LINGGYM: How Far Are LLMs from Thinking Like Field Linguists?

Changbing Yang[Œ], Franklin Ma[Œ], Freda Shi[ɯ,ʊ], Jian Zhu[Œ]
[Œ]University of British Columbia
[ɯ]University of Waterloo [ʊ]Vector Institute
cyang33@mail.ubc.ca, franklin.ma@ubc.ca
fhs@uwaterloo.ca, jian.zhu@ubc.ca

Abstract

This paper introduces LINGGYM, a new benchmark that evaluates LLMs' capacity for metalinguistic reasoning using Interlinear Glossed Text (IGT) and grammatical descriptions extracted from 18 typologically diverse reference grammars. Unlike previous work that focuses on specific downstream tasks, we assess whether LLMs can generalize linguistic inference across low-resource languages and structures not seen during training. We present a controlled evaluation task: Word-Gloss Inference, in which the model must infer a missing word and gloss from context using varying levels of linguistic information (e.g., glosses, grammatical explanations, translations). Our results show that incorporating structured linguistic cues leads to consistent improvements in reasoning performance across all models. This work highlights both the promise and current limitations of using LLMs for typologically informed linguistic analysis and low-resource language documentation.

1 Introduction

In recent years, researchers have been actively exploring how large language models (LLMs; OpenAI et al., 2024; Yang et al., 2024a; Grattafiori et al., 2024) can assist and accelerate scientific discoveries in various disciplines (e.g., Romera-Paredes et al., 2024; Merchant et al., 2023; Fawzi et al., 2022; Hayes et al., 2025; Zhang et al., 2024c). However, exploration on how LLMs can assist social sciences is relatively limited (Grossmann et al., 2023; Bail, 2024; Ziems et al., 2024). In particular, LLMs with the capacity of reasoning about metalinguistic knowledge have the potential to become powerful tools for language documentation, linguistic hypothesis testing, and typological research. For example, by generalizing linguistic structures such as morphology, syntax, and word order across languages, they can suggest morpheme segmentations and glosses, identify patterns or counterexamples to test hypotheses, and compare structural features across different languages.

Predicative adjectives

When adjectives function predicatively, they may receive copular morphology, although this is not obligatory (neither for adjectives nor for nouns). These predicative adjectives occur clause-finally (the position held prototypically by verbs).

Orthography: *Mï anmapïna.*
Segmentation: mï anma=p-na
Gloss: 3SG.SUBJ good=COP-IRR
Translation: 'It will be good.'

Figure 1: An excerpt explaining predicative adjectives in Ulwa, with an associated example as IGT from *A Grammar of Ulwa* (*Papua New Guinea*; Barlow, 2023, p. 166). The example IGT is represented with the Leipzig Glossing rules (Comrie et al., 2017). The text in red highlights the emphasized word discussed in the grammar explanation. The gloss consists of a third-person singular subject marker (3SG.SUBJ) for "it," and an irrealis copula (COP-IRR) marker for "will be."

On the other hand, while recent advances have shown impressive performance in high-resource languages like English, our understanding of their effectiveness on typologically diverse and underrepresented languages remains limited (Alhanai et al., 2025), mainly due to the overwhelming dominance of English and other high-resource languages in their training data (Blasi et al., 2022; Khade et al., 2025; Li et al., 2024; Wu and Dredze, 2020).

To explore how well LLMs can understand low-resource languages when provided with structured linguistic input, we turn to reference grammars (Mosel, 2006; Chelliah, 2013), which aim to comprehensively describe the structure of individual languages. Reference grammars offer two valuable types of information:

1. **Interlinear glossed text (IGT)**, a standard text format used by field linguists to present linguistic data, which is useful for tasks like morphological analysis, syntactic structure identification. IGT typically consists of four lines: a phonological or orthographic transcription, a segmentation of words into morphemes, corresponding grammatical glosses, and a free English translation. Conventions include hyphens to mark morpheme boundaries, equals signs for clitic boundaries, and periods to separate multiple glossing elements for a single morpheme (Comrie et al., 2017), as illustrated in Figure 1.
2. **Grammatical terms and explanations** embedded throughout the text, where important linguistic terms (e.g., tense markers, case particles, verb classes) are defined and contextualized within the grammar.

Together, these resources reflect the approach taken by human linguists, who analyze unfamiliar languages by studying structured descriptions rather than relying on raw corpora. Thanks to decades of documentation efforts, such materials are available for many endangered and low-resource languages, presenting a valuable opportunity to test LLMs' ability to reason over structured linguistic knowledge curated by experts. Unlike the unstructured web-scale corpora typically used to train LLMs, descriptive grammars hold a unique advantage by offering systematic and interpretable accounts of a language's morphology and syntax. In addition to serving human language learners and linguists, these structured frameworks that encode rich metalinguistic knowledge also offer a valuable resource for evaluating LLMs. By drawing on this explicit information, we can design targeted evaluation tasks that probe model performance across diverse linguistic phenomena and typological patterns.

In this work, we design a task-oriented approach: for each target sentence, the LLM receives the utterance or the utterance paired with its glosses, augmented by targeted grammatical cues (e.g., rules about verb conjugation or case marking). We evaluate the model's comprehension through a controlled task (Figure 2): **word-gloss inference**, where the model anwsers a multiple-choice question to infer a missing word or its corresponding gloss based on the linguistic context.

Our contributions are as follows: first, we present a cleaned and structured dataset of IGT examples drawn from 18 endangered and low-resource languages—these examples are extracted from publicly available reference grammars and subsequently verified by hand (§3.2, §4). Second, we develop an evaluation framework grounded in descriptive linguistic resources to assess how well LLMs can interpret and infer in low-resource languages using IGT data and grammatical rules (§4.2). Third, we benchmark multiple state-of-the-art LLMs on our proposed tasks and provide a typologically informed analysis of their performance, highlighting both capabilities and limitations when processing structured linguistic knowledge (§5.1, §5.2). We release the benchmark on GitHub.[1]

2 Related Work

2.1 NLP for Low-Resource languages

The value of language models as tools to assist language documentation and revitalization has been well recognized (Bird, 2020) in both linguistics and natural language processing (NLP) communities. These models enable a variety of applications, including automatic transcriptions of speech (Dunbar et al., 2017; Li et al., 2020; Samir et al., 2025; Zhu et al., 2025), low-resource speech synthesis (Kazantsevaa et al., 2024; Wang et al., 2025), automatic interlinear glossing (Moeller et al., 2020; He et al., 2024; Yang et al., 2024b), grapheme-to-phoneme conversion (Li et al., 2022; Zhu et al., 2022), and more (Gessler and Von Der Wense, 2024). Most existing work formulates a specific subtask in language documentation as an established NLP task with standard evaluation metrics.

While these directions have led to many low-resource NLP technologies, there are still many limitations (Bird, 2020). First, many models are trained on specific languages where training data is available, and are usually not generalizable to unseen languages. Second, many technologies are developed in highly artificial settings with well-defined tasks and clean data. As a result, they are unable to solve many linguistic tasks in real language documentation that are more complex, noisy, and subjective.[2] To bridge the gap, in this work, we present a benchmark in real language documentation scenarios and use it to assess the reasoning capabilities of general-purpose LLMs.

[1] https://github.com/changbingY/LingGym

[2] Language documentation is often subjective because linguists have different habits, preferences, and theoretical approaches to analysis, and there is no universally standardized method for representing or annotating linguistic data.

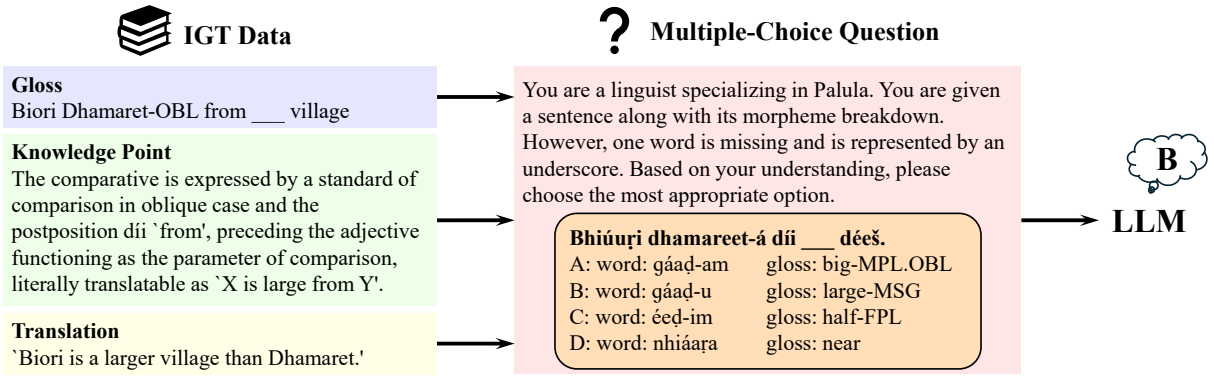

Figure 2: An illustration of how IGT data is transformed into a multiple-choice question for evaluating an LLM. One word is masked with an underscore in the provided sentence. An LLM takes the constructed prompt and the most contextually appropriate answer based on the linguistic information.

2.2 Assessing Linguistic Knowledge in LMs

Assessing the linguistic knowledge of LMs has long been a central topic in computational linguistics. Early studies focused on assessing the implicit linguistic knowledge of language models emerging from training, such as general syntactic knowledge (Gulordava et al., 2018; Goldberg, 2019; Hu et al., 2020; Wilcox et al., 2018), dependency structure (Hewitt and Manning, 2019; Manning et al., 2020), natural language inference (McCoy et al., 2019), and psycholinguistics judgments (Warstadt et al., 2020; Ettinger, 2020). This line of work centers mostly on English and only measure the implicit linguistic knowledge through proxies like probes, logits, and perplexity.

As LLMs' capacities continue to increase, research has shown that LLMs can follow explicit meta-linguistic concepts or learn languages from explicit meta-linguistic descriptions (Tanzer et al., 2024; Bean et al., 2024; Zhang et al., 2024b; Spencer and Kongborrirak, 2025; Zhang et al., 2024a; Ramos et al., 2024; Beguš et al., 2023), and can infer the underlying grammatical rules through concrete examples during in-context learning (Ginn et al., 2024). Yet, there is still large room for improvement in terms of linguistic reasoning, even for the state-of-the-art LLMs (Bean et al., 2024)—most importantly, existing approaches only deal with a handful of low-resource languages and mostly on machine translation tasks. It remains unclear if LLMs can perform abstract meta-linguistic reasoning across low-resource languages that are not seen during training. In this paper, we evaluate the extent to which LLMs can perform linguistic reasoning across a wide range of structural phenomena and generalize to unseen low-resource languages. The ability to make correct inferences would demonstrate the models' potential to support the analysis of previously understudied languages.

