Linguistic Minimal Pairs Elicit Linguistic Similarity in Large Language Models

Xinyu Zhou[1]* **Delong Chen**[2]*
Samuel Cahyawijaya[2,4] **Xufeng Duan**[1] **Zhenguang G. Cai**[1,3]
[1]Department of Linguistics and Modern Languages, CUHK
[2]Department of Electronic and Computer Engineering, HKUST
[3]Brain and Mind Institute, CUHK
[4]Cohere

Abstract

We introduce a novel analysis that leverages linguistic minimal pairs to probe the internal linguistic representations of Large Language Models (LLMs). By measuring the similarity between LLM activation differences across minimal pairs, we quantify the *linguistic similarity* and gain insight into the linguistic knowledge captured by LLMs. Our large-scale experiments, spanning 100+ LLMs and 150k minimal pairs in three languages, reveal properties of linguistic similarity from four key aspects: consistency across LLMs, relation to theoretical categorizations, dependency to semantic context, and cross-lingual alignment of relevant phenomena. Our findings suggest that 1) linguistic similarity is significantly influenced by training data exposure, leading to higher cross-LLM agreement in higher-resource languages. 2) Linguistic similarity strongly aligns with fine-grained theoretical linguistic categories but weakly with broader ones. 3) Linguistic similarity shows a weak correlation with semantic similarity, showing its context-dependent nature. 4) LLMs exhibit limited cross-lingual alignment in their understanding of relevant linguistic phenomena. This work demonstrates the potential of minimal pairs as a window into the neural representations of language in LLMs, shedding light on the relationship between LLMs and linguistic theory.[2]

1 Introduction

The categorization of linguistic phenomena[3] based on their relevance has been a long-standing endeavor, dating back to Aristotle Aristotle [350 BC]. This has led to the widely accepted theoretical linguistic consensus of a hierarchical categorization of language structure encompassing syntax, semantics, morphology, etc., which provides a structured way to understand the intricate nature of language, and allows linguists to investigate the interrelationships and commonalities among these linguistic domains Comorovski [2013], Li [2004].

Alongside the theoretical discussions of linguistic phenomena, a growing body of research on *quantitative measurement of similarities* based on statistical modeling on large-scale corpora has been observed in computational linguistics. Examples include lexical similarity Holman et al. [2011], syntactic similarity Boghrati et al. [2018], Schoot et al. [2016], semantic similarity Pennington et al.

[*] Joint first authors. Xinyu Zhou is now affiliated with Université Paris Cité and Sorbonne Université.

[2] For the detailed and complete version of this paper, please refer to the preprint on arXiv: https://arxiv.org/abs/2409.12435.

[3] Linguistic phenomena refer to observable patterns or features in language use. For example, subject-verb agreement is a linguistic phenomenon where verbs must agree with subjects in number and person. An example would be: *"The dog barks"* (correct) instead of *"*The dog bark"* (incorrect).

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

[2014], Reimers and Gurevych [2019], among others. These examples showcase the possibilities of understanding the nature of language through purely statistical methodologies. However, there has been limited research on quantitatively measuring the relationships between different linguistic phenomena. Given that language is a complex system composed of numerous interrelated linguistic phenomena, addressing this gap could lead to a more comprehensive understanding of language structure and its underlying mechanisms.

In this work, we aim to uncover and analyze the internal linguistic knowledge of Large Language Models (LLMs) when presented with a wide range of linguistic phenomena. LLMs are large-scale unsupervised language learners without any prior linguistic knowledge, and have demonstrated human-level language capability, as evidenced by their leading performance on language understanding benchmarks and impressive language generation fluency Zhao et al. [2023], Bang et al. [2023]. More specifically, we are interested in how LLMs represent different linguistic phenomena, and whether linguistically similar phenomena are represented similarly in LLMs.

To elicit such representations, we examine the activations in LLMs in response to linguistic minimal pairs Warstadt et al. [2020]. As shown in Fig. 1, these pairs consist of sentences that differ only in a word/phrase, with one being grammatical and the other ungrammatical. Since minimal pairs differ *only* in one particular linguistic phenomenon, information about other aspects (such as topic and semantic meaning) will be canceled out through subtraction. We interpret the remaining differences as the LLMs' internal representation of a specific linguistic phenomenon. By calculating the similarity between multiple such representations, we derive a measure of *linguistic similarity* between linguistic minimal pairs.

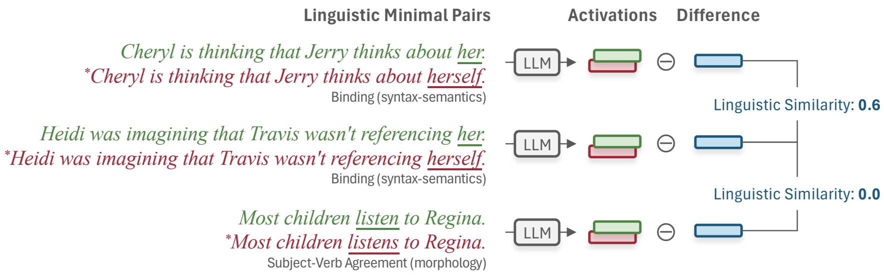
Linguistic Minimal Pairs
Activations
Difference
Cheryl is thinking that Jerry thinks about her.
*Cheryl is thinking that Jerry thinks about herself.
Binding (syntax-semantics)
LLM
Linguistic Similarity: 0.6
Heidi was imagining that Travis wasn't referencing her.
*Heidi was imagining that Travis wasn't referencing herself.
Binding (syntax-semantics)
LLM
Linguistic Similarity: 0.0
Most children listen to Regina.
*Most children listens to Regina.
Subject-Verb Agreement (morphology)
LLM

Figure 1: **The process of measuring linguistic similarity in an LLM**. We extract LLM activations for sentences in linguistic minimal pairs and compute their differences. Since the sentences differ solely in a specific linguistic phenomenon, the resulting difference only contains information about that phenomenon. We then measure the similarity between these activation differences, which we refer to as *linguistic similarity*.

We then conduct an extensive analysis of linguistic similarity in LLMs. Our experiment encompasses 100+ LLMs of varying scales and pretraining corpora, utilizing 150,000 linguistic minimal pairs across 3 different languages. We report our observations correspond to the following key questions:

1) How consistent is linguistic similarity across different LLMs? LLMs have the highest alignment of linguistic similarity in English, which is the most widely used language for LLM pertaining, while the alignments are comparatively weaker in Chinese and Russian. We further visualized the relationships among these LLMs with UMAP McInnes et al. [2018]. On Chinese samples, we observed a distinct clustering pattern: bilingual and multilingual LLMs formed one cluster, while English-only models formed another. The above results suggest that the language distribution in the training data influences the linguistic similarity in LLMs.

2) Does linguistic similarity align with theoretical linguistic categorizations? We compared linguistic similarity across three levels of theoretical linguistic categorizations. Our analysis revealed that fine-grained classifications exhibit significantly higher intra-class similarities compared to inter-class similarities. However, this disparity diminishes considerably at higher categorization levels. Meanwhile, we can also observe some highly correlated phenomena pairs that are not classified to the same theoretical categorization.

3) To what extent does linguistic similarity correlate with semantic similarity? We showed a weak correlation between semantic similarity and linguistic similarity, despite many existing samples with low linguistic similarity and high semantic similarity, and conversely, high in linguistic and low in semantic. The weak correlation indicates that linguistic similarity in LLMs has a *context-dependent* nature.

4) Whether relevant phenomena in different languages enjoy higher linguistic similarities? We compare the linguistic similarity of the shared three linguistic phenomena in English and Chinese. Our UMAP visualization revealed that while English phenomena are clustered within a shared region, they are "attracted" by their relevant phenomena in Chinese.

We hope this paper sparks new exploration into LLMs' internal linguistic representations, uncovering deeper insights into their inner workings and potentially informing linguistic theory. To facilitate future research, the activation differences of the 100+ LLMs, pre-computed sample-level linguistic similarities, and all the codes are made publicly available at `https://github.com/ChenDelong1999/Linguistic-Similarity`.

2 Measuring Linguistic Similarity in Large Language Models

2.1 Definition

Let x^+ denote a grammatically correct natural language sentence, and x^- represent an ungrammatical sentence derived from x^+ with a minimal modification affecting a specific linguistic phenomenon. The pair $\langle x^+, x^- \rangle$ is referred to as a *linguistic minimal pair*. Let f_{LLM} denote an LLM that takes sentences as input and generates corresponding hidden activations, *i.e.,* $z^+ = f_{\text{LLM}}(x^+)$ and $z^- = f_{\text{LLM}}(x^-)$, where $z^+, z^- \in \mathbb{R}^n$ and n is the dimensionality of the hidden representations.

We compute the difference between the hidden activations: $\Delta z = z^+ - z^-$. While z^+ and z^- individually encode rich information about the input sentences, including both semantic and linguistic properties, their difference Δz primarily captures the representation of the specific linguistic phenomenon that distinguishes the minimal pair, as other aspects of the sentences are guaranteed to be identical and thus will be canceled out. Formally, given two linguistic minimal pairs $\langle x_1^+, x_1^- \rangle$ and $\langle x_2^+, x_2^- \rangle$, we define their linguistic similarity as $\texttt{sim}(\Delta z_1, \Delta z_2)$, where $\texttt{sim}(\cdot)$ is a similarity metric and we used cosine similarity in this work.

2.2 Implementation

Data. We utilized linguistic minimal pairs from three existing datasets: BLiMP Warstadt et al. [2020], SLING Song et al. [2022], and RuBLiMP Taktasheva et al. [2024], which consist of 67, 38, and 45 linguistic phenomena in English, Chinese, and Russian, respectively. Each linguistic phenomenon is associated with 1,000 corresponding linguistic minimal pairs, yielding a total of 150,000 pairs.

