Linguini: A benchmark for language-agnostic linguistic reasoning

Eduardo Sánchez[⋆ †] **Belen Alastruey**[⋆] **Christophe Ropers**[⋆] **Arina Turkatenko**[⋆]
Pontus Stenetorp[†] **Mikel Artetxe**[‡] **Marta R. Costa-jussà**[⋆]
[⋆]Meta [†]University College London
[‡]University of the Basque Country (UPV/EHU)
{eduardosanchez, alastruey, chrisropers, costajussa}@meta.com
p.stenetorp@cs.ucl.ac.uk mikel.artetxe@ehu.eus

Abstract

We propose a new benchmark to measure a language model's linguistic reasoning skills without relying on pre-existing language-specific knowledge. The test covers 894 questions grouped in 160 problems across 75 (mostly) extremely low-resource languages, extracted from the International Linguistic Olympiad corpus. To attain high accuracy on this benchmark, models don't need previous knowledge of the tested language, as all the information needed to solve the linguistic puzzle is presented in the context. We find that, while all analyzed models rank below 25% accuracy, there is a significant gap between open and closed models, with the best-performing proprietary model scoring 24.05% and the best-performing open model 8.84%.

1 Introduction

Recently, language models have shown impressive multilingual skills (Xu et al., 2024), achieving state of the art results in several tasks, such as machine translation (OpenAI, 2024), bilingual lexicon induction (Brown et al., 2020) and cross-lingual classification (Xue et al., 2021). However, the sometimes steep increase in performance of these tasks has led to saturation of popular benchmarks, such as MMLU (Hendrycks et al., 2021), where SotA performance has gone from 60% in December 2021 (Rae et al., 2022) to 90% in December 2023 (Gemini Team, 2024), providing diminishing returns when it comes to quantifying differences between models.

Moreover, in the case of linguistic reasoning, the task of evaluating a model's linguistic skills is often tied to the comprehensive knowledge a model has of a certain language (most commonly, English), making it difficult to evaluate a model's underlying linguistic skills beyond language-specific knowledge.

To address these issues, we introduce Linguini[1], a linguistic reasoning benchmark. Linguini consists of linguistic problems which require meta-linguistic awareness and deductive reasoning capabilities to be solved instead of pre-existing language proficiency. Linguini is based on problems extracted from the International Linguistic Olympiad (IOL)[2], a secondary school level contest where participants compete in solving Rosetta Stone-style problems (Derzhanski and Payne, 2010) relying solely on their understanding of linguistic concepts. An example of the type of challenges and the reasoning steps needs to solve it can be seen in Figure 2.

[1]The dataset is available at https://github.com/facebookresearch/linguini

[2]The problems are shared only for research purposes under the license CC-BY-SA 4.0. The problems are copyrighted by ©2003-2024 International Linguistics Olympiad

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

We evaluate a list of open and proprietary models on Linguini, showing a major gap between open and closed language models, in favor of the latter. We also conduct a series of experiments aiming at understanding the impact of contextual information in the accuracy obtained in the benchmark, performing both form (transliteration) and content (removing context) ablations, with results showing a main reliance on the context to solve the problems, minimizing the impact of language or task contamination in the models' training sets.

2 Related Work

There has been an increasing number of articles focusing on evaluating reasoning in language models (Chang et al., 2024). In the area of mathematical reasoning, Qin et al. (2023) analyze models' arithmetic reasoning, while Frieder et al. (2023) leverage publicly-available problems to build GHOSTS, a comprehensive mathematical benchmark in natural language. Bang et al. (2023) include symbolic reasoning in their multitask, multilingual and multimodal evaluation suite. Wu et al. (2024) and Hartmann et al. (2023) show that current language models have profound limitations when performing abstract reasoning, but Liu et al. (2023) indicate promising logical reasoning skills; however, performance is limited on out-of-distribution data. Multi-step reasoning is assessed by Chain-of-Thought Hub (Fu et al., 2023) and ThoughtSource (Ott et al., 2023), pointing out the limitations of language models in complex reasoning tasks.

Coverage of linguistic reasoning, which can be defined as the ability to understand and operate under the rules of language, has been limited in evaluation datasets for language models. One of the earliest examples is PuzzLing Machines (Şahin et al., 2020), which presents 7 different patterns from the Rosetta Stone paradigm Bozhanov and Derzhanski (2013) for models to perform exclusively machine translation. Chi et al. (2024) replicate Şahin et al. (2020)'s approach, manually creating a number of examples to avoid data leakage. Recently, some approaches have leveraged long context capabilities of language models to include in-context linguistic information (e.g. a grammar book (Tanzer et al., 2024) and other domain-specific sources (Zhang et al., 2024)) to solve different linguistic tasks. For large-scale linguistic reasoning evaluation, Big-Bench (Lewkowycz et al., 2022) includes a task linguistic mappings[3], relying on arbitrary artificial grammars to perform logical deduction. This approach is limited by its reliance on constructed languages instead of natural languages, which overlooks more complex underlying properties of languages, (e.g., voicing rules). Moreover, Waldis et al. (2024) present Holmes, a comprehensive benchmark for linguistic competence in English language. Finally, Bean et al. (2024) concurrently introduced a linguistic benchmark based on the UK Linguistic Olympiad, with a language coverage of around 90 high and low resource languages and a limited script (Latin-only) and language family coverage with respect to IOL.

3 Benchmarking linguistic reasoning

To overcome the previous limitations, we built a dataset where, in most cases, a model has no information about task language outside of the given context. To achieve this, we worked with problems extracted from the International Linguistic Olympiad.

3.1 IOL

The International Linguistic Olympiad (IOL)[4] is a contest for students up to secondary school level, where contestants must compete solving problems based on their understanding of linguistics (Derzhanski and Payne, 2010). The presented problems are formulated following the Rosetta Stone paradigm and present participants with challenges related to a variety of (mainly) extremely low-resource languages that students are not expected to be familiar with. The goal is for participants to leverage their linguistic skills rather than their foreign language knowledge. The IOL has been held yearly since 2003 (with the exception of 2020), and every year includes 5 short problems (to be solved individually) and 1 long, multipart problem (to be solved in groups). Problems are formulated in English and in several languages (up to 25 languages for the 2023 edition). The IOL corpus is available on their website in different formats of PDF with questions and correct answers,

[3] https://github.com/google/BIG-bench/blob/main/bigbench/benchmark_tasks/linguistic_mappings/

