XCOMPS: A Multilingual Benchmark of Conceptual Minimal Pairs

Linyang He[*1] **Ercong Nie**[*2, 3]
Sukru Samet Dindar[1] **Arsalan Firoozi**[1] **Adrian Florea**[1]
Van Nguyen[3] **Corentin Puffay**[5] **Riki Shimizu**[1] **Haotian Ye**[2,3]
Jonathan Brennan[4] **Helmut Schmid**[3] **Hinrich Schütze**[2, 3†] **Nima Mesgarani**[1†]
[1]Columbia University [2]Munich Center for Machine Learning
[3]LMU Munich [4]University of Michigan [5]KU Leuven
linyang.he@columbia.edu nie@cis.lmu.de
hinrich@hotmail.com nima@ee.columbia.edu

Abstract

In this work, we introduce XCOMPS, a multilingual conceptual minimal pair dataset that covers 17 languages. Using this dataset, we evaluate LLMs' multilingual conceptual understanding through metalinguistic prompting, direct probability measurement, and neurolinguistic probing. We find that: 1) LLMs exhibit weaker conceptual understanding for low-resource languages, and accuracy varies across languages despite being tested on the same concept sets. 2) LLMs excel at distinguishing concept-property pairs that are visibly different but exhibit a marked performance drop when negative pairs share subtle semantic similarities. 3) More morphologically complex languages yield lower concept understanding scores and require deeper layers for conceptual reasoning. The dataset is publicly available at: https://github.com/LinyangHe/XCOMPS/.

1 Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across various natural language understanding (NLU) tasks. Recent advances, such as GPT-4 (Achiam et al., 2023) and Llama 3 (Dubey et al., 2024), have shown that LLMs can produce human-like outputs and handle complex linguistic phenomena. However, whether LLMs genuinely understand semantics or merely rely on shallow statistical correlations is disputable (Lake and Baroni, 2018; Elazar et al., 2021; Huang et al., 2023). One fundamental aspect of human conceptual understanding is that it is not dependent on specific linguistic forms or modalities (Carey, 2000; Mandler, 2004). When humans learn and reason about concepts, they do not require the knowledge to be tied to a particular medium, such as text, images, or video, nor do they rely on a specific language. This raises an important question: *Does LLMs' conceptual-property reasoning remain stable across languages, or is it language-specific?*

To explore this, Misra et al. (2023) introduced the COMPS dataset, designed to probe the semantic reasoning abilities of LLMs through minimal pairs in English. However, COMPS only evaluates monolingual conceptual-property reasoning, leaving open the question of whether LLMs generalize such reasoning across languages. In this work, we introduce XCOMPS, a multilingual extension of COMPS, to assess whether LLMs' semantic reasoning is universally consistent across languages. XCOMPS covers 17 languages, including analytic, inflectional, and agglutinative languages, ensuring a broad representation of linguistic structures.

Beyond dataset expansion, evaluating LLMs' reasoning abilities has increasingly relied on prompt engineering, often referred to as metalinguistic prompting (Hu and Levy, 2023). However, recent work (Hu and Levy, 2023; He et al., 2024b) suggests that metalinguistic prompting primarily assesses performance—that is, how well a model produces correct outputs—rather than its underlying competence in conceptual understanding. This distinction is crucial, as models may perform well on explicit prompts but lack true conceptual representations (Piantadosi and Hill, 2022). To investigate LLMs' multilingual capabilities and determine whether they genuinely encode conceptual knowledge across languages, we adopt a three-pronged evaluation approach: *Metalinguistic prompting*, *Neurolinguistic probing*, and *Direct probability measurement*. Our experimental results reveal several insights into the multilingual conceptual reasoning capabilities of LLMs: 1) Conceptual understanding is not consistently maintained across languages. Even when models perform well in English, their reasoning ability deteriorates significantly in low-resource languages; the extent of deterioration also varies across different low-resource languages. 2) Models perform well when concep-

[*] Equal contribution.
[†] Corresponding authors.