3 Data

We construct our benchmark from a collection of low-resource languages documented in publicly available reference grammars published by *Language Science Press* (LSP),[3] an open-access publisher of high-quality linguistic research. We select books from their *Studies in Diversity Linguistics* and *Comprehensive Grammar Library* series.

3.1 Reference Grammars

Reference grammars are comprehensive, systematic descriptions of individual languages, often based on fieldwork and long-term collaboration with native speakers (Mosel, 2006; Chelliah, 2013). Their goal is to capture linguistic intricacies through various examples and discussions of language use in diverse contexts. Typically, they address all major linguistic domains, including phonology, morphology, syntax, semantics, and pragmatics, providing valuable resources for theoretical research, typological comparison, language learning, documentation, and revitalization.

For instance, Carol J. Pebley and Thomas E. Payne authored *A Grammar of Kagayanen*: a Western Austronesian language spoken by around 30,000 people in the Philippines (Pebley and Payne, 2024). The work adopts a typologically informed descriptive framework inspired by Dixon's Basic Linguistic Theory (Dixon, 2009).

All grammar books used in this study are publicly available under the Creative Commons Attri-

[3] https://langsci-press.org/

bution 4.0 International License (CC BY 4.0).[4]

3.2 Data Preprocessing

Parsing the LaTeX source files. We retrieve the LaTeX source code of 18 reference grammars from their publicly available GitHub sites. To ensure the utility of each grammar for our benchmark, we filter each chapter's raw LaTeX source file against our criteria. Specifically, we retain languages that (i) include labelled sections (via `\label` tags) that correspond to grammatical rules or descriptive content, and (ii) contain IGT examples that are explicitly linked to these rule explanations. For each selected IGT instance, we require that a target keyword (typically a word, morpheme or form under discussion) be highlighted within the example, thereby allowing us to align example sentences with specific grammatical features.

We begin by converting the raw LaTeX files from each grammar into plain text, removing all formatting commands while retaining boldfaced keywords that indicate grammatical focus. Chapters that lack IGT examples, such as acknowledgments and appendices, are excluded in certain languages.

Categorizing individual chapters. Chapters are manually categorized into either phonology, morphology, syntax, semantics, pragmatics, or other linguistic subfields based on their introductory content (see Appendix F.2). We exclude chapters related to phonetics (if applicable) due to the lack of IGT content and inconsistent formatting of the symbols from the International Phonetic Alphabet.

Extracting IGT instances. After cleaning the LaTeX syntax, we extract structured IGT examples from each chapter, along with their preceding **knowledge points (KPs)**, which we define as explanatory paragraphs containing an IGT label tag and the grammatical rationale for the associated example (see Figure 1). In addition, we record the hierarchical metadata for each example, which includes the chapter title, section heading, and subsection heading. Each IGT instance is filtered based on structural markers such as label tags, transcription lines (when applicable), morpheme segmentation lines, glossing lines, and free translations. Figure 3 shows an example of Pichi (Yakpo, 2019) from the cleaning process. After this automated extraction, we perform manual cleaning

[4] https://creativecommons.org/licenses/by/4.0/ This license permits use, distribution, and adaptation of the materials, provided appropriate credit is given to the original authors and source.

```
  \label{ex:key:127}
  \gll Dí  gɛ́l  pikín  \textbf{ova}-\textbf{dráy}  ó.\\
this girl child over.\textsc{cpd}{}-be.dry \textsc{sp}\\

\glt 'This girl is really too lean.' [dj07ae 207]
\z
```

↓

```
\label{ex:key:127}
\gll   Dí    gɛ́l   pikín  \textbf{ova}-\textbf{dráy} ó.
\gls   this  girl   child  over.CPD-be.dry           SP
\glt   'This girl is really too lean.'
```

Figure 3: The top portion shows the raw LaTeX source of an IGT example in Pichi (Yakpo, 2019), where individual morphemes and glosses are annotated using various commands. The bottom portion shows the cleaned version after processing: it converts the gloss line into three aligned components—the morpheme line, gloss line, and translation line.

to ensure that all examples have a complete and aligned IGT structure. We then verify that the number of words (separated by spaces) matches the number of glosses. In each word, we also ensure one-to-one alignment between morphemes (separated by hyphens) and glosses.

In this study, we use the morpheme-segmented line as standardized input across languages. This choice reflects finer-grained grammatical units and ensures better alignment with glosses.

4 The LINGGYM Benchmark

The high-level characteristics of our LINGGYM dataset are summarized in Table 1. In total, we process 18 reference grammars from LSP, spanning 8 language families, and yield 19,612 IGT examples aligned with relevant KPs after data filtering and cleaning. Most languages in LINGGYM are from the African and Pacific regions, areas that have traditionally been underrepresented in the NLP community. A summary of the dataset's distribution across linguistic subfields is shown in Table 2: the benchmark covers all aspects of the linguistic subfields commonly used to describe the structures of languages, with a strong focus on syntax in the questions reflecting the typical emphasis in language documentation practices.

4.1 Word-Gloss Inference

We introduce a multiple-choice, cloze-style word/word-gloss inference task, which can be used to evaluate whether LLMs can infer grammatical information from structured linguistic data. Each question presents an IGT example in which a sin-

Language	Family	Examples
Pichi	Atlantic-Congo	2,846
Gyeli	Atlantic-Congo	691
Moloko	Atlantic-Congo	439
Fwe	Atlantic-Congo	147
Papuan Malay	Austronesian	3,766
Rapa Nui	Austronesian	1,709
Kagayanen	Austronesian	550
Vamale	Austronesian	67
Komnzo	Trans-New Guinea	709
Mauwake	Trans-New Guinea	1,787
Kalamang	Trans-New Guinea	656
Ulwa	Trans-New Guinea	1,851
Palula	Indo-European	1,674
Tuatschin	Indo-European	1,113
Japhug	Sino-Tibetan	358
Yauyos Quechua	Quechuan	1,143
Mehweb	Northeast Caucasian	85
Ik	Nilo-Saharan	21
	Total	**19,612**

Table 1: Number of KP-IGT pairs for the LINGGYM dataset. In total, 18 reference grammars from 8 language families are processed.

Linguistic Subfield	# Examples	% of Total
Morphology	1,410	7.19%
Phonology	71	0.36%
Pragmatics	139	0.71%
Semantics	967	4.93%
Syntax	16,747	85.39%
Other	278	1.42%
Total	**19,612**	**100%**

Table 2: Distribution of examples by linguistic subfield.

gle word or a single word plus its gloss has been masked. The model must identify the correct word or word-gloss pair from four options, based on the sentence context, grammatical structure, and accompanying explanation. More details of the task can be found in §4.3.

4.2 Question Generation

We generate these questions using examples drawn from our cleaned data. For each instance, if a word or any of its morphemes is marked with a `\textbf` tag, we identify that word as the target. To create the set of four answer choices (one correct answer and three distractors), we employ three strategies to generate plausible distractors:

- **Form-based distractor (LCS-based)**: We find a distractor gloss that shares the longest common substring (LCS) with the correct gloss but differs in grammatical function. For example, given the correct word-gloss pair walk–PST (walk-ed), we generate a distractor like walk–PROG (walk-ing). This shares the root *walk* (via the LCS) but fulfills a different grammatical function.
- **Semantics-based distractor**: We compute the semantic similarity between glosses by embedding them using Sentence-BERT (Reimers and Gurevych, 2019). The gloss that has the highest semantic similarity with, but is not identical to, the correct gloss is selected and mapped back to the corresponding word in the dataset. This approach introduces subtle meaning contrasts to test deeper grammatical understanding.
- **Chapter-local distractor**: To promote lexical and structural diversity, we randomly sample a word-gloss pair from the same grammar chapter, ensuring that the distractor does not overlap in form or gloss with any of the other options. This approach adds noise that reflects the topic domain but avoids trivial elimination.

All distractor candidates are also ensured not to overlap with each other. To prevent positional bias in candidate answers, we randomly assign the correct answer to one of the four choice positions in each question. This randomization is applied uniformly across all examples, ensuring that each position (A–D) contains the correct answer approximately 25% of the time.

To construct each question, we mask all correct choice words in the gloss and knowledge point lines. The masking in the surface line is always performed at the word level, ensuring consistent granularity across examples. This masking approach preserves the context while clearly signalling the missing element to the model. However, masking in the free translation line presents a challenge, as translations often paraphrase or use semantically related expressions rather than a direct lexical equivalent of the source word/morpheme. As a result, the corresponding segment in the translation cannot always be reliably identified or removed without altering the naturalness or interpretability of the sentence—this introduces a limitation in our masking approach: while the surface and gloss lines are systematically masked, the translation may still contain indirect cues about the target word.

4.3 Difficulty Levels

To evaluate the impact of different types of linguistic information, we design our prompts to include the following types of knowledge:

- Original sentence (S): the morpheme-segmented sentence in target language.

Prompt Template

You are a linguist specializing in `{language}`. You are given a sentence along with its morpheme breakdown, gloss, and translation. Words are separated by spaces, and morphemes are separated by hyphens. However, a word and its gloss are missing and represented by an underscore. Based on your understanding, please choose the most appropriate option.

Sentence (with missing item): `{sentence}`

Gloss (with missing item): `{gloss}`

The English translation of this sentence is: `{translation}`

Here is a relevant knowledge point for this example, with the related morphemes and glosses masked: `{knowledgePoint}`

Options:

A: `{wordA}` gloss: `{glossA}`
B: `{wordB}` gloss: `{glossB}`
C: `{wordC}` gloss: `{glossC}`
D: `{wordD}` gloss: `{glossD}`

Please only return the letter (A–D). Do not output anything else

Figure 4: The prompt template used across different difficulty levels.

- Gloss information (G): The glosses for the given words.
- Knowledge points (KP): The relevant knowledge points in the grammar book.
- English translations (T): The English translation.

We conduct our main experiments with four difficulty levels based on data availability: **S**, **S+G**, **S+G+KP**, and **S+G+KP+T**. All prompts follow the template displayed in Figure 4, and an example is shown in Figure 2.

5 Experiments and Results

5.1 Experimental Setup

We evaluate a diverse set of publicly available LLMs, covering a range of sizes and model families. Our evaluation includes models from four major families: **Qwen2.5** (Yang et al., 2024a), **Gemma 3** (Team et al., 2025), **DeepSeek-R1** (Guo et al., 2025), and **LLaMA3** (Dubey et al., 2024). For Qwen2.5, we include the 7B and 32B models; for Gemma3, we evaluate the 4B, 12B, and 27B variants; for DeepSeek-R1, we test the 7B and 32B; and for LLaMA3, we assess the 8B and AWQ-quantized 70B models. We use the AWQ quantization (Lin et al., 2024) for larger 70B models due to limited computing resources. All models are instruction-tuned and are accessed via open-source platforms (i.e., HuggingFace Hub). Inference was performed through `vLLM` (Kwon et al., 2023) and `transformers` (Wolf et al., 2020). All experiments were run on A6000 Ada GPUs, with more details provided in Appendix A. For evaluation, since the word-gloss inference task is formulated as multiple-choice questions with balanced choice distributions, we report standard accuracy as the primary evaluation metric.

