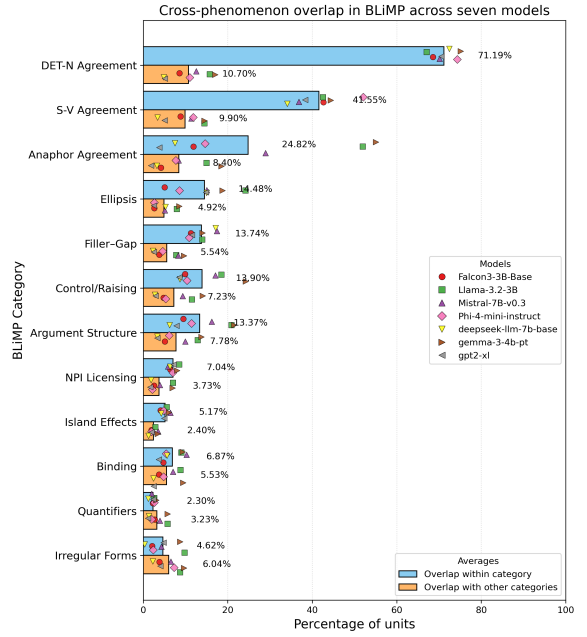


(b) Replicated BLiMP results.

Figure 9: **Comparison of the original and replicated two-fold localisation results for BLiMP.** Bars show the mean percentage overlap between the top-1% unit sets identified from two independent halves of each paradigm. In the replication, black markers show the results for the individual evaluated models. Bar colours denote the corresponding BLiMP syntactic categories, and the grey bar marks the 1% random expectation.



(a) Original BLiMP cross-overlap results.



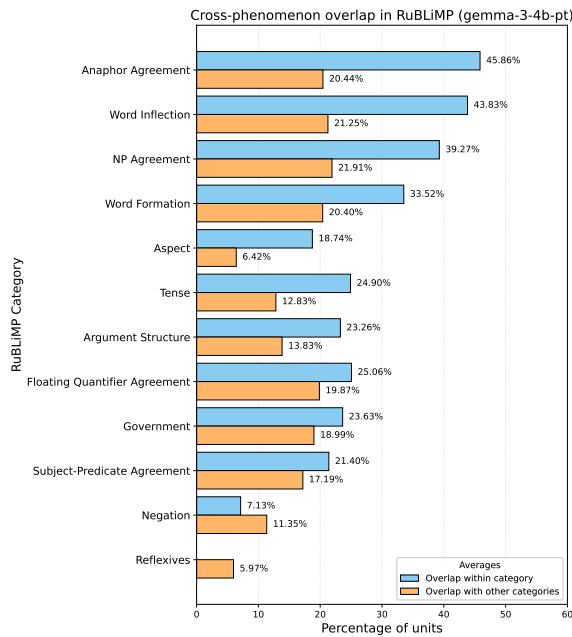
(b) Replicated BLiMP cross-overlap results.

Figure 10: **Comparison of the original and replicated within- and cross-category unit overlap results for BLiMP.** For each BLiMP category, blue bars show the mean overlap among paradigms within the category, while orange bars show the mean overlap with paradigms outside the category. Markers report the corresponding values for individual models.

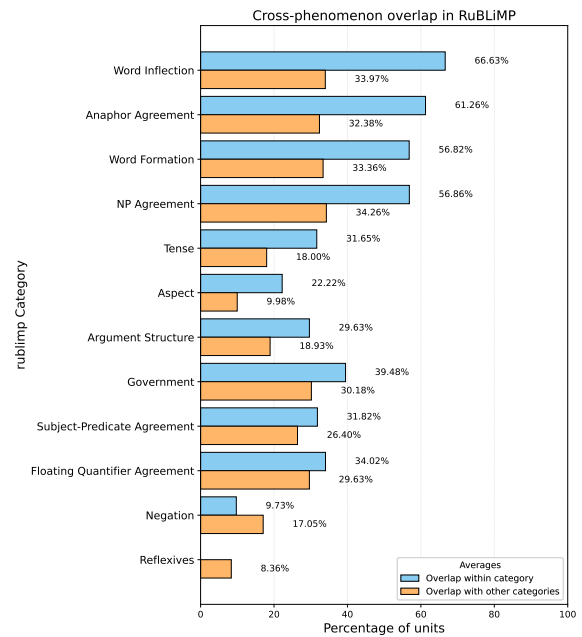
A.2 Russian

We also reproduce the original RuBLiMP analyses to assess whether the reported localisation patterns are recoverable under an independent implementation. The figures below compare the original and replicated

cross-validation, cross-overlap, and agreement-overlap results. Across analyses, we recover the same overall patterns with only minor numerical differences, supporting the reproducibility of the original unit-localisation method for Russian.

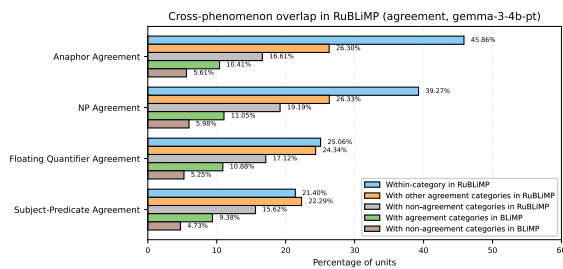


(a) Original RuBLiMP cross-overlap results.

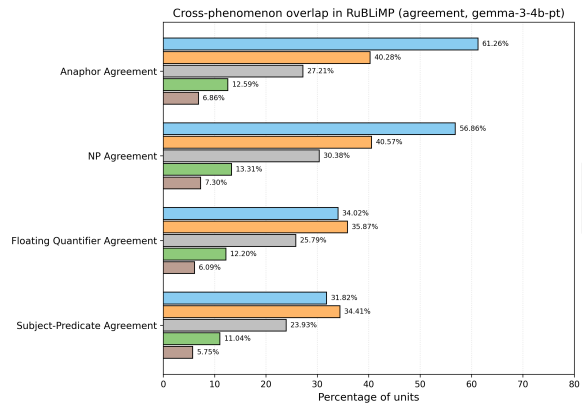


(b) Replicated RuBLiMP cross-overlap results.

Figure 11: **Within- and cross-category unit overlap.** Results are shown for original RuBLiMP (left) and replicated RuBLiMP (right), both using the Gemma model.



(a) Original RuBLiMP agreement cross-overlap results.



(b) Replicated RuBLiMP agreement cross-overlap results.

Figure 12: **Cross-lingual overlap of units localised for agreement phenomena.** Results are shown for original RuBLiMP (left) and replicated RuBLiMP (right), both using the Gemma model. Bars report mean overlap as a percentage of the top-1% localised units: within the same agreement category (blue), across other agreement categories in the same language (orange), with non-agreement phenomena in the same language (grey), with agreement categories in English BLiMP (green), and with non-agreement categories in English BLiMP (brown).