Proceedings of the 7th Workshop on Research in Computational Linguistic Typology and Multilingual NLP (SIGTYP 2025), pages 75–81
August 1, 2025 ©2025 Association for Computational Linguistics

Type	Language	Acceptable Sentence	Unacceptable Sentence
Taxonomic	Spanish	*Tostadora* se utiliza para calentar alimentos. (A toaster is used for heating food.)	*Cafetera* se utiliza para calentar alimentos. (A coffee maker is used for heating food.)
Overlap	Vietnamese	*Máy nướng bánh mì được* sử dụng để hâm nóng thực phẩm. (A toaster is used for heating food.)	*Tủ lạnh được* sử dụng để hâm nóng thực phẩm. (A refrigerator is used for heating food.)
Co-occurrence	Hungarian	*Kenyérpirító* ételek melegítésére használják. (A toaster is used for heating food.)	*Vízforraló* ételek melegítésére használják. (A kettle is used for heating food.)
Random	Dutch	*Broodrooster* wordt gebruikt om voedsel te verwarmen. (A toaster is used for heating food.)	*Winterkoning* wordt gebruikt om voedsel te verwarmen. (A wren is used for heating food.)

Table 1: XCOMPS examples, illustrating each linguistic variant pairs an acceptable sentence (positively matched property) with an unacceptable counterpart (negatively matched property).

tual relationships are highly distinct but struggle with subtle semantic distinctions. 3) Languages with higher morphological complexity (agglutinative > inflected > analytic) yield lower concept-reasoning scores. These results suggest that LLMs' semantic reasoning may not generalize universally across linguistic boundaries.

2 Language Performance vs. Competence

As suggested in He et al. (2024b), LLMs can be evaluated through three methods: *metalinguistic prompting*, which assesses *performance* based on explicit responses; direct probability measurement, which provides an intermediate evaluation by comparing model-generated probabilities; and *neurolinguistic probing*, which directly examines *competence* by analyzing internal activation patterns[1].

Metalinguistic Prompting for Performance This method involves explicitly querying the model about linguistic expressions, often in a comparative or multiple-choice format. By asking the model to choose between minimal pairs (e.g., "Which sentence is more grammatically correct?"), researchers can evaluate how well the model retrieves and verbalizes knowledge. Using prompting, researchers have revealed new classes of emergent abilities such as arithmetic, instruction-following, grounded conceptual mappings, and sentence acceptability judgments (Brown et al., 2020; Wei et al., 2022; Patel and Pavlick, 2021; Dentella et al., 2023). Because the responses are influenced by prompt engineering and surface-level cues, this method primarily reflects performance rather than deep conceptual competence.

Direct Probability Measurement Instead of relying on explicit responses, this method examines the model's probability assignment to different sentences within minimal pairs. For example, a model should assign a higher probability to 'A robin can fly' than to 'A penguin can fly'. This approach offers a more objective evaluation than metalinguistic prompting and captures implicit model preferences, placing it between performance and competence. Researchers have designed syntactic, semantic/conceptual, and discourse inference tasks using the probability assignment method, offering different insights into LLMs' capabilities compared to metalinguistic prompting (Futrell et al., 2019; Gauthier et al., 2020; Hu et al., 2020; Warstadt et al., 2020; Beyer et al., 2021; Misra et al., 2023; Kauf et al., 2023). However, it still relies on external outputs and does not fully reveal how the model internally represents concepts.

Neurolinguistic Probing for Competence This approach goes beyond external outputs by analyzing internal activation patterns across different layers of the model (He et al., 2024a,b). Using diagnostic classifiers, researchers can probe whether LLMs inherently encode conceptual-property relationships or simply rely on statistical correlations. Since it provides a direct measure of competence, neurolinguistic probing is more reliable for assessing the depth of linguistic understanding.

3 XCOMPS

3.1 Concept Selection

To ensure that XCOMPS maintains conceptual alignment with COMPS while extending its scope to multiple languages, we use the same 521 concepts and their negative samples from COMPS. As shown in Table 1, these negative samples can be categorized into three types. *Taxonomy-based* negative samples are selected based on hierarchical relationships among concepts. Negative samples come from the same broad category as the positive concept but differ in key property attributions. *Property norm-based (overlap)* negative samples are chosen based on shared semantic properties with the positive concept while lacking the specific

[1] For simplicity, we refer to these three methods as Meta, Direct, Neuro.

property under evaluation. *Co-occurrence-based samples* are selected from concepts that frequently appear in similar contexts but do not share the target property. XCOMPS also has additional *random negative concepts* from the set of concepts that do not possess the property of the original positive concept.