5.2 Results

Our main results (Table 3) present accuracies for all evaluated LLMs across four difficulty levels. More detailed results are provided in Appendix B, with a concrete prompt example (S+G+KP+T) and model prediction results shown in Figure 8.

The meta-linguistic reasoning benchmark is challenging to LLMs despite data contamination issues. Data contamination is a common issue in many LLM benchmarks (Sainz et al., 2023; Deng et al., 2024), as LLMs are trained on almost all found data on the Internet. All reference grammar books we processed are subject to this risk, as they are openly accessible as LaTeX source code hosted on GitHub. To clarify the potential impact of data contamination, we test the LLM's performance only with the raw sentences in evaluation languages, without providing any additional information—if LLMs perform above the chance level (25%), it is likely that they have seen some of the language during pretraining.

Indeed, we find evidence of potential data contamination (first row in Table 3): all LLMs have above-chance performance even when provided only the original sentences. Larger models tend to memorize even more, evidenced by higher performance. However, the overall performance is still far from perfect, suggesting that the memorization effect is limited; that is, our dataset serves as a meaningfully challenging benchmark in a highly specialized domain.

	Qwen2.5		Gemma 3			DeepSeek-R1		LLaMA3		GPT-4
Difficulty	7B	32B	4B	12B	27B	7B	32B	8B	70B	o4-mini
S	33.04	38.66	32.74	43.63	41.48	33.39	39.62	29.62	34.46	**41.74**
S+G	41.64	46.75	38.88	47.03	**48.17**	35.24	48.16	30.83	42.37	46.02
S+G+KP	56.08	60.97	49.76	59.47	61.83	46.18	**65.50**	39.44	59.64	57.28
S+G+KP+T	71.09	78.29	63.92	73.97	77.02	54.39	**81.17**	50.32	78.25	73.88

Table 3: Accuracies for all languages across input settings and models.

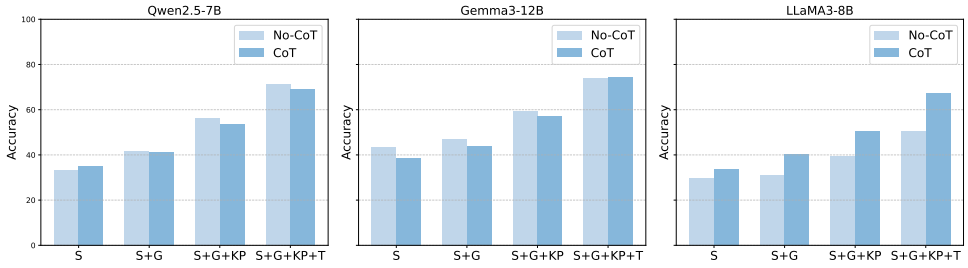

Figure 5: Accuracies of CoT vs. No-CoT prompting across different models and input settings.

Difficulty	Qwen2.5-7B	Gemma3-12B	DeepSeek-R1-7B	LLaMA3-8B
S	33.04	43.63	33.39	29.62
S+G	41.64	47.03	35.24	30.83
S+T	48.65	68.52	53.91	48.65
S+KP	52.78	55.76	45.22	46.22
S+G+KP	56.08	59.47	46.18	39.44
S+G+T	66.59	69.43	54.39	54.39
S+KP+T	59.12	72.95	58.79	59.12
S+G+KP+T	71.09	73.97	54.39	50.32

Table 4: Accuracies across all information permutations for selected models. Full results are shown in Appendix C.

KPs improves performance across all conditions. In line with earlier work (Tanzer et al., 2024; Zhang et al., 2024b), we find that adding KPs brings consistent improvements across LLM families and parameters (Table 3), suggesting that LLMs possess some abilities to comprehend the linguistic concepts in KPs and associate them with concrete language examples. As expected, larger models outperform smaller models by a large margin. The best performing model, DeepSeek-R1 32B, reaches around 81% accuracy, suggesting that LLMs show remarkable capabilities in meta-lingusitic reasoning that is independent of languages.

To further validate this effect, we conduct controlled experiments on selected models by testing LLMs across all difficulty condition permutations. Our ablation results from Table 4 indicate that gloss information, English translations, and knowledge points each contribute to the meta-linguistic reasoning, independent of each other. Yet none of the LLMs achieve perfect accuracy on these tasks, suggesting a large room for improvement.

CoT does not bring clear improvement to the performance. Chain-of-Thought (CoT) prompting (Wei et al., 2022) has been shown to effectively improve performance on reasoning tasks, although the improvement is mainly limited to math and symbolic reasoning tasks (Sprague et al., 2025). As shown in Figure 5, we do not find conclusive evidence that meta-linguistic reasoning benefits much from CoT across all LLMs from different families.

Reasoning models like DeepSeek-R1 and o4-mini are also not competitive with non-reasoning models. The only exception is DeepSeek-R1 32B (Guo et al., 2025), a reasoning model trained to perform long CoT. Although DeepSeek-R1 32B dominates in almost all conditions, DeepSeek-R1 7B does not exhibit such an advantage.

LLM performance is relatively similar across individual languages, language families, and linguistic subfields. The full table by language and models can be found in Appendix E. Figure 6 indicates that the performance is relatively stable across benchmarks, despite some minor variations. This further validates the efficacy of our benchmark, indicating that our benchmark is representative and balanced within and across each language.

As LINGGYM is sourced from the whole reference grammar books, it covers structural descriptions in all linguistic subfields that are considered necessary to describe a language. As shown in Figure 7, performance does not vary substantially across linguistic subfields, aside from minor variations. This suggests that LLMs are able to reason—at least to some extent—across linguistic subfields

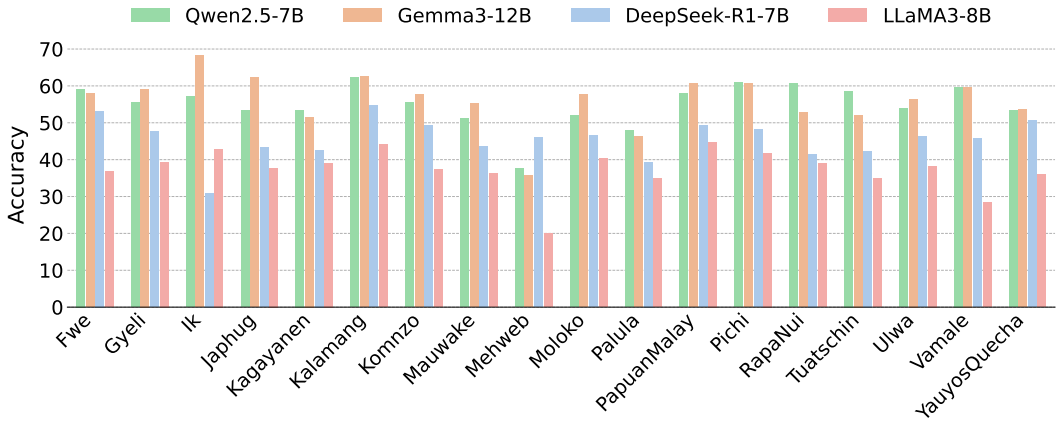

Figure 6: Weighted average accuracy scores across languages under the S+G+KP setting for select models.

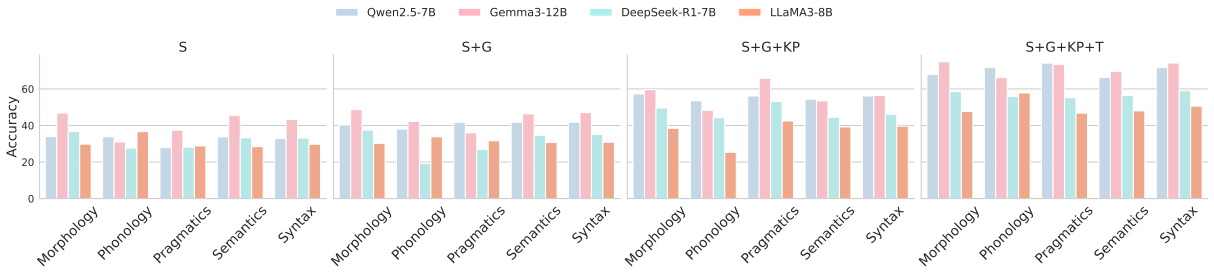

Figure 7: Weighted average accuracies of selected language models across five linguistic subfields: morphology, phonology, pragmatics, semantics, and syntax, under four levels of input difficulty.

represented in these reference grammars.

5.3 Error Analysis

We further investigate errors made under our strongest configuration, DeepSeek-R1 32B, with sentence, gloss, knowledge points, plus translation lines (S+G+KP+T) based on the model's predictions. Most failures can be categorized into three main types:

Abbreviation-heavy items with opaque gloss tags. When correctness depends on understanding dense sequences of gloss abbreviations (such as `PP4-CON=DEM.I7`[5] and `OBJ.LINK-PL`[6]), the model appears to treat the tags as uninterpreted symbols and guess among look-alike forms. For example, in a Moloko sentence where the prompt gloss already encodes number and possession ("`children=Pl`[7] _____ `POSS=1S.POSS=Pl`[8]", with the translation "These particular children here belong to me."), the model prefers `DEM=Pl`[9] over the correct bare `DEM`. Similarly, in Komnzo for the phrase ("Do it here / with the tools here."), the model chooses the bare *zane* (glossed as `DEM:PROX`[10]), whereas the correct answer is *zane=me* (glossed as `DEM:PROX=INS`[11]). The instrumental clitic `=me` is the decisive element that the model fails to recognize. This indicates that comprehensive explanations of these abbreviations need to be incorporated into the prompts as well. We will treat this enhancement as future work.

Semantically similar distractors. The model often selects an option that is plausible in English but morphosyntactically ill-formed in the target language. In Kalamang, for the sentence "I particularly like doing that thing.", the model chooses *great* over the gold *gladly*. Both convey positive affect in English, but only *gladly* fits the required collocational/morphological slot. A similar error occurs in Moloko: given the gloss "`1S+IFV-see`[12]

[5] This gloss is extracted from Fwe. `PP4` = pronominal prefix with agreement set 4; `CON` = connective; `DEM.I7` = demonstrative (series I, form 7; grammar-specific).

[6] This gloss is from Gyeli. `OBJ.LINK` = object linking H tone; `PL` = plural marker.

[7] `Pl` = plural noun clitic.