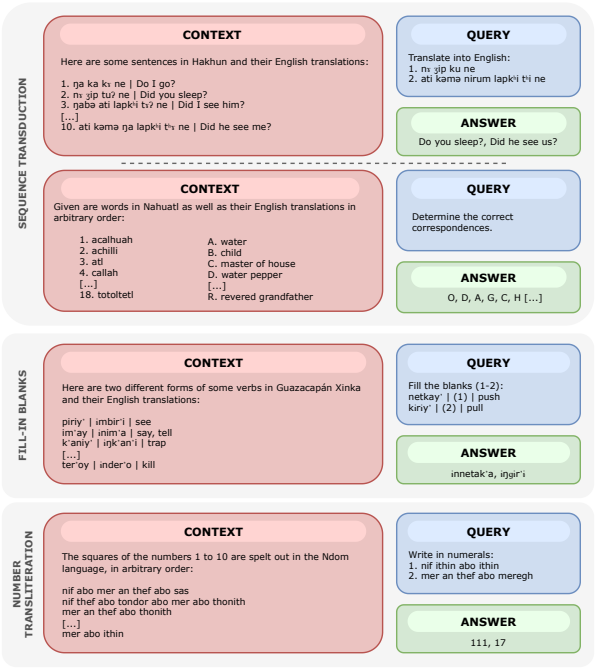
[4] https://ioling.org

explanations of some answers and total marks for each problem. Beyond IOL, there are regional contests (e.g. Asia Pacific Linguistic Olympiad[5] and The Australian Computational and Linguistics Olympiad[6]) that award places for the IOL.

3.2 Selecting problems

To select the types of questions for the dataset, we built a taxonomy exploring the IOL from 2003 to 2023. We excluded all instances for which their category only appears once; those where the question includes an image or those where the response is only an explanation. The remaining problems require solving different linguistic reasoning tasks, such as morphosyntactic segmentation (eg., verb conjugation), morphosemantic alignment (e.g., noun negation), derivation (e.g., finding cognates in related languages), morphophonological segmentation (e.g., pluralization) or graphophonemic transcription (e.g., transcription from one script to another). In total, Linguini is composed by 894 questions grouped in 160 problems across 75 (mostly) extremely low-resource language. A question denotes an item that corresponds to a single answer, usually related to other items by a common context. A list of languages can be found in Appendix B. We classify the problems included in Linguini into the three categories according to their content: sequence transduction, fill-in-blanks and number transliteration. Figure 1 shows one example of each.

Figure 1: Examples of Linguini entries covering the three problems included in the dataset: sequence transduction, fill-in-blanks, number transliteration.



Sequence transduction This category includes sequence production (identified in the benchmark as `'translation'`) and sequence matching (identified as `'match_letter'`). The problems require the model to transform a sequence into a different space (e.g., language, phonetic representation, script) based on few examples. In some cases, basic phonetic/phonological knowledge is needed. For example, the model should be able to reason over principles of voicing and their implementation in situations of coarticulation. Some problems require to know that consonants come in voiced-voiceless pairs, and that one element of the pair may in some cases be a substitute for the other element in the pair under certain circumstances.

[5] `https://aplo.asia`
[6] `https://ozclo.org.au`

Fill-in blanks Fill-in blanks are mainly morphophonological derivation tasks, and they are identified in the benchmark as `'fill_blanks'`. Models need to understand what are the morphophonological rules that make it possible to go from the first form of a word to its second form. This can usually be applied to verbal (e.g., verb tense conjugation), nominal or adjectival (e.g., case declension) derivation. It involves understanding affixation rules and morpheme swapping rules, which often come with phonological rules if there are different coarticulation phenomena with different affixes or phonotactic phenomena such as consonantal mutations.

Digit/text number transliteration These problems are identified by the labels `'text_to_num'` and `'num_to_text'`. In them, models have to produce a digit or text equivalent, respectively. They require a model's understanding of morphological analysis and morpheme order.

Figure 2: A subset of the context of a problem in Terenâ language and the reasoning steps needed to solve it. To correctly answer the question, the model must notice that (a) voiced *d* mutates to voiceless paired sound *t* (fortition), (b) *n* is dropped because there are no voiceless nasal alveolar sounds and (c) an epenthetic vowel has to be added between the mutation consonant and the rest of the word (a root), and that the vowel that gets added matches the aperture of the vowel in the root. If the aperture is closed, the epenthetic vowel is the closed front vowel *i*; if the aperture is mid, the epenthetic vowel is the mid front vowel *e*.

4 Experiments

We perform zero-shot to few-shot (0-5 in-context examples) evaluation across the whole dataset for an array of open and proprietary LLMs. Given the size of the benchmark, we employ a leave-one-out cross-validation scheme to maximize the number of in-context candidates per task. For every given inference, we include examples of the same format (e.g., `'translation'`, `'match_letter'`), but we exclude in-content examples of the same language to avoid language contamination.

Setup and Models We prompt models with an instruction, a context that provides information to unambiguously solve the linguistic problem and the problem itself. Scores of answers to each item of a problem are averaged to provide a single score (0-100) per task. We evaluate several major open LLMs and commercially available (behind API) SotA LLMs at the publication of this work. For open models, we conduct inference experiments in an 8 A100 GPUs node. An exhaustive list can be found in Appendix C.

Evaluation We use exact match (accuracy) as main evaluation criterion. Given the almost null performance on exact match of certain models, we also include chrF (Popović, 2015) as a *softer* metric. A low ChrF score indicates extremely low performance models, e.g. not understanding the domain of the task at hand.

5 Results and Discussion

Table 1 shows there's a gap between the best performing open model and the best performing proprietary model, with several tiers of proprietary models above the best open model (*llama-3-70b*). We also find mixed impact of in-context examples (ICEs) in the performance of the models. While some models benefit from it (such as *Llama 3 70b Instruct*), other models' performance degrades as the number of examples increases (such as *Claude 3 Opus*). This disparity might be due to the two factors introduced by the ICEs: from one side, they set an answer format that could be useful for models that can't infer it directly from a single natural language instruction and, from another side, they introduce tokens of languages potentially unrelated to the evaluated problem. It is possible that for models more capable of instruction following, only the second factor plays a role in the model's performance. Results of reasoning models reported by the community (Kazemi et al., 2025) on Linguini are also reproduced (marked with a $\star$). Overall, performance remains firmly below best reported results in IOL contests (above 82 points for every year). Although reasoning models (Besta et al., 2025) have a higher average performance than regular LLMs, they don't bring about a phase change for the task of linguistic reasoning, remaining comparable to the best-performing non-reasoning LLMs. We include results with chrF in Appendix E for reference.

Table 1: Exact match results with Linguini for 0-5 ICEs. Models marked with a $\star$ were reported by the community in Kazemi et al. (2025).