LLMs. Each sentence was input into various LLMs without any prompts, and hidden states were extracted from five evenly sampled layers. We specifically extracted the activations of the last-but-two token, as previous tokens remain visible, while the final tokens correspond to the `<end-of-sentence>` token. A comprehensive range of 104 LLMs from Huggingface was employed; the complete list is provided in Appendix C.

Computation. All inferences on the LLMs were conducted using half-precision (`float-16`). Given the extensive volume of linguistic minimal pairs and the number of LLMs, we optimized storage by saving both the activation differences and the computed pairwise linguistic similarity matrix in `int8` precision. Under this configuration, the activation differences for `Llama-2-7B` required 1.3 GB of storage (a tensor of 67,000 samples × 5 layers × 4096 neurons), while the similarity matrix of 67,000 × 67,000 necessitated 4.2 GB of storage.

3 Result and Discussion

To quantitatively assess the consistency of linguistic similarity across different LLMs, we adopted the *mutual k-nearest neighbors* metric as proposed in Huh et al. [2024]. Specifically, for a given linguistic minimal pair, we retrieved the top-k closest neighbors based on linguistic similarity from two distinct LLMs and calculated the percentage of overlapping samples among the retrieved sets as the alignment score (see Huh et al., 2024 for further details).

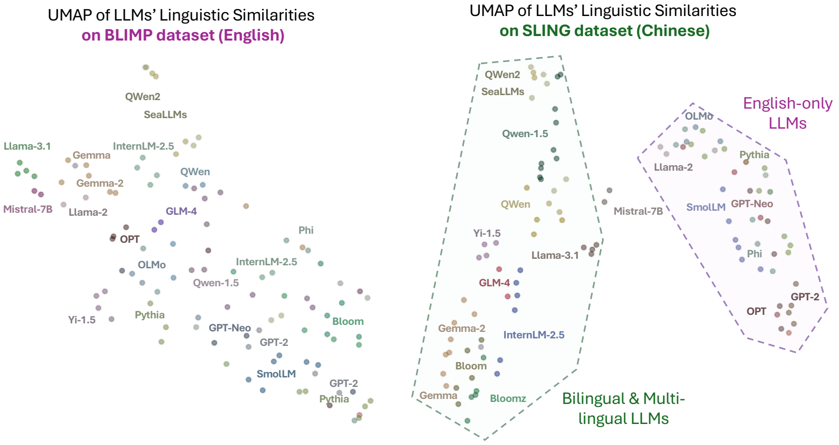

Figure 2: **The relationship between linguistic similarity across LLMs**. In English, LLMs form a single cluster, while in Chinese, two distinct clusters emerge: one for bilingual and multilingual LLMs, and another for English-only models. Detailed visualizations can be found in Appendix E.1.

We computed pairwise alignment scores for all 104 LLMs across three datasets. For computational efficiency, we randomly sampled 10% of the total samples from each dataset (*i.e.,* 6,700 for BLiMP, 3,800 for SLING, and 4,500 for RuBLiMP) and set k to 1% of the sample pool size for retrieval (*i.e.,* 67, 38, and 45). Notably, the BLiMP dataset (English) exhibited the highest alignment, with an average score of 0.471 (*i.e.,* 47.1% of the top-1% similar minimal pairs are shared across LLMs on average), whereas the SLING (Chinese) and RuBLiMP (Russian) demonstrated average alignment scores of 0.414 and 0.139, respectively.

We also observed that the distribution of alignment scores for English and Russian datasets is unimodal, while the Chinese dataset exhibits a bimodal distribution. To further investigate this phenomenon, we employed UMAP McInnes et al. [2018] to embed the LLMs to 2D plane based on a distance metric of $\texttt{distance} = -\log(\texttt{alignment score})$, following Huh et al. [2024]. As illustrated in Fig. 2, each dot represents an LLM, with closer dots indicating similar linguistic similarities. LLMs from the same family (*e.g.,* models from the same creator with different sizes or base/chat versions) are clustered together, indicated by the same color. Interestingly, in Chinese, we observed two distinct clusters: the left cluster (marked with a green dotted line) predominantly consists of bilingual (English-Chinese) and multilingual models, while the right cluster marked with purple is primarily composed of English-only trained LLMs. These results suggest a correlation between alignment scores and LLMs' pertaining language.

3.1 Alignment between Linguistic Similarity and Theoretical Categorizations

To investigate the alignment between linguistic similarity and theoretical linguistic categorizations, we compared intra-class and inter-class similarities at different levels of linguistic classification for both English (BLiMP) and Chinese (SLING) datasets. BLiMP and SLING provide hierarchical linguistic classifications, with the 1st level being the most fine-grained "phenomena", the 2nd level capturing broader categories of "terms", and the 3rd level representing the most general "fields" (examples are shown in leftmost and rightmost columns in Fig. 6).

The analysis in this section is based on averaged linguistic similarities across the 104 LLMs. We focused on BLiMP and SLING datasets due to their high linguistic similarity alignment across LLMs, as demonstrated in the previous Section 3. We excluded the RuBLiMP (Russian) dataset with weaker consensus across LLMs.

As shown in Fig. 6, our analysis revealed that at the lowest level, intra-class similarities were significantly higher than inter-class similarities for both BLiMP and SLING datasets. This suggests that linguistic similarity effectively captures the nuanced distinctions within these detailed categories. This clear separation indicates a strong alignment between linguistic similarity and these fine-grained theoretical linguistic categorizations. However, as we moved to higher levels of classification, the gap between intra-class and inter-class similarities diminished considerably. For both datasets, the disparity between inner and inter-class similarities approximately halved with each ascent in the

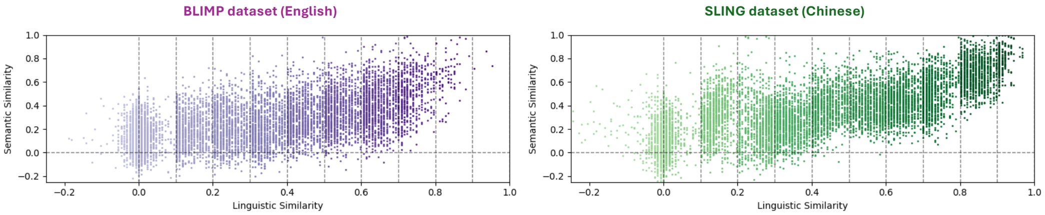

Figure 3: **Joint distribution of linguistic similarity and semantic similarity**. Each dot in the plots represents a pair of linguistic minimal pairs. We observed a weak correlation between the two similarity measurements, showing that linguistic similarity has a context-dependent nature.

categorization hierarchy. Higher-level linguistic categories exhibit greater interconnectedness and mutual influence than previously recognized, which may explain the diminishing differentiation in linguistic similarity at broader classification levels.

We further visualize the phenomenon-level linguistic similarity matrix in Fig. 5 (visualizations for other datasets can be found in Appendix E.2). The observations here confirm our above conclusions from Fig. 6. The clearly distinguishable diagonal entries represent the lowest linguistic phenomenon-level similarity, which enjoys significant inter/intra separation. For higher 2nd-level classification, as noted by dashed black lines, there exist both homogeneous (*e.g.,* anaphor agreement, determiner none agreement) and heterogeneous (*e.g.,* irregular forms and many others) ones. We cannot observe any clear clustering effects for the 3rd-level classification as separated by bold black lines, which also aligns with our findings from Fig. 6.

Interestingly, we found that there are some pairs of phenomena that do not belong to the exact same theoretical categorization, but also enjoy considerably high similarities. For example, *"sentential negation NPI licensor present"* in the *"NPI licensing"* term of *"semantics"* field has very high similarity with *"sentential negation NPI scope"* in the *"NPI licensing"* term of *"syntax-semantics"* field; and *"principle A domain 3"* in the *"binding"* term of *"syntax-semantics"* filed has considerably high similarities to the phenomena in *"anaphor agreement"* term of the *"morphology"* field. These observations underscore the potential of linguistic similarity as a valuable tool for refining our understanding of language structure and organization.

3.2 The Relationship Between Linguistic Similarity and Semantic Similarity

To further investigate the nature of linguistic similarity captured by our method, we compare it with the semantic similarity between minimal pairs. We employed a multilingual Sentence Transformer Reimers and Gurevych [2019] model[4] to generate sentence embeddings for the correct sentence in the minimal pairs. Cosine similarity between these embeddings served as our measure of semantic similarity following default practice. We sampled 1k pairs of minimal pairs that have linguistic similarity within each range of (0.9, 1.0), (0.8, 0.9), ..., (0, 0.1), $(-\infty, 0)$, and plotted their semantic similarity against their linguistic similarity.

Fig. 3 provides the results for the BLiMP and SLING. We can observe a weak correlation between linguistic and semantic similarities for both datasets, suggesting that the linguistic similarity in LLMs is *context dependent* to some extent. However, linguistic and semantic similarities can vary independently. In English, two minimal pairs from the Binding (syntax-semantics) phenomenon show high linguistic similarity (>0.6) but low semantic similarity (<0.3), indicating that while these pairs involve the same linguistic structures, the semantic impact of the changes differs substantially. Conversely, we found cases of low linguistic similarity (<0.3) but high semantic similarity (>0.6) between minimal pairs from Binding (syntax-semantics) and Argument Structure (syntax). We also observed cases where both linguistic and semantic similarities are consistently high or low, which often emerges when minimal pairs share similar vocabulary or address completely unrelated linguistic features. The result of SLING reveals similar patterns to those observed in BLiMP.