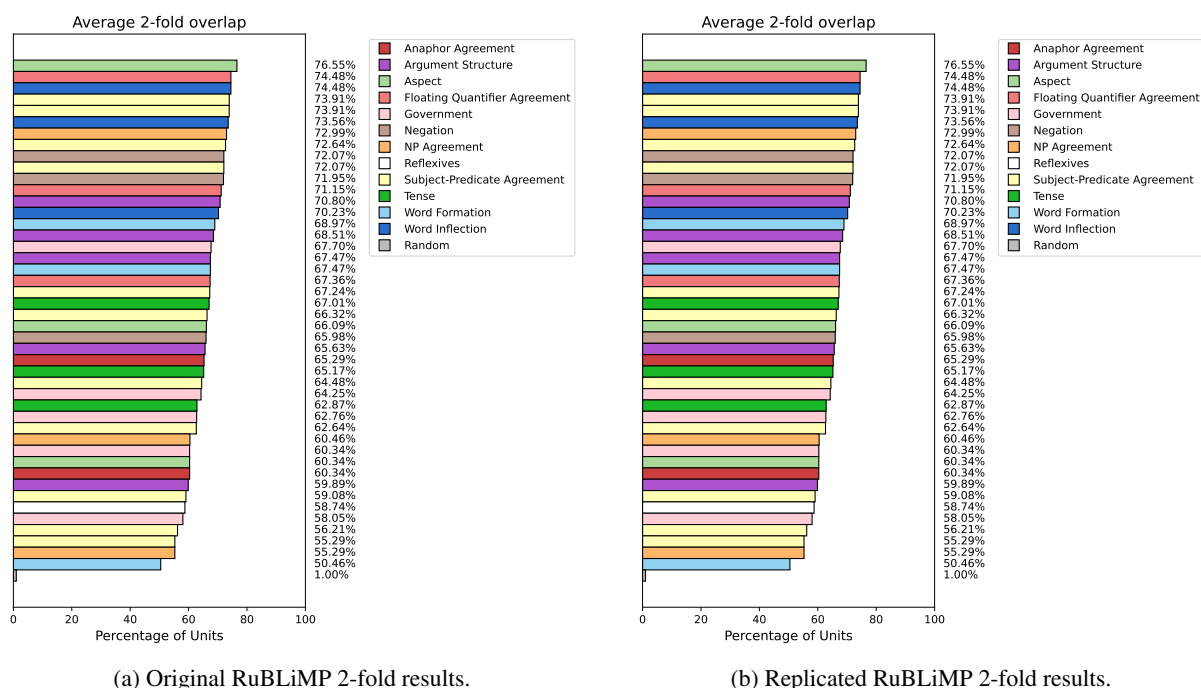


Figure 13: **Two-fold cross-validation.** targeting the top 1% of units. Results are shown for the original RuBLiMP analysis (left) and the replicated RuBLiMP analysis (right).

A.3 Chinese

For Chinese, we evaluate whether the SLING localisation patterns reported in the original paper persist under replication. The figures below present the original and reproduced 2-fold cross-validation, cross-phenomenon overlap and agreement-specific overlap results. The replicated analyses preserve the same broad structure, with only small numerical deviations, indicating that the reported SLING localisation patterns are robust to reimplementations.

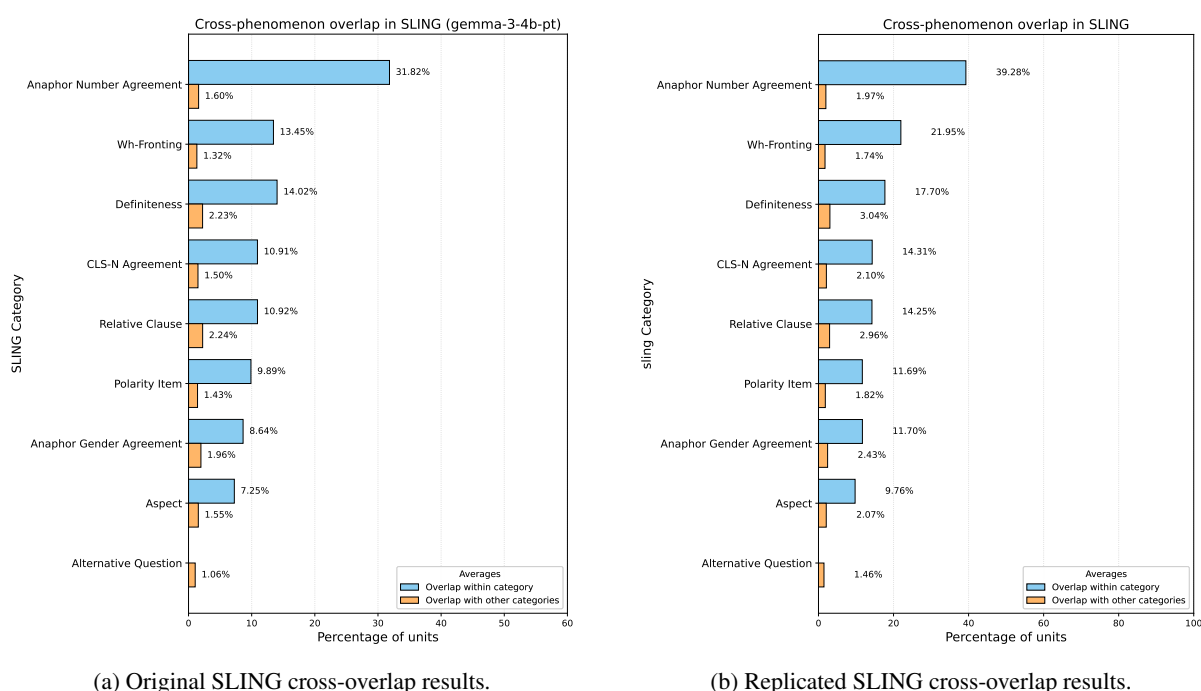
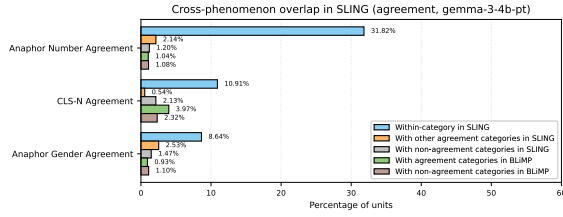
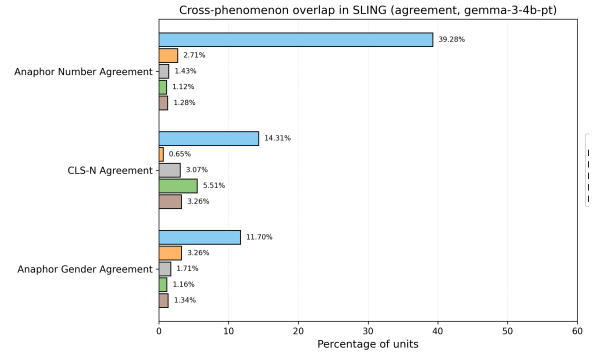


Figure 14: **Within- and cross-category unit overlap.** Results are shown for original SLING (left) and replicated SLING (right), both using the Gemma model.

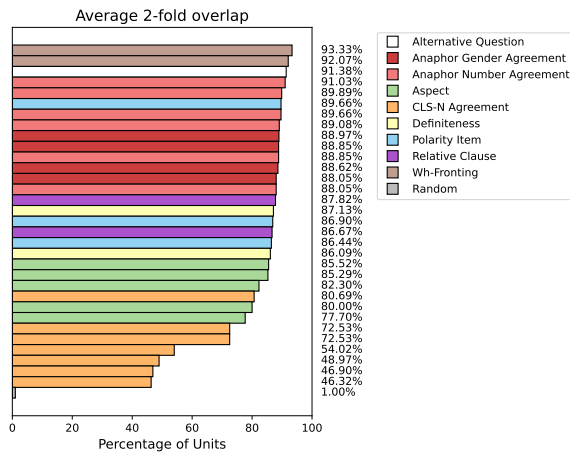


(a) Original SLING agreement cross-overlap results.

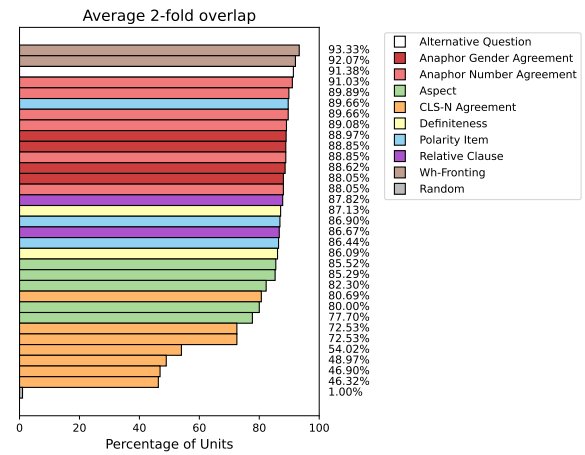


(b) Replicated SLING agreement cross-overlap results.

Figure 15: **Cross-lingual overlap of units localised for agreement phenomena.** Results are shown for original SLING (left) and replicated SLING (right), both using the Gemma model.