Model	0	1	2	3	4	5	Best(↑)
Claude 3 Opus	24.05	20.58	21.36	19.91	17.00	15.1	24.05
Gemini 2.5 Flash	23.15	-	-	-	-	-	23.15
DeepSeek R1 $\star$	19.50	-	-	-	-	-	19.50
o3-mini $\star$	17.00	-	-	-	-	-	17.00
Gemini 2.0 Flash $\star$	15.50	-	-	-	-	-	15.50
GPT-4o	14.65	12.98	13.87	12.98	13.98	13.76	14.65
GPT-4	6.38	9.96	11.52	12.98	11.74	13.31	13.31
Claude 3 Sonnet	12.30	8.95	10.29	10.40	9.28	8.72	12.30
Llama 4 Maverick	11.96	-	-	-	-	-	11.96
GPT-4 Turbo	8.72	9.40	9.96	7.49	8.61	9.96	9.96
Llama 3 70b	8.17	5.93	7.72	8.84	8.72	6.60	8.84
Llama 3 70b Instruct	4.81	5.93	7.16	7.38	6.82	8.39	8.39
Claude 3 Haiku	6.04	7.61	4.36	6.04	6.94	7.05	7.61
Distill R1 Qwen 32b $\star$	6.00	-	-	-	-	-	6.00
Llama 4 Scout	5.03	-	-	-	-	-	5.03
Llama 2 70b	4.70	2.24	2.57	3.24	3.36	3.58	3.58
Mistral 0.1 8x7b	2.46	3.47	3.91	3.02	3.24	3.47	3.91
Llama 2 70b Instruct	0.89	1.45	2.80	3.02	3.13	2.80	3.13
Gemma 2b	0.34	2.01	1.90	1.34	1.45	1.90	2.01
Qwen 1.5 110b Instruct	1.45	1.23	1.34	1.45	1.45	1.68	1.68

In addition to our main experiments, we performed a series of ablation studies to get a better insight of how language models perform linguistic reasoning.

5.1 No-Context Prompting

Given that we don't have information about training data for the majority of the analyzed models, we performed a series of experiments to study the degree in which models rely on the given context to provide correct answers. Models that have not been trained on any data of the task language should have a null-adjacent performance when not given the context necessary to solve the task. We analyze the impact of ignoring the context provided in the benchmark as a proxy of possible data contamination. The results are shown in Table 2.

We find steep performance drops for every model, which points towards a low likelihood of the language (or the training examples) being present in the models' training sets.

Table 2: No-context results. Δ denotes the differential vs regular zero-shot evaluation.

Model	ZS	No ctx	Δ
Claude 3 Opus	24.05	1.23	-22.82
GPT-4o	14.65	1.45	-13.20
Claude 3 Sonnet	12.30	2.01	-10.29
GPT-4 Turbo	8.72	1.45	-7.27
Llama 3 70b	8.17	1.67	-6.50
GPT-4	6.38	1.34	-5.04
Claude 3 Haiku	6.04	1.12	-4.92
Llama 3 70b Instruct	4.81	1.12	-3.69
Llama 2 70b	4.70	1.07	-3.63
Qwen 1.5 110b Instruct	1.45	0.43	-1.02
Mistral 0.1 8x7b	2.46	1.98	-0.48
Llama 2 70b Instruct	0.89	0.56	-0.33
Gemma 2b	0.34	0.09	-0.25

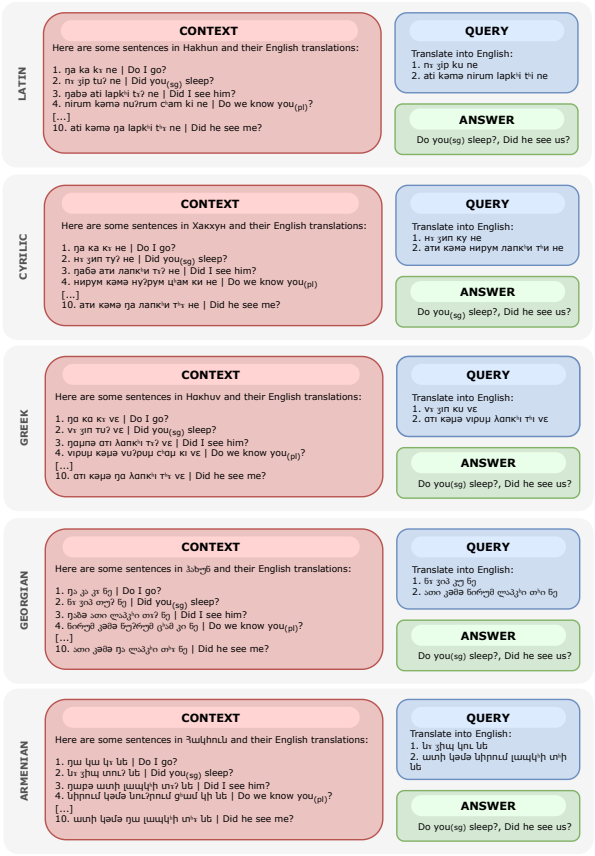


Figure 3: Example of transliteration of a problem into Cyrillic, Greek, Georgian and Armenian scripts.

5.2 Character-wise substitution

Since most problems are presented in Latin script, we wanted to understand whether the script in which the task languages are presented impact the performance on Linguini. But given that all information needed to solve the task is present in the context, the script should not have a major impact on the performance beyond encoding constraints. In other words, if the model doesn't rely on instances of the language (or the problem) in its training set, it should be able to solve the task in a

non-Latin script as well. We selected the best performing model (*Claude 3 Opus*) and transcribed the best performing problems (those where the *accuracy* >= 75) into 4 non-Latin alphabetical scripts (Cyrilic, Greek, Georgian and Armenian)[7]. An example of a transliterated problem can be found in Figure 3.

Given the difficulty of uniformly transcribing a diverse set of orthographic systems and diacritics, we opted for performing a character/bi-character-wise substitution of the standard Latin alphabet character, leaving non-standard characters with their original Unicode symbol. We filtered 17 well performing problems, and excluded one with a non-Latin script task language (English Braille). We performed transcriptions on the remaining 16 problems.

Table 3: Scores of selected problems with different language scripts for *claude-3-opus*.