[4] `https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2`

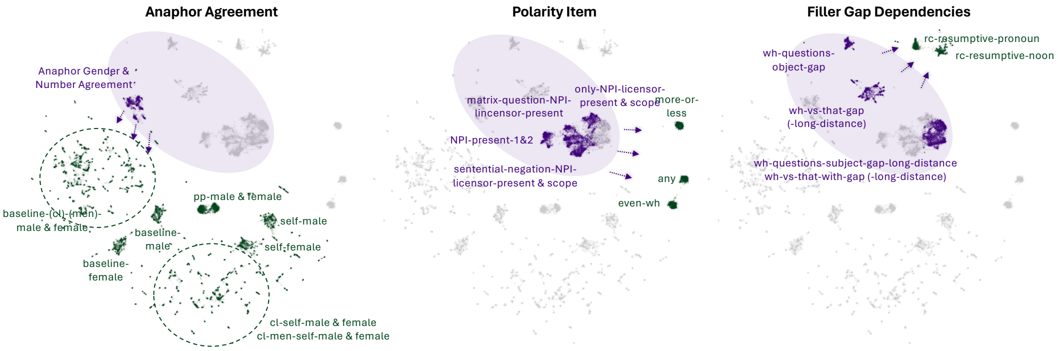

Figure 4: **UMAP visualization of minimal pairs in same categories (terms) but different languages**. Linguistic phenomena are dominantly grouped by their language in multilingual LLMs: English samples are all clustered within purple-shaded areas. While relevant linguistic phenomena in different languages are not fully overlapped, LLM does capture some relationships. As indicated by purple arrows, English samples seem to be "attracted" by the corresponding Chinese samples.

3.3 Linguistic Similarity Across Different Languages

We are interested in how LLMs represent similar linguistic phenomena across different languages–for relevant phenomena in different languages, whether they exhibit higher linguistic similarity? We conducted a multi-lingual analysis focusing on three key linguistic terms: anaphor agreement, polarity items, and filler-gap dependency, on both BLiMP (English) and SLING (Chinese) datasets. Linguistic phenomena within each of these terms are considered relevant, and a total of 39 phenomena are involved, leading to a total of 39k minimal pairs.

To explore the relationship between those phenomena, we used a representative state-of-the-art multilingual LLM `Llama-3.1`[5], and employed UMAP dimensionality reduction based on pair-wise linguistic similarity.

Interestingly, we also observe that relevant phenomena across different languages often exhibit closer proximity than unrelated phenomena within the same language. This cross-lingual correlation is illustrated by the purple arrows in Figure 4, which *"attracts"* English samples to Chinese samples. Quantitatively, we calculated the average pair-wise linguistic similarity of the three terms in BLiMP and SLING.

4 Conclusion

Our comprehensive analysis of linguistic similarity in LLMs, spanning over 100 models and 150k minimal pairs across three languages, reveals several key insights. We found that linguistic similarity consistency across LLMs is strongly influenced by pertaining data, with high-resource languages showing greater alignment. LLMs' internal representations align well with fine-grained linguistic categorizations, but this alignment weakens at broader levels. The weak correlation between linguistic and semantic similarities suggests that LLMs' representation of linguistic phenomena is context-dependent. Cross-lingually, while LLMs tend to group phenomena by language, they do capture some relationships between relevant phenomena across languages.

These findings contribute to our understanding of LLMs' internal language processing, potentially bridging the gap between neural language models and linguistic theory. As LLMs continue to advance, this work provides a foundation for future research into their linguistic representations, informing model development and offering insights into both artificial and human language processing.

[5] `https://huggingface.co/meta-llama/Meta-Llama-3.1-8B`

Appendix

A Related Work

Linguistic Minimal Pairs. Acceptability judgments Chomsky [1957] have long served as a proxy for grammaticalness in generative syntax, relying on native speakers' intuitions to evaluate whether sequences generated by a grammar are perceived as acceptable. In the past decade, grammatical acceptability judgment tasks Linzen et al. [2016], Futrell et al. [2018], Wilcox et al. [2019], Warstadt et al. [2019], Gauthier et al. [2020] have been commonly used to evaluate language models by comparing output probabilities, providing a direct linguistic measure of sentence acceptability. Among these benchmarks, BLiMP Warstadt et al. [2020] introduced a large-scale linguistic minimal pair benchmark in English. This work was followed by many studies in other languages, namely CLiMP Xiang et al. [2021] and SLING Song et al. [2022] for Chinese, JBLiMP Someya and Oseki [2023] for Japanese, BHASA Leong et al. [2023] for Indonesian, RuBLiMP Taktasheva et al. [2024] for Russian and BLiMP-NL Suijkerbuijk et al. [2024] for Dutch.

Linguistic and Language Representation in LLMs. A growing body of research is investigating the linguistic mechanisms within LLMs through probing and interventional strategies Arora et al. [2024], He et al. [2024], Weber et al. [2024]. These studies mostly focus on specific linguistic phenomena, such as subject-verb agreement Giulianelli et al. [2018], plurality Hanna [2022], long-distance agreement Li et al. [2023], negative polarity items DeCarlo et al. [2023], and adjective order Jumelet et al. [2024], see Millière [2024] for a comprehensive review. Furthermore, recent studies have also identified linguistic regions Zhang et al. [2024] and language-specific neurons Tang et al. [2024], Kojima et al. [2024] that contribute to multilingual capabilities, and further suggests that LLMs exhibit layer-wise specialization, with intermediate layers processing information in a common "language" concept space and final layers generating responses in the specific language Wendler et al. [2024], Zhong et al. [2024].

B Limitations and Future Works

Our study's findings are inherently dependent on the quality and scope of existing linguistic minimal pair datasets, which may have the possibility of violating the assumption of having *only* difference of individual phenomena. Additionally, our analysis is limited to three languages, which cannot fully represent global linguistic diversity. Despite examining 150 linguistic phenomena, this still captures only a fraction of language's complexity. Future directions include developing more high-quality minimal pairs, broadening language coverage to include more diverse linguistic families, and increasing the range of phenomena studied.

C The List of 104 LLMs

The following is a list of all LLMs that are adopted in our analysis, grouped by model families:

Llama-2

- `NousResearch/Llama-2-7b-chat-hf`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `NousResearch/Llama-2-13b-chat-hf`: Total 41 layers, 5120 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.

Llama-3 & Llama-3.1

- `meta-llama/Meta-Llama-3-8B`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `meta-llama/Meta-Llama-3-8B-Instruct`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `meta-llama/Meta-Llama-3.1-8B`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.

- `meta-llama/Meta-Llama-3.1-8B-Instruct`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.

Mistral-7B-v0.3

- `mistralai/Mistral-7B-v0.3`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `mistralai/Mistral-7B-Instruct-v0.3`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.

OLMo

- `allenai/OLMo-1B-hf`: Total 17 layers, 2048 neurons per layer. Sampled layers: 2, 5, 8, 11, 14.
- `allenai/OLMo-7B-hf`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `allenai/OLMo-7B-Instruct-hf`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.

Qwen, Qwen-1.5, & Qwen-2

- `Qwen/Qwen-7B`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `Qwen/Qwen-7B-Chat`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `Qwen/Qwen-14B`: Total 41 layers, 5120 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.
- `Qwen/Qwen-14B-Chat`: Total 41 layers, 5120 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.
- `Qwen/Qwen1.5-0.5B`: Total 25 layers, 1024 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- `Qwen/Qwen1.5-0.5B-Chat`: Total 25 layers, 1024 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- `Qwen/Qwen1.5-1.8B`: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- `Qwen/Qwen1.5-1.8B-Chat`: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- `Qwen/Qwen1.5-4B`: Total 41 layers, 2560 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.
- `Qwen/Qwen1.5-4B-Chat`: Total 41 layers, 2560 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.
- `Qwen/Qwen1.5-7B`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `Qwen/Qwen1.5-7B-Chat`: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- `Qwen/Qwen1.5-14B`: Total 41 layers, 5120 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.
- `Qwen/Qwen1.5-14B-Chat`: Total 41 layers, 5120 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.
- `Qwen/Qwen2-0.5B`: Total 25 layers, 896 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- `Qwen/Qwen2-0.5B-Instruct`: Total 25 layers, 896 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- `Qwen/Qwen2-1.5B`: Total 29 layers, 1536 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- `Qwen/Qwen2-1.5B-Instruct`: Total 29 layers, 1536 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- `Qwen/Qwen2-7B`: Total 29 layers, 3584 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- `Qwen/Qwen2-7B-Instruct`: Total 29 layers, 3584 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.

SeaLLMs

- `SeaLLMs/SeaLLMs-v3-1.5B`: Total 29 layers, 1536 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- `SeaLLMs/SeaLLMs-v3-1.5B-Chat`: Total 29 layers, 1536 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- `SeaLLMs/SeaLLMs-v3-7B`: Total 29 layers, 3584 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- `SeaLLMs/SeaLLMs-v3-7B-Chat`: Total 29 layers, 3584 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.

SmolLM

- HuggingFaceTB/SmolLM-135M: Total 31 layers, 576 neurons per layer. Sampled layers: 5, 10, 15, 20, 25.
- HuggingFaceTB/SmolLM-135M-Instruct: Total 31 layers, 576 neurons per layer. Sampled layers: 5, 10, 15, 20, 25.
- HuggingFaceTB/SmolLM-360M: Total 33 layers, 960 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- HuggingFaceTB/SmolLM-360M-Instruct: Total 33 layers, 960 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- HuggingFaceTB/SmolLM-1.7B: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- HuggingFaceTB/SmolLM-1.7B-Instruct: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.

TinyLlama

- TinyLlama/TinyLlama_v1.1: Total 23 layers, 2048 neurons per layer. Sampled layers: 3, 7, 11, 15, 19.
- TinyLlama/TinyLlama_v1.1_chinese: Total 23 layers, 2048 neurons per layer. Sampled layers: 3, 7, 11, 15, 19.
- TinyLlama/TinyLlama_v1.1_math_code: Total 23 layers, 2048 neurons per layer. Sampled layers: 3, 7, 11, 15, 19.

