Evaluating Large Language Models for In-Context Learning of Linguistic Patterns In Unseen Low Resource Languages

Hongpu Zhu and **Yuqi Liang** and **Wenjing Xu** and **Hongzhi Xu**
Institute of Corpus Studies and Applications
Shanghai International Studies University
{zhuhp,yuqiliang,xuwenjing,hxu}@shisu.edu.cn

Abstract

This paper investigates the ability of large language models (LLMs) to capture linguistic patterns from unseen languages and apply them to translation between the languages and English within an in-context learning framework. Inspired by the International Linguistic Olympiad (IOL), we create test data consisting of translation puzzles between 40 low-resource languages and English. We test the LLMs in two different strategies: direct prompting and step-by-step prompting. In the latter, the puzzles are manually decomposed into intermediate steps to allow LLMs to learn and apply linguistic rules incrementally. The results show that this strategy can significantly improve the performance of LLMs, achieving results comparable or slightly superior to humans when translating the unseen languages into English. However, LLMs still struggle with translating English into the unseen languages, typically with complex syntactic rules. We further observe that LLMs cannot deal with languages with object-subject and noun-adjective word order compared to others, reflecting the potential impact imposed by typological features of languages in training data. We have released our dataset on a public repository (Appendix A).

1 Introduction

Large Language Models (LLMs) have demonstrated impressive capabilities for in-context and few-shots learning tasks in natural language processing (Brown, 2020). Furthermore, they seem to exhibit reasoning abilities in areas such as mathematics and coding (Ahn et al., 2024). Despite these successes, LLMs still rely on large amounts of training data and computational resources to achieve practical performance. Like many other NLP systems, their applications in low-resource (LR) languages have been limited due to the scarcity of training data (Joshi et al., 2020). We are thus interested in how we can leverage their in-context learning and reasoning abilities to process LR language with minimal data.

Existing studies have explored ways of teaching LLMs to comprehend new languages through in-context learning and prompt engineering by providing supplementary linguistic knowledge (Cahyawijaya et al., 2024; Zhang et al., 2024) or retrieving extra examples from large corpora (Ginn et al., 2024). However, these methods remain insufficient, and LLMs still consistently underperform humans in various tasks. In addition, there is no systematic evaluation of how LLMs can generalize their linguistic skills to LR languages that are absent and typologically different from training data.

The current study investigates whether LLMs can learn and apply different linguistic rules (phonology, morpho-syntax, etc.) via in-context learning, and assesses how well they perform on translation tasks between English and LR languages with diverse typological features. LLMs are expected to rely on their intrinsic linguistic reasoning abilities rather than external knowledge or large corpora.

Inspired by the International Linguistics Olympiad (IOL) and its regional variants, we create a dataset covering 40 LR languages, which contains 168 manually constructed puzzles. The puzzles follow a "Rosetta Stone" format, where test-takers are given 10-15 exemplar sentences in a foreign language that is previously unknown to them, along with corresponding translations in their native language. Test-takers need to deduce linguistic rules from the examples and apply them by translating new ones.

Previous studies (Bean et al., 2024; Şahin et al., 2020) have shown that these puzzles from IOL are challenging for LLMs, and prompt engineering techniques such as chain-of-thought provides little improvement (Lin et al., 2023). We posit that the original puzzles might be too complicated for LLMs because several different rules are often in-

Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM 2025), pages 414–426
January 20, 2025. ©2025 Association for Computational Linguistics

volved in one puzzle, and the complexity prevents LLMs from recognizing meaningful patterns. Such complexity also limits detailed analysis of LLMs' strengths and weaknesses in linguistic reasoning.

To mitigate that, we take a step-by-step approach to let LLMs learn linguistic rules incrementally. The original puzzles are broken down into a series of smaller, more manageable ones, each targeting one specific linguistic rule. The principle is to start with simple sentences where LLMs can learn vocabulary and basic syntax, which are followed by sentences centering on morpho-syntax features such as tense or agreement, and finally complicated sentences where they need to combine all the rules together. We evaluate five state-of-the-art LLMs and compare their performance with that of 16 human testers with linguistic training.

LLMs have shown strong meta-linguistic competence, defined by Chomsky et al. (1976) as 'the knowledge of the characteristics and structures of language' in the major languages that they are trained in. However, it is not clear whether they can transfer such linguistic knowledge to unseen languages, and our approach aims to address that. We believe that our results can potentially facilitate research in LLMs and LR languages. If humans can benefit from meta-linguistic abilities when learning new languages, we shall expect the same for LLMs when dealing with LR languages as well, thus providing a future possibility of using LLMs in research of LR languages, such as annotation of LR data, producing glosses for linguists, developing machine translation systems, and so on.

In the following sections, we review previous work on evaluating LLMs' linguistic abilities and their performance in LR languages. We then describe our dataset and experiments, followed by a presentation and analysis of our results.

2 Related Works

One of the primary focuses of research in the field of LLMs is concerned with their reasoning capabilities. They have shown significant improvements over earlier counterparts, achieving promising performance on tasks such as mathematics (Yuan et al., 2023), geometry (Chen et al., 2022), automated theorem proving (Wu et al., 2023), code generation (Chen et al., 2021), and so on. In addition, they can perform tasks that they are not explicitly trained for, via in-context learning. This ability, first identified in GPT-3 (Brown, 2020), allows LLMs to learn and execute new tasks with just a few examples. While some studies suggest that LLMs may not truly "learn" and instead exploit superficial patterns in input examples (Min et al., 2022; Mirzadeh et al., 2024), the potential for generalizing beyond training data presents a new possibility for processing LR languages, where data scarcity has long been a challenge.

In light of such abilities, recent studies have explored the possibility of using LLMs as an alternative to fine-tuned models for machine translation in LR languages. For example, Tanzer et al. (2023) present Machine Translation from One Book, where LLMs are tasked to translate between English and Kalamang, an endangered language, using a grammar book as the primary resource. The authors show that LLMs can generate reasonable translations given lexical and grammatical descriptions, but are considerably inferior to humans in terms of grammatical consistency.