Problem code & language	**Latn**	**Cyrl**	**Grek**	**Geor**	**Armn**
012023010100 (qda-gua)	75.00	100.00	75.00	100.00	0.00
012021020500 (zun)	100.00	0.00	100.00	0.00	0.00
012012030100 (eus)	78.57	7.14	92.86	0.00	0.00
012018020100 (nst-hkn)	83.33	83.33	66.67	83.33	100.00
012007050100 (tur)	75.00	75.00	50.00	37.50	50.00
012006020100 (cat)	75.00	50.00	50.00	58.33	33.33
012003030200 (eus)	100.00	100.00	75.00	100.00	100.00
012004010100 (txu)	100.00	100.00	66.67	66.67	33.33
012007030100 (kat)	80.00	13.33	6.67	100.00	0.00
012009050100 (nci)	83.33	83.33	83.33	83.33	50.00
012015020100 (kbd-bes)	100.00	66.67	100.00	66.67	83.33
012012050100 (rtm)	100.00	100.00	100.00	100.00	100.00
012011040200 (nci)	100.00	50.00	75.00	75.00	0.00
012013010200 (yii)	100.00	100.00	100.00	75.00	100.00
012012030200 (eus)	100.00	50.00	0.00	0.00	0.00
012012030300 (eus)	100.00	50.00	100.00	0.00	0.00
Average	85.71	56.12	65.31	63.27	38.78

Table 3 shows that the model retains the capacity to perform linguistic reasoning even after changing scripts, which backs the hypothesis of the model relying mainly on the presented context and not on spurious previous knowledge. The fact that for 13 our of 16 of the given problems there's at least one non-Latin script in which the model can solve the problem with greater or equal performance than with Latin script further supports this claim. Performance disparity among scripts could be related to either the difference in tokenization of different scripts or to the inherent limitations of our transliteration strategy (e.g. the Armenian script might lack a specific consonant cluster that needs to be developed to provide the right answer, and character/bi-character-wise substitution doesn't take this nuance into account).

5.3 Language resourcefulness and accuracy

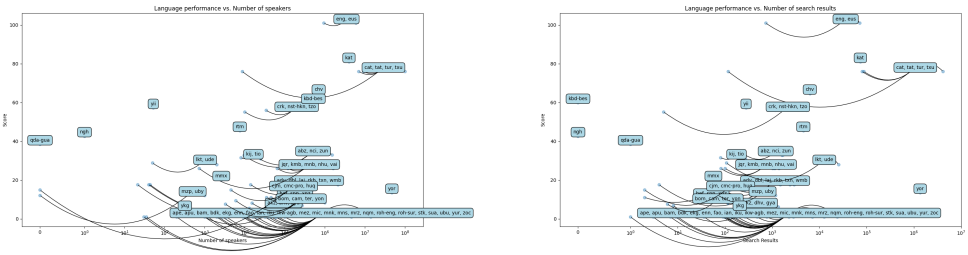
We were also interested in assessing whether higher-resource languages perform, on average, better than lower-resource languages. We use two metrics as proxies of language resourcefulness: number of speakers (Figure 4a) and online presence (Figure 4b), measured by Google searches).

We find the distribution to follow a uniform trend with respect to both metrics of language resourcefulness, which suggests that the accuracy isn't largely correlated to to its likelihood of being included in the training set. Notable exceptions to this trend are a number of very high-resource languages (e.g., Catalan, Euskera, Georgian, Turkish), which due to their institutional status are very likely to be included in the model's training set.

5.4 One-Book Prompting

Previous studies (Tanzer et al., 2024) have shown the capacity of language models to acquire some proficiency in the task of machine translation for an unseen language only through an in-context textbook. We leverage publicly available textbooks to scale Tanzer et al. (2024)'s analysis in number

[7]The mappings from Latin script to the rest can be found at `https://github.com/barseghyanartur/transliterate/`



(a) Accuracy vs. number of speakers. Data points are clustered for readability.

(b) Accuracy vs. number of Google searches. Data points are clustered for readability.

Figure 4: Accuracy as a function of (a) number of speakers and (b) number of Google searches.

of languages and types of tasks. We convert the textbooks in PDF format to raw text using the pdftotext library[8] and include them as context without any pre-processing. A list of employed textbooks can be found in Appendix D.

Table 4: Scores for a subset of examples evaluated with no problem context, with context, with a textbook and with a combination of both.

Language code	No-context	Context	Textbook	Context + Textbook
akz	0.00	5.13	0.00	3.85
apu	0.00	0.00	0.00	16.67
mnk	0.00	0.00	0.00	0.00
Average	0.00	1.71	0.00	6.84

Even thought in many cases the orthography of the task language greatly varies from the textbook to the problem and the PDF to text conversion introduces errors for highly diacritical text (as shown in Figure 5), the results in Table 4 show that a model can learn to model linguistic phenomena relying on a single in-context textbook.

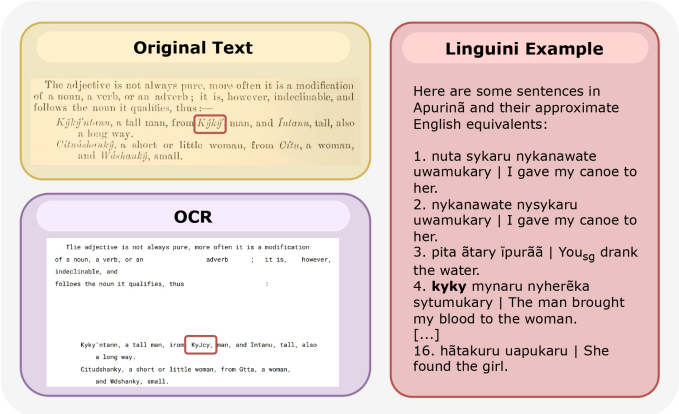


Figure 5: Example of transliteration of a problem into Cyrillic, Greek, Georgian and Armenian scripts. The discrepancies between the term *kyky* (English: *man*) in the original document (a scan from a 1894 grammar book of Apurinã language), its OCR conversion and the text of a problem in the benchmark are highlighted. In spite of the noise introduced by different orthographies and imperfect OCR, performance for Apurinã increases from 0% 16.67% with the full OCR text in-context.

[8] https://github.com/jalan/pdftotext

5.5 Human Evaluation

Given the potential limitations of automatic evaluation metrics, we performed human evaluations on the outputs of three models (*Claude 3 Opus*, *Llama 3 70b* and *GPT 4o*) with three annotators, asking them to rate the correctness of models' outputs in a scale of 0-4. The guidelines for human evaluation can be found in Appendix G.

Table 5: Average human evaluations, presented alongside average automatic evaluations.

Model	Annotator 1	Annotator 2	Annotator 3	Exact match	chrF
Claude 3 Opus	42.03	43.91	40.00	24.05	63.96
GPT-4o	30.00	35.47	25.63	14.65	58.99
Llama 3 70b	22.03	28.44	23.28	8.84	41.92

For the three selected models, the ranking provided by human evaluations is consistent both among each other and with the two automatic metrics (exact match and chrF). Since the human and automatic scores are not directly comparable (see Appendix G for the human evaluation guidelines), we look at their correlations to assess the validity of our selected automatic metrics.

Table 6: Correlation of metrics to rest of humans. The values were obtained by averaging the correlation of each metric verus the rest ofg human metrics (e.g., Human 1 is compared to Human 2 and 3, Exact match is compared to Human 1, 2 and 3). Full set of values in Figure 9 in Appendix F.