(a) Original RuBLiMP 2-fold results.



(b) Replicated SLING 2-fold results.

Figure 16: **Two-fold cross-validation, targeting the top 1% of units.** Results are shown for the original SLING analysis (left) and the replicated SLING analysis (right).

B Selected BLiMP-style Phenomena

An overview of the experimental materials used in our methodological extensions are provided below. The tables present the 12 paradigms selected for evaluation, alongside representative acceptable and unacceptable examples, and summarise the models included in our experiments.

Phenomenon	UID	Acceptable example	Unacceptable example
Anaphor Agreement	anaphor_gender_agreement	Katherine cannot help herself .	Katherine cannot help himself .
Argument Structure	animate_subject_passive	Amanda was respected by some waitresses .	Amanda was respected by some picture .
Binding	principle_A_c_command	A lot of actresses healed themselves .	A lot of actresses healed herself .
Control/Raising	existential_there_object_raising	William has declared there to be no guests getting fired.	William has obliged there to be no guests getting fired.
Determiner-Noun Agr.	determiner_noun_agreement_1	Craig explored that grocery store .	Craig explored that grocery stores .
Ellipsis	ellipsis_n_bar_1	Brad passed one big museum and Eva passed several.	Brad passed one museum and Eva passed several big.
Filler-Gap	wh_questions_object_gap	Joel discovered the vase that Patricia might take.	Joel discovered what Patricia might take the vase.
Irregular Forms	irregular_past_participle_adjectives	The hidden cactus was out in the open.	The hid cactus was out in the open.
Island Effects	adjunct_island	Who has Colleen aggravated before kissing Judy?	Who has Colleen aggravated Judy before kissing?
Quantifiers	existential_there_quantifiers_1	There aren't many lights darkening.	There aren't all lights darkening.
Subject-Verb Agr.	distractor_agreement_relational_noun	A sketch of lights doesn't appear.	A sketch of lights don't appear.
	regular_plural_subject_verb_agreement_2	The guests aren't cooking.	The guest aren't cooking.

Table 1: Examples of the 12 paradigms selected from the BLiMP benchmark (Warstadt et al., 2020) and used in our methodological extensions.

Phenomenon	PID	Acceptable example	Unacceptable example
Word Formation	add_new_suffix	Priekhala staren'kaya , malen'kaya, khuden'kaya zhenshchina.	Priekhala starsken'kaya , malen'kaya, khuden'kaya zhenshchina.
Word Inflection	change_verb_conjugation	I nikomu uzhe ne dokazhesh' , chto ty – eto ty, – tak on dumal.	I nikomu uzhe ne dokazhish' , chto ty – eto ty, – tak on dumal.
Government	adposition_government	Vpervye kosmonavt spal v nevesomosti .	Vpervye kosmonavt spal v nevesomost'yu .
	nominalization_case	Pakhotnye ploshchadi podverzheny smyvu i vyduvaniyu vetrom .	Pakhotnye ploshchadi podverzheny smyvu i vyduvaniyu veter .
	verb_acc_object	My opisem nashu aksiomatizatsiyu putem opisaniya struktury formul.	My opisem nashu aksiomatizatsiya putem opisaniya struktury formul.
Anaphor Agreement	anaphor_agreement_gender	Tekhnika, kotoruyu on izobrel, poluchila nazvanie "skul'ptura sveta".	Tekhnika, kotoryj on izobrel, poluchila nazvanie "skul'ptura sveta".
Aspect	change_duration_aspect	Pri etom vopros avtorstva dolgo ostavalsya otkrytym.	Pri etom vopros avtorstva dolgo ostalsya otkrytym.
Floating Quantifier Agreement	floating_quantifier_agreement_case	Ego samogo uzhe malo kto priznaet avtoritetom.	Ego samomu uzhe malo kto priznaet avtoritetom.
Subject-Predicate Agreement	genitive_subj_predicate_agreement_person	Detey u Magnusa i Elizavety ne bylo .	Detey u Magnusa i Elizavety ne budu .
Negation	negative_concord	I konechno, nikto ne toropilsya vzyat' vinu na sebya.	I konechno, nikto toropilsya vzyat' vinu ne na sebya.
Noun Phrase Agreement	np_agreement_case	Zoloto bylo obnaruzheno v etom rajone v 1923 godu.	Zoloto bylo obnaruzheno v etogo rajone v 1923 godu.
Tense	tense_marker	Tonnel' na Sinopskoy naberezhnoy otkroyut na budushchey nedele.	Tonnel' na Sinopskoy naberezhnoy otkroyut na minuvshyey nedele.

Table 2: Examples of the 12 paradigms selected from the RuBLiMP benchmark (Taktasheva et al., 2024) and used in our methodological extensions.

C Attention Head results

C.1 BLiMP

The figures below present the full attention-head analysis for the 12 selected BLiMP paradigms across all five models. Although the individual heads are not directly comparable, recurring phenomenon-level relationships emerge in each model. In particular, Principle A c-command shows consistently elevated overlap with Anaphor Gender Agreement and Animate Subject Passive constructions. These patterns suggest that attention heads sensitive to related agreement and dependency-based computations may be reused across syntactic phenomena.

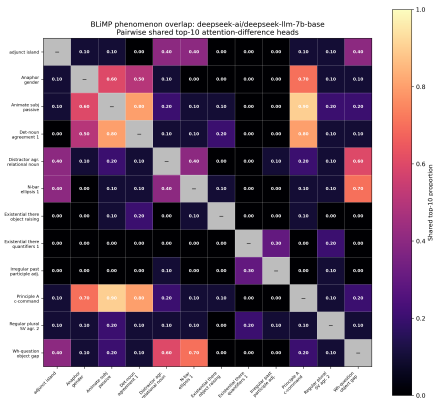


Figure 17: **Pairwise attention-head overlap across selected BLiMP phenomena** for DeepSeek-LLM-7B-Base. For each phenomenon, the ten heads with the largest grammatical–ungrammatical attention difference are selected. Each off-diagonal cell reports the proportion of these heads shared by a pair of phenomena; the diagonal is omitted.

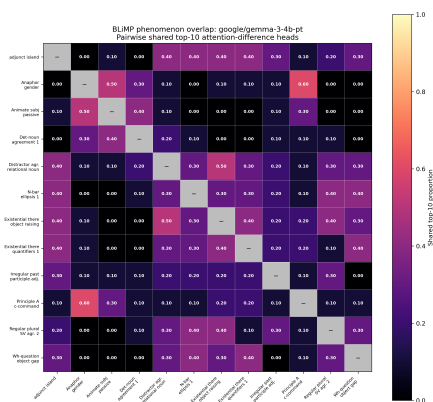


Figure 18: **Pairwise attention-head overlap across selected BLiMP phenomena** for Gemma-3 4B. Cells report pairwise overlap among the ten most syntax-responsive heads; the diagonal is omitted.

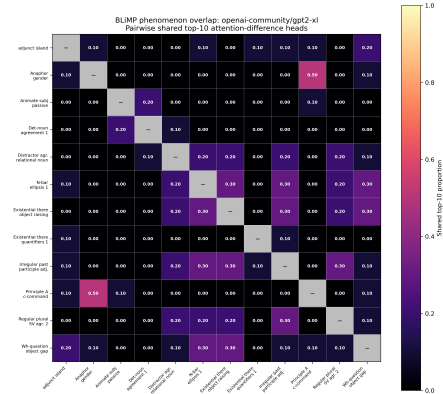


Figure 19: **Pairwise attention-head overlap across selected BLiMP phenomena** for GPT-2 XL. Cells report pairwise overlap among the ten most syntax-responsive heads; the diagonal is omitted.