Yi

- 01-ai/Yi-1.5-6B: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- 01-ai/Yi-1.5-6B-Chat: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- 01-ai/Yi-1.5-9B: Total 49 layers, 4096 neurons per layer. Sampled layers: 8, 16, 24, 32, 40.
- 01-ai/Yi-1.5-9B-Chat: Total 49 layers, 4096 neurons per layer. Sampled layers: 8, 16, 24, 32, 40.

Bloom & Bloomz

- bigscience/bloom-560m: Total 25 layers, 1024 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- bigscience/bloomz-560m: Total 25 layers, 1024 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- bigscience/bloom-1b1: Total 25 layers, 1536 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- bigscience/bloomz-1b1: Total 25 layers, 1536 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- bigscience/bloom-1b7: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- bigscience/bloomz-1b7: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- bigscience/bloom-3b: Total 31 layers, 2560 neurons per layer. Sampled layers: 5, 10, 15, 20, 25.
- bigscience/bloomz-3b: Total 31 layers, 2560 neurons per layer. Sampled layers: 5, 10, 15, 20, 25.
- bigscience/bloom-7b1: Total 31 layers, 4096 neurons per layer. Sampled layers: 5, 10, 15, 20, 25.
- bigscience/bloomz-7b1: Total 31 layers, 4096 neurons per layer. Sampled layers: 5, 10, 15, 20, 25.

Gemma & Gemma-2

- google/gemma-2b: Total 19 layers, 2048 neurons per layer. Sampled layers: 3, 6, 9, 12, 15.
- google/gemma-2b-it: Total 19 layers, 2048 neurons per layer. Sampled layers: 3, 6, 9, 12, 15.
- google/gemma-7b: Total 29 layers, 3072 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- google/gemma-7b-it: Total 29 layers, 3072 neurons per layer. Sampled layers: 4, 9, 14, 19, 24.
- google/gemma-2-2b: Total 27 layers, 2304 neurons per layer. Sampled layers: 4, 9, 13, 18, 22.
- google/gemma-2-2b-it: Total 27 layers, 2304 neurons per layer. Sampled layers: 4, 9, 13, 18, 22.
- google/gemma-2-9b: Total 43 layers, 3584 neurons per layer. Sampled layers: 7, 14, 21, 28, 35.

- google/gemma-2-9b-it: Total 43 layers, 3584 neurons per layer. Sampled layers: 7, 14, 21, 28, 35.

GLM

- THUDM/glm-4-9b: Total 41 layers, 4096 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.
- THUDM/glm-4-9b-chat: Total 41 layers, 4096 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.

GPT-Neo

- EleutherAI/gpt-neo-125m: Total 13 layers, 768 neurons per layer. Sampled layers: 2, 4, 6, 8, 10.
- EleutherAI/gpt-neo-1.3B: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- EleutherAI/gpt-neo-2.7B: Total 33 layers, 2560 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- EleutherAI/gpt-neox-20B: Total 45 layers, 6144 neurons per layer. Sampled layers: 7, 15, 22, 30, 37.

GPT-2

- openai-community/gpt2: Total 13 layers, 768 neurons per layer. Sampled layers: 2, 4, 6, 8, 10.
- openai-community/gpt2-medium: Total 25 layers, 1024 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- openai-community/gpt2-large: Total 37 layers, 1280 neurons per layer. Sampled layers: 6, 12, 18, 24, 30.
- openai-community/gpt2-xl: Total 49 layers, 1600 neurons per layer. Sampled layers: 8, 16, 24, 32, 40.

InternLM-2.5

- internlm/internlm2_5-1_8b: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- internlm/internlm2_5-1_8b-chat: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- internlm/internlm2_5-7b: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- internlm/internlm2_5-7b-chat: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- internlm/internlm2_5-20b: Total 49 layers, 6144 neurons per layer. Sampled layers: 8, 16, 24, 32, 40.
- internlm/internlm2_5-20b-chat: Total 49 layers, 6144 neurons per layer. Sampled layers: 8, 16, 24, 32, 40.

OPT

- facebook/opt-125m: Total 13 layers, 768 neurons per layer. Sampled layers: 2, 4, 6, 8, 10.
- facebook/opt-1.3b: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- facebook/opt-2.7b: Total 33 layers, 2560 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- facebook/opt-6.7b: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- facebook/opt-13b: Total 41 layers, 5120 neurons per layer. Sampled layers: 6, 13, 20, 27, 34.

Phi

- microsoft/phi-1: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- microsoft/phi-1_5: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- microsoft/phi-2: Total 33 layers, 2560 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- microsoft/Phi-3-mini-128k-instruct: Total 33 layers, 3072 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.

- microsoft/Phi-3-small-128k-instruct: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.

Pythia

- EleutherAI/pythia-1.4b: Total 25 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- EleutherAI/pythia-12b: Total 37 layers, 5120 neurons per layer. Sampled layers: 6, 12, 18, 24, 30.
- EleutherAI/pythia-14m: Total 7 layers, 128 neurons per layer. Sampled layers: 1, 2, 3, 4, 5.
- EleutherAI/pythia-160m: Total 13 layers, 768 neurons per layer. Sampled layers: 2, 4, 6, 8, 10.
- EleutherAI/pythia-1b: Total 17 layers, 2048 neurons per layer. Sampled layers: 2, 5, 8, 11, 14.
- EleutherAI/pythia-2.8b: Total 33 layers, 2560 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- EleutherAI/pythia-410m: Total 25 layers, 1024 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- EleutherAI/pythia-6.7b: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- EleutherAI/pythia-70m: Total 7 layers, 512 neurons per layer. Sampled layers: 1, 2, 3, 4, 5.

Others

- anas-awadalla/mpt-1b-redpajama-200b: Total 24 layers, 2048 neurons per layer. Sampled layers: 4, 8, 12, 16, 20.
- CohereForAI/aya-23-8B: Total 33 layers, 4096 neurons per layer. Sampled layers: 5, 11, 16, 22, 27.
- distilbert/distilgpt2: Total 7 layers, 768 neurons per layer. Sampled layers: 1, 2, 3, 4, 5.

D List of Linguistic Phenomena used in Section 3.3

BLIMP Dataset (English)

- Anaphor agreement
  - anaphor gender agreement
  - anaphor number agreement
- Polarity item
  - matrix question npi licensor present
  - npi present 1
  - npi present 2
  - only npi licensor present
  - only npi scope
  - sentential negation npi licensor present
  - sentential negation npi scope
- Filler gap dependency
  - wh questions object gap
  - wh questions subject gap
  - wh questions subject gap long distance
  - wh vs that no gap
  - wh vs that no gap long distance
  - wh vs that with gap
  - wh vs that with gap long distance

SLING Dataset (Chinese)

- Anaphor agreement

  - self male
  - cl men self female
  - pp male
  - baseline male
  - baseline cl men female
  - baseline cl female
  - cl self male
  - cl men self male
  - self female
  - pp female
  - baseline cl men male
  - menself female
  - menself male
  - cl self female
  - baseline cl male
  - baseline female
  - baseline men female
  - baseline men male

- Polarity item
  - any
  - more or less
  - even wh

- Filler gap dependency
  - rc resumptive noun
  - rc resumptive pronoun

E Additional Visualizations

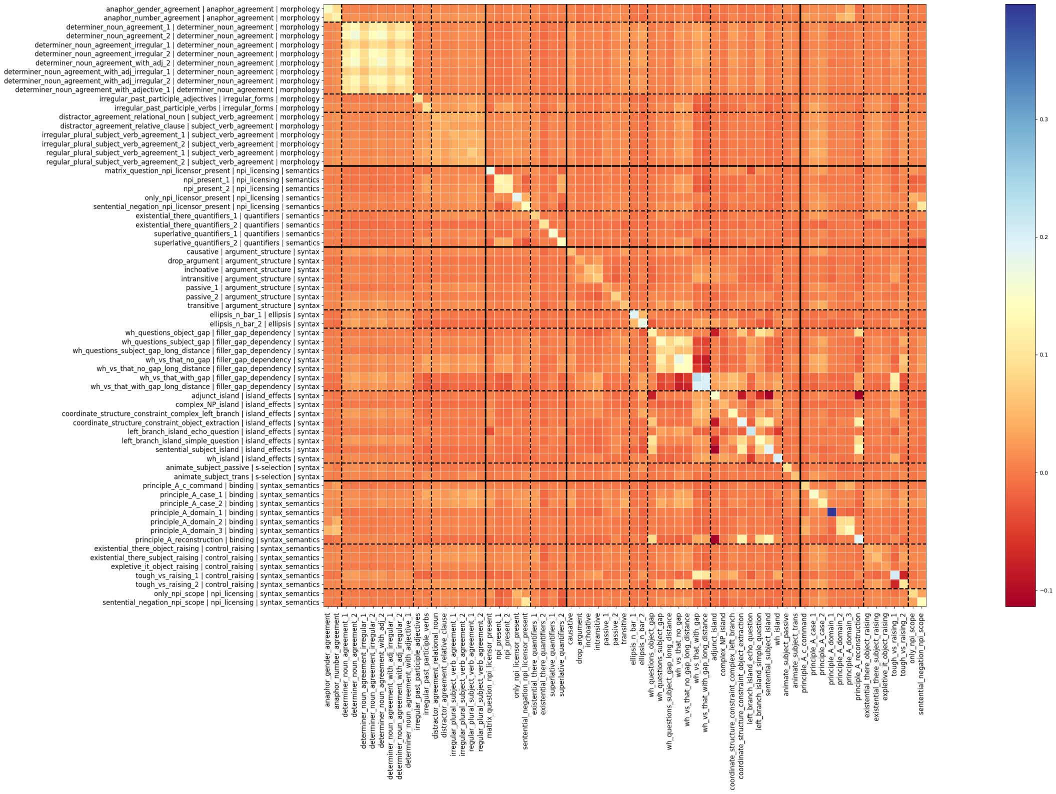

Figure 5: **Phenomena-level linguistic similarity matrix of BLiMP**. Each grid corresponds to the average similarity between two linguistic phenomena. The categorization in the 2nd-level (linguistic terms) and the 3rd-level are respectively separated by dashed and bold black lines. On the left, we provide label of the 1st to 3rd levels of linguistic classifications, separated by "|". Visualizations of SLING and RuBLiMP can be found in Appendix E.2.