Efforts to improve LLM performance in this area are centered around prompt engineering techniques, such as providing LLMs with external linguistic knowledge. For example, when dealing with several LR languages, Su et al. (2024) show that LLMs prompted with grammatical description of the languages can sometimes outperform fine-tuned transformer models. Zhang et al. (2024) prompt LLMs with extra morphological gloss information, a dictionary, and a grammar book, when translating unseen LR languages to English, boosting the performance in few-shots translation from near 0 to around 10 in BLEU scores. More elaborate works have attempted retrieval-based methods. Ginn et al. (2024) use LLMs to produce interlinear gloss for LR languages, with examples retrieved from a corpus with carefully designed strategies. Although LLMs have not beaten SOTA supervised methods, they outperform basic fine-tuned transformer models. Similarly, Guo et al. (2024) build a framework to construct dedicated textbooks for LLMs, and retrieve vocabulary and syntactic patterns to teach LLMs unseen LR languages, achieving notable improvements on translation tasks.

As LLMs are increasingly applied to LR languages, understanding how well they generalize their meta-linguistic abilities becomes crucial, especially when dealing with languages that are typologically different from those in the training data. Many recent studies are focused on evaluating LLMs' linguistic skills across various phenomena. For instance, Waldis et al. (2024) introduce the Holmes benchmark to assess language models' understanding of syntax, morphology, semantics, and discourse. However, the study only examines English, leaving the question of how well LLMs generalize in cross-lingual situations unanswered.

To address this gap, several researchers turn to linguistic puzzles from IOL, which offer an opportunity to test LLMs' ability to infer and apply linguistic rules in unfamiliar languages. Şahin et al. (2020) propose the PuzzLing Machines dataset, with around 100 Rosetta Stone puzzles from IOL covering 81 languages. While statistical and neural models at the time scored near zero on these problems, GPT 3.5 achieved significantly better results. Prompting strategies, such as tree of thought, provide little improvement (Lin et al., 2023). Chi et al. (2024) create the MODELING dataset, featuring 48 puzzles across 19 LR languages. They handcraft these puzzles instead of using puzzles from IOL. Their problems focus on four features, namely basic word order, noun-adjective order, possession, and mapping vocabulary. Bean et al. (2024) present the LINGOLY benchmark with puzzles in diverse formats and categories from the UK Linguistic Olympiad, while Sánchez et al. (2024) introduce Linguini, covering 75 LR languages with various puzzle types collected from IOL. Results also show that larger, proprietary models generally outperform smaller, open-source ones.

Prior in-context learning framework of LR languages have mostley relied on external knowledge or corpora. Evaluation of intrinsic abilities of LLMs using IOL puzzles consistently report a low accuracy of 25-30% across all models, and prompting strageties show little improvements. These IOL puzzles often involve linguistic features in one puzzle, and models have to process semantic, morphology and syntax patterns at the same time, Our approach differs by decomposing such puzzles into smaller, more manageable ones focusing one rule at a time. We will show that by doing so, LLMs can take better advantage of their in-context learning and reasoning abilities, and the performance of translation tasks between unseen languages and English can be significantly improved.

3 Data

Our study builds upon the previous efforts and is aimed at addressing the limitations of existing approaches. We propose a step-by-step framework for linguistic reasoning that where LLMs learn linguistic rules one at a time over a multi-round conversation. Unlike the original IOL puzzles, which involves processing multiple linguistic rules across different levels (semantics, phonology, morphology, syntax) at the same time, our framework is built upon puzzles that focus on one rule at a time. This allows LLMs to learn the patterns incrementally and also allows for a more detailed analysis of LLMs' strengths and weaknesses in linguistic reasoning for unseen LR languages.

3.1 Data Source

We collect language puzzles in "Rosetta Stone" format from IOL and its regional variants, including the UK Linguistic Olympiads, the North America Computational Linguistics Open Competition, and the Asia-Pacific Linguistics Olympiads. These competitions are held annually for secondary students around the world. They expose students to a diverse range of rarely known languages and linguistic phenomena with puzzles in various formats. Their educational value in linguistics has been widely appreciated (Derzhanski and Payne, 2010).

A typical Rosetta Stone puzzle provides test-takers with 10-15 pairs of sentences in a foreign language and their mother tongue. The task is to observe these sentences, map the vocabulary, derive grammar rules, and then apply these patterns to translate new sentences (Bozhanov and Derzhanski, 2013; Littell et al., 2013). A full example is provided in Appendix B. These puzzles generally adhere to a few design principles:

- **Genuine**: All puzzles use authentic linguistic data from natural human languages.

- **Self-contained**: Each puzzle provides all the necessary information, and only the necessary information for solution.

- **Reasoning**: Solutions require at least one intermediate step of reasoning and cannot be acquired by simple analogy or intuition alone.

The original dataset collected from the above sources consist of 40 puzzles, representing 40 LR languages from 20 language families. A comprehensive list of languages is provided in Appendix D. The dataset includes a total of 525 training sentences and 335 testing sentences.

3.2 A step-by-step approach

Inspired by the chain-of-thought strategy (Wei et al., 2022), we develop a step-by-step approach, where LLMs learn one linguistic rule in one round of conversation as a "step". In each step, LLMs receive a simplified version of Rosseta Stone puzzle, and its training sentences are designed specifically for this rule. For example, for a puzzle targeting tense, the training sentences may describe the same action occuring at different time. In a multi-round conversation, LLMs go through many such steps to learn a complex set of linguistic rules. These steps follow a specific order described below:

1. **Lexical semantics and word order**: In the first step, puzzles involve goals of developing a vocabulary of the given language and understanding its basic syntax, such as word order. The training sentences consist of simple subject-verb-object sentences, and avoid variation in tense, person, etc. as much as possible.

2. **Phonology**: The second step involves phonological rules such as vowel harmony, tone changes, and allomorph. We create training examples consisting of base and derived forms of words, and models must deduce the phonological rules behind these derivations.

3. **Morpho-syntax**: This set of puzzles are concerned with rules about person, number, gender, agreement, tense, etc. Sentences are carefully constructed to provide sufficient information to represent the rules. Each puzzle focuses on only one particular rule or a few closely related rules.

4. **Syntax**: This set consists of puzzles with more complicated syntactic structures, including negation, questions, and clauses. They require the combination of all that have been learned in the previous steps.