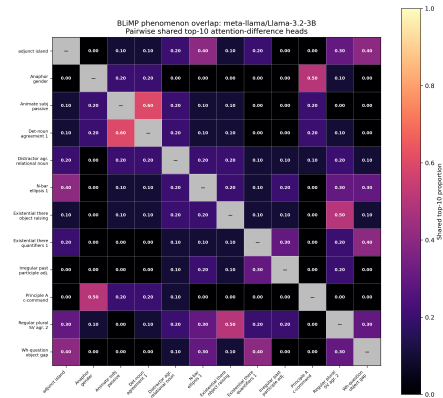


Figure 20: **Pairwise attention-head overlap across selected BLiMP phenomena** for Llama-3.2 3B. Cells report pairwise overlap among the ten most syntax-responsive heads; the diagonal is omitted.

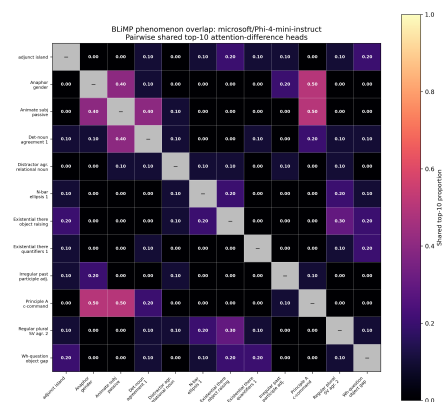


Figure 21: **Pairwise attention-head overlap across selected BLiMP phenomena** for Phi-4-mini-instruct. Cells report pairwise overlap among the ten most syntax-responsive heads; the diagonal is omitted.

C.2 RuBLiMP

We extend the analysis cross-linguistically to Russian using RuBLiMP, selecting 12 paradigms spanning a range of morphosyntactic phenomena. Gemma and Llama exhibit relatively dense overlap matrices, suggesting broad sharing of attention heads across phenomena involving agreement, case, tense and aspect. This may partly reflect Russian’s rich inflectional morphology, in which these features interact closely. Phi, however, shows a substantially sparser overlap structure, indicating that the extent and organisation of morphosyntactic sharing are model-dependent.

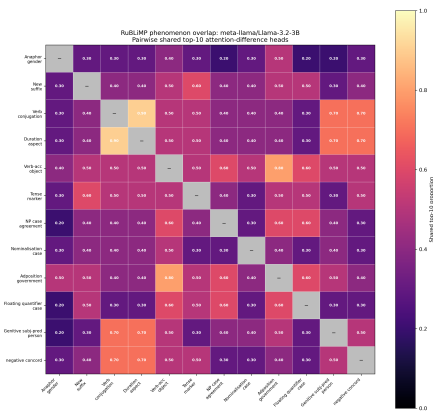


Figure 22: **Pairwise attention-head overlap across selected RuBLiMP phenomena** for Llama-3.2 3B. Cells report pairwise overlap among the ten most syntax-responsive heads; the diagonal is omitted.

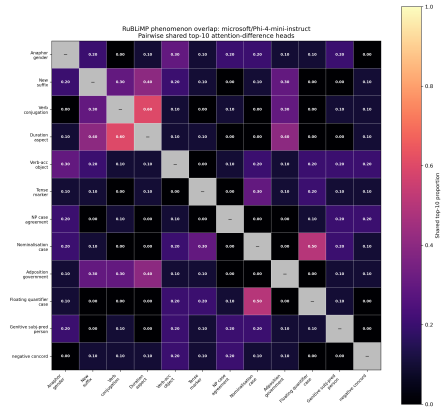


Figure 23: **Pairwise attention-head overlap across selected RuBLiMP phenomena** for Phi-4-mini-instruct. Cells report pairwise overlap among the ten most syntax-responsive heads; the diagonal is omitted.

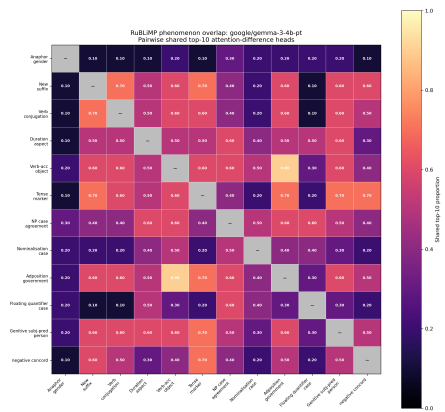


Figure 24: **Pairwise attention-head overlap across selected RuBLiMP phenomena** for Gemma-3 4B. Cells report pairwise overlap among the ten most syntax-responsive heads; the diagonal is omitted.

D DBM and DAS Localisation Results

We extend the localisation methodology to DBM and DAS, explicated in Section 2.1. We carry out the original 1% ablation experiments on a subset of BLiMP for 5 models, and cross-validation and cross-overlap on Gemma, Llama, and Phi. Our findings match those of Kryvosheieva et al. (2026), however we additionally show differences in ablation effects across models and methodologies.

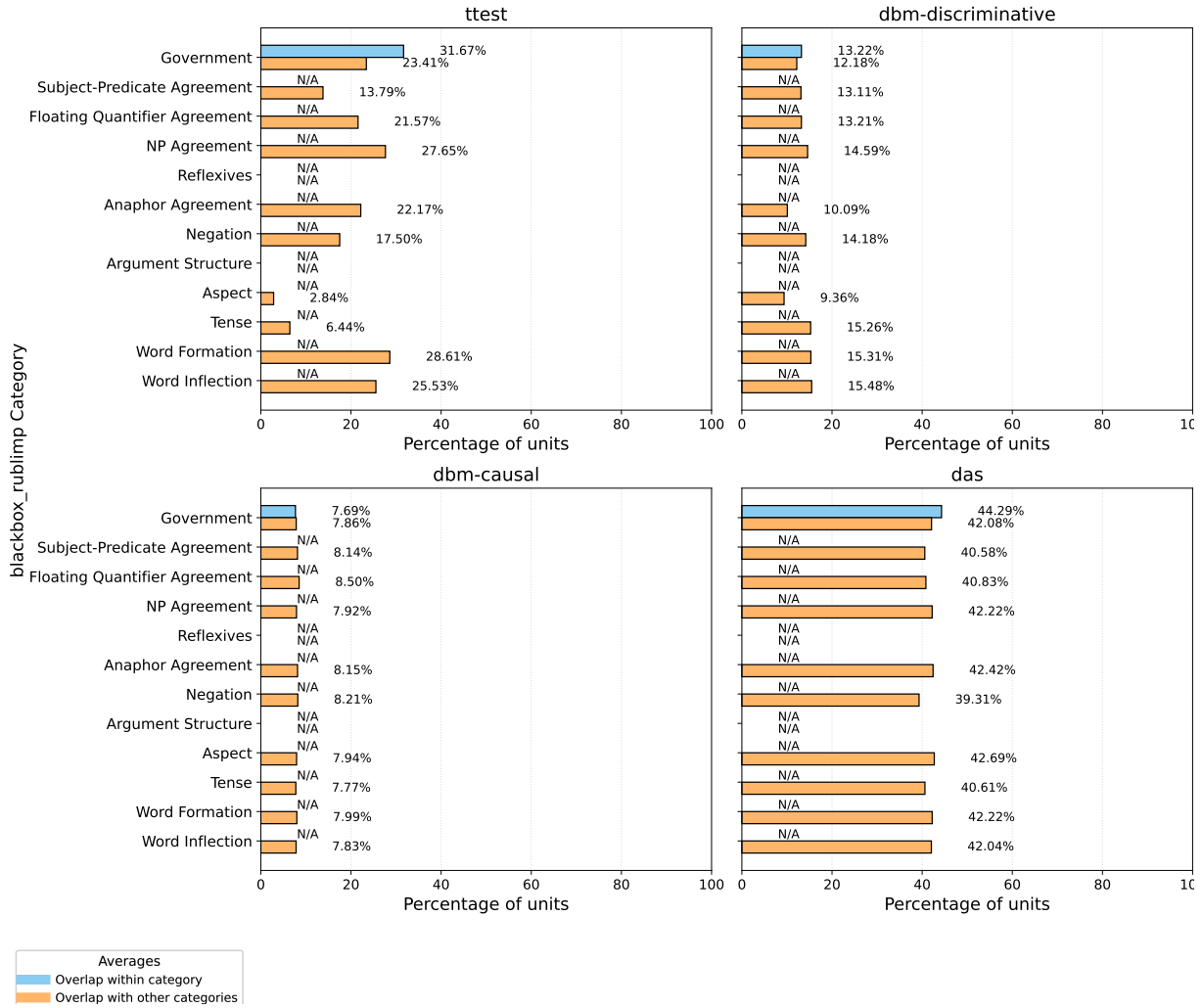


Figure 25: Cross-overlap effects on a subset of RuBLiMP aggregated across Gemma-3-4B, Llama-3.2-3B, and Phi-4 Mini. The original methodology is illustrated in the ‘ttest’ method, and are compared with the results of the alternative functional localisation methods we use.