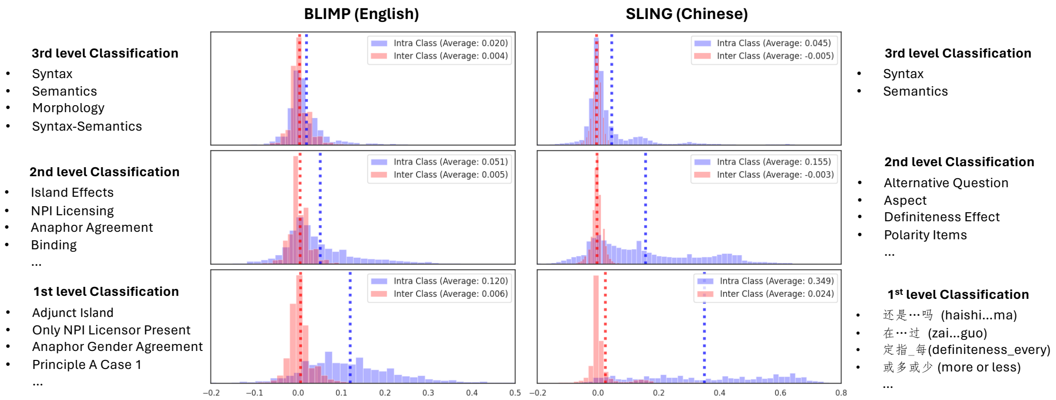

Figure 6: **Intra-class and inter-class linguistic similarities at different levels of linguistic classification**. At the most fine-grained level (1st level), intra-class similarities are significantly higher than inter-class similarities, indicating a strong alignment with detailed theoretical linguistic categorizations. As we move to broader categories (2nd and 3rd levels), the gap between inner and inter-class similarities narrows notably.

E.1 Consistency of Linguistic Similarity Across 104 LLMs

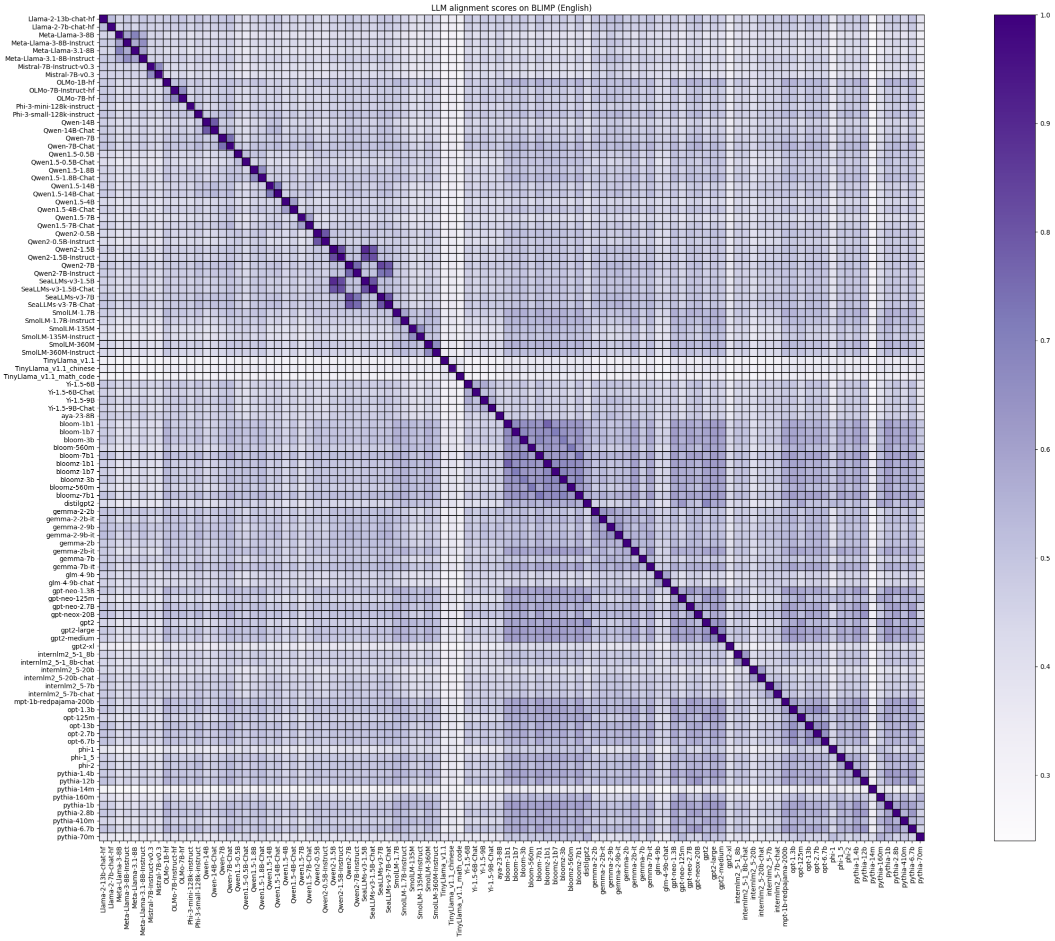

Figure 7: Pair-wise alignment scores of 104 LLMs in BLiMP dataset (English), corresponding to Fig. 2 left.

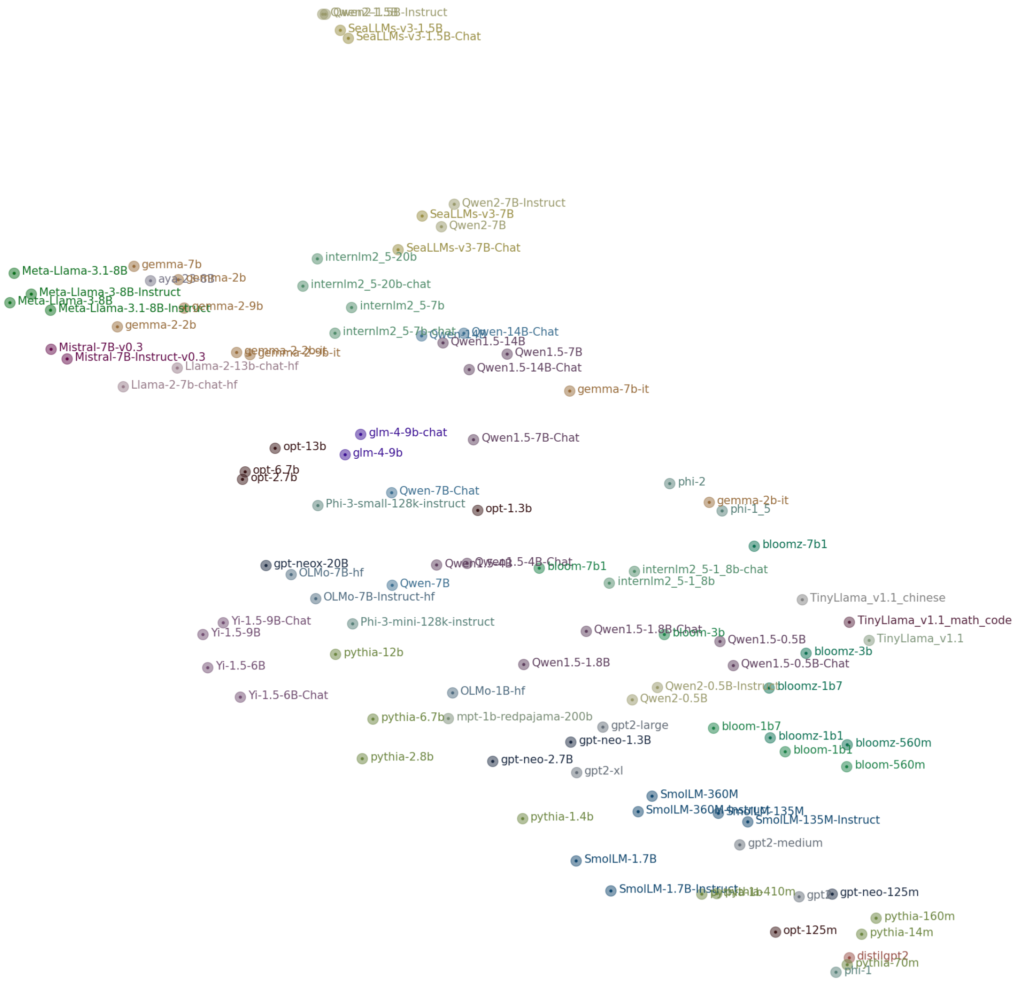

Figure 8: UMAP visualization based on LLM alignment scores on BLiMP dataset (English), corresponding to Fig. 2 left.