[8] `POSS` = possessive pronoun; "." stacks features inside one tag; `1S` = first person singular.

[9] `DEM` = demonstrative.

[10] `PROX` = proximal demonstrative.

[11] `INS` = instrumental case.

[12] `1S` = first person singular; `IFV` = imperfective aspect.

goat=1S.POSS=Pl[13] three 2S[14] _____" with its translation "I see my three goats that you gave to me", the model predicts *amə-vəl=ɔkʷ* (glossed as DEP-give=2S.IO[15]), whereas the correct form is *amə-vəl=aw* (glossed as DEP-give=1S.IO[16]). This suggests that the model struggles to differentiate between synonyms or semantically related terms when precise morphological constraints are involved. The underlying issue appears to be that the model relies on semantic similarity rather than understanding the specific grammatical requirements of the target language structure.

Fine-grained form differences (tones, vowels). The model also struggles when answer choices differ only by minimal morpho-orthographic features such as tone marks or single vowels. In Ulwa, when selecting a word meaning "also", the model chooses *maweka* despite the correct form being *moweka*. Both forms have the same meaning, but the single vowel difference completely changes the correctness of the answer. This indicates that the model has insufficient knowledge of phonological and orthographic variants, resulting in treating morphologically distinct forms as interchangeable alternatives.

6 Conclusion

We present LINGGYM, a comprehensive benchmark to assess the meta-lingusitic reasoning ability of LLMs in 18 languages across all linguistic aspects. Our analyses show that LLMs exhibit some capabilities to perform meta-linguistic reasoning, highlighting the potential of using LLMs to assist linguistic analysis.

Our benchmark emphasizes mapping abstract linguistic rules to concrete sentences. Yet in actual fieldwork, it is also important to induce linguistic rules from linguistic samples, which might be assisted with LLMs (Spencer and Kongborrirak, 2025). In the future, we will extend our work to cover more diverse and in-depth use cases for linguistic analysis, especially for low-resource and endangered languages.

Limitations

Linguistic analysis is inherently theory-laden and value-laden (Bird, 2020). Our benchmark is still limited in scope. The grammatical analyses from most reference grammars follow the structuralist framework, which is only one of the many theoretical frameworks in linguistics.

Linguistic analysis is a complex task. The immediately preceding KP often does not paint the full picture of a given grammatical construction (i.e., extracted KPs often make references to parent subsections or sections), though they still constitute a good starting point.

Our study only analyzes 18 languages. While these languages are understudied within the NLP community, they only represent a tiny fraction of human languages. Grambank, a linguistic typological database, records reference grammar books or papers for around 2400 languages (Skirgård et al., 2023), and we will continue to expand our analyses to more languages.

Our dataset is imbalanced across linguistic subfields, reflecting the natural skew of reference grammars, which devote disproportionate attention to syntax. In principle, subdividing the syntax bucket into finer categories (e.g., word/constituent order, agreement, clause structure) would yield more diagnostic analyses. In practice, however, the specific syntactic topics covered vary substantially across languages and sources, which makes a uniform subcategorization scheme difficult to apply consistently and limits cross-language comparability. We therefore report results at a coarse level—extending to stable and finer-grained syntactic categories is a valuable direction for our future work.

While we have attempted to evaluate LLMs across model families and parameter counts, due to limited budget, we were not able to evaluate on the larger state-of-the-art models like DeepSeek-R1-671B, o4, and Gemini 2.5 Pro. These models might demonstrate stronger abilities than the models reported.

Ethics Statement

We only selected the reference grammar books that are publicly available with permissive Creative Commons licenses, allowing us to reprocess and redistribute the dataset.

Our study falls into the scope of fundamental research in natural language processing and linguistics, with the goal of assisting language documentation with LLMs. There is no direct harm associated with this type of research. We expect this work to contribute to the analysis and docu-

[13] POSS = possessive; Pl = plural noun clitic.

[14] 2S = Second person singular.

[15] DEP = dependent form of the verb; 2S.IO = 2nd-person singular indirect object pronominal.

[16] 1S.IO = first person singular indirect object pronominal.

mentation of endangered languages.

Acknowledgements

We thank three anonymous reviewers and the area chairs for their thoughtful comments on the original manuscript. We would also like to extend our gratitude to the field workers and the community members documenting these languages. This research was enabled in part through the computational resources provided by Advanced Research Computing at the University of British Columbia and the Digital Research Alliance of Canada. The research activities were also supported by the NSERC Discovery Grant and the CFI-JELF Grant awarded to JZ, and a Canada CIFAR AI Chair Award to FS.

References

Tuka Alhanai, Adam Kasumovic, Mohammad M Ghassemi, Aven Zitzelberger, Jessica M Lundin, and Guillaume Chabot-Couture. 2025. Bridging the gap: Enhancing llm performance for low-resource african languages with new benchmarks, fine-tuning, and cultural adjustments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 27802–27812.

Christopher A Bail. 2024. Can generative ai improve social science? *Proceedings of the National Academy of Sciences*, 121(21):e2314021121.

Russell Barlow. 2023. *A grammar of Ulwa (Papua New Guinea)*. Number 6 in Comprehensive Grammar Library. Language Science Press, Berlin.

Andrew Michael Bean, Simeon Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Andrew Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Gašper Beguš, Maksymilian Dąbkowski, and Ryan Rhodes. 2023. Large linguistic models: Analyzing theoretical linguistic abilities of llms. *arXiv preprint arXiv:2305.00948*.

Liisa Berghäll. 2015. *A grammar of Mauwake*. Number 4 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Steven Bird. 2020. Decolonising speech and language technology. In *28th International Conference on Computational Linguistics, COLING 2020*, pages 3504–3519. Association for Computational Linguistics (ACL).

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

Shobhana Chelliah. 2013. Fieldwork for language description. *Research methods in linguistics*, pages 51–73.

Bernard Comrie, Martin Haspelmath, and Balthasar Bickel. 2017. The leipzig glossing rules.

Michael Daniel, Nina Dobrushina, and Dmitry Ganenkov, editors. 2019. *The Mehweb language*. Number 1 in Languages of the Caucasus. Language Science Press, Berlin.

Chunyuan Deng, Yilun Zhao, Xiangru Tang, Mark Gerstein, and Arman Cohan. 2024. Investigating data contamination in modern benchmarks for large language models. 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 8706–8719, Mexico City, Mexico. Association for Computational Linguistics.

R M W Dixon. 2009. Basics. In *Basic Linguistic Theory*. Oxford University Press.

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

Ewan Dunbar, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *2017 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. IEEE.

Christian Döhler. 2018. *A grammar of Komnzo*. Number 22 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Alhussein Fawzi, Matej Balog, Aja Huang, Thomas Hubert, Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Francisco J R Ruiz, Julian Schrittwieser, Grzegorz Swirszcz, et al. 2022. Discovering faster matrix multiplication algorithms with reinforcement learning. *Nature*, 610(7930):47–53.

Dianne Friesen. 2017. *A grammar of Moloko*. Number 3 in African Language Grammars and Dictionaries. Language Science Press, Berlin.

Luke Gessler and Katharina Von Der Wense. 2024. Nlp for language documentation: Two reasons for the gap between theory and practice. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (Americas-NLP 2024)*, pages 1–6.

Michael Ginn, Mans Hulden, and Alexis Palmer. 2024. Can we teach language models to gloss endangered languages? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5861–5876, Miami, Florida, USA. Association for Computational Linguistics.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *arXiv preprint arXiv:1901.05287*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, et al. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Nadine Grimm. 2021. *A grammar of Gyeli*. Number 2 in Comprehensive Grammar Library. Language Science Press, Berlin.

Igor Grossmann, Matthew Feinberg, Dawn C Parker, Nicholas A Christakis, Philip E Tetlock, and William A Cunningham. 2023. Ai and the transformation of social science research. *Science*, 380(6650):1108–1109.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Hilde Gunnink. 2022. *A grammar of Fwe*. Number 6 in African Language Grammars and Dictionaries. Language Science Press, Berlin.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q Tran, Jonathan Deaton, Marius Wiggert, et al. 2025. Simulating 500 million years of evolution with a language model. *Science*, page eads0018.

Taiqi He, Kwanghee Choi, Lindia Tjuatja, Nathaniel Robinson, Jiatong Shi, Shinji Watanabe, Graham Neubig, David Mortensen, and Lori Levin. 2024. Wav2Gloss: Generating interlinear glossed text from speech. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 568–582, Bangkok, Thailand. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Guillaume Jacques. 2021. *A grammar of Japhug*. Number 1 in Comprehensive Grammar Library. Language Science Press, Berlin.

Ross Krekoskid Kazantsevaa, Roland Kuhna, Samuel Larkina, Patrick Littella, Delaney Lothiana, Korin Richmondc Akwiratékha'Martina, Marc Tessiera, Cassia Valentini-Botinhaoc, Dan Wellsc, and Junichi Yamagishib. 2024. Speech generation for indigenous language education. *Computer Speech & Language*.

Omkar Khade, Shruti Jagdale, Abhishek Phaltankar, Gauri Takalikar, and Raviraj Joshi. 2025. Challenges in adapting multilingual LLMs to low-resource languages using LoRA PEFT tuning. In *Proceedings of the First Workshop on Challenges in Processing South Asian Languages (CHiPSAL 2025)*, pages 217–222, Abu Dhabi, UAE. International Committee on Computational Linguistics.

Paulus Kieviet. 2017. *A Grammar of Rapa Nui*. Number 12 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Angela Kluge. 2017. *A grammar of Papuan Malay*. Number 11 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pages 611–626.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R Mortensen, Graham Neubig, Alan W Black, et al. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253. IEEE.

Xinjian Li, Florian Metze, David R Mortensen, Shinji Watanabe, and Alan W Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115.

Zihao Li, Yucheng Shi, Zirui Liu, Fan Yang, Ali Payani, Ninghao Liu, and Mengnan Du. 2024. Language ranker: A metric for quantifying llm performance across high and low-resource languages. *Preprint*, arXiv:2404.11553.

Henrik Liljegren. 2016. *A grammar of Palula*. Number 8 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Wei-Ming Chen, Wei-Chen Wang, Guangxuan Xiao, Xingyu Dang, Chuang Gan, and Song Han. 2024. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *Proceedings of Machine Learning and Systems*, 6:87–100.

Christopher D Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. 2020. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences*, 117(48):30046–30054.

Philippe Maurer-Cecchini. 2021. *A grammar of Tuatschin*. Number 3 in Comprehensive Grammar Library. Language Science Press, Berlin.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85.

Sarah Moeller, Ling Liu, Changbing Yang, Katharina Von Der Wense, and Mans Hulden. 2020. Igt2p: From interlinear glossed texts to paradigms. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5251–5262.