3.2 Properties of Concepts

In XCOMPS, the properties assigned to concepts are inherited from COMPS, ensuring alignment across languages while maintaining the original conceptual-property relationships. These properties in COMPS were originally derived from the XCSLB dataset, an extended version of the CSLB property norm dataset (Devereux et al., 2014), which captures human-annotated perceptual, functional, and categorical attributes of concepts. Additionally, taxonomic relationships from resources like WordNet (Miller, 1995) were used to infer properties through hierarchical inheritance, ensuring that general category attributes (e.g., "mammals have fur") are systematically applied to their subcategories. Some properties also reflect real-world associations observed in corpus-based co-occurrence statistics.

3.3 Multilingual Data Construction

To construct XCOMPS, which covers 17 languages (Table 2 in Appendix A), we adopted a human-LLM interactive translation pipeline, leveraging both human expertise and the multilingual generation capabilities of LLMs. The language set for XCOMPS aligns with the prior knowledge probing benchmarks, such as BMLAMA-17 (Qi et al., 2023) and KLAR (Wang et al., 2025), ensuring consistency in multilingual evaluation. The highly structured nature of conceptual minimal pair datasets, where positive and negative sentences primarily consist of two components–concepts and properties–enabled us to design a multi-step translation process that ensures high-quality multilingual data.

The construction process consists of four stages. We use the GPT-4o model (GPT-4o-2024-08-06) via the OpenAI API as the translation assistant in the pipeline. In the first stage, we manually translated the original concepts and properties from English into German and Chinese using language experts. We used German and Chinese as additional seed languages to further reduce ambiguity, This multilingual seed data helped disambiguate concepts that might otherwise be unclear in translation. For example, the English word "bat" could refer to either the flying animal or the sports equipment. By including the German term "Schläger" and the Chinese term "球拍", which both unambiguously refer to the sports equipment, we ensured that the intended concept was accurately captured during translation.

In the second stage, we used LLMs to expand the seed data into the remaining 15 languages. LLMs were tasked with translating the concepts and properties, leveraging their multilingual machine translation capabilities. By providing seed data in three languages (English, German, and Chinese), we enhanced the LLMs' ability to generate accurate translations, as the additional context reduced the likelihood of semantic errors.

In the third stage, human experts for each target language manually reviewed and corrected the translated concepts and properties. This step ensured that the translations were accurate, culturally appropriate, and semantically aligned with the original dataset. Human intervention was particularly critical for low-resource languages, where LLMs often struggle with semantic precision in translation tasks.

Finally, in the fourth stage, LLMs were employed to generate complete sentences based on the verified concepts and properties. This step involved formulating positive and negative sentence pairs, which can be viewed as a straightforward language manipulation task. By providing the translated concepts and properties as input, we enabled the LLMs to focus on generating fluent and grammatically correct sentences, leveraging their strengths in multilingual text generation. This approach ensured that the most challenging aspect of the task–accurate translation of concepts and properties–was already resolved, allowing the LLMs to produce high-quality outputs.

By splitting the process into property translation and sentence generation, using multilingual seed data to reduce ambiguity, and combining human expertise with LLM capabilities, we ensured the quality and consistency of the XCOMPS dataset. This human-LLM interactive translation pipeline demonstrates how LLMs' multilingual understanding and generation capabilities can be effectively harnessed to construct high-quality multilingual benchmarks.

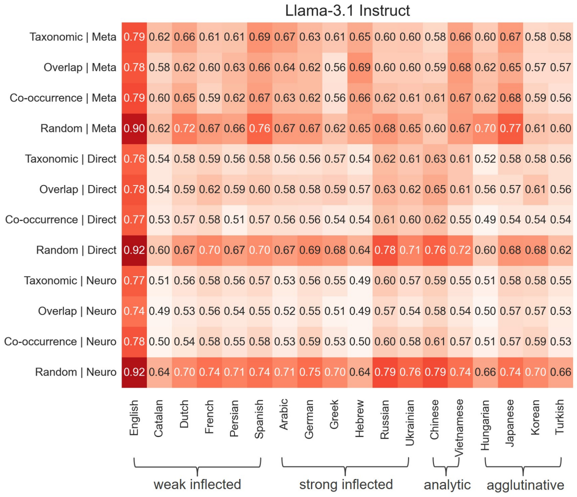

Figure 1: Metalinguistic prompting (meta), direct probability measurement (direct), and minimal pair probing (neuro) results on XCOMPS. The meta method evaluates LLMs' language performance; the neuro method evaluates LLMs' language competence, and the direct method falls in between. Languages are grouped according to morphological typology. Neuro-probing is a layer-wise method, and here we use the max value across all layers to compare with Meta and Direct.