We decompose each original IOL puzzle into 4-5 smaller ones following this order and handcraft new training sentences for them. Compared with the original ones, they are equivalent in terms of linguistic difficulty, but are significantly less complex. They require LLMs to deduce the same set of rules with a similar amount of limited samples (around 5-6 sentences for each rule), but allow LLMs to learn each one separately without interference from other rules. Figure 1 illustrates the genral idea and a full example is provided in Appendix C.

We also ensure that all the puzzles have only one possible solution. Sentences that can be interpreted in more than one possible ways are either not included or disambiguated. Since the original puzzles are available online, all the sentences in our constructed puzzles are different from those in the original ones, just in case that they might be present in LLMs' training data.

In the constructed dataset, the original 40 IOL puzzles are decomposed into 168 puzzles. Each puzzle comes with its own training and testing sentences and in total there are 1058 training sentences and 379 test sentences. Table 1 is the statistical information of our constructed dataset.

Category	Count
Lexical semantics and word order	40
Phonology	9
Morpho-syntax	93
Syntax	26

Table 1: The number of puzzles in our constructed dataset under each category.

4 Experiments

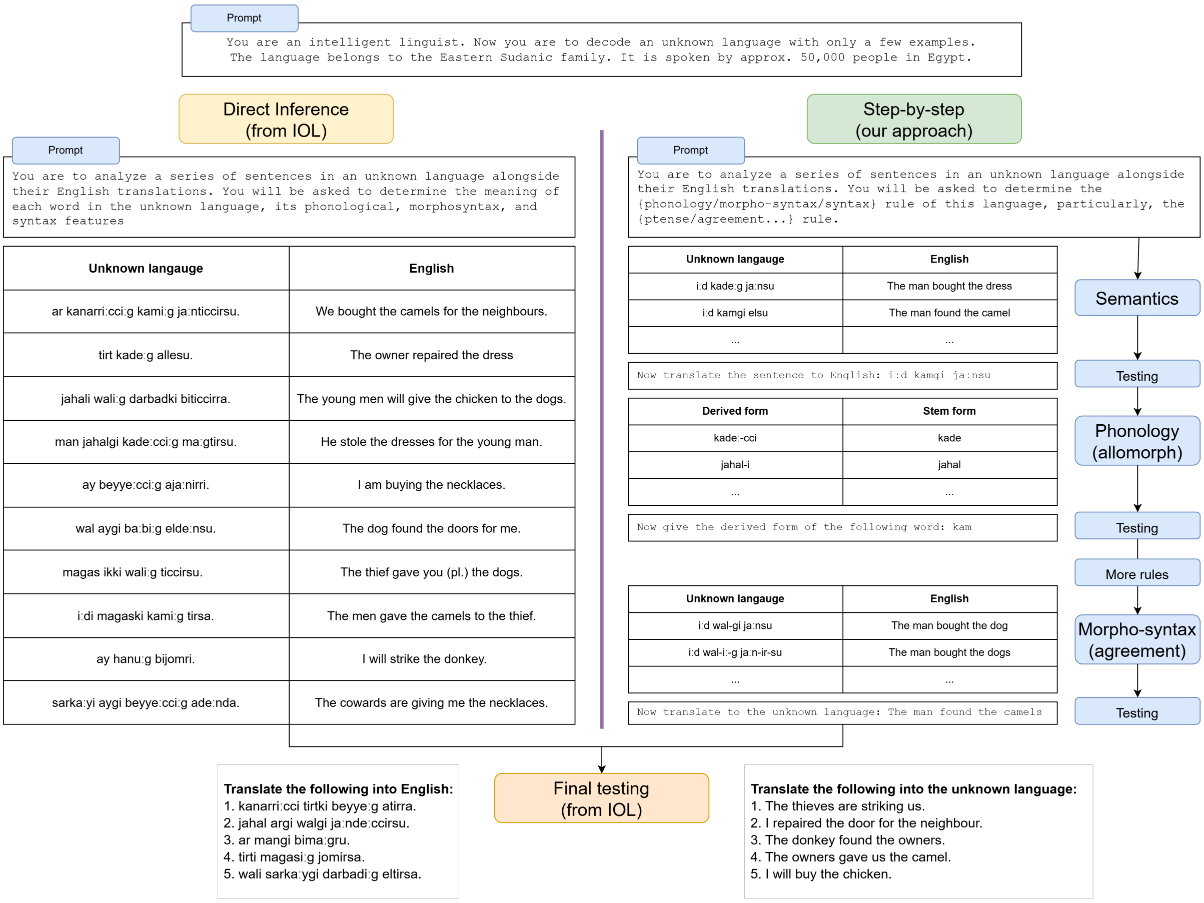
4.1 Tested models

We test 5 state-of-the-art LLMs with our dataset, including Claude 3.5 Sonnet (20240620), GPT-4o (20240816), Llama 3.1 405B, Llama 3.2 90 B, and Deepseek V2.5, covering both proprietary and open source models. Each model is provided with an introductory prompt explaining the task, as well as a brief description of the language, which includes its genealogical taxonomy, number of speakers and orthography explanations. The name of the language is omitted to prevent data leakage.

The LLMs are tested in two different settings, namely *step-by-step* and *direct-inference*. Let p and t represent training and testing data of an original IOL puzzle, and $p_1, \ldots, p_n$ and $t_1, \ldots, t_n$ stand for the step-by-step puzzles corresponding to the same original puzzle, the two experimental settings can be described as:

- **Direct-inference** The original puzzle including p and t are directly used as prompts for the LLMs. This setting serves as a baseline for comparison.