Ulrike Mosel. 2006. Grammaticography: The art and craft of writing grammars. *TRENDS IN LINGUISTICS STUDIES AND MONOGRAPHS*, 167:41.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Carol J. Pebley and Thomas E. Payne. 2024. *A grammar of Kagayanen*. Number 8 in Comprehensive Grammar Library. Language Science Press, Berlin.

Rita Ramos, Everlyn Asiko Chimoto, Maartje ter Hoeve, and Natalie Schluter. 2024. Grammamt: Improving machine translation with grammar-informed in-context learning. *arXiv preprint arXiv:2410.18702*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Jean Rohleder. 2024. *A grammar of Vamale*. Number 9 in Comprehensive Grammar Library. Language Science Press, Berlin.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Oscar Sainz, Jon Campos, Iker García-Ferrero, Julen Etxaniz, Oier Lopez de Lacalle, and Eneko Agirre. 2023. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 10776–10787, Singapore. Association for Computational Linguistics.

Farhan Samir, Emily P. Ahn, Shreya Prakash, Márton Soskuthy, Vered Shwartz, and Jian Zhu. 2025. A comparative approach for auditing multilingual phonetic transcript archives. *Transactions of the Association for Computational Linguistics*, 13:595–612.

Terrill Schrock. 2017. *The Ik language*. Number 1 in African Language Grammars and Dictionaries. Language Science Press, Berlin.

Aviva Shimelman. 2017. *A grammar of Yauyos Quechua*. Number 9 in Studies in Diversity Linguistics. Language Science Press, Berlin.

Hedvig Skirgård, Hannah J Haynie, Damián E Blasi, Harald Hammarström, Jeremy Collins, Jay J Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, et al. 2023. Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16):eadg6175.

Piyapath T. Spencer and Nanthipat Kongborrirak. 2025. Can LLMs help create grammar?: Automating grammar creation for endangered languages with in-context learning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 10214–10227, Abu Dhabi, UAE. Association for Computational Linguistics.

Zayne Rea Sprague, Fangcong Yin, Juan Diego Rodriguez, Dongwei Jiang, Manya Wadhwa, Prasann Singhal, Xinyu Zhao, Xi Ye, Kyle Mahowald, and Greg Durrett. 2025. To cot or not to cot? chain-of-thought helps mainly on math and symbolic reasoning. In *The Thirteenth International Conference on Learning Representations*.

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. In *The Twelfth International Conference on Learning Representations*.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Eline Visser. 2022. *A grammar of Kalamang*. Number 4 in Comprehensive Grammar Library. Language Science Press, Berlin.

Shenran Wang, Changbing Yang, Michael l Parkhill, Chad Quinn, Christopher Hammerly, and Jian Zhu. 2025. Developing multilingual speech synthesis system for Ojibwe, mi'kmaq, and maliseet. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pages 817–826, Albuquerque, New Mexico. Association for Computational Linguistics.

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

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.

Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. 2018. What do RNN language models learn about filler–gap dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 211–221, Brussels, Belgium. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual BERT? In *Proceedings of the 5th Workshop on Representation Learning for NLP*, pages 120–130, Online. Association for Computational Linguistics.

Kofi Yakpo. 2019. *A grammar of Pichi*. Number 23 in Studies in Diversity Linguistics. Language Science Press, Berlin.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024a. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Changbing Yang, Garrett Nicolai, and Miikka Silfverberg. 2024b. Multiple sources are better than one: Incorporating external knowledge in low-resource glossing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4537–4552, Miami, Florida, USA. Association for Computational Linguistics.

Chen Zhang, Xiao Liu, Jiuheng Lin, and Yansong Feng. 2024a. Teaching large language models an unseen language on the fly. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 8783–8800, Bangkok, Thailand. Association for Computational Linguistics.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024b. 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, Bangkok, Thailand. Association for Computational Linguistics.

Qiang Zhang, Keyang Ding, Tianwen Lyv, Xinda Wang, Qingyu Yin, Yiwen Zhang, Jing Yu, Yuhao Wang, Xiaotong Li, Zhuoyi Xiang, et al. 2024c. Scientific large language models: A survey on biological & chemical domains. *arXiv preprint arXiv:2401.14656*.

Jian Zhu, Farhan Samir, Eleanor Chodroff, and David R. Mortensen. 2025. ZIPA: A family of efficient models for multilingual phone recognition. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 19568–19585, Vienna, Austria. Association for Computational Linguistics.

Jian Zhu, Cong Zhang, and David Jurgens. 2022. ByT5 model for massively multilingual grapheme-to-phoneme conversion. In *Proc. Interspeech 2022*, pages 446–450.

Caleb Ziems, William Held, Omar Shaikh, Jiaao Chen, Zhehao Zhang, and Diyi Yang. 2024. Can large language models transform computational social science? *Computational Linguistics*, 50(1):237–291.

A Sampling Parameters of LLMs

Parameter	Value / Description
Temperature	0.7
Top-p	0.9
Max Tokens	2048
Repetition Penalty	1.1
Decoding Strategy	Sampling-based decoding

Table 5: Sampling parameters used for LLM generation.

B Full Results

		Qwen2.5		Gemma 3			DeepSeek-R1		LLaMA3		GPT-4
Language	**Difficulty**	7B	32B	4B	12B	27B	7B	32B	8B	70B	o4-mini
Atlantic-Congo Language Family											
Pichi	S	31.48	37.91	33.45	44.27	42.83	30.97	41.43	30.29	36.33	42.45
	S+G	46.03	48.31	43.57	49.51	51.05	37.15	53.03	35.10	48.91	47.79
	S+G+KP	61.00	65.32	57.34	62.02	65.43	48.25	68.75	41.78	66.90	60.40
	S+G+KP+T	73.86	79.86	69.15	75.86	79.23	60.26	81.24	53.51	82.31	75.47
Gyeli	S	26.48	29.52	30.10	34.44	30.68	32.88	27.49	30.68	28.51	32.13
	S+G	35.75	39.13	35.31	39.94	39.94	34.99	38.14	20.09	34.15	40.09
	S+G+KP	55.57	60.78	46.89	60.35	63.53	47.54	63.50	39.36	58.61	57.60
	S+G+KP+T	64.25	72.21	56.44	67.29	73.95	57.16	76.06	47.03	71.92	70.62
Moloko	S	25.97	29.84	28.93	41.46	33.26	29.76	28.34	22.78	27.56	34.40
	S+G	38.29	36.90	33.71	46.47	47.38	30.68	36.83	28.70	35.31	43.51
	S+G+KP	51.94	61.05	47.84	59.68	63.55	46.42	64.45	40.32	64.01	59.45
	S+G+KP+T	67.88	77.90	65.15	71.75	78.36	57.14	80.59	45.79	81.55	72.89
Fwe	S	35.37	40.82	37.41	42.86	36.73	29.55	33.33	29.93	38.78	36.73
	S+G	44.22	47.62	43.54	40.14	40.82	32.35	50.00	29.25	34.01	41.50
	S+G+KP	59.18	63.27	49.66	57.14	63.27	53.44	69.44	36.73	61.09	63.27
	S+G+KP+T	65.31	77.55	63.27	69.39	73.47	55.47	78.32	48.98	71.43	68.71
Austronesian Language Family											
Papuan Malay	S	39.80	49.76	42.30	50.82	49.52	38.37	49.20	33.30	42.14	54.30
	S+G	43.12	51.73	43.63	52.89	50.77	38.46	52.08	32.58	45.42	54.12
	S+G+KP	57.89	63.52	52.52	63.78	65.37	48.28	68.30	44.80	62.42	62.83
	S+G+KP+T	77.51	82.97	68.91	80.64	81.41	64.38	84.73	60.12	81.86	81.12
Rapa Nui	S	38.68	37.62	28.91	43.89	40.26	28.38	43.10	31.01	40.78	39.56
	S+G	51.84	53.00	41.95	46.58	48.51	35.19	53.57	34.87	53.31	52.08
	S+G+KP	60.62	63.39	46.99	57.28	55.59	41.53	65.62	38.99	64.97	53.01
	S+G+KP+T	73.96	82.16	62.79	70.39	72.62	52.59	81.98	47.51	80.68	70.39
Kagayanen	S	30.36	38.18	33.09	39.82	39.82	30.26	38.64	25.64	34.73	40.91
	S+G	43.64	46.73	39.82	41.82	45.82	32.45	48.70	32.18	47.09	45.45
	S+G+KP	53.45	55.64	50.73	53.64	57.45	42.70	65.06	39.09	56.00	54.18
	S+G+KP+T	66.73	72.41	61.64	69.82	72.36	55.70	80.22	52.73	79.05	71.27
Vamale	S	31.34	29.85	35.82	43.28	32.84	39.06	38.81	31.34	35.82	38.81
	S+G	34.33	47.76	44.78	50.57	38.81	38.46	48.44	23.88	29.85	43.28
	S+G+KP	59.70	55.22	56.72	50.75	56.72	45.31	67.16	28.36	46.27	55.22
	S+G+KP+T	68.66	80.60	74.63	71.64	77.61	63.64	82.09	50.75	74.63	73.13
Trans-New Guinea Language Family											
Komnzo	S	34.41	34.27	31.88	41.47	42.88	32.68	31.23	26.09	27.22	33.99
	S+G	43.16	42.74	39.63	43.86	50.49	35.61	42.69	26.66	35.54	40.62
	S+G+KP	55.43	60.08	51.20	59.94	65.59	49.62	67.58	37.24	59.76	58.39
	S+G+KP+T	68.41	77.29	65.87	73.34	75.46	55.39	83.69	46.83	75.32	73.20
Mauwake	S	32.40	41.63	33.41	48.68	46.22	35.48	39.83	26.52	30.78	42.59
	S+G	38.61	45.57	38.05	50.87	51.32	34.36	46.20	28.32	37.16	44.15
	S+G+KP	51.20	57.54	46.22	59.54	61.00	43.74	61.54	36.34	51.65	56.41
	S+G+KP+T	64.63	73.52	59.60	71.24	77.11	56.23	77.25	43.59	73.05	71.80
Kalamang	S	28.81	38.26	31.40	39.33	38.11	34.69	37.85	27.44	32.47	39.48
	S+G	35.37	45.73	36.74	41.92	47.41	37.58	46.62	28.35	37.04	45.73
	S+G+KP	62.20	65.19	52.74	64.94	67.99	54.71	68.77	44.05	63.11	61.59
	S+G+KP+T	75.30	82.16	64.18	77.90	83.08	65.42	86.52	54.42	80.55	78.20
Ulwa	S	30.25	34.47	26.09	41.22	39.17	34.17	34.88	27.61	30.09	36.03
	S+G	34.25	42.86	30.09	45.87	48.68	33.50	39.23	27.55	36.20	40.46
	S+G+KP	53.92	61.78	44.41	60.08	64.34	46.28	66.39	38.25	57.16	55.11
	S+G+KP+T	65.42	73.62	56.94	72.18	75.36	58.20	77.64	45.43	73.80	68.99
Indo-European Language Family											
Palula	S	28.73	31.06	27.42	35.36	35.19	32.66	32.71	30.23	31.54	35.96
	S+G	38.47	40.28	37.22	40.20	43.25	32.87	43.73	29.45	39.14	36.68
	S+G+KP	47.91	50.06	45.40	47.97	50.42	39.41	54.72	34.95	52.15	47.79
	S+G+KP+T	67.03	72.87	59.98	68.04	69.71	54.16	75.89	44.92	73.78	67.74
Tuatschin	S	29.29	35.13	29.02	41.96	33.69	31.60	40.53	28.75	30.37	34.95
	S+G	44.65	49.51	35.94	44.65	46.09	33.77	58.27	29.83	45.64	47.62
	S+G+KP	58.58	62.61	48.07	60.74	61.90	42.36	71.30	34.86	63.43	57.86
	S+G+KP+T	72.87	78.43	62.53	73.41	75.83	57.94	83.77	47.26	79.34	73.94
Other Language Families											
Japhug	S	27.65	31.01	31.28	43.30	45.81	32.34	36.49	27.37	32.68	36.87
	S+G	38.83	51.40	40.50	47.21	53.91	30.33	48.86	24.86	49.16	48.60
	S+G+KP	53.35	67.04	51.68	63.13	68.44	42.99	64.37	37.71	61.90	61.17
	S+G+KP+T	66.20	81.28	67.88	75.14	80.73	54.27	83.10	44.41	79.05	73.18
Yauyos Quechua	S	32.98	38.06	31.06	41.29	39.11	31.64	35.19	30.36	28.35	43.74
	S+G	37.27	42.91	31.41	44.01	40.86	33.43	43.25	28.52	33.07	39.81
	S+G+KP	53.42	57.62	47.51	57.13	57.66	50.66	62.79	35.96	48.56	53.11
	S+G+KP+T	71.92	80.40	64.36	76.03	78.65	61.34	84.82	48.56	77.17	75.24
Mehweb	S	30.59	25.88	20.00	25.88	30.59	25.00	34.12	22.35	30.59	24.71
	S+G	31.76	32.94	25.88	35.29	44.71	21.43	40.24	27.06	24.71	24.71
	S+G+KP	37.65	32.94	35.29	34.12	51.76	45.68	40.00	20.00	30.59	31.76
	S+G+KP+T	60.00	63.53	44.71	52.94	70.59	55.00	65.48	40.00	69.41	51.76
Ik	S	28.57	28.57	33.33	38.10	23.81	40.00	33.33	28.57	19.05	38.10
	S+G	42.86	47.62	57.14	38.10	42.86	25.00	47.62	38.10	42.86	47.62
	S+G+KP	57.14	57.14	57.14	57.14	66.67	30.00	80.95	42.86	66.67	61.90
	S+G+KP+T	90.48	95.24	85.71	71.43	90.48	76.47	95.24	61.90	100.00	90.48