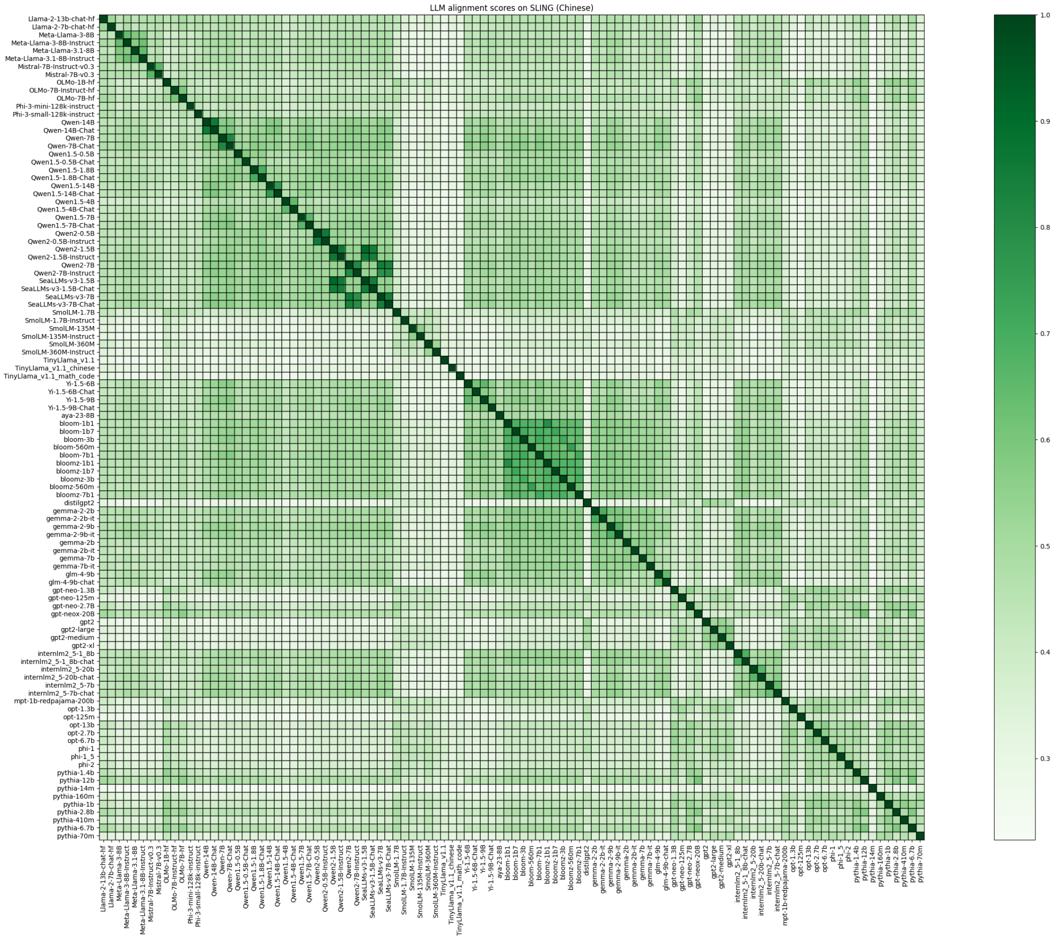

Figure 9: Pair-wise alignment scores of 104 LLMs in SLING dataset (Chinese), corresponding to Fig. 2 right.

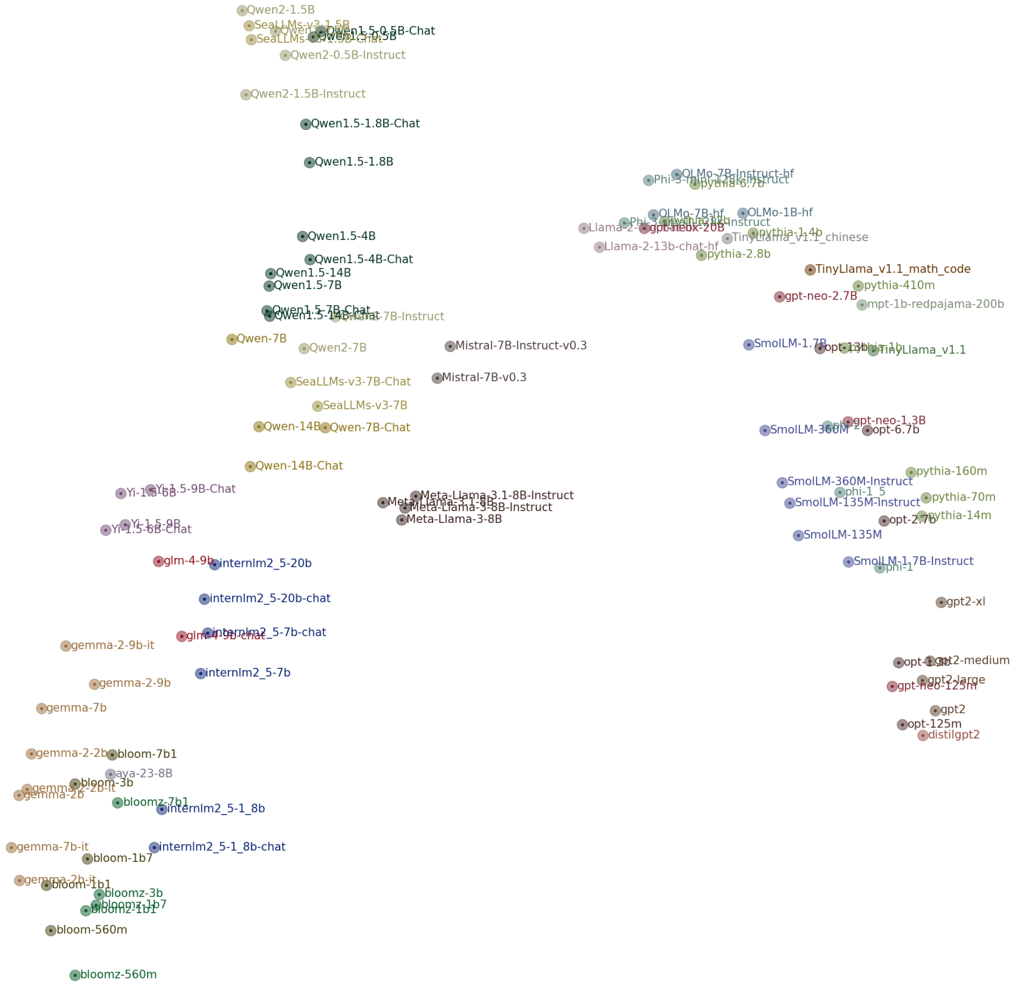

Figure 10: UMAP visualization based on LLM alignment scores on SLING dataset (Chinese), corresponding to Fig. 2 right.

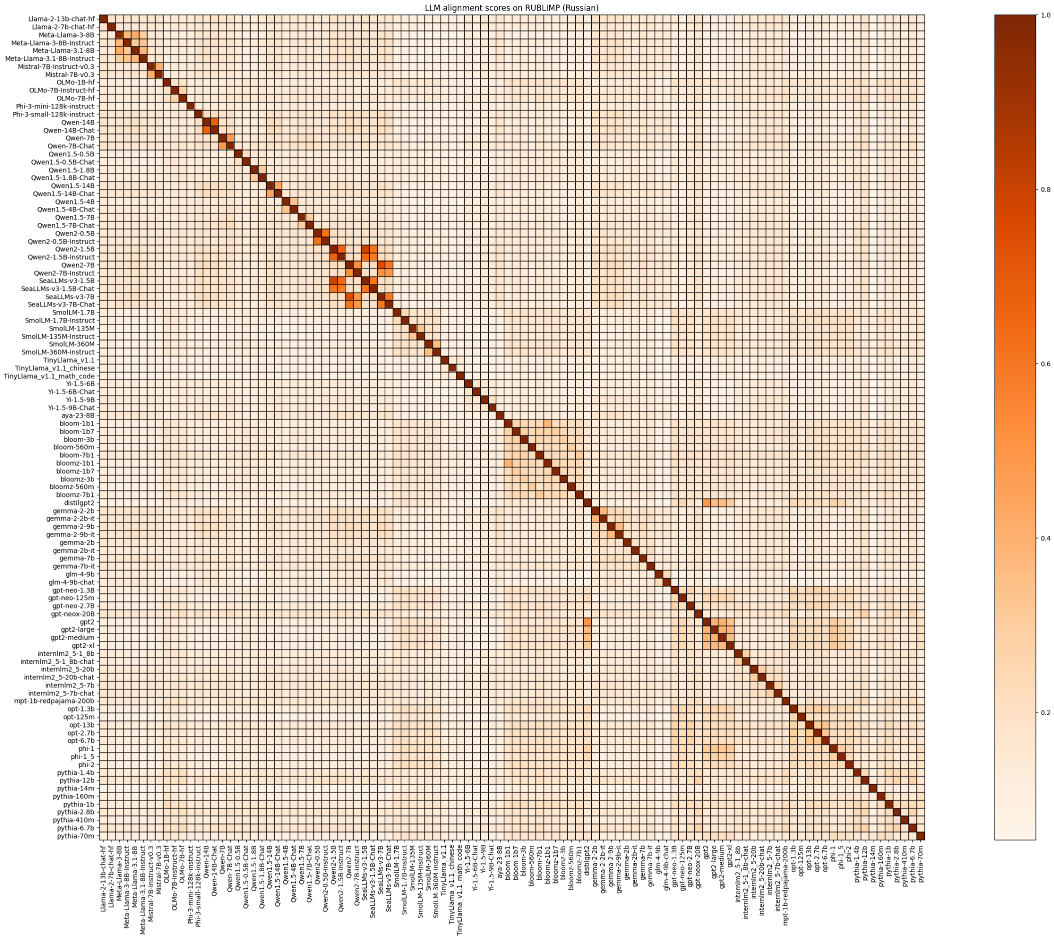

Figure 11: Pair-wise alignment scores of 104 LLMs in RuBLiMP dataset (Russian).