- **Step-by-step** For each original puzzle, the training examples of the corresponding small



```
You are an intelligent linguist. Now you are to decode an unknown language with only a few examples.
The language belongs to the Eastern Sudanic family. It is spoken by approx. 50,000 people in Egypt.
```

```
You are to analyze a series of sentences in an unknown language alongside
their English translations. You will be asked to determine the meaning of
each word in the unknown language, its phonological, morphosyntax, and
syntax features
```

Unknown langauge	English
ar kanarri:cci:g kami:g ja:nticcirsu.	We bought the camels for the neighbours.
tirt kade:g allesu.	The owner repaired the dress
jahali wali:g darbadki biticcirra.	The young men will give the chicken to the dogs.
man jahalgi kade:cci:g ma:gtirsu.	He stole the dresses for the young man.
ay beyye:cci:g aja:nirri.	I am buying the necklaces.
wal aygi ba:bi:g elde:nsu.	The dog found the doors for me.
magas ikki wali:g ticcirsu.	The thief gave you (pl.) the dogs.
i:di magaski kami:g tirsa.	The men gave the camels to the thief.
ay hanu:g bijomri.	I will strike the donkey.
sarka:yi aygi beyye:cci:g ade:nda.	The cowards are giving me the necklaces.

```
You are to analyze a series of sentences in an unknown language alongside
their English translations. You will be asked to determine the
{phonology/morpho-syntax/syntax} rule of this language, particularly, the
{ptense/agreement...} rule.
```

Unknown langauge	English
i:d kade:g ja:nsu	The man bought the dress
i:d kamgi elsu	The man found the camel
...	...

```
Now translate the sentence to English: i:d kamgi ja:nsu
```

Derived form	Stem form
kade:-cci	kade
jahal-i	jahal
...	...

```
Now give the derived form of the following word: kam
```

Unknown langauge	English
i:d wal-gi ja:nsu	The man bought the dog
i:d wal-i:-g ja:n-ir-su	The man bought the dogs
...	...

```
Now translate to the unknown language: The man found the camels
```

Translate the following into English:
1. kanarri:cci tirtki beyye:g atirra.
2. jahal argi walgi ja:nde:ccirsu.
3. ar mangi bima:gru.
4. tirti magasi:g jomirsa.
5. wali sarka:ygi darbadi:g eltirsa.

Translate the following into the unknown language:
1. The thieves are striking us.
2. I repaired the door for the neighbour.
3. The donkey found the owners.
4. The owners gave us the camel.
5. I will buy the chicken.

Figure 1: Illustration of our step-by-step approach and experimental settings.

puzzles, $p_1, \ldots, p_n$, are fed into the LLMs one by one in different rounds of the same session to let the LLMs learn patterns from them. Finally, the testing sentences of the original puzzle t are provided to test the LLMs in the same session.

4.2 Human performance

To examine if LLMs can achieve comparable performance to humans, we recruit 16 students with linguistic training to complete the test. To qualify, they must answer an example test puzzle correctly. Human participants follow the same procedure as the LLMs in the step-by-step setting.

4.3 Evaluation metrics

Since the performance is evaluated with translation tasks, we use three metrics commonly applied in machine translation evaluations:

- **BLEU-2**: We use bi-grams to calculate the BLEU scores. It is computed at the corpus level over the whole test set.

- **ChrF**: As many puzzles include morphological and phonological variations, we include ChrF as a character-level assessment. It is computed at the corpus level for each language and averaged across languages.

- **Exact Match (EM)** Exact matches are counted when the two sentences are exactly the same except for the punctuations and cases. This metric serves as a straightforward measure of accuracy.

5 Results and Discussion

5.1 Performance on the original IOL test set

We compare the models' performance in the two experimental settings on the original IOL test puzzles. The results shown in Table 2 and Figure 2 indicate that our step-by-step approach significantly boosts the performance in both translation directions across all LLMs, support our hypothesis that breaking down complex linguistic rules into steps allows LLMs to acquire these rules more effectively. Also, LLMs perform better in translating

Setting	Model	To English			To LR languages		
		BLEU	**ChrF**	**EM (%)**	BLEU	**ChrF**	**EM (%)**
Step-by-step	Claude 3.5 Sonnet	**76.374**	**82.010**	**41.493**	**62.352**	**77.238**	**27.463**
	GPT-4o	63.296	69.625	22.687	46.459	64.258	11.343
	Llama 3.1	58.452	65.648	15.224	45.842	64.928	12.836
	Llama 3.2	58.777	65.736	16.716	42.383	62.367	9.254
	Deepseek V2.5	59.751	66.819	18.209	45.288	62.720	10.448
	Human	68.351	73.608	35.220	54.605	68.289	21.590
Direct inference	Claude 3.5 Sonnet	**66.825**	**73.715**	**26.866**	**60.665**	**57.227**	**11.343**
	GPT-4o	42.972	53.260	6.866	31.303	48.470	2.687
	Llama 3.1	38.690	49.089	5.373	27.737	45.214	0.896
	Llama 3.2	36.639	46.356	4.985	24.201	38.460	1.216
	Deepseek V2.5	39.603	49.138	4.477	23.798	41.325	0.000

Table 2: LLM and human performance on the IOL puzzle test set in the two experimental settings.

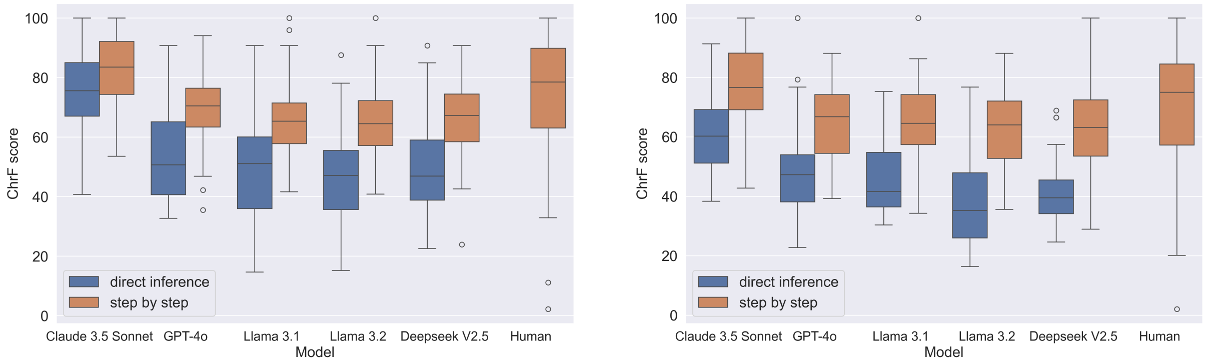


Figure 2: ChrF scores on the original IOL puzzle test set in the two experimental settings. Left: translating to English; right: translating to LR language

LR language to English than translation English to the LR languages.

In the step-by-step setting, Claude 3.5-Sonnet consistently outperforms other LLMs, and also surpasses human performance. Other models still lag behind humans considerably. Performance under the direct-inference setting is notably lower for all models, especially in exact match scores. Claude 3.5 Sonnet shows the smallest performance gap between the two settings and also the smallest gap between the two translation directions. In the direct-inference setting, while the performance of other models drops to near zero in terms of accuracy, Claude maintains scores comparable to other models in the step-by-step setting.