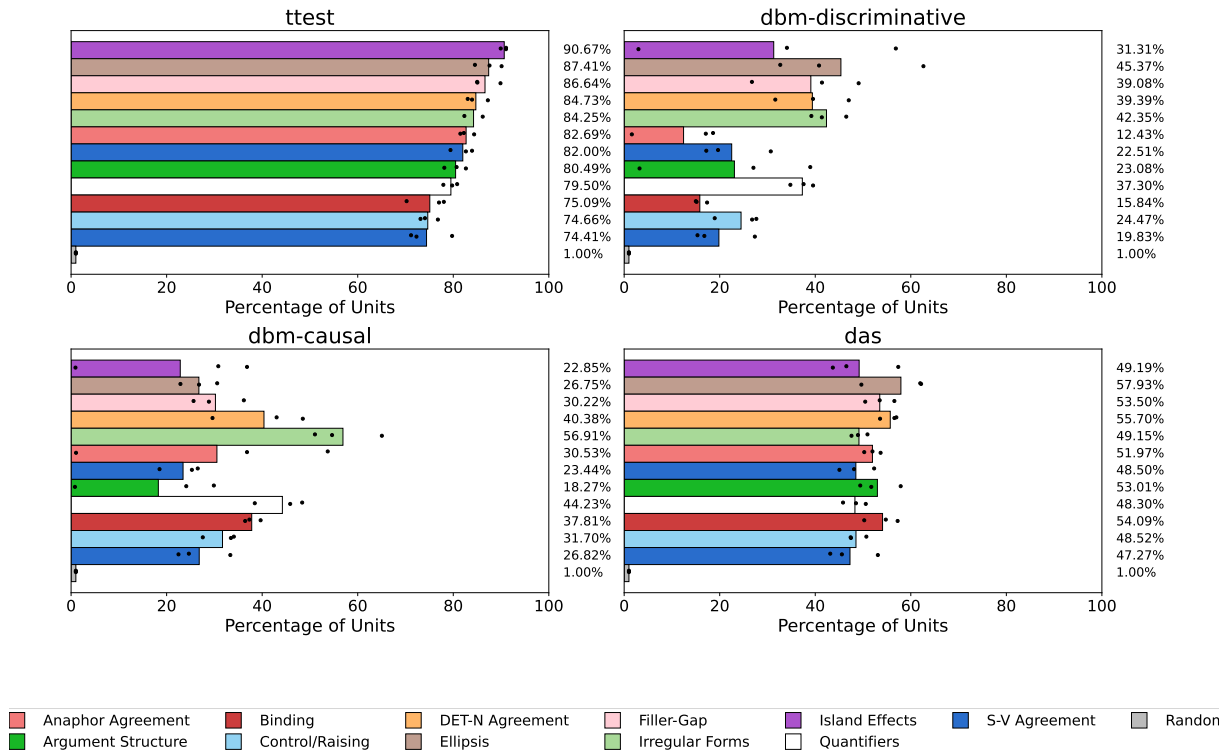


Figure 26: Cross-validation effects on a subset of BLiMP aggregated across Gemma-3-4B, Llama-3.2-3B, and Phi-4 Mini. The original methodology of Kryvosheieva et al. (2026) is illustrated in the ‘ttest’ method, and are compared with the results of the alternative functional localisation methods we use.

Model	Original	DBM-discriminative	DBM-causal	DAS
Gemma-3-4B-PT	-17.44%	-2.79%	-42.75%	-27.60%
Phi-4-mini-instruct	0.82%	4.50%	-62.42%	-21.90%
Llama-3.2-3B	-4.60%	-0.09%	-58.89%	-7.46%
GPT2-XL	-2.51%	-1.74%	-61.78%	-13.17%
DeepSeek-LLM-7B-Base	-1.71%	-4.52%	-51.50%	-16.88%

Table 3: 1% ablation effects on a subset of BLiMP aggregated across Gemma-3-4B, Llama-3.2-3B, Phi-4 Mini, GPT2-XL, and DeepSeek-7B. Δ between performance change following random ablation and performance change following ablation of localised units is illustrated. The original methodology of Kryvosheieva et al. (2026) is compared with the results of the alternative functional localisation methods we use. Differences in ablation effect are shown across different models, with high variance across methodologies, without a clear pattern per model.

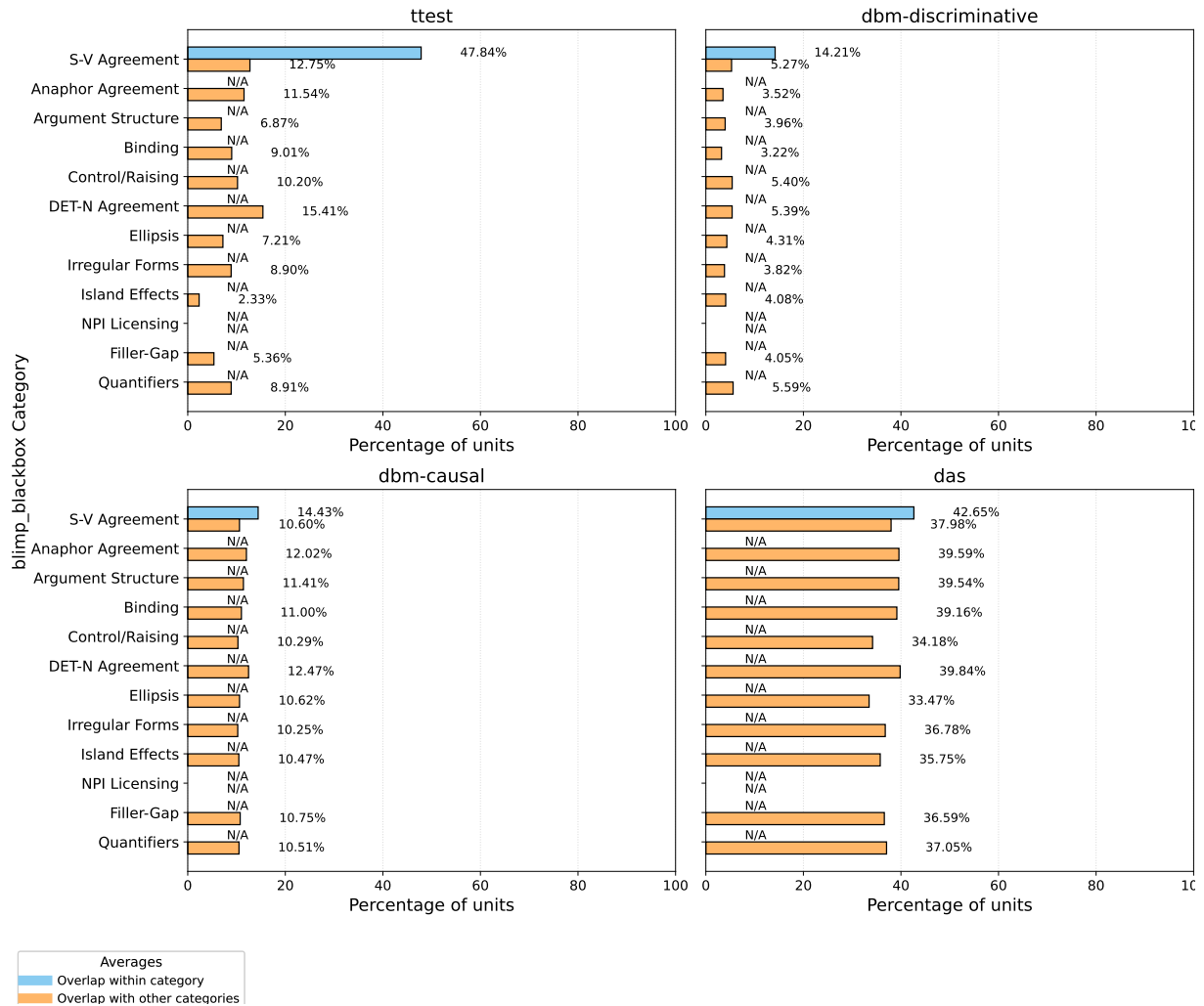


Figure 27: Cross-overlap effects on a subset of BLiMP aggregated across Gemma-3-4B, Llama-3.2-3B, and Phi-4 Mini. The original methodology of Kryvosheieva et al. (2026) is illustrated in the ‘ttest’ method, and are compared with the results of the alternative functional localisation methods we use. N/A here denotes instances where no within-category overlap data was available.