Table 6: Accuracy scores across languages and difficulties for all models.

C CoT and Non-CoT Prompting

Language	Difficulty	Qwen2.5-7B	Gemma3-12B	LLaMA3-8B
Atlantic-Congo Language Family				
Pichi	S	31.48	44.27	30.97
	S+G	46.03	49.51	37.15
	S+G+KP	61.00	62.02	48.25
	S+G+KP+T	73.86	75.86	60.26
	CoT-S	36.47	40.02	34.80
	CoT-S+G	43.68	45.71	45.52
	CoT-S+G+KP	58.14	60.70	55.37
	CoT-S+G+KP+T	71.10	77.20	70.28
Gyeli	S	26.48	34.44	32.88
	S+G	35.75	39.94	34.99
	S+G+KP	55.57	60.35	47.54
	S+G+KP+T	64.25	67.29	57.16
	CoT-S	24.60	30.54	27.06
	CoT-S+G	31.40	39.22	35.12
	CoT-S+G+KP	54.27	56.67	53.48
	CoT-S+G+KP+T	62.23	69.18	66.81
Moloko	S	25.97	41.46	29.76
	S+G	38.29	46.47	30.68
	S+G+KP	51.94	59.68	46.42
	S+G+KP+T	67.88	71.75	57.14
	CoT-S	25.06	35.76	30.30
	CoT-S+G	36.45	40.09	38.27
	CoT-S+G+KP	50.57	53.76	49.20
	CoT-S+G+KP+T	69.25	74.49	67.88
Fwe	S	35.37	42.86	29.55
	S+G	44.22	40.14	32.35
	S+G+KP	59.18	57.14	53.44
	S+G+KP+T	65.31	69.39	55.47
	CoT-S	33.33	38.78	25.85
	CoT-S+G	39.46	36.05	40.41
	CoT-S+G+KP	53.74	53.06	48.63
	CoT-S+G+KP+T	62.59	72.11	60.54
Austronesian Language Family				
Papuan Malay	S	39.80	50.82	38.37
	S+G	43.12	52.89	38.46
	S+G+KP	57.89	63.78	48.28
	S+G+KP+T	77.51	80.64	64.38
	CoT-S	44.90	48.33	38.97
	CoT-S+G	45.62	49.55	42.36
	CoT-S+G+KP	56.51	62.79	51.77
	CoT-S+G+KP+T	76.21	79.94	72.49
Rapa Nui	S	38.68	43.89	28.38
	S+G	51.84	46.58	35.19
	S+G+KP	60.62	57.28	41.53
	S+G+KP+T	73.96	70.39	52.59
	CoT-S	34.82	34.11	35.25
	CoT-S+G	44.12	42.36	44.05
	CoT-S+G+KP	53.63	51.84	50.18
	CoT-S+G+KP+T	65.75	69.20	64.51
Kagayanen	S	30.36	39.82	30.26
	S+G	43.64	41.82	32.45
	S+G+KP	53.45	53.64	42.70
	S+G+KP+T	66.73	69.82	55.70
	CoT-S	32.91	38.00	33.64
	CoT-S+G	39.45	42.91	40.55
	CoT-S+G+KP	51.45	52.91	43.69
	CoT-S+G+KP+T	66.18	69.64	64.84
Vamale	S	31.34	43.28	39.06
	S+G	34.33	50.57	38.46
	S+G+KP	59.70	50.75	45.31
	S+G+KP+T	68.66	71.64	63.64
	CoT-S	19.40	31.34	35.82
	CoT-S+G	29.85	37.31	38.81
	CoT-S+G+KP	46.27	47.76	43.28
	CoT-S+G+KP+T	61.19	68.66	70.15
Trans-New Guinea Language Family				
Komnzo	S	34.41	41.47	32.68
	S+G	43.16	43.86	35.61
	S+G+KP	55.43	59.94	49.62
	S+G+KP+T	68.41	73.34	55.39
	CoT-S	35.68	35.26	27.22
	CoT-S+G	36.67	40.54	37.29
	CoT-S+G+KP	53.17	57.97	48.30
	CoT-S+G+KP+T	65.87	71.79	65.54
Mauwake	S	32.40	48.68	35.48
	S+G	38.61	50.87	34.36
	S+G+KP	51.20	59.54	43.74

(*Continued on next page*)

(*Table 7 continued from previous page*)

Language	Difficulty	Qwen2.5-7B	Gemma3-12B	LLaMA3-8B
	S+G+KP+T	64.63	71.24	56.23
	CoT-S	33.30	40.12	33.35
	CoT-S+G	40.87	44.21	37.33
	CoT-S+G+KP	50.14	57.58	47.56
	CoT-S+G+KP+T	65.30	71.18	62.81
Kalamang	S	28.81	39.33	34.69
	S+G	35.37	41.92	37.58
	S+G+KP	62.20	64.94	54.71
	S+G+KP+T	75.30	77.90	65.42
	CoT-S	34.76	33.84	32.01
	CoT-S+G	39.63	37.96	37.00
	CoT-S+G+KP	58.23	62.04	56.10
	CoT-S+G+KP+T	74.09	78.96	73.02
Ulwa	S	30.25	41.22	34.17
	S+G	34.25	45.87	33.50
	S+G+KP	53.92	60.08	46.28
	S+G+KP+T	65.42	72.18	58.20
	CoT-S	32.41	35.66	31.37
	CoT-S+G	37.33	40.36	36.26
	CoT-S+G+KP	52.24	57.05	50.00
	CoT-S+G+KP+T	65.59	71.91	64.25
Indo-European Language Family				
Palula	S	28.73	35.36	32.66
	S+G	38.47	40.20	32.87
	S+G+KP	47.91	47.97	39.41
	S+G+KP+T	67.03	68.04	54.16
	CoT-S	28.67	33.09	30.92
	CoT-S+G	38.05	39.96	37.84
	CoT-S+G+KP	45.10	45.94	46.24
	CoT-S+G+KP+T	64.28	70.19	64.44
Tuatschin	S	29.29	41.96	31.60
	S+G	44.65	44.65	33.77
	S+G+KP	58.58	60.74	42.36
	S+G+KP+T	72.87	73.41	57.94
	CoT-S	31.09	38.85	32.43
	CoT-S+G	45.01	46.45	42.59
	CoT-S+G+KP	53.73	56.33	51.40
	CoT-S+G+KP+T	67.30	73.05	68.10
Other Language Families				
Japhug	S	27.65	43.30	32.34
	S+G	38.83	47.21	30.33
	S+G+KP	53.35	63.13	42.99
	S+G+KP+T	66.20	75.14	54.27
	CoT-S	33.80	37.43	31.01
	CoT-S+G	41.06	48.60	39.39
	CoT-S+G+KP	55.03	65.36	51.13
	CoT-S+G+KP+T	70.11	77.37	64.71
Yauyos Quechua	S	32.98	41.29	31.64
	S+G	37.27	44.01	33.43
	S+G+KP	53.42	57.13	50.66
	S+G+KP+T	71.92	76.03	61.34
	CoT-S	34.12	35.61	33.83
	CoT-S+G	39.37	39.63	34.33
	CoT-S+G+KP	49.82	53.98	46.92
	CoT-S+G+KP+T	69.73	77.87	66.55
Mehweb	S	30.59	25.88	25.00
	S+G	31.76	35.29	21.43
	S+G+KP	37.65	34.12	45.68
	S+G+KP+T	60.00	52.94	55.00
	CoT-S	36.90	25.88	29.41
	CoT-S+G	30.12	34.12	32.94
	CoT-S+G+KP	39.76	36.47	38.82
	CoT-S+G+KP+T	62.35	55.29	45.88
Ik	S	28.57	38.10	40.00
	S+G	42.86	38.10	25.00
	S+G+KP	57.14	57.14	30.00
	S+G+KP+T	90.48	71.43	76.47
	CoT-S	23.81	19.05	33.33
	CoT-S+G	42.86	42.86	47.62
	CoT-S+G+KP	55.00	52.38	57.14
	CoT-S+G+KP+T	90.48	76.19	90.48

Table 7: Accuracy scores with and without CoT across languages and difficulties for select models.