Model	Human 1	Human 2	Human 3	Exact match	chrF
Correlation to other humans	0.67	0.61	0.63	0.57	0.41

We find that the exact match score closely resembles the correlation to human average of each of the human annotators (Table 6), proving it a more appropriate metric to evaluate linguistic reasoning on Linguini problems with respect to chrF. The full correlation values among each human annotator and both evaluation metrics can be found in Figure 9 in Appendix F.

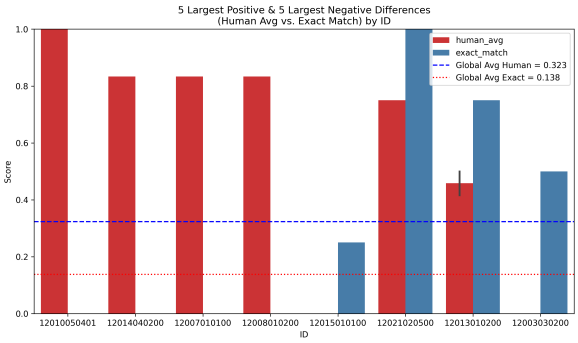
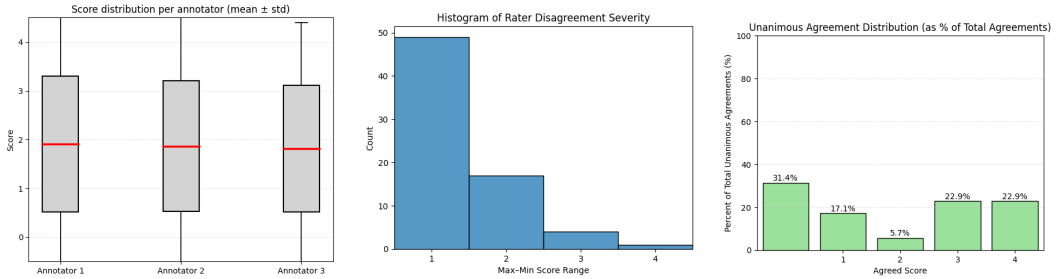


Figure 6: Largest score differentials between human and exact match evaluations. This figure shows the largest score mismatches between average human scores and Claude 3 Opus' exact match score.

We performed a qualitative analysis of problems with largest score differentials between human and exact match, the best performing automatic evaluation (Figure 6). Most relevant sources of disagreement included issues with diacritica, insertions of a single character or encoding issues (Braille script).

5.5.1 Annotator Agreement Analysis

We analyzed the consistency of scores among annotators. Figure 7a shows the score distributions for the three annotators, represented by mean-centered boxes and standard-deviation intervals. Although the general scoring trends are similar, some individual differences emerge, particularly around intermediate scores, suggesting variation in how partial correctness is interpreted. Across the 106



(a) Per-annotator score distributions (b) Disagreement severity histogram (c) Unanimous agreement distribution

Figure 7: Human-rater agreement analysis. (a) Score distributions per annotator, (b) overall disagreement severity, and (c) distribution of unanimous judgments by score.

```
Yagujba nyungu.
Gulugbi ga ngunybulugi.
Juwa gu bardba gijilulunguji.
Ngirrajba gunu janji.
Ngajbi ngiyinya alangani.
```

(a) Reference

```
Yagujba nyu ngiya.
Gulugbi nginyu ngunybulugi.
Bardbi gu juwani gijilulunguji.
Gijilulu gu.
Ngajbi ginyi alanga.
```

(b) Model output

Figure 8: Example of inter-annotator disagreement. (a) Reference text and (b) model output from *Claude 3 Opus*, rated as 0, 1, and 3 by the three annotators.

evaluated items, the mean inter-annotator range (*max–min*) was 0.93 ± 0.85. As shown in Figure 7b, most examples exhibited relatively low disagreement: 79.3% of items differed by at most one point, while only 20.8% differed by two or more, and 4.7% by three or more points. This suggests that, while exact unanimity is limited, raters generally stay within a narrow band of interpretation. Complete agreement (identical scores from all three annotators) occurred in 33.0% of examples. The distribution of these unanimous cases, shown in Figure 7c, is polarized: most agreements occurred at score 0 (31.4%), while the highest two scores (3 and 4) each accounted for 22.9%. Middle-ground consensus was rare (only 5.7% of unanimous cases were at score 2), indicating that raters found it easier to agree on outputs that were either clearly incorrect or clearly correct.

A representative case illustrating high disagreement is shown in Figure 8. For problem 012015040200, the model *Claude 3 Opus* received scores of 0, 1, and 3 from the three annotators. Two annotators appear to have emphasized exact lexical overlap with the reference, while the third rated the output more generously, likely rewarding partial correspondences across lines. This case illustrates how different evaluative emphases (lexical fidelity versus partial semantic adequacy) can lead to diverging scores, even under a well-defined rubric. Such patterns highlight the inherent difficulty of assigning a single "correct" human judgment to borderline outputs and the limitations of using gold human references for nuanced linguistic evaluation.

6 Conclusions

We presented Linguini, a new linguistic reasoning evaluation dataset. Our experiments show that Linguini provides a compact and effective benchmark to assess linguistic reasoning without relying on a substrate of existing language-specific knowledge. Subsequent experiments also show very low likelihood of dataset contamination in the analyzed models. Linguini can be relevant to the research community in the degree it is a benchmark in which high school-level humans are able to score >80, but frontier language models fail to score above 25, which means that linguistic reasoning is an axis of human intelligence not yet covered by generative models. Linguini shows a positive but weak correlation with other general-purpose benchmarks, with Claude 3 Opus scoring slightly less than GPT4-Turbo in the LMArena (1323 vs. 1324, at the time of writing in July 31st, 2025), but more than doubling its performance in Linguini (24.04 vs. 9.96). This suggests that skills related to linguistic reasoning are poorly represented in major LLM benchmarks.

References

AI@Meta. 2024. Llama 3 model card.

Anthropic AI. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. 2023. Qwen technical report.

Yejin Bang, Samuel Cahyawijaya, Nayeon Lee, Wenliang Dai, Dan Su, Bryan Wilie, Holy Lovenia, Ziwei Ji, Tiezheng Yu, Willy Chung, Quyet V. Do, Yan Xu, and Pascale Fung. 2023. A multitask, multilingual, multimodal evaluation of chatgpt on reasoning, hallucination, and interactivity.

Andrew M. Bean, Simi Hellsten, Harry Mayne, Jabez Magomere, Ethan A. Chi, Ryan Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. Lingoly: A benchmark of olympiad-level linguistic reasoning puzzles in low-resource and extinct languages.