Among the LLMs, Claude 3.5 Sonnet demonstrates the highest performance across all metrics, followed by GPT-4o, which outperforms all the open-source models. Llama 3.1 (405B) outperforms its smaller counterpart, Llama 3.2 (90B). Deepseek V2.5, another open-source model, performs similarly to Llama 3.1.

5.2 Performance on step-by-step test set

5.2.1 Overall performance

To better analyze the strengths and weaknesses of LLMs on the task, we also report their performance on the test set of the 168 decomposed puzzles. Table 3 presents the overall performance of different models and humans. Again, translation quality to English consistently surpasses that of translation to LR languages. The best-performing LLM, Claude 3.5 Sonnet, achieves comparable and even better results compared to humans, while human testers consistently outperform all other LLMs.

Figure 3 shows the performance of the models and human with respect to each step in the reasoning task. As the number of steps increases, both the context length of the conversation and the complexity of the linguistic problem increase. For LLMs, this seems to impact their linguistic abilities more when translating English to the LR languages (represented by dashed lines), where performance declines as the steps increase. Conversely, when translating LR languages into English (solid lines),

Model	To English			To LR languages		
	BLEU	ChrF	EM (%)	BLEU	ChrF	EM (%)
Claude 3.5 Sonnet	**87.816**	**90.410**	**68.144**	**71.181**	**84.128**	**48.549**
GPT-4o	81.447	85.464	56.510	65.128	78.301	38.522
Llama 3.1 405B	80.320	84.480	58.449	62.667	75.550	38.259
Llama 3.2 90B	73.785	80.189	51.801	53.692	68.345	31.398
Deepseek 2.5	80.007	84.200	55.679	61.977	73.311	35.620
Human	86.204	88.840	66.040	69.368	81.827	**52.604**

Table 3: Overall performance of models and humans on test set in our contracted step-by-step dataset

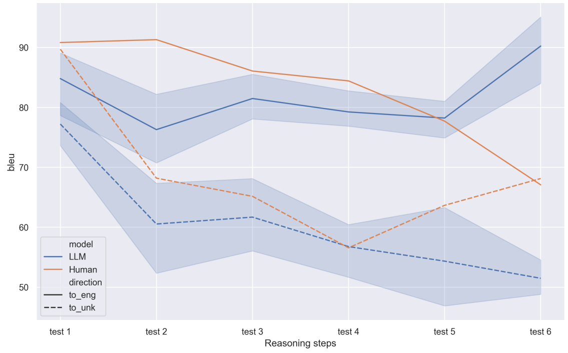


Figure 3: Average BLEU scores of humans and LLMs on test sets of each step.

the models demonstrate more resilience, with performance remaining relatively stable as complexity increases. This difference implies that LLMs are better equipped to handle familiar languages in linguistic reasoning. For humans, though the ChrF score generally decreases as complexity increases, the overall trend seems to be more robust.

5.2.2 Performance on puzzles in different categories

Figure 4 shows the ChrF scores of LLMs and humans across different categories of problems in our dataset. Full performance table in available in appendix F. When translating to English (left), LLMs generally perform well on simpler tasks like word semantics, and they demonstrate stronger reasoning abilities in morpho-syntax puzzles than in syntax puzzles. Humans show better performance than LLMs in syntax puzzles, and demonstrate similar performance in morpho-syntax and syntax puzzles.

When translating English to LR languages (right), both LLMs and humans achieve the highest scores in semantic problems, followed by syntax and morpho-syntax tasks, with phonological problems presenting the greatest challenge. Actually, the best model, Claude, score the lowest in terms of ChrF scores when dealing with phonological rules, and other models also underperform humans. In addition, LLM performance seems to show larger variance compared to humans in both translation directions.

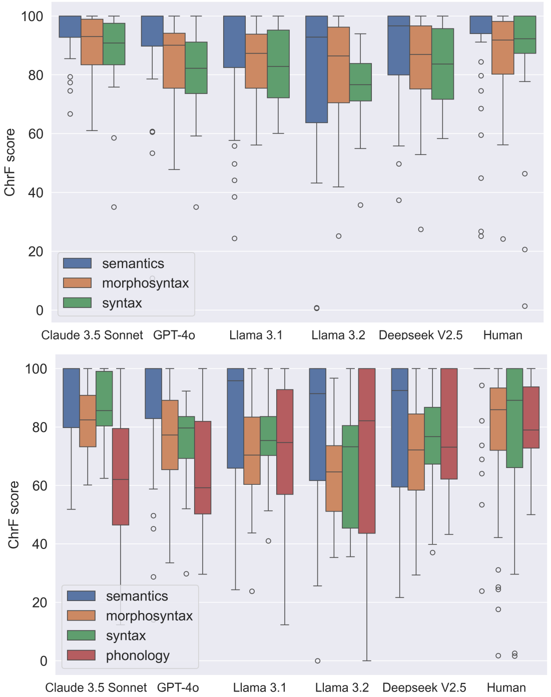


Figure 4: ChrF scores on puzzles of different categories in our test set. Up: to English, down: to LR languages

In terms of typological features, we discover an interesting phenomenon that LLMs struggle in certain word orders. Specifically, all models except Claude perform significantly worse in Object-Subject (O-S) languages than in Subject-Object (S-O) languages (see Figure 5) when translating to English, and three of the models, GPT-4o, Llama 3.2, and Deepseek also perform poorly when translating English to LR languages. Humans do not seem to show the same discrepancy with different orders. Additionally, both LLMs and humans tend to struggle with languages that follow a Noun-Adjective (N-A) order instead of an Adjective-Noun (A-N)