Suite	Original	DBM-discriminative	DBM-causal	DAS
principle_A_c_command	-11.34%	-1.94%	-47.16%	-11.78%
regular_plural_subject_verb_agreement_2	-8.26%	-4.22%	-29.14%	-11.00%
adjunct_island	-8.88%	-0.36%	-55.70%	-29.58%
ellipsis_n_bar_1	-5.70%	-6.94%	-57.52%	-24.92%
determiner_noun_agreement_1	-8.46%	0.16%	-31.52%	-18.94%
animate_subject_passive	-7.04%	-0.32%	-41.56%	-11.34%
distractor_agreement_relational_noun	-6.28%	1.20%	-51.68%	-21.28%
anaphor_gender_agreement	-4.60%	-6.78%	-64.62%	-12.98%
existential_there_object_raising	-5.94%	0.44%	-58.28%	-17.50%
existential_there_quantifiers_1	-3.18%	-1.32%	-73.42%	-7.52%
wh_questions_object_gap	-0.88%	-0.58%	-66.34%	-19.30%
irregular_past_participle_adjectives	9.50%	9.52%	-88.70%	-22.66%

Table 4: 1% ablation effects on a subset of BLiMP aggregated across Gemma-3-4B, Llama-3.2-3B, Phi-4 Mini, GPT2-XL, and DeepSeek-7B. Δ between performance change following random ablation and performance change following ablation of localised units is illustrated. The original methodology of Kryvosheieva et al. (2026) is compared with the results of the alternative functional localisation methods we use. Differences in ablation effect are shown across different phenomena, with no clear cross-method pattern as to which phenomena show the strongest localisation effects.

Model	Original	DBM-discriminative	DBM-causal	DAS
Gemma-3-4B-PT	-30.26%	-2.39%	-20.77%	-36.90%
Phi-4-mini-instruct	-17.13%	-5.25%	-19.56%	-31.48%
Llama-3.2-3B	-2.57%	-0.37%	-22.13%	-11.70%

Table 5: 1% ablation effects on a subset of RuBLiMP aggregated across Gemma-3-4B, Llama-3.2-3B, and Phi-4 Mini. Δ between performance change following random ablation and performance change following ablation of localised units is illustrated. The original methodology of Kryvosheieva et al. (2026) is compared with the results of the alternative functional localisation methods we use. Differences in ablation effect are shown across different models, with high variance across methodologies, without a clear pattern per model.

Suite	Original	DBM-discriminative	DBM-causal	DAS
np_agreement_case	-28.47%	-0.07%	-5.00%	-32.20%
anaphor_agreement_gender	-27.73%	-4.70%	-30.83%	-22.03%
change_verb_conjugation	-24.00%	-5.87%	-21.80%	-31.47%
add_new_suffix	-17.97%	1.60%	-7.60%	-28.33%
verb_acc_object	-19.27%	-0.67%	-8.67%	-25.13%
nominalization_case	-20.47%	-4.23%	-4.97%	-33.17%
floating_quantifier_agreement_case	-19.53%	-0.03%	-21.27%	-22.27%
adposition_government	-16.63%	-12.50%	-18.17%	-26.97%
negative_concord	-12.23%	0.17%	-16.07%	-30.07%
genitive_subj_predicate_agreement_person	-6.80%	-1.03%	-46.87%	-20.63%
tense_marker	-5.47%	-3.97%	-25.87%	-21.57%
change_duration_aspect	-1.27%	-0.73%	-42.73%	-26.47%

Table 6: 1% ablation effects on a subset of RuBLiMP aggregated across Gemma-3-4B, Llama-3.2-3B, Phi-4 Mini. Δ between performance change following random ablation and performance change following ablation of localised units is illustrated. The original methodology of [Kryvosheieva et al. \(2026\)](#) is compared with the results of the alternative functional localisation methods we use. Differences in ablation effect are shown across different phenomena, with no clear cross-method pattern as to which phenomena show the strongest localisation effects.

E BLiMP-NL

We provide the full BLiMP-NL results underlying the analysis in Section 3.2.1. The figures report cross-phenomenon overlap across all syntactic categories, 2-fold consistency for the 84 individual paradigms, and overlap within and across selected A'-dependency phenomena. These analyses allow the variation in phenomenon-specificity and localisation stability across the Dutch benchmark to be examined in greater detail.

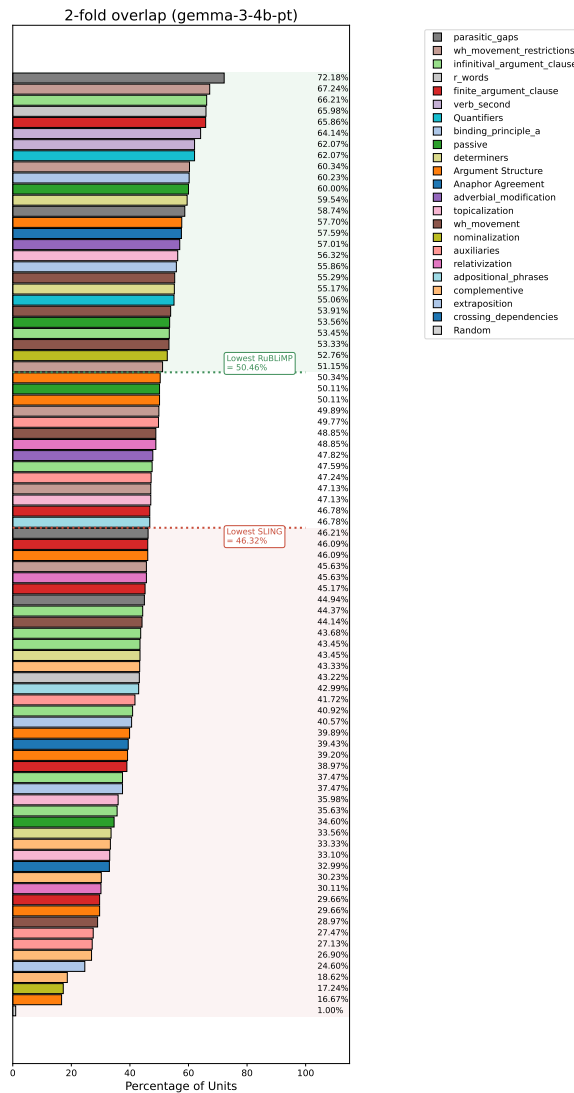


Figure 28: Two-fold localisation consistency for the 84 BLiMP-NL paradigms in Gemma-3 4B. Each bar shows the percentage overlap between top-1% unit sets identified using two independent halves of a paradigm. Paradigms are ordered from highest to lowest overlap and coloured by syntactic category; the grey bar marks the analytical random expectation of 1%. The dotted lines indicate the lowest value in RuBLiMP and SLING.