4 Experiment Setup

4.1 Model

We use meta-llama/Llama-3.1-8B-Instruct from Hugging Face in our experiment, which applies instruction tuning to the base model for more intuitive user-prompt handling. During the inference, we adopt float16 precision to minimize computational resource consumption while maintaining model performance.

4.2 Evaluation

For **Meta**, we present both sentences of a minimal pair within a single prompt. We convert the target property into a question and compare the probabilities assigned to acceptable vs. unacceptable concepts. Figure 2 in Appendix A shows the prompts used in the experiment. For **Direct**, we compute sentence probabilities directly from the model's logits. A prediction is considered correct if the model assigns a higher probability to the valid sentence within each minimal pair. For **Neuro**, we adopt last-token pooling to represent each sentence, extracting the final token's hidden state from every layer. This approach ensures coverage of all preceding tokens (Meng et al., 2024). We then apply a logistic regression classifier for probing, using the F1 score (averaged over five cross-validation folds) as our primary evaluation metric.

4.3 Results and Analysis

Cross-linguistic variability in conceptual reasoning. From Figure 1, we observe that the model can perform relatively well on English conceptual tasks but show marked declines for low-resource languages. Notably, some languages with limited training data (e.g., Hungarian, Catalan) exhibit greater deterioration than others, indicating that cross-linguistic generalization of conceptual understanding is far from uniform. Even within the low-resource category, the degree of performance drop varies, underscoring that LLMs' semantic reasoning is neither universally stable nor equally supported by existing multilingual corpora. These patterns reinforce the idea that conceptual capabilities learned in English do not necessarily transfer seamlessly to languages that differ typologically or have weaker representations in training data.

Models excel at distinct conceptual contrasts but falter with subtler differences. High scores all appear in Random rows, where the negative concept is clearly distinct (e.g., "toaster" vs. "wren"), and the model easily detects mismatches. In Taxonomic, Overlap, or Co-occurrence rows, however, performance drops because the negative concepts share subtle semantic similarities (e.g., "toaster" vs. "coffee maker"). This indicates that the models may rely on conspicuous cues rather than true conceptual reasoning.

Direct and neuro convergence. By comparing direct and neuro results in Figure 1, and from Figure 3 in Appendix A, we see high correlations across all negative types, indicating that direct measurements closely track the models' internal representations.

Higher morphological complexity, lower conceptual reasoning. Figure 4 in Appendix A shows that languages with greater morphological complexity (moving from Analytic to Inflected to Agglutinative) tend to yield lower concept-reasoning scores. This indicates that, as linguistic structure becomes more complex, it becomes harder for the models to capture concept-property relationships consistently.

5 Conclusion

In this work, we introduce the XCOMPS benchmark, which provides a multilingual conceptual minimal pair dataset for evaluating the language model's semantic understanding across 17 languages. This work reveals that while LLMs demonstrate surface-level multilingual capabilities, they lack a universal semantic reasoning mechanism that transcends language boundaries.

Limitation

While XCOMPS significantly advances the evaluation of multilingual conceptual understanding, certain limitations remain. First, although the dataset covers 17 typologically diverse languages, it does not encompass all linguistic families or low-resource languages, which may limit its generalizability to underrepresented languages. Second, the reliance on human-LLM interaction for data construction ensures high quality but introduces potential inconsistencies due to variations in human expertise and model outputs. Lastly, while XCOMPS focuses on conceptual understanding, it does not explicitly address other challenges in multilingual NLP, such as pragmatics or contextual reasoning. Despite these limitations, XCOMPS provides a robust foundation for assessing and improving LLMs' multilingual capabilities, and future work can extend its scope to address these areas.

Acknowledgement

We thank the anonymous reviewers for their valuable advice and feedback. This research was partially supported by DFG (German Research Foundation) grant SCHU 2246/14-1 and Munich Center for Machine Learning (MCML).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. *arXiv preprint arXiv:2105.03495*.

Tom B Brown, Benjamin Mann, Nick Ryder, Melvin Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33.