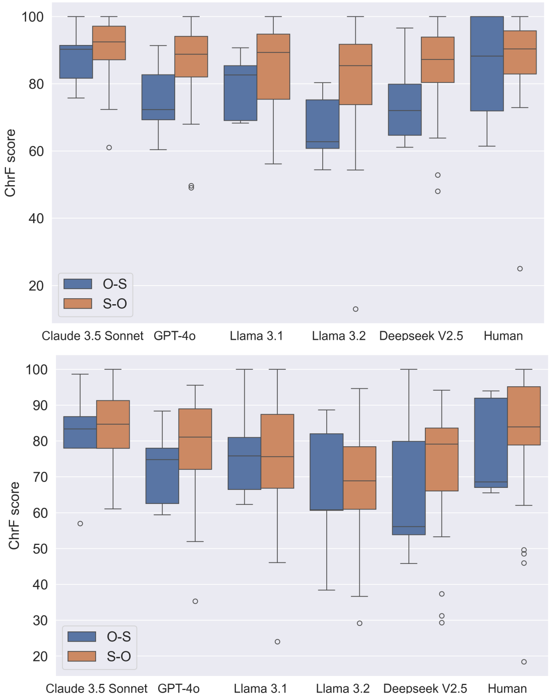


Figure 5: ChrF scores in O-S and S-O order languages. Up: to English; down: to LR languages

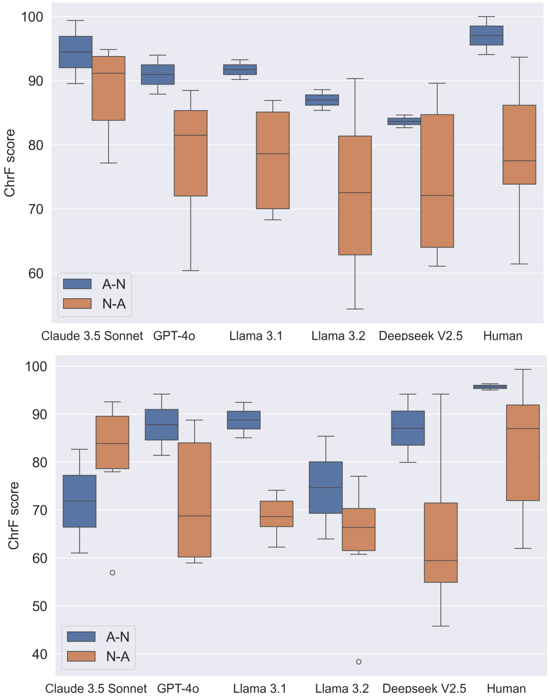


Figure 6: ChrF scores in N-A and A-N order languages. Up: to English; down: to LR languages

order, except for Claude when translating English to LR languages (Figure 6). This difference indicates a possible deficiency of processing certain word orders in some LLMs when comprehending LR languages.

5.3 Discussion

Our experiments reveal key insights into the linguistic reasoning capabilities of LLMs when dealing with diverse linguistic structures. Firstly, larger and proprietary models outperform smaller and open-source models. Secondly, all their performances decline as the complexity of the reasoning task increases. Thirdly, translation of English to LR languages presents a bigger challenge than the opposite direction. These findings are in line with the findings of previous studies. A probable cause of better performance in English is that LLMs are always able to generate coherent English sentences, regardless of whether they fully understand the rules in LR languages, but it is not the case for LR languages.

Overall, our step-by-step approach significantly enhances LLM performance in translating unseen LR languages to English. We show that they can infer linguistic rules from carefully constructed data with their intrinsic meta-linguistic abilities. In fact, the best model, Claude, even slightly surpasses human performance. Currently our approach relies on human-curated data, and this process might be automated in the future by formally describing the linguistic rules and the .

It is also shown that LLMs have different strengths and weaknesses compared to humans in terms of dealing with different categories of linguistic features. When translating to English, LLMs perform relatively well on simpler tasks such as word semantics, and they handle morpho-syntactic tasks more effectively than syntax. When translating to LR languages, both LLMs and humans achieve their highest scores on lexical semantic tasks, followed by syntax and morpho-syntax, with the worst performance on phonological tasks. An intriguing bias of LLMs is also revealed in our study, that they seem to have trouble processing O-S order and N-A order. The deficiency in processing O-S language is possibly attributed to a bias in the training data. However, the training data in fact contain N-A languages, like French, which are able to provide experience with this feature. This deficiency in N-A languages will need future investigations.

6 Conclusion

In general, this paper presents an evaluation of LLMs' ability to learn and apply complex linguistic rules across diverse language structures. Inspired by linguistic puzzles from IOL, we design a step-by-step approach for LLMs to learn linguistic rules in-context with their intrinsic meta-linguistic abilities. It involves creating a series of puzzles that allows LLMs to learn complex linguistic rules incrementally. The results show that our approach significantly boosts model performance in translation tasks, and the best model can match human level performance. We hope our dataset provides a starting point for future studies to further improve LLM performance and promote LLM applications in LR languages.

Limitations

While our approach provides insights into the linguistic reasoning capabilities of LLMs when dealing with unseen LR languages, several limitations may require further investigations. First, we have not conducted a systematic examination of how specific typological features affect model performance. We report preliminary findings with certain word orders, but further studies are needed to understand these biases, potentially using a wider variety of typological features. Also, a more detailed error analysis of the models' reasoning processes and translation results might further provide insights into their performance. We have relied on automatic evaluation metrics for measuring performance. If the translation results could be further annotated for types of different errors, it might be able to discover recurring patterns in these errors, thus revealing specific weaknesses in LLMs' linguistic reasoning abilities. Our results will also benefit from more extensive human testing and comparison with traditional machine translation systems, generic chain-of-thought prompting, or LLMs specifically desgined for reasoning, such as the O1 model.

Acknowledgment

This work is supported by funding from the Graduate Student Innovation Research Program of Institute of Corpus Studies and Applications at Shanghai International Studies University.

References

Janice Ahn, Rishu Verma, Renze Lou, Di Liu, Rui Zhang, and Wenpeng Yin. 2024. Large language models for mathematical reasoning: Progresses and challenges. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 225–237.

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. *arXiv preprint arXiv:2406.06196*.

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

Tom B Brown. 2020. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*.

Samuel Cahyawijaya, Holy Lovenia, and Pascale Fung. 2024. Llms are few-shot in-context low-resource language learners. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 405–433.

Jiaqi Chen, Tong Li, Jinghui Qin, Pan Lu, Liang Lin, Chongyu Chen, and Xiaodan Liang. 2022. Unigeo: Unifying geometry logical reasoning via reformulating mathematical expression. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3313–3323.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Nathan A Chi, Teodor Malchev, Ryan A Riley Kong, Lucas Huang, Ethan A Chi, R Thomas McCoy, and Dragomir Radev. 2024. Modeling: A novel dataset for testing linguistic reasoning in language models. *SIGTYP 2024*, page 113.