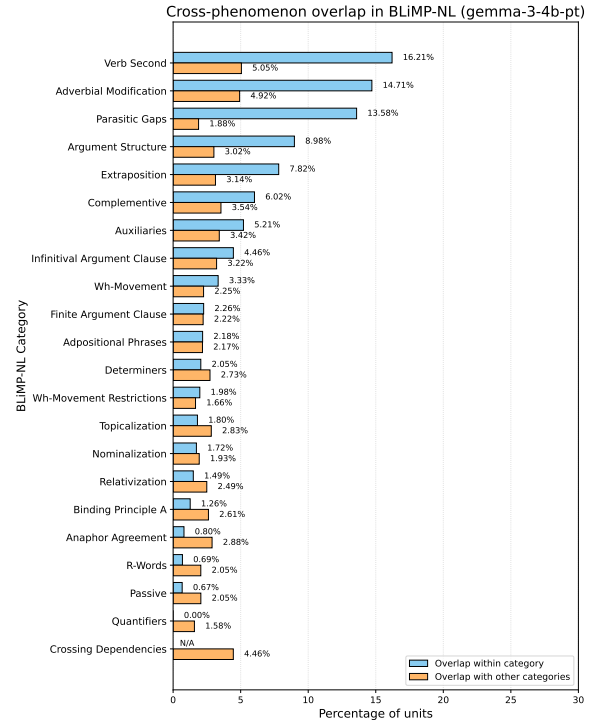


Figure 29: Full Within- and cross-category unit overlap in BLiMP-NL for Gemma-3 4B. For each syntactic category, blue bars show the mean pairwise overlap between top-1% unit sets localised for paradigms within that category, while orange bars show the mean overlap with paradigms from other categories. N/A indicates that no within-category paradigm pair was available.

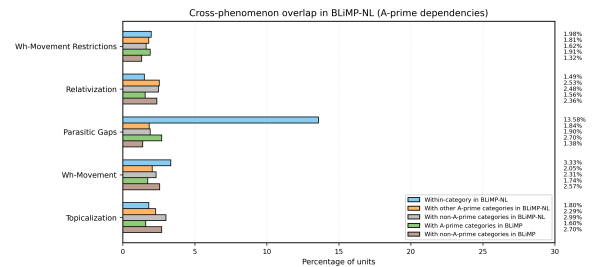


Figure 30: Cross-phenomenon overlap for selected A'-dependency categories in BLiMP-NL. For each category, bars compare overlap within the category, with other A' categories in BLiMP-NL, with non-A' categories in BLiMP-NL, with A' categories in English BLiMP, and with non-A' categories in English BLiMP. Values report the percentage of shared top-1% localised units in Gemma-3 4B.

F Turkish

We present the supplementary TurBLiMP results for Gemma-3 4B. The figures report 2-fold cross validation and cross-phenomenon overlap centred on Subject Agreement, alongside the full pairwise overlap matrix across all TurBLiMP phenomena. Together, these analyses provide a more detailed view of the extent to which agreement-related constructions recruit shared units and how these patterns relate to broader morphosyntactic overlap across the dataset.

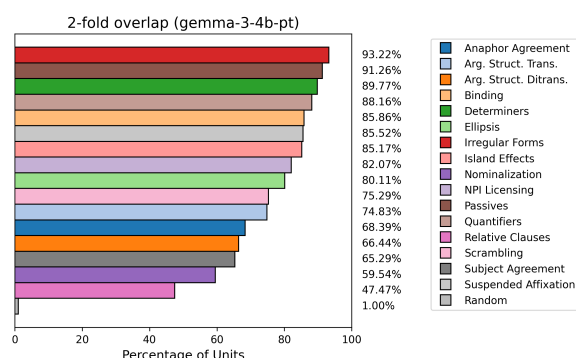


Figure 31: Two-fold localisation consistency across TurBLiMP phenomena for Gemma-3 4B. Each bar shows the percentage overlap between the top-1% unit sets identified using two independent halves of the examples for a phenomenon. The grey bar indicates the analytical random expectation of 1%.

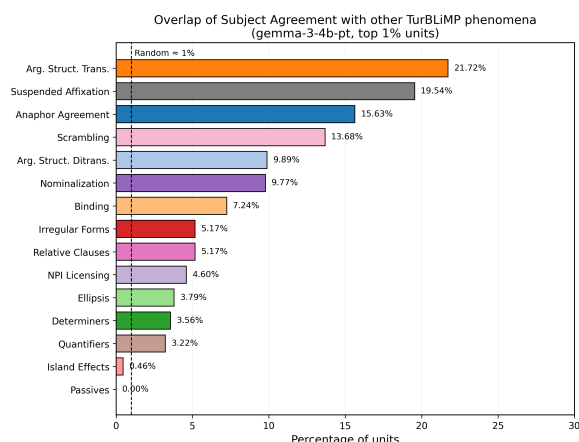


Figure 32: **Cross-phenomenon overlap** between Subject Agreement and the remaining TurBLiMP phenomena for Gemma-3 4B.

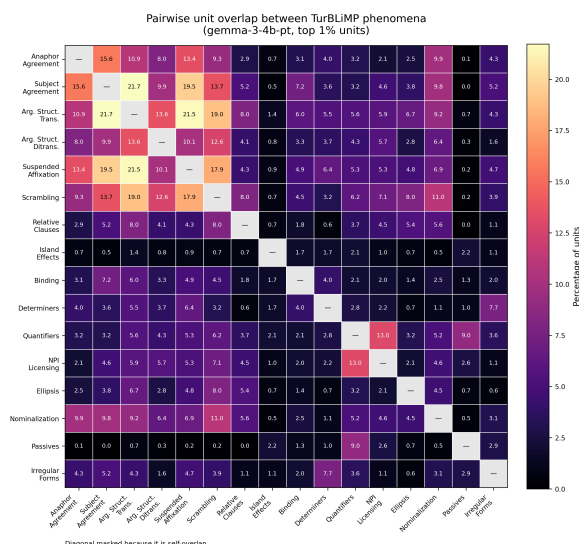


Figure 33: Pairwise unit overlap among TurBLiMP phenomena for Gemma-3 4B. Each off-diagonal cell reports the percentage of top-1% localised units shared by a pair of phenomena. The diagonal, corresponding to self-overlap, is omitted; the analytical random expectation is 1%.

G CLAMS: Russian

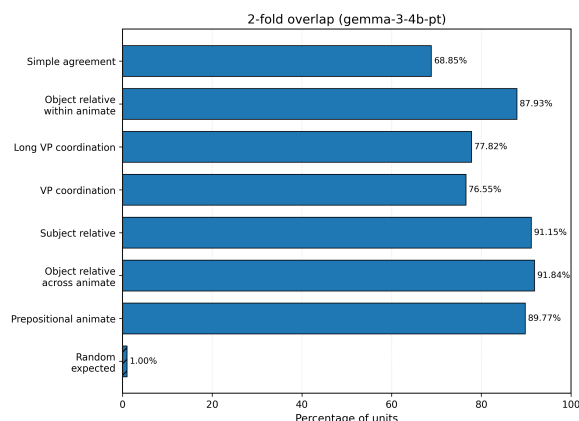


Figure 34: Two-fold localisation consistency across the CLAMS-Russian subject-verb agreement paradigms for Gemma-3 4B. Each bar shows the percentage overlap between top-1% unit sets identified using two independent halves of a paradigm. The grey bar marks the analytical random expectation of 1%.

Figure 34 presents the 2-fold overlap for each subject-verb agreement paradigm in CLAMS-Russian. The overlap between units localised using two independent halves of each paradigm was consistently high, ranging from 68.85% for simple agreement to 91.84% for object-relative agreement across animate nouns. All paradigms substantially exceeded the random expectation of 1%, indicating that unit localisation was stable within paradigms

and was not strongly dependent on the particular subset of examples used.