Susan Carey. 2000. The origin of concepts. *Journal of Cognition and Development*, 1(1):37–41.

Vittoria Dentella, Elliot Murphy, Gary Marcus, and Evelina Leivada. 2023. Testing ai performance on less frequent aspects of language reveals insensitivity to underlying meaning. *arXiv preprint arXiv:2302.12313*.

Barry J Devereux, Lorraine K Tyler, Jeroen Geertzen, and Billi Randall. 2014. The centre for speech, language and the brain (cslb) concept property norms. *Behavior research methods*, 46:1119–1127.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. *arXiv preprint arXiv:1903.03260*.

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.

Linyang He, Peili Chen, Ercong Nie, Yuanning Li, and Jonathan R Brennan. 2024a. Decoding probing: Revealing internal linguistic structures in neural language models using minimal pairs. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 4488–4497.

Linyang He, Ercong Nie, Helmut Schmid, Hinrich Schütze, Nima Mesgarani, and Jonathan Brennan. 2024b. Large language models as neurolinguistic subjects: Identifying internal representations for form and meaning. *arXiv preprint arXiv:2411.07533*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A systematic assessment of syntactic generalization in neural language models. *arXiv preprint arXiv:2005.03692*.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event knowledge in large language models: the gap between the impossible and the unlikely. *Cognitive Science*, 47(11):e13386.

Brenden Lake and Marco Baroni. 2018. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR.

Jean Matter Mandler. 2004. *The foundations of mind: Origins of conceptual thought*. Oxford University Press.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. 2024. Sfr-embedding-mistral:enhance text retrieval with transfer learning. Salesforce AI Research Blog.

George A Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. Comps: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2920–2941.

Roma Patel and Ellie Pavlick. 2021. Mapping language models to grounded conceptual spaces. In *International conference on learning representations*.

Steven T Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-lingual consistency of factual knowledge in multilingual language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 10650–10666, Singapore. Association for Computational Linguistics.

Mingyang Wang, Heike Adel, Lukas Lange, Yihong Liu, Ercong Nie, Jannik Strötgen, and Hinrich Schütze. 2025. Lost in multilinguality: Dissecting cross-lingual factual inconsistency in transformer language models. *arXiv preprint arXiv:2504.04264*.

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.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

A Appendix

Table 2 shows the detailed information of the languages covered by XCMOPS. Figure 2 displays the prompt templates of different languages used for metalinguistic prompting evaluation. Figures 3 and 4 show detailed experimental results.

lid	**language**	**Typology**	**Family**
ar	Arabic	Inflectional	Semitic
ca	Catalan	Inflectional	Indo-European (Romance)
de	German	Inflectional	Indo-European (Germanic)
el	Greek	Inflectional	Indo-European (Hellenic)
es	Spanish	Inflectional	Indo-European (Romance)
fa	Persian	Inflectional	Indo-European (Iranian)
fr	French	Inflectional	Indo-European (Romance)
he	Hebrew	Inflectional	Semitic
hu	Hungarian	Agglutinative	Uralic
ja	Japanese	Agglutinative	Isolate
ko	Korean	Agglutinative	Isolate
nl	Dutch	Inflectional	Indo-European (Germanic)
ru	Russian	Inflectional	Indo-European (Slavic)
tr	Turkish	Agglutinative	Turkic
uk	Ukrainian	Inflectional	Indo-European (Slavic)
vi	Vietnamese	Analytic	Austroasiatic
zh	Chinese	Analytic	Sino-Tibetan