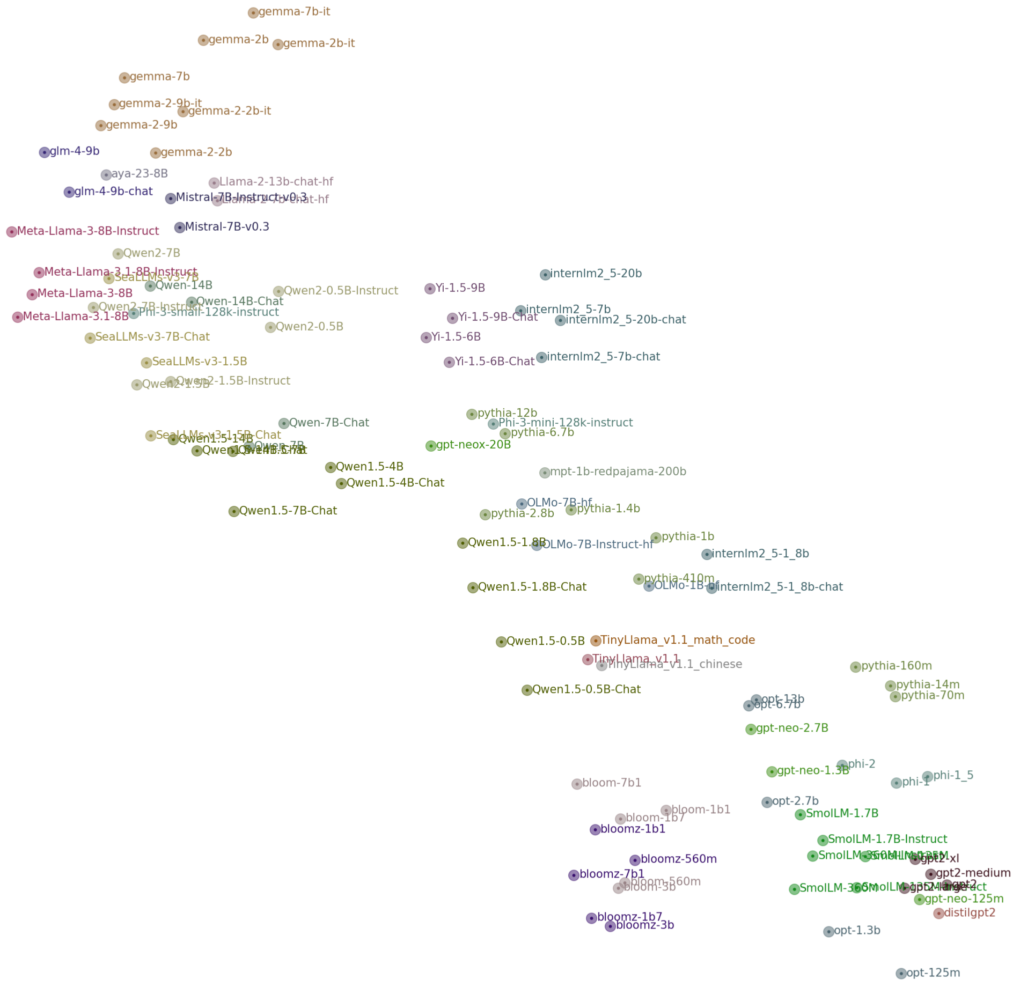

Figure 12: UMAP visualization based on LLM alignment scores on RuBLiMP dataset (Russian).

E.2 Alignment between Linguistic Similarity and Theoretical Categorizations

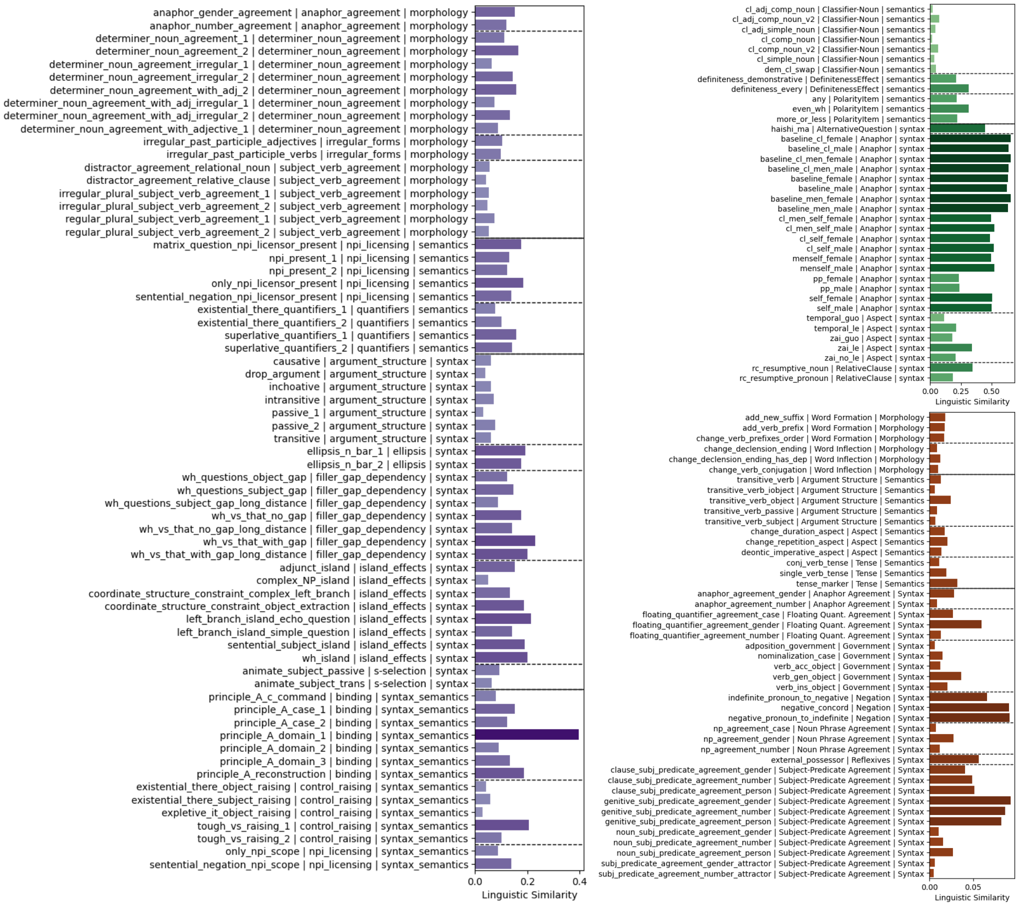

Figure 13: Self-similarities of linguistic phenomena in English, Chinese, and Russian