Maciej Besta, Julia Barth, Eric Schreiber, Ales Kubicek, Afonso Catarino, Robert Gerstenberger, Piotr Nyczyk, Patrick Iff, Yueling Li, Sam Houliston, Tomasz Sternal, Marcin Copik, Grzegorz Kwaśniewski, Jürgen Müller, Łukasz Flis, Hannes Eberhard, Hubert Niewiadomski, and Torsten Hoefler. 2025. Reasoning language models: A blueprint.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8, Sofia, Bulgaria. Association for Computational Linguistics.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2024. A survey on evaluation of large language models. *ACM Trans. Intell. Syst. Technol.*, 15(3).

Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. ModeLing: A novel dataset for testing linguistic reasoning in language models. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 113–119, St. Julian's, Malta. Association for Computational Linguistics.

Ivan Derzhanski and Thomas Payne. 2010. The linguistics olympiads: Academic competitions in linguistics for secondary school students. *Linguistics at school: language awareness in primary and secondary education*, pages 213–26.

D Eberhard, G Simons, and C Fennig. 2020. Ethnologue: Languages of the world. twenty-third edition. dallas, texas: Sil international. online version:[inter-net]. ethnologue.

Simon Frieder, Luca Pinchetti, Alexis Chevalier, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Christian Petersen, and Julius Berner. 2023. Mathematical capabilities of chatgpt.

Yao Fu, Litu Ou, Mingyu Chen, Yuhao Wan, Hao Peng, and Tushar Khot. 2023. Chain-of-thought hub: A continuous effort to measure large language models' reasoning performance.

Gemini Team. 2024. Gemini: A family of highly capable multimodal models.

Gemini Team. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities.

Gemma Team. 2024. Gemma: Open models based on gemini research and technology.

Jochen Hartmann, Jasper Schwenzow, and Maximilian Witte. 2023. The political ideology of conversational ai: Converging evidence on chatgpt's pro-environmental, left-libertarian orientation.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K. Jain, Virginia Aglietti, Disha Jindal, Peter Chen, Nishanth Dikkala, Gladys Tyen, Xin Liu, Uri Shalit, Silvia Chiappa, Kate Olszewska, Yi Tay, Vinh Q. Tran, Quoc V. Le, and Orhan Firat. 2025. Big-bench extra hard.

Aitor Lewkowycz, Ambrose Slone, Anders Andreassen, Daniel Freeman, Ethan S Dyer, Gaurav Mishra, Guy Gur-Ari, Jaehoon Lee, Jascha Sohl-dickstein, Kristen Chiafullo, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. Technical report, Technical report.

Hanmeng Liu, Ruoxi Ning, Zhiyang Teng, Jian Liu, Qiji Zhou, and Yue Zhang. 2023. Evaluating the logical reasoning ability of chatgpt and gpt-4.

Llama 4 Team. 2025. The llama 4 herd: The beginning of a new era of natively multimodal ai innovation. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`. Accessed: 2025-10-23.

Karen Jacque Lupardus. 1982. *The language of the Alabama Indians*. University of Kansas.

OpenAI. 2024. Gpt-4 technical report.

Simon Ott, Konstantin Hebenstreit, Valentin Liévin, Christoffer Egeberg Hother, Milad Moradi, Maximilian Mayrhauser, Robert Praas, Ole Winther, and Matthias Samwald. 2023. Thoughtsource: A central hub for large language model reasoning data. *Scientific Data*, 10(1).

Jacob Evert Resyek Polak. 1894. *A Grammar and a Vocabulary of the Ipuriná Language*. 1. Published for the Fund By Kegan Paul, Trench, Trübner.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang. 2023. Is chatgpt a general-purpose natural language processing task solver?

Jack W. Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, Eliza Rutherford, Tom Hennigan, Jacob Menick, Albin Cassirer, Richard Powell, George van den Driessche, Lisa Anne Hendricks, Maribeth Rauh, Po-Sen Huang, Amelia Glaese, Johannes Welbl, Sumanth Dathathri, Saffron Huang, Jonathan Uesato, John Mellor, Irina Higgins, Antonia Creswell, Nat McAleese, Amy Wu, Erich Elsen, Siddhant Jayakumar, Elena Buchatskaya, David Budden, Esme Sutherland, Karen Simonyan, Michela Paganini, Laurent Sifre, Lena Martens, Xiang Lorraine Li, Adhiguna Kuncoro, Aida Nematzadeh, Elena Gribovskaya, Domenic Donato, Angeliki Lazaridou, Arthur Mensch, Jean-Baptiste Lespiau, Maria Tsimpoukelli, Nikolai Grigorev, Doug Fritz, Thibault Sottiaux, Mantas Pajarskas, Toby Pohlen, Zhitao Gong, Daniel Toyama, Cyprien de Masson d'Autume, Yujia Li, Tayfun Terzi, Vladimir Mikulik, Igor Babuschkin, Aidan Clark, Diego de Las Casas, Aurelia Guy, Chris Jones, James Bradbury, Matthew Johnson, Blake Hechtman, Laura Weidinger, Iason Gabriel, William Isaac, Ed Lockhart, Simon Osindero, Laura Rimell, Chris Dyer, Oriol

Vinyals, Kareem Ayoub, Jeff Stanway, Lorrayne Bennett, Demis Hassabis, Koray Kavukcuoglu, and Geoffrey Irving. 2022. Scaling language models: Methods, analysis & insights from training gopher.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

Richard Alan Spears. 1965. *The Structure of Faranah-Maninka*. Indiana University.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: Benchmark the linguistic competence of language models.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks.

Yuemei Xu, Ling Hu, Jiayi Zhao, Zihan Qiu, Yuqi Ye, and Hanwen Gu. 2024. A survey on multilingual large language models: Corpora, alignment, and bias.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions.

A Limitations, further work and broader impact

Evaluation of long in-context learning for linguistic reasoning is limited in this paper to a few languages, given the difficulties of finding publicly available grammar books. We plan to scale up the number of covered languages in further versions of the benchmark to perform a better encompassing analysis of long in-context learning.

Our dataset also lacks a curated list of explanations for each problem, which could be used as a basis to run chain-of-thought experiments and improve lingusitic reasoning skills of language models. We intend to engage with linguists and IOL organizers to fill this gap.

This benchmark intends to address and quantify the root of multilingualism, which in turn can impact the support of language models for the majority of world languages.

This paper includes the work of human annotators. Annotators were paid a fair rate. Each of the annotators signed a consent form agreeing to the usage of their annotations.