Table 2: Detailed information of the languages covered by XCOMPS.

```
en: Which concept is most likely to have the following property: "{property}", "{word1}" or "{word2}"? Answer: "
ar: أي مفهوم من المرجح أن يكون لديه الخاصية التالية: "{property}"، "{word1}" أو "{word2}"؟ الإجابة: "
he: איזה מושג סביר ביותר שיש לו את המאפיין הבא: "{property}", "{word1}" או "{word2}"? תשובה: "
fa: کدام مفهوم به احتمال زیاد دارای ویژگی زیر است: "{property}"، "{word1}" یا "{word2}"؟ پاسخ: "
de: Welches Konzept hat am wahrscheinlichsten die folgende Eigenschaft: "{property}", "{word1}" oder "{word2}"? Antwort: "
zh: 哪个概念最有可能有如下特征: "{property}", "{word1}" 还是 "{word2}"? 回答: "
fr: Quel concept est le plus susceptible d'avoir la propriété suivante : "{property}", "{word1}" ou "{word2}" ? Réponse : "
nl: Welk concept heeft waarschijnlijk de volgende eigenschap: "{property}", "{word1}" of "{word2}"? Antwoord: "
es: ¿Qué concepto es más probable que tenga la siguiente propiedad: "{property}", "{word1}" o "{word2}"? Respuesta: "
ja: どの概念が次の特性を持つ可能性が最も高いですか: "{property}"、"{word1}" または "{word2}"? 答え: "
ko: 어떤 개념이 다음 속성을 가질 가능성이 가장 높습니까: "{property}", "{word1}" 또는 "{word2}"? 답변: "
vi: Khái niệm nào có khả năng nhất có thuộc tính sau: "{property}", "{word1}" hoặc "{word2}"? Câu trả lời: "
el: Ποια έννοια είναι πιο πιθανό να έχει την ακόλουθη ιδιότητα: "{property}", "{word1}" ή "{word2}"? Απάντηση: "
hu: Melyik fogalomnak van a legnagyobb esélye, hogy rendelkezik a következő tulajdonsággal: "{property}", "{word1}" vagy "{word2}"? Válasz: "
tr: Hangi kavramın şu özelliğe sahip olma olasılığı daha yüksektir: "{property}", "{word1}" veya "{word2}"? Cevap: "
ca: Quin concepte és més probable que tingui la següent propietat: "{property}", "{word1}" o "{word2}"? Resposta: "
uk: Яке поняття найімовірніше має таку властивість: "{property}", "{word1}" чи "{word2}"? Відповідь: "
ru: Какое понятие с наибольшей вероятностью обладает следующим свойством: "{property}", "{word1}" или "{word2}"? Ответ: "
```

Figure 2: Prompt templates of different languages used for metalinguistic prompting.

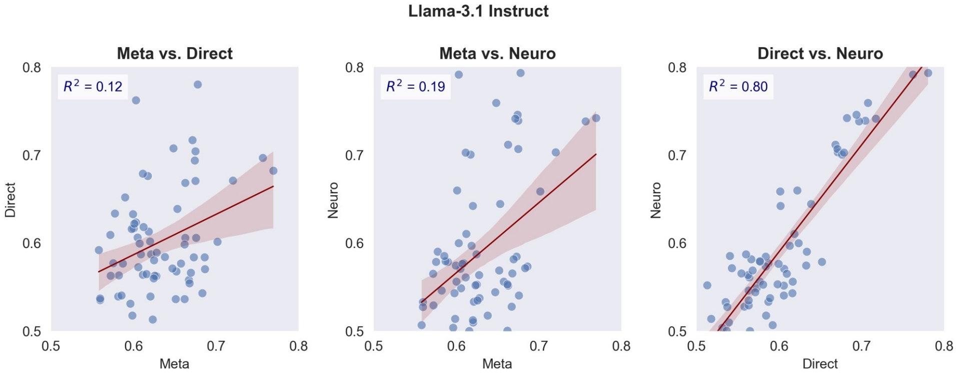

Figure 3: Linear correlation among meta, direct, and neuro evaluation results for all four tasks.

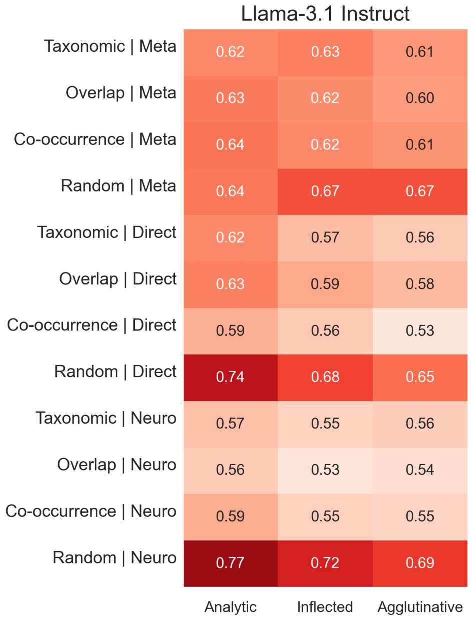

Figure 4: Averaged results across different language types. English results are dropped to make the comparison more reliable among low-resource languages.