D Ablation Study

Language	Difficulty	Qwen2.5-7B	Gemma3-12B	DeepSeek-R1-7B	LLaMA3-8B
Atlantic-Congo Language Family					
Pichi	S	31.48	44.27	30.97	30.29
	S+G	46.03	49.51	37.15	35.10
	S+KP	59.17	60.64	45.98	52.26
	S+T	54.80	72.63	54.81	54.80
	S+G+KP	61.00	62.02	48.25	41.78
	S+G+T	71.54	72.49	57.30	50.67
	S+KP+T	64.40	76.53	61.52	64.40
	S+G+KP+T	73.86	75.86	60.26	53.51
Gyeli	S	26.48	34.44	32.88	30.68
	S+G	35.75	39.94	34.99	20.09
	S+KP	52.03	59.29	47.77	48.99
	S+T	44.04	60.38	49.62	44.04
	S+G+KP	55.57	60.35	47.54	39.36
	S+G+T	58.61	61.22	48.94	38.64
	S+KP+T	54.00	69.68	54.83	54.00
	S+G+KP+T	64.25	67.29	57.16	47.03
Moloko	S	25.97	41.46	29.76	22.78
	S+G	38.29	46.47	30.68	28.70
	S+KP	50.00	57.54	41.18	42.73
	S+T	48.06	68.38	51.54	48.06
	S+G+KP	51.94	59.68	46.42	40.32
	S+G+T	62.64	71.53	53.10	42.60
	S+KP+T	61.78	76.12	57.11	61.78
	S+G+KP+T	67.88	71.75	57.14	45.79
Fwe	S	35.37	42.86	29.55	29.93
	S+G	44.22	40.14	32.35	29.25
	S+KP	50.34	57.86	46.62	46.94
	S+T	41.50	59.59	43.07	41.50
	S+G+KP	59.18	57.14	53.44	36.73
	S+G+T	55.78	57.82	40.56	42.86
	S+KP+T	56.16	65.25	51.47	56.16
	S+G+KP+T	65.31	69.39	55.47	48.98
Austronesian Language Family					
Papuan Malay	S	39.80	50.82	38.37	33.30
	S+G	43.12	52.89	38.46	32.58
	S+KP	58.55	60.64	48.03	49.43
	S+T	53.72	74.51	60.01	53.72
	S+G+KP	57.89	63.78	48.28	44.80
	S+G+T	72.94	76.90	60.64	54.38
	S+KP+T	65.09	79.68	64.19	65.09
	S+G+KP+T	77.51	80.64	64.38	60.12
Rapa Nui	S	38.68	43.89	28.38	31.01
	S+G	51.84	46.58	35.19	34.87
	S+KP	49.21	52.46	39.64	43.75
	S+T	46.20	65.96	47.28	46.20
	S+G+KP	60.62	57.28	41.53	38.99
	S+G+T	73.03	64.95	49.21	44.70
	S+KP+T	55.98	68.60	51.65	55.98
	S+G+KP+T	73.96	70.39	52.59	47.51
Kagayanen	S	30.36	39.82	30.26	25.64
	S+G	43.64	41.82	32.45	32.18
	S+KP	46.00	51.59	41.70	44.34
	S+T	47.64	66.61	50.93	47.64
	S+G+KP	53.45	53.64	42.70	39.09
	S+G+T	66.00	66.00	52.06	47.82
	S+KP+T	58.06	68.65	57.56	58.06
	S+G+KP+T	66.73	69.82	55.70	52.73
Vamale	S	31.34	43.28	39.06	31.34
	S+G	34.33	50.57	38.46	23.88
	S+KP	49.25	59.09	43.94	46.27
	S+T	39.39	60.61	61.90	39.39
	S+G+KP	59.70	50.75	45.31	28.36
	S+G+T	55.22	65.67	61.29	31.34
	S+KP+T	61.19	57.63	60.61	54.34
	S+G+KP+T	68.66	71.64	63.64	50.75
Trans-New Guinea Language Family					
Komnzo	S	34.41	41.47	32.68	26.09
	S+G	43.16	43.86	35.61	26.66
	S+KP	52.55	57.58	45.98	45.03
	S+T	45.66	61.71	51.69	45.66
	S+G+KP	55.43	59.94	49.62	37.24
	S+G+T	62.06	67.14	51.41	37.52
	S+KP+T	53.69	73.09	57.47	53.69
	S+G+KP+T	68.41	73.34	55.39	46.83

(*Continued on next page*)

(*Table 8 continued from previous page*)

Language	Prompt	Qwen2.5-7B	Gemma3-12B	DeepSeek-R1-7B	LLaMA3-8B
Mauwake	S	32.40	48.68	35.48	26.52
	S+G	38.61	50.87	34.36	28.32
	S+KP	49.41	55.14	42.64	44.23
	S+T	42.45	66.25	53.30	42.45
	S+G+KP	51.20	59.54	43.74	36.34
	S+G+T	60.88	68.72	54.34	40.91
	S+KP+T	53.17	70.53	55.05	53.17
	S+G+KP+T	64.63	71.24	56.23	43.59
Kalamang	S	28.81	39.33	34.69	27.44
	S+G	35.37	41.92	37.58	28.35
	S+KP	59.60	62.96	54.21	55.12
	S+T	48.62	68.75	55.37	48.62
	S+G+KP	62.20	64.94	54.71	44.05
	S+G+T	64.02	67.07	58.82	42.23
	S+KP+T	63.89	79.44	64.84	63.89
	S+G+KP+T	75.30	77.90	65.42	54.42
Ulwa	S	30.25	41.22	34.17	27.61
	S+G	34.25	45.87	33.50	27.55
	S+KP	51.11	56.35	48.55	46.04
	S+T	45.26	64.61	51.70	45.26
	S+G+KP	53.92	60.08	46.28	38.25
	S+G+T	58.18	66.18	48.50	37.06
	S+KP+T	54.34	71.19	57.92	54.34
	S+G+KP+T	65.42	72.18	58.20	45.43
Indo-European Language Family					
Palula	S	28.73	35.36	32.66	30.23
	S+G	38.47	40.20	32.87	29.45
	S+KP	43.84	46.43	40.09	37.67
	S+T	49.01	66.01	52.27	49.01
	S+G+KP	47.91	47.97	39.41	34.95
	S+G+T	62.25	65.47	51.98	40.38
	S+KP+T	55.66	67.16	55.85	55.66
	S+G+KP+T	67.03	68.04	54.16	44.92
Tuatschin	S	29.29	41.96	31.60	28.75
	S+G	44.65	44.65	33.77	29.83
	S+KP	49.41	52.09	41.84	43.54
	S+T	45.35	65.52	53.14	45.35
	S+G+KP	58.58	60.74	42.36	34.86
	S+G+T	66.67	67.12	52.66	39.98
	S+KP+T	57.78	69.88	56.89	57.78
	S+G+KP+T	72.87	73.41	57.94	47.26
Other Language Families					
Japhug	S	27.65	43.30	32.34	27.37
	S+G	38.83	47.21	30.33	24.86
	S+KP	50.56	57.53	43.90	41.01
	S+T	46.22	63.71	50.00	46.22
	S+G+KP	53.35	63.13	42.99	37.71
	S+G+T	65.92	68.44	50.59	38.27
	S+KP+T	58.43	72.73	51.80	58.43
	S+G+KP+T	66.20	75.14	54.27	44.41
Yauyos Quechua	S	32.98	41.29	31.64	30.36
	S+G	37.27	44.01	33.43	28.52
	S+KP	51.49	53.76	50.23	41.01
	S+T	46.45	72.00	56.27	46.45
	S+G+KP	53.42	57.13	50.66	35.96
	S+G+T	66.23	69.82	55.28	43.39
	S+KP+T	58.91	78.65	63.61	58.91
	S+G+KP+T	71.92	76.03	61.34	48.56
Mehweb	S	30.59	25.88	25.00	22.35
	S+G	31.76	35.29	21.43	27.06
	S+KP	27.71	34.72	27.85	36.90
	S+T	34.12	59.04	46.25	34.12
	S+G+KP	37.65	34.12	45.68	20.00
	S+G+T	56.47	55.29	50.62	43.53
	S+KP+T	43.53	56.06	46.84	43.53
	S+G+KP+T	60.00	52.94	55.00	40.00
Ik	S	28.57	38.10	40.00	28.57
	S+G	42.86	38.10	25.00	38.10
	S+KP	61.90	70.59	47.62	61.90
	S+T	66.67	80.95	61.90	66.67
	S+G+KP	57.14	57.14	30.00	42.86
	S+G+T	80.95	80.95	65.00	47.62
	S+KP+T	94.74	82.35	80.00	94.74
	S+G+KP+T	90.48	71.43	76.47	61.90

Table 8: Accuracy scores across languages across all information permutations for selected models.

E Linguistic Subfields

		Qwen2.5		Gemma 3			DeepSeek-R1		LLaMA3		GPT-4
Subfield	**Difficulty**	7B	32B	4B	12B	27B	7B	32B	8B	70B	o4-mini
Morphology	S	33.90	41.20	34.47	46.81	44.04	36.65	40.86	29.79	31.77	41.49
Morphology	S+G	40.28	44.36	39.22	48.65	46.95	37.47	47.06	30.21	37.87	42.69
Morphology	S+G+KP	57.23	62.24	56.45	59.59	63.90	49.55	66.66	38.46	56.23	59.01
Morphology	S+G+KP+T	67.87	76.72	64.04	74.82	77.23	58.55	80.40	47.66	76.03	73.05
Phonology	S	33.80	35.21	42.25	30.99	28.17	27.61	44.49	36.62	39.44	36.62
Phonology	S+G	38.03	46.48	40.84	42.25	36.62	19.32	53.01	33.80	46.48	39.44
Phonology	S+G+KP	53.52	57.75	64.79	48.29	56.34	44.27	75.81	25.35	45.07	45.07
Phonology	S+G+KP+T	71.83	83.10	59.16	66.20	74.65	55.84	85.53	57.75	77.47	73.24
Pragmatics	S	28.06	38.13	30.21	37.41	39.57	28.12	38.13	28.78	32.37	33.10
Pragmatics	S+G	41.73	44.60	35.97	35.97	44.61	26.99	39.47	31.65	43.89	41.01
Pragmatics	S+G+KP	56.11	62.59	65.47	65.79	64.03	53.07	70.50	42.44	64.03	58.99
Pragmatics	S+G+KP+T	74.10	74.68	65.47	73.38	75.54	55.17	78.97	46.76	73.82	75.54
Semantics	S	33.71	38.78	34.44	45.40	39.81	33.24	38.22	28.44	32.99	40.95
Semantics	S+G	41.78	47.47	39.50	46.33	49.23	34.61	46.74	30.71	41.26	45.71
Semantics	S+G+KP	54.39	58.14	56.46	53.49	62.77	44.52	63.91	39.19	55.62	57.81
Semantics	S+G+KP+T	66.29	73.52	58.64	69.60	74.77	56.48	78.05	47.98	74.10	71.04
Syntax	S	32.87	38.46	32.43	43.30	41.33	33.18	39.44	29.75	34.67	41.88
Syntax	S+G	41.76	47.01	38.80	47.05	48.25	35.18	48.36	30.84	42.85	46.32
Syntax	S+G+KP	56.13	61.09	60.28	56.39	61.62	46.08	65.36	39.60	60.21	57.19
Syntax	S+G+KP+T	71.63	78.70	64.23	74.15	77.12	58.99	81.40	50.62	78.65	74.11

Table 9: Accuracy scores across linguistic subfields and select difficulties.