B Languages of Linguini

Table 7: Languages and their characteristics

Lang. Code	Language	No. Speakers[9]	No. Search Results[10]	Language Family	Script
abz	Abui	16,000	263	Trans-New Guinea	Latin
ady	Adyghe	425,000	2,370	Abkhaz-Adyghe	Latin
akz	Alabama	370	1,350	Muskogean	Latin
abz	Mountain Arapesh	16,000	98	Torricelli	Latin
apu	Apurinã	2800	264	Maipurean	Latin
bam	Bambara	14000000	7150	Niger-Congo	N'Ko
bdk	Budukh	200	126	Nakh-Daghestanian	Latin
bef	Bena Bena	45000	107	Trans-New Guinea	Latin
bom	Birom	1000000	115	Niger-Congo	Latin
cam	Cemuhî	3300	6	Austronesian	Latin
cat	Catalan	9200000	87100	Indo-European	Latin
chv	Chuvash	700000	6260	Turkic	Latin
cjm	Phan Rang Cham	491448	2	Austronesian	Latin
cmc-pro[11]	Proto-Chamic	0	267	Austronesian	Latin
crk	Plains Cree	34000	5290	Algic	Latin
dbl	Dyirbal	21	2900	Australian	Latin
dhv	Drehu	13,000	216	Austronesian	Latin
ekg	Ekari	100000	141	Trans-New Guinea	Latin
eng	English Braille	6000000	728	Indo-European	Latin
enn	Engenni	20000	185	Niger-Congo	Latin
eus	Basque	936,812	71100	Isolate	Latin
fao	Faroese	69000	23800	Indo-European	Latin
gya	Northwest Gbaya	267000	8	-	Latin
huq	Tsat	4500	128	Austronesian	Latin
ian	Iatmül	46000	9	Papua New Guinea	Latin
iku	Inuktitut	39,000	12500	Eskimo-Aleut	Latin
ikw-agb[11]	Agbirigba	30	1	Niger-Congo	Latin
jqr	Jaqaru	725	101	Aymaran	Latin
kat	Georgian	4000000	73700	Kartvelian	Latin
kbd-bes[11]	Besleney Kabardian	516000	0	Abkhaz-Adyghe	Latin
kij	Kilivila	25000	271	Austronesian	Latin
kmb	Kimbundu	1600000	1130	Niger-Congo	Latin
laj	Lango	2100000	1490	Nilo-Saharan	Latin
lkt	Lakhota	2000	25300	Siouan-Catawban	Latin
mez	Menominee	2000	2240	Algic	Latin
mic	Micmac	11000	774	Algic	Latin
mmx	Madak	2600	57	Austronesian	Latin
mnb	Muna	270000	1020	Austronesian	Latin
mnk	Maninka	4600000	478	Niger-Congo	N'Ko
mns	Mansi	2229	1490	Uralic	Latin
mrz	Coastal Marind	9000	100	Trans-New Guinea	Latin
mzp	Movima	1000	72	Isolate	Latin
nci	Classical Nahuatl	1500000	1690	Uto-Aztecan	Latin
ngh	Nluuki	1	0	Tuu	Latin
nhu	Nooni	64000	82	Niger-Congo	Latin
nqm	Ndom	1200	154	Trans-New Guinea	Latin
nst-hkn[11]	Hakhun	10000	5	Sino-Tibetan	Latin
qda-gua[11]	Guazacapán Xinka	0	1	Xincan	Latin
rkb	Rikbaktsa	40	54	Isolate	Latin
roh-eng[10]	Engadine	60000	7	Indo-European	Latin
roh-sur[11]	Sursilvan	60000	3	Indo-European	Latin
rtm	Rotuman	7500	4560	Austronesian	Latin
spp	Supyire	460000	45	Niger-Congo	Latin
stk	Arammba	1000	36	South-Central Papuan	Latin
sua	Sulka	3500	107	Isolate	Latin
tat	Tatar	7000000	79700	Turkic	Latin
ter	Terêna	15,000	115	Maipurean	Latin
tio	Teop	8000	81	Austronesian	Latin

Lang. Code	Language	No. Speakers	No. Search Results	Language Family	Script
tur	Turkish	100000000	4130000	Turkic	Latin
txn	West Tarangan	14,000	4	Austronesian	Latin
txu	Kayapo	8600	116	Jean	Latin
tzo	Tzotzil	550000	1160	Mayan	Latin
ubu	Umbu-Ungu	32,000	90	Trans-New Guinea	Latin
uby	Ubykh	0	1180	Abkhaz-Adyghe	Latin
ude	Udihe	50	108	Tungusic	Latin
vai	Vai	120000	1380	Niger-Congo	Latin
wmb	Wambaya	43	112	Australian	Latin
xnz	Kunuz Nubian	35000	2	Nilo-Saharan	Latin
yii	Yidiny	52	280	Australian	Latin
ykg	Tundra Yukaghir	320	206	Yukaghir	Latin
yon	Yonggom	6,000	48	Trans-New Guinea	Latin
yor	Yoruba	47000000	1360000	Niger-Congo	Latin
yur	Yurok	35	2830	Algic	Latin
zoc	Copainalá Zoque	10000	10	Mixe-Zoquean	Latin
zun	Zuni	9500	1610	Isolate	Latin

C Models

Table 8: Overview of Large Language Models

Name	API	Org	Size[12]	Open Weights	Ref
Claude 3 Opus	claude-3-opus-20240229	Anthropic	-	✗	Anthropic AI (2024)
Gemini 2.5 Flash	gemini-2.5-flash	Google	-	✗	Gemini Team (2025)
GPT-4o	gpt-4o-2024-05-13	OpenAI	-	✗	OpenAI (2024)
GPT-4	gpt-4-0125-preview	OpenAI	-	✗	OpenAI (2024)
Claude 3 Sonnet	claude-3-sonnet-20240229	Anthropic	-	✗	Anthropic AI (2024)
Llama 4 Maverick	-	Meta	400	✓	Llama 4 Team (2025)
GPT-4 Turbo	gpt-4-turbo-2024-04-09	OpenAI	-	✗	OpenAI (2024)
Llama 3 70b	-	Meta	70.6	✓	AI@Meta (2024)
Llama 3 70b Instruct	-	Meta	70.6	✓	AI@Meta (2024)
Claude 3 Haiku	claude-3-haiku-20240307	Anthropic	-	✗	Anthropic AI (2024)
Llama 4 Scout	-	Meta	109	✓	Llama 4 Team (2025)
Llama 2 70b	-	Meta	69.0	✓	Touvron et al. (2023)
Mistral 0.1 8x7b	-	Mistral	46.7	✓	Jiang et al. (2024)
Llama 2 70b Instruct	-	Meta	69.0	✓	Touvron et al. (2023)
Gemma 2b	-	Google	2.5	✓	Gemma Team (2024)
Qwen 1.5 110b Instruct	-	Alibaba	111.0	✓	Bai et al. (2023)

D Books

Table 9: Overview of Grammar Books

Lang	Title	Ref
akz	The Language of the Alabama Indians	Lupardus (1982)
apu	A Grammar and a Vocabulary of the Ipuriná Language	Polak (1894)
mnk	The Structure of Faranah-Maninka	Spears (1965)

Table 10: chrF results with Linguini for 0-5 ICEs. The order respects exact match ranking.