H Unit Localisation Methodologies

H.1 Experimental Model Selection

Model	Attribution	Param.	L	H	Lang.
GPT2-XL	Radford et al. (2019)	1.5B	48	1600	1
Llama-3.2-3B	Meta (2024)	3.2B	28	3072	8
Phi-4-mini-instruct	Microsoft (2025)	3.8B	32	3072	23
Gemma-3-4B-PT	DeepMind (2025)	4.3B	34	2560	140+
DeepSeek-LLM-7B-Base	DeepSeek (2024)	6.9B	30	4096	2

Table 7: Models used in our experiments **attention heads, DBM, and DAS**, sorted according to parameter count (Param.). For each model, we additionally report the number of layers (L), hidden size (H), and the number of languages included during training (Lang.)

H.2 Methodology: Differential Binary Masking: (DBM)

To localize the specific network units responsible for syntactic phenomena, we replace the univariate Welch t-test baseline (Kryvosheieva et al., 2026) with Differential Binary Masking (DBM). Rather than scoring units independently, DBM learns a sparse binary mask over units end-to-end via gradient descent. To overcome the non-differentiability of discrete binary masks, we use the Hard-Concrete distribution (Louizos et al., 2018), which acts as a differentiable relaxation of an L_0 penalty.

For each layer l , the model learns a set of continuous parameters α_l . During training, the model samples differentiable gates $z \in [0, 1]$ using the reparameterisation trick, allowing us to exact a continuous L_0 penalty. We use two distinct objectives for optimizing these gates: a computationally efficient *discriminative* proxy, and a direct *causal* intervention.

H.2.1 The Discriminative Objective

The discriminative objective identifies units that together separate the phenomenon condition from a control condition. Operating entirely on pre-computed representations, this approach serves as an efficient multivariate alternative to the baseline t-test.

Following the intuition of the baseline, we operate on the absolute, standardized activations of the units, $|X|$. This ensures the classifier captures units that signal importance via increased variance (magnitude of activation) rather than strictly shifting their mean. To prevent severe overparameterisation given the limited number of benchmark sentences,

we train independent gated linear classifiers per layer rather than a single joint classifier.

The loss for layer l combines Binary Cross-Entropy (BCE) with an expected L_0 sparsity penalty and L_2 weight decay:

$$\mathcal{L}_l = \text{BCE}(y, \sigma(|X_l| \cdot z_l)W_l + b_l) + \lambda_l E[L_0(z_l)] + \lambda_{\text{decay}} \|W_l\|_2^2 \quad (1)$$

H.2.2 The Causal Objective

While the discriminative objective relies on a correlational proxy, the causal objective directly optimizes the mask against the actual intervention (ablation) used during final evaluation.

We attach custom forward hooks to the frozen language model. During the forward pass, the sampled gates z act as physical soft-ablations on the hidden states h :

$$h_{\text{ablated}} = h \odot (1 - z) \quad (2)$$

Initialization Strategy: In standard L_0 pruning, $z = 1$ denotes retaining a weight, so gates are initialized open to begin training from an intact network. In our causal formulation, $z = 1$ denotes *ablating* a unit. If initialized using standard parameters, the model would begin training with $\approx 97\%$ of its units ablated, severely degrading the gradient signal. Instead, we analytically compute the initial parameters $\log \alpha_0$ such that the initial ablation probability is close to the target sparsity threshold, allowing the mask to grow toward harmful units rather than shrinking from total network collapse.

The causal loss directly minimizes the model’s post-ablation grammaticality margin (the log-probability difference between positive and negative sentences) alongside the sparsity penalty:

$$\mathcal{L} = \mathbb{E} \left[\log P(x_{\text{pos}} | h_{\text{ablated}}) - \log P(x_{\text{neg}} | h_{\text{ablated}}) \right] + \sum_l \lambda_l \mathbb{E}[L_0(z_l)] \quad (3)$$

H.2.3 Dynamic Sparsity Control

Because different layers exhibit varying degrees of recoverable signal, applying a single, global sparsity penalty λ often leads to under-masking in critical layers and over-masking in irrelevant ones. We implement a layer-wise dual-ascent optimization strategy. Every 20 steps, each layer’s penalty weight λ_l is independently updated based on its

distance from the global target sparsity fraction p_{target} :

$$\lambda_l^{(t+1)} = \lambda_l^{(t)} \exp\left(\eta \cdot (p_l^{(t)} - p_{\text{target}})\right) \quad (4)$$

This ensures all layers converge robustly to the desired sparsity constraint without requiring manual hyperparameter tuning for the penalty bounds. Finally, for both objectives, the continuous importance scores $\log \alpha$ are converted into binary form to extract the exact top- k units for downstream evaluation.

H.3 Methodology: Distributed Alignment Search (DAS)

While the univariate t-test and Differential Binary Masking (DBM) methods evaluate units independently, they rely on the assumption that causally relevant computation aligns with the network’s standard basis (e.g. individual neurons). To relax this assumption, we implement Distributed Alignment Search (DAS) (Geiger et al., 2024), which searches for a distributed, dense causal representation. Instead of selecting discrete units, DAS learns an orthogonal rotation of a layer’s hidden space and applies an interchange intervention to test whether a small, fixed-width subspace of the rotated coordinates carries the phenomenon’s causal effect.

H.3.1 Low-Rank Subspace Intervention

The original formulation of DAS (Geiger et al., 2024) parameterises a full $D \times D$ orthogonal rotation matrix R (where D is the hidden dimension), slices the first k columns to form the aligned subspace, and applies the inverse rotation. In practice, representing R requires a matrix-exponential of a skew-symmetric generator, resulting in an $\mathcal{O}(D^3)$ computational cost per forward pass. For modern language models with large hidden dimensions, this becomes a severe bottleneck.

To address this, we adopt the low-rank subspace parameterisation proposed by Huang et al. (2024). Rather than a full rotation, this directly parameterises the k orthogonal vectors spanning the aligned subspace using a matrix $W \in \mathbb{R}^{D \times k}$ with orthonormal columns (i.e., $W^T W = I_k$). The interchange intervention II , which patches the hidden states of a “base” (ungrammatical) sentence with those of a “source” (grammatical) sentence, is implemented as a rank- k projection:

$$II(h_{\text{base}}, h_{\text{source}}) = (I - WW^T)h_{\text{base}} + WW^T h_{\text{source}} \quad (5)$$

Equivalently, this is computed without instantiating the full $D \times D$ projection matrix via:

$$h_{\text{int}} = h_{\text{base}} + W(W^T(h_{\text{source}} - h_{\text{base}})) \quad (6)$$

This reduces the computational cost of the forward hook to $\mathcal{O}(D \cdot k)$ while remaining algebraically identical to the original full-rotation formulation. Because the intervention relies on precise position-wise alignment, we restrict this analysis strictly to sentence pairs that tokenize to equal lengths.

H.3.2 Optimization Objective

DAS fits one orthogonal subspace independently per layer. During training, the base model’s weights remain frozen. For each sequence pair, we capture the unperturbed hidden states of the grammatical sentence (h_{source}). In a subsequent forward pass on the ungrammatical sentence (h_{base}), we apply the interchange intervention using our learned subspace W .

The parameter W is optimized to maximize the teacher-forced log-probability of the ungrammatical sentence’s tokens under the intervention. By maximizing this grammaticality margin, DAS identifies the subspace that maximally “rescues” the model’s likelihood, confirming its causal contribution to the phenomenon – the trained result is a quantity to be reported, rather than a binary confirmation.

H.3.3 Bridging DAS to a Per-Unit Mask

Standard evaluation pipelines require a binary mask indicating unit importance, which fundamentally differs from the continuous subspace W learned by DAS. To translate the learned subspace back to the standard basis for direct comparison with DBM and t-test baselines, we compute a participation score for each individual unit. The importance of the i -th unit in layer l is calculated as the L_1 norm of its corresponding row in W (its loading across the k basis vectors):

$$\text{score}_{l,i} = \|W_{l,i,:}\|_1 = \sum_{j=1}^k |W_{l,i,j}| \quad (7)$$

These continuous scores are then hard-binarized by selecting the exact top- k values globally across the network, yielding a binary mask directly comparable to those produced by standard unit-level localization methods.