Noam Chomsky et al. 1976. *Reflections on language*. Temple Smith London.

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.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? *arXiv preprint arXiv:2406.18895*.

Ping Guo, Yubing Ren, Yue Hu, Yunpeng Li, Jiarui Zhang, Xingsheng Zhang, and He-Yan Huang. 2024. Teaching large language models to translate on low-resource languages with textbook prompting. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources* *and Evaluation (LREC-COLING 2024)*, pages 15685–15697.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293.

Zheng-Lin Lin, Chiao-Han Yen, Jia-Cheng Xu, Deborah Watty, and Shu-Kai Hsieh. 2023. Solving linguistic olympiad problems with tree-of-thought prompting. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 262–269.

Patrick Littell, Lori Levin, Jason Eisner, and Dragomir Radev. 2013. Introducing computational concepts in a linguistics olympiad. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 18–26.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11048–11064.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. 2024. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*.

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.

Eduardo Sánchez, Belen Alastruey, Christophe Ropers, Pontus Stenetorp, Mikel Artetxe, and Marta R Costa-jussà. 2024. Linguini: A benchmark for language-agnostic linguistic reasoning. *arXiv preprint arXiv:2409.12126*.

Jim Su, Justin Ho, George Broadwell, Sarah Moeller, and Bonnie Dorr. 2024. A comparison of fine-tuning and in-context learning for clause-level morphosyntactic alternation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 179–187.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: Benchmark the linguistic competence of language models. *arXiv preprint arXiv:2404.18923*.

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

Yiran Wu, Feiran Jia, Shaokun Zhang, Hangyu Li, Erkang Zhu, Yue Wang, Yin Tat Lee, Richard Peng, Qingyun Wu, and Chi Wang. 2023. An empirical study on challenging math problem solving with gpt-4. *arXiv preprint arXiv:2306.01337*.

Zheng Yuan, Hongyi Yuan, Chuanqi Tan, Wei Wang, and Songfang Huang. 2023. How well do large language models perform in arithmetic tasks? *arXiv preprint arXiv:2304.02015*.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: learning endangered languages in llms with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15654–15669.

A Data release

Our constructed dataset is available at URL: https://github.com/Zhurp2020/LR_LLM_Eval

B Example of a Rosetta Stone puzzle

Below you see romanised sentences in the Lakhota language and their English translations:

Lakhota	**English**
akhota ki wičhakte	The Indian killed them.
matho ki wakte	I killed the bear.
lakhota ki mačho	The Indian called me.
tuwa ničho he	Who called you?
wičhaša ki tuwa kte	The person killed someone.
tuwa hi he	Who came?
matho ki wičhačho	He called the bears.
yahi čha hi	You came, and he came.
matho ki hipi ną lakhota ki čhopi	The bears came and called the Indian.
yahi čha hokšila ki nikte	You came, and the boy killed you.
lakhota ki wačho ną hokšila ki wakte	I called the Indian and killed the boy.
hokšila ki wakte čha tuwa lakhota ki wičhačho	I killed the boy, and someone called the Indians.
lakhota ki hipi čha mayačho	The Indians came, and you called me.

Assignment 1. Translate into English:

1. wahi čha lakhota ki matho ki wičhačhopi
2. wičhaša ki nikte ną mačho
3. wičhaša ki nikte čha mačho
4. nikte

Assignment 2. Translate into English in all possible ways:

1. tuwa kte he

Assignment 3. Translate into Lakhota:

1. The Indians killed the boy, and the bear came.
2. You came and killed the Indian.
3. Whom did I call?
4. The people came, and someone killed them.

Note. The Lakhota (Dakota) language is of the Sioux family. It is spoken by 6000 people in the Midwest of the USA. š, č, h, y, w, ą are specific sounds of the Lakhota language.

C Example of puzzles in our step-by-step approach

Lakhota	**English**
train 1 (semantics)	
lakhota ki matho ki kte	The Indian killed the bear.
wičhaša ki hokšila ki kte	The man killed the boy.
lakhota ki hokšila ki čho	The Indian called the boy.
wičhaša ki hi	The man came.
test 1	
matho ki hokšila ki čho	The bear called the boy.
hokšila ki matho ki čho	The boy called the bear.

train 2 (morpho-syntax/object agreement)	
ma-kte	He killed me.
ni-kte	He killed you.
matho ki čho	He called the bear.
matho ki wičha-čho	He called the bears.

test 2	
ma-čho	He called me.
ni-čho	He called you.
matho ki kte	He killed the bear.
matho ki wičha-kte	He killed the bears.

train 3 (morpho-syntax/subject agreement)	
wa-kte	I killed him.
ya-kte	You killed him.
matho ki čho-pi	They called the bear.
test 3	
wa-čho	I called him.
ya-čho	You called him.
matho ki kte-pi	They killed the bear.

train 4(morpho-syntax/subject and object agreement)	
ma-ya-kte	You killed me.
ma-kte-pi	They killed me.
matho ki wičha-wa-kte	I killed the bears.
matho ki wičha-kte-pi	They killed the bears.
test 4	
ni-wa-kte	I killed you.
ni-kte-pi	They killed you.
matho ki wičha-ya-kte	You killed the bears.
matho ki wičha-kte	He killed the bears.

train 5 (syntax/interrogative and clause)	
ya-hi čha matho ki kte	You came, and he killed the bear.
ya-hi čha ma-čho	You came, and he called me.
matho ki ya-kte ną ma-ya-čho	You killed the bear and called me.
tuwa matho ki kte ną ma-čho	Someone killed the bear and called me.
tuwa ni-čho he	Who called you?
tuwa ya-čho he	Whom did you call?
test 5	
wa-čho čha hi	I called him and he came.
ma-čho ną hi	He called me and came.
tuwa kte-pi he?	Whom did they kill?

D Language list

See Table 4.