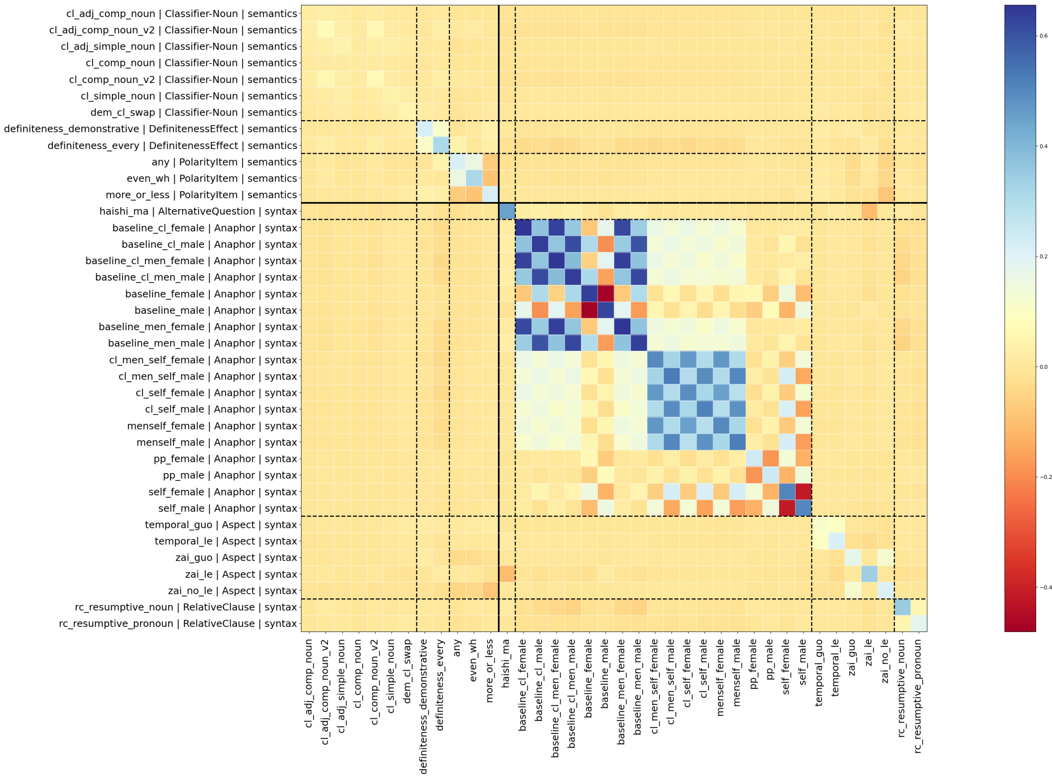

Figure 14: Phenomena-level linguistic similarity matrix of SLING

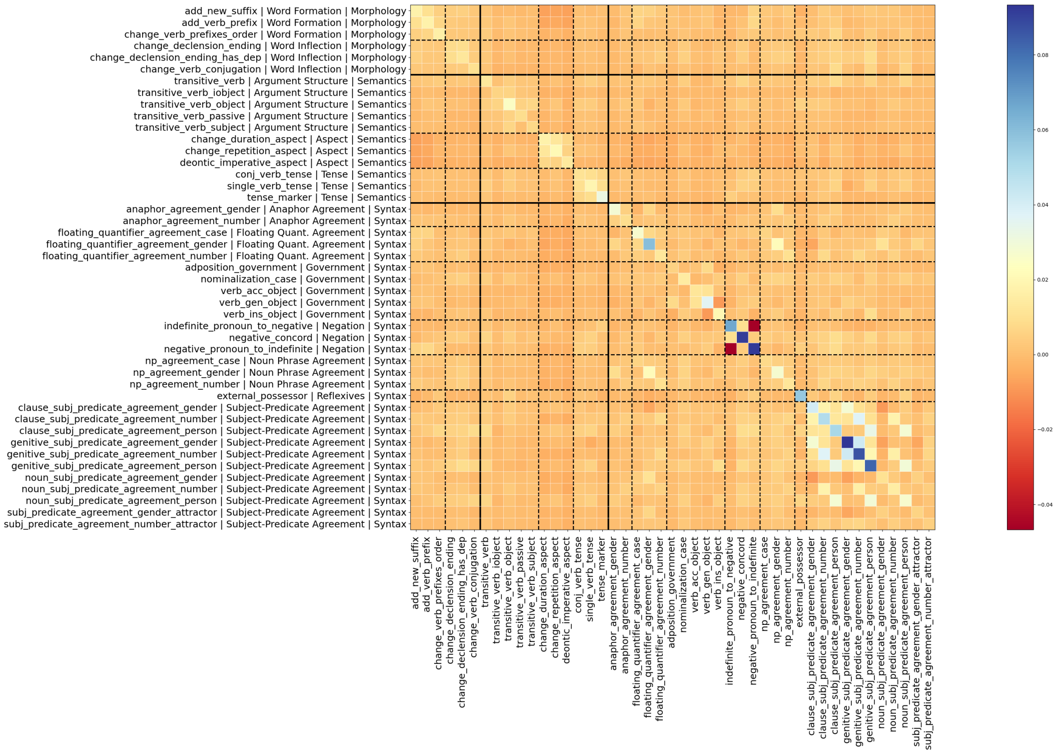

Figure 15: Phenomena-level linguistic similarity matrix of RUBLiMP

E.3 Linguistic Similarity Across Different Languages

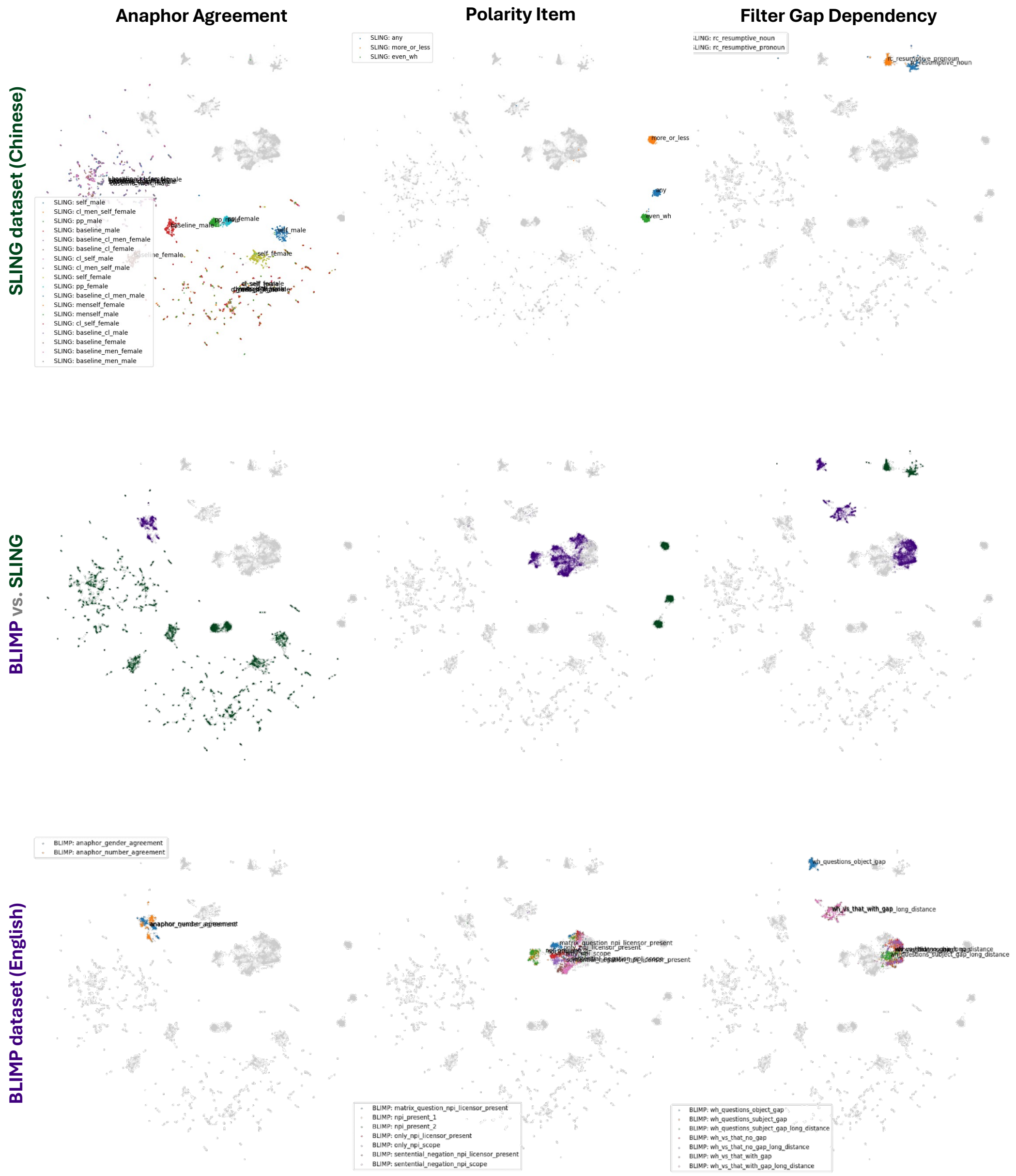

Figure 16: Extended detailed visualization of Fig. 4.

References

Aristotle. Categories. 350 BC.

Aryaman Arora, Dan Jurafsky, and Christopher Potts. Causalgym: Benchmarking causal interpretability methods on linguistic tasks. *arXiv preprint arXiv:2402.12560*, 2024.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, et al. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 675–718, 2023.

Reihane Boghrati, Joe Hoover, Kate M Johnson, Justin Garten, and Morteza Dehghani. Conversation level syntax similarity metric. *Behavior research methods*, 50:1055–1073, 2018.

Noam Chomsky. *Syntactic structures*. Mouton de Gruyter, 1957.

Ileana Comorovski. *Interrogative phrases and the syntax-semantics interface*, volume 59. Springer Science & Business Media, 2013.

Deanna DeCarlo, William Palmer, Michael Wilson, and Bob Frank. Npis aren't exactly easy: Variation in licensing across large language models. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 332–341, 2023.

Richard Futrell, Ethan Wilcox, Takashi Morita, and Roger Levy. Rnns as psycholinguistic subjects: Syntactic state and grammatical dependency. *arXiv preprint arXiv:1809.01329*, 2018.

Jon Gauthier, Jennifer Hu, Ethan Wilcox, Peng Qian, and Roger Levy. 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, 2020.

Mario Giulianelli, Jacqueline Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. Under the hood: Using diagnostic classifiers to investigate and improve how language models track agreement information. *arXiv preprint arXiv:1808.08079*, 2018.

Michael Hanna. Investigating large language models' representations of plurality through probing interventions. 2022.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. *arXiv preprint arXiv:2403.17299*, 2024.

Eric W Holman, Cecil H Brown, Søren Wichmann, André Müller, Viveka Velupillai, Harald Hammarström, Sebastian Sauppe, Hagen Jung, Dik Bakker, Pamela Brown, et al. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, 52(6):841–875, 2011.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis. *arXiv preprint arXiv:2405.07987*, 2024.

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

Takeshi Kojima, Itsuki Okimura, Yusuke Iwasawa, Hitomi Yanaka, and Yutaka Matsuo. On the multilingual ability of decoder-based pre-trained language models: Finding and controlling language-specific neurons. *arXiv preprint arXiv:2404.02431*, 2024.

Wei Qi Leong, Jian Gang Ngui, Yosephine Susanto, Hamsawardhini Rengarajan, Kengatharaiyer Sarveswaran, and William Chandra Tjhi. Bhasa: A holistic southeast asian linguistic and cultural evaluation suite for large language models. *arXiv preprint arXiv:2309.06085*, 2023.

Bingzhi Li, Guillaume Wisniewski, and Benoît Crabbé. Assessing the capacity of transformer to abstract syntactic representations: A contrastive analysis based on long-distance agreement. *Transactions of the Association for Computational Linguistics*, 11:18–33, 2023.

Yafei Li. *Xo: A theory of the morphology-syntax interface*. MIT Press, 2004.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535, 2016.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Raphaël Millière. Language models as models of language. *arXiv preprint arXiv:2408.07144*, 2024.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL `http://arxiv.org/abs/1908.10084`.

Lotte Schoot, Evelien Heyselaar, Peter Hagoort, and Katrien Segaert. Does syntactic alignment effectively influence how speakers are perceived by their conversation partner? *PloS one*, 11(4): e0153521, 2016.

Taiga Someya and Yohei Oseki. Jblimp: Japanese benchmark of linguistic minimal pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, 2023.

Yixiao Song, Kalpesh Krishna, Rajesh Bhatt, and Mohit Iyyer. Sling: Sino linguistic evaluation of large language models. *arXiv preprint arXiv:2210.11689*, 2022.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Jelle Zuidema, and Stefan L Frank. Blimp-nl: A corpus of dutch minimal pairs and grammaticality judgements for language model evaluation. 2024.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, and Ekaterina Artemova. Rublimp: Russian benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2406.19232*, 2024.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*, 2024.

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

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

Lucas Weber, Jaap Jumelet, Elia Bruni, and Dieuwke Hupkes. Interpretability of language models via task spaces. *arXiv preprint arXiv:2406.06441*, 2024.

Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in english? on the latent language of multilingual transformers. *arXiv preprint arXiv:2402.10588*, 2024.

Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. Structural supervision improves learning of non-local grammatical dependencies. *arXiv preprint arXiv:1903.00943*, 2019.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. Climp: A benchmark for chinese language model evaluation. *arXiv preprint arXiv:2101.11131*, 2021.

Zhihao Zhang, Jun Zhao, Qi Zhang, Tao Gui, and Xuanjing Huang. Unveiling linguistic regions in large language models. *arXiv preprint arXiv:2402.14700*, 2024.

Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. Beyond english-centric llms: What language do multilingual language models think in? *arXiv preprint arXiv:2408.10811*, 2024.