Model	0	1	2	3	4	5
Claude 3 Opus	63.96	58.26	58.5	53.17	49.01	46.55
GPT-4o	57.68	58.13	57.32	58.86	58.99	58.22
GPT-4	44.62	55.05	58.47	57.36	57.62	58.18
Claude 3 Sonnet	54.97	45.32	50.91	47.35	46.51	42.06
GPT-4 Turbo	52.89	50.82	50.03	50.94	49.98	51.79
Llama 3 70b	37.25	36.04	41.83	41.21	41.92	41.63
Llama 3 70b Instruct	45.35	42.65	43.89	45.99	48.07	51.08
Claude 3 Haiku	47.74	50.75	41.02	45.38	42.32	41.83
Llama 2 70b	45.3	35.39	34.06	35.54	36.21	36.44
Mistral 0.1 8x7b	42.0	34.8	38.01	37.57	37.64	37.63
Llama 2 70b Instruct	43.55	41.42	39.73	41.42	39.69	39.34
Gemma 2b	33.72	27.19	24.62	26.04	27.04	27.63
Qwen 1.5 110b Instruct	2.57	0.0	0.22	0.78	1.12	2.8

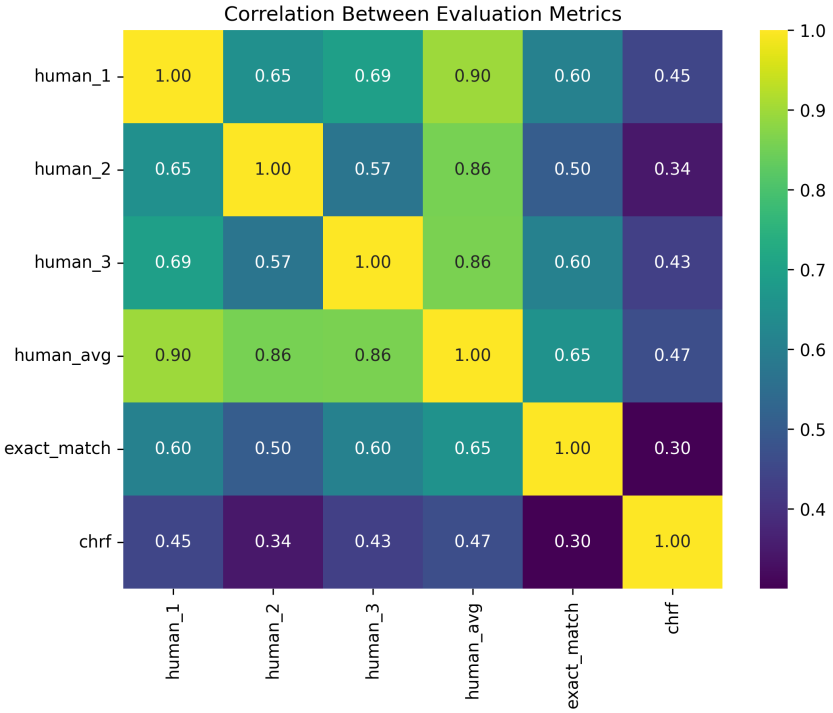


Figure 9: Correlations between human evaluations and automatic metrics.

E chrF Results

F Metric correlation

G Human evaluation guidelines

Guidelines were presented to human evaluators with the following content:

[9] According to Eberhard et al. (2020)
[10] Number of search results of the exact string "<Language name> language" using Google Seach API
[11] Language code not in ISO-639-3
[12] in billion parameter

G.1 Objective

We would like to know how well a machine learning model can solve a linguistic problem. For that, we need to obtain human opinion: we will provide the problems, the answer key and the answer given by the model.

G.2 Project Context

A linguistic problem is essentially a question about a low resource, rare language: one has to answer the question by finding patterns and links in the given language data.

G.3 Languages and volume

All problems will be IN English and will contain information about different languages. It is NOT expected that the evaluator speaks or even is aware of these languages. There are 160 problems in total. Each problem will need evaluation from 3 different people.

G.4 Annotator proficiency requirements

All annotators must meet ALL of the following requirements: Native English speaker OR excellent command of English (C2) High school graduate or higher

G.5 Task

You will be given a spreadsheet containing the following information: The problem itself with some context if needed The correct answer to the problem Output of a machine learning model A Output of a machine learning model B Output of a machine learning model C

You will need to compare the outputs A, B and C to the correct answer. To give your judgement, you will need to use a numerical Likert scale of 0-4 where 0: nothing in the machine output matches the correct answer. The output is completely wrong 1: only a very limited part of the machine output matches the correct answer. The correct parts seem accidental, the machine did not figure out any patterns or links, or only a few of them. 2: approximately half of the machine output is correct 3: Most parts of the machine output are correct, but there are some mistakes as well. The machine "understood" the logic in most cases, drew the correct conclusions and followed the correct patterns. 4: machine output and the correct answer match completely.

You will need to evaluate each model (A, B, and C) on this scale. You will also be asked to provide a short (one sentence is enough) explanation of your judgement.

G.6 Information on linguistic problems and their types

A linguistic problem can be seen as a little game, or puzzle, where one needs to understand the links and the patterns in the given data and use them to give an answer. The problems we will use for this task can be split into following categories:

G.6.1 "Fill-in-the-blanks" tasks

This type of problem usually entails the following: some data is given and some items correspond to others. The machine needs to find these links and fill in the blanks using the same logic as in the given data.

G.6.2 Match letters

This means that this problem will provide the solver only with two lists of words (or phrases): one in the target language, and one in English. The machine will need to match the translations correctly, having no additional info.

G.6.3 Match translations

These puzzles will first provide the solver with some information on how some words in a rare language are translated. Using this info, the machine will need to match the correct translations to the words in question.

G.6.4 Numbers to text

The machine will need to spell out the given numbers in a particular language, using the info given at the start of the task.

G.6.5 Text to numbers

The machine will need to understand which numbers are spelled out in the given language in this task, also using the info given at the start of the task.

G.6.6 Translation tasks

These tasks are asking for translating words or phrases, from English into the given language or the other way round.

As you can see, to solve these puzzles successfully, one needs to find patterns and links in the data. With this evaluation, we are trying to understand how well different ML models can do that. None of the models used was specifically trained to solve such tasks. We conceal the names of the models used to prevent bias - and also because they do not really add anything which could be useful for this evaluation.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Section 5

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section A

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: N/A

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: Section 1

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: Section 4

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: Computationally expensive

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Section 4

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Section A

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Section A

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: N/A

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Section 1

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: Section 1

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [Yes]

Justification: Section G (contractual contraints prevent compensation details from being released)

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: No personal information is collected from humans and no risks exist

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

Answer: [NA]

Justification: N/A

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.