Language	Language family
Adyghe	Northwest Caucasian
Ainu	Isolate
Apurinã	Arawakan
Coastal Marind	Anim
Dyirbal	Pama–Nyungan
Engenni	Niger–Congo
Gilbertese	Austronesian
Hakhun	Sino-Tibetan
Inanwatan	Trans–New Guinea
Inuktitut	Eskaleut
Jarawara	Arawakan
K’iche’	Mayan
Kayapo	Macro-Jê
Kilivila	Austronesian
Kimbundu	Niger–Congo
Kombai	Trans–New Guinea
Kunuz Nubian	Nilo-Saharan
Lakhota	Siouan
Mairasi	Mairasi
Mee	Trans–New Guinea
Miskito	Misumalpan
Muklom	Sino-Tibetan
Muna	Austronesian
Nuuki	Tuu
Nahuatl	Uto-Aztecan
Niuean	Austronesian
Nooni	Niger–Congo
Panará	Macro-Jê
Pitjantjatjara	Pama–Nyungan
Sandawe	isolate
Taa	Tuu
Teop	Austronesian
Tutuba	Austronesian
Tzotzil	Mayan
Walman	Torricelli
Wambaya	Mirndi
Yonggom	Trans–New Guinea
Zou	Sino-Tibetan

Table 4: The list of 40 languages in our dataset

Category	Linguistic feature	Count
word order		40
Phonology	allomorph	6
	vowel harmony	3
	tone change	7
Morpho-syntax	alignment	15
	indirect object	7
	noun class	5
	noun gender	6
	noun number	11
	animate	2
	definitiveness	1
	proper name	1
	subject agreement	28
	object agreement	16
	focus	5
	possessive	19
	tense	25
	mood	2
	word derivation	6
	demonstrative	3
	causative	2
	locative	3
	reflective	3
	case	2
	adverb	6
	adjective	9
Syntax	interrogative	14
	negative	11
	expletive	3
	clause	5
	conjunction	2
	secondary order	1

Table 5: The list of 33 linguistic features covered in our data

E List of linguistic features covered in our data

See Table 5.

F Model performance in different categories of puzzles and word orders

See Table 6 and Table 7.

Category	Model	**To English**			**To LR languages**		
		BLEU	**ChrF**	**EM (%)**	**BLEU**	**ChrF**	**EM (%)**
Semantics	Claude 3.5 Sonnet	**94.164**	**95.414**	**81.707**	**83.640**	**90.261**	**67.073**
	GPT-4o	89.801	91.139	69.512	82.120	87.146	65.854
	Llama 3.1	86.542	87.435	67.073	76.803	83.011	59.756
	Llama 3.2	74.506	79.208	57.317	71.888	78.222	56.098
	Deepseek V2.5	88.690	88.676	67.073	77.022	79.366	53.659
	Human	92.008	89.981	82.927	91.221	93.691	82.927
Phonology	Claude 3.5 Sonnet				32.763	61.370	38.889
	GPT-4o				0.000	65.411	33.333
	Llama 3.1				29.369	69.233	50.000
	Llama 3.2				21.508	65.306	55.556
	Deepseek V2.5				0.000	76.377	50.000
	Human				35.355	78.774	55.556
Morphosyntax	Claude 3.5 Sonnet	86.020	89.277	66.063	65.962	81.718	42.986
	GPT-4o	79.594	84.179	57.014	59.602	76.592	33.484
	Llama 3.1	77.772	84.131	58.371	55.472	71.749	31.674
	Llama 3.2	74.857	81.362	54.299	45.086	63.433	21.267
	Deepseek V2.5	77.004	83.047	54.751	55.769	70.634	29.412
	Human	84.508	86.347	63.793	62.185	76.416	42.857
Syntax	Claude 3.5 Sonnet	83.844	86.859	56.897	74.530	86.357	46.552
	GPT-4o	76.986	80.354	36.207	64.772	74.105	20.690
	Llama 3.1	79.059	82.927	46.552	67.999	76.447	29.310
	Llama 3.2	69.738	75.865	34.483	60.532	66.880	27.586
	Deepseek V2.5	77.995	82.106	44.828	65.305	73.939	29.310
	Human	82.871	83.349	53.333	67.142	76.345	42.857

Table 6: Model performance in different categories of linguistic rules.

Word order	Model	**To English**			**To LR languages**		
		BLEU	**ChrF**	**EM (%)**	**BLEU**	**ChrF**	**EM (%)**
O-S	Claude 3.5 Sonnet	**83.161**	87.8183	45.255	58.201	80.75	32.374
	GPT-4o	66.923	75.2078	21.168	42.145	72.613	8.6331
	Llama 3.1	70.126	79.2061	26.277	46.778	77.108	15.108
	Llama 3.2	53.758	66.6992	19.708	33.551	66.089	11.511
	Deepseek V2.5	66.346	74.8481	18.248	37.049	67.133	11.511
	Human	77.622	84.323	39.583	61.863	77.413	30.612
S-O	Claude 3.5 Sonnet	88.52	90.3878	46.197	73.442	84.236	29.56
	GPT-4o	84.093	86.2946	30.649	69.09	78.575	19.89
	Llama 3.1	82.474	84.7276	27.293	65.392	74.817	18.681
	Llama 3.2	78.264	81.2239	25.889	57.364	68.052	15.265
	Deepseek V2.5	82.686	85.0603	28.3	66.445	73.62	16.923
	Human	87.832	86.726	54.027	70.747	82.004	39.185
A-N	Claude 3.5 Sonnet	91.591	94.4654	44.444	55.572	71.852	9.2593
	GPT-4o	88.059	90.9486	29.63	89.466	87.798	25.926
	Llama 3.1	88.614	91.7183	25.926	88.842	88.751	29.63
	Llama 3.2	84.38	86.9885	22.222	68.275	74.687	18.519
	Deepseek V2.5	80.042	83.6529	18.519	87.105	87.066	25.926
	Human	97.068	97.0304	50	89.789	95.678	38.889
N-A	Claude 3.5 Sonnet	85.774	88.3955	35.583	62.677	80.932	19.76
	GPT-4o	71.255	77.8957	15.951	50.399	71.962	8.3832
	Llama 3.1	69.977	77.8152	13.497	44.762	68.747	6.5868
	Llama 3.2	63.232	72.295	9.816	41.489	63.298	7.1856
	Deepseek V2.5	69.312	74.1964	14.724	46.202	64.639	6.5868
	Human	75.273	78.6403	25	64.186	82.652	26.667

Table 7: Model performance in different word orders.