F Overview of Languages Processed

F.1 Language Families

Reference Grammar Title	Language Family	Citation
A grammar of Pichi	Atlantic-Congo	Yakpo, 2019
A grammar of Gyeli	Atlantic-Congo	Grimm, 2021
A grammar of Moloko	Atlantic-Congo	Friesen, 2017
A grammar of Fwe	Atlantic-Congo	Gunnink, 2022
A grammar of Papuan Malay	Austronesian	Kluge, 2017
A grammar of Rapa Nui	Austronesian	Kieviet, 2017
A grammar of Kagayanen	Austronesian	Pebley and Payne, 2024
A grammar of Vamale	Austronesian	Rohleder, 2024
A grammar of Komnzo	Trans-New Guinea	Döhler, 2018
A grammar of Mauwake	Trans-New Guinea	Berghäll, 2015
A grammar of Kalamang	Trans-New Guinea	Visser, 2022
A grammar of Ulwa (Papua New Guinea)	Trans-New Guinea	Barlow, 2023
A grammar of Palula	Indo-European	Liljegren, 2016
A grammar of Tuatschin	Indo-European	Maurer-Cecchini, 2021
A grammar of Japhug	Sino-Tibetan	Jacques, 2021
A grammar of Yauyos Quechua	Quechuan	Shimelman, 2017
The Mehweb language	Northeast Caucasian	Daniel et al., 2019
The Ik language	Nilo-Saharan	Schrock, 2017

Table 10: Reference grammars and their language families: We process reference grammars from the *Studies in Diversity Linguistics* and *Comprehensive Grammar Library* series published by *Language Science Press*. Each language's corresponding language family and each chapter's linguistic subfield are determined by reading the relevant sections (shown in F.2).

F.2 Chapter Categorization

Table 11: Overview of extracted chapters by language and linguistic subfield, in the order they appear in their respective reference grammar.

Language	Chapter	Subfield
Pichi	Introduction	Other
Pichi	Segmental phonology	Phonology
Pichi	Suprasegmental phonology	Phonology
Pichi	Morphology	Syntax
Pichi	The nominal system	Syntax
Pichi	The verbal system	Syntax
Pichi	The clause	Syntax
Pichi	Spatial and temporal relations	Syntax
Pichi	Grammatical relations	Syntax
Pichi	Clause linkage	Syntax
Pichi	Multiverb constructions	Syntax
Pichi	Pragmatic elements and routines	Pragmatics
Pichi	Pichi and Spanish in contact	Other
Gyeli	Introduction	Other
Gyeli	Phonology	Phonology
Gyeli	Parts of speech	Syntax

Continued on next page

Language	Chapter	Subfield
Gyeli	Morphology	Morphology
Gyeli	The noun phrase	Syntax
Gyeli	The verbal complex	Syntax
Gyeli	Simple clauses	Syntax
Gyeli	Complex clauses	Syntax
Moloko	Clause	Syntax
Moloko	The na marker and na constructions	Syntax
Moloko	Clause combining	Syntax
Moloko	Grammatical classes	Syntax
Moloko	Noun morphology	Morphology
Moloko	Noun phrase	Syntax
Moloko	The verb complex	Syntax
Moloko	Verb phrase	Syntax
Moloko	Verb types and transitivity	Syntax
Fwe	Mood	Semantics
Fwe	Negation	Semantics
Fwe	Syntax and information structure	Syntax
Fwe	Nominal morphology	Morphology
Fwe	Minor word categories	Syntax
Fwe	Verbal derivation	Syntax
Fwe	Tense	Semantics
Papuan Malay	Introduction	Other
Papuan Malay	Phonology	Phonology
Papuan Malay	Word-formation	Morphology
Papuan Malay	Reduplication	Morphology
Papuan Malay	Word classes	Syntax
Papuan Malay	Personal pronouns	Syntax
Papuan Malay	Demonstratives and locatives	Syntax
Papuan Malay	Noun phrases	Syntax
Papuan Malay	Adnominal possessive relations	Syntax
Papuan Malay	Prepositions and the prepositional phrase	Syntax
Papuan Malay	Verbal clauses	Syntax
Papuan Malay	Nonverbal clauses	Syntax
Papuan Malay	Negative, interrogative, and directive clauses	Syntax
Papuan Malay	Conjunctions and constituent combining	Syntax
Rapa Nui	Introduction	Other
Rapa Nui	Nouns and verbs	Syntax
Rapa Nui	Closed word classes	Syntax
Rapa Nui	Noun phrase	Syntax
Rapa Nui	Possession	Syntax
Rapa Nui	Verb phrase	Syntax
Rapa Nui	Verbal clause	Syntax
Rapa Nui	Nonverbal clauses	Syntax
Rapa Nui	Mood	Semantics
Rapa Nui	Combining clauses	Syntax
Kagayanen	Voice	Syntax
Kagayanen	Pragmatically marked structures	Pragmatics
Kagayanen	Clause combining	Syntax
Kagayanen	Referring expressions	Semantics

Continued on next page

Language	Chapter	Subfield
Kagayanen	Modification	Semantics
Kagayanen	Non-verbal clauses	Syntax
Kagayanen	Verb structure and inflection	Syntax
Kagayanen	Stem-forming morphological processes	Morphology
Kagayanen	Morphosyntactically defined verb classes	Syntax
Kagayanen	Semantically motivated verb classes	Semantics
Vamale	Noun phrases	Syntax
Vamale	Nouns	Syntax
Vamale	Verb phrases	Syntax
Vamale	Verbs	Syntax
Vamale	Voice	Syntax
Vamale	Word classes	Syntax
Komnzo	Word classes	Syntax
Komnzo	Nominal morphology	Morphology
Komnzo	Verb morphology	Morphology
Komnzo	Tense, aspect and mood	Semantics
Komnzo	Syntax of the noun phrase	Syntax
Komnzo	Clausal syntax	Syntax
Komnzo	Complex syntax	Syntax
Komnzo	Aspects of the lexicon	Semantics
Mauwake	Introduction	Other
Mauwake	Morphology	Morphology
Mauwake	Phrase level syntax	Syntax
Mauwake	Clause	Syntax
Mauwake	Functional domains	Semantics
Mauwake	Sentence types	Syntax
Mauwake	Clause combinations	Syntax
Mauwake	Theme, topic, and focus	Semantics
Kalamang	Morphological units and processes	Morphology
Kalamang	Word classes	Syntax
Kalamang	Nouns, noun phrases and postpositional phrases	Syntax
Kalamang	Pronouns and person reference and address	Syntax
Kalamang	Quantifiers	Semantics
Kalamang	Possessive and associative constructions	Semantics
Kalamang	Demonstratives	Semantics
Kalamang	Verbs	Syntax
Kalamang	The clause	Syntax
Kalamang	Complex predicates	Syntax
Kalamang	Clausal modification	Syntax
Kalamang	Multiclausal constructions	Syntax
Kalamang	Information structure	Syntax
Kalamang	Other topics	Other
Ulwa	Adjectives	Syntax
Ulwa	Clause-level syntax	Syntax
Ulwa	Complex sentences	Syntax
Ulwa	Determiners	Syntax
Ulwa	The structural consequences of language loss	Syntax
Ulwa	The Maruat-Dimiri-Yaul dialect of Ulwa	Other
Ulwa	Nouns	Syntax

Continued on next page

Language	Chapter	Subfield
Ulwa	Other word classes	Syntax
Ulwa	A grammatical overview of Ulwa	Syntax
Ulwa	Phrase-level syntax	Syntax
Ulwa	Predicates	Syntax
Ulwa	Pronouns	Syntax
Ulwa	Topics in semantics	Semantics
Ulwa	Additional topics in syntax	Syntax
Ulwa	Verbs	Syntax
Palula	Typological overview	Other
Palula	Nouns	Syntax
Palula	Pronouns	Syntax
Palula	Adjectives and quantifiers	Syntax
Palula	Adverbs and postpositions	Syntax
Palula	Verbs	Syntax
Palula	Verbal categories	Syntax
Palula	Noun phrases and non-verbal agreement	Syntax
Palula	Grammatical relations	Syntax
Palula	Simple clauses and argument structure	Syntax
Palula	Complex constructions	Syntax
Palula	Sentence modification	Syntax
Tuatschin	Phonology	Phonology
Tuatschin	Noun phrase	Syntax
Tuatschin	Verb phrase	Syntax
Tuatschin	Simple sentences	Syntax
Tuatschin	Complex sentences	Syntax
Tuatschin	Morphological processes	Morphology
Japhug	A grammatical sketch	Syntax
Japhug	Phonology	Phonology
Japhug	Nominal morphology	Morphology
Japhug	Pronouns	Syntax
Japhug	Postpositions and relator nouns	Syntax
Japhug	The noun phrase	Syntax
Japhug	Expressive words and sentence final particles	Syntax
Japhug	Non-concatenative verbal morphology	Morphology
Japhug	Tense, aspect, modality and evidentiality	Semantics
Japhug	Simple clauses	Syntax
Japhug	Relative clauses	Syntax
Japhug	Complement clauses	Syntax
Japhug	Other types of multiclausal constructions	Syntax
Japhug	Degree and comparison	Semantics
Yauyos Quechua	Substantives	Syntax
Yauyos Quechua	Verbs	Syntax
Yauyos Quechua	Particles	Syntax
Yauyos Quechua	Enclitics	Syntax
Yauyos Quechua	Syntax	Syntax
Yauyos Quechua	Further analysis of evidential modifiers	Syntax
Mehweb	Phonology	Phonology
Mehweb	Mood of Mehweb	Semantics
Mehweb	Causatives	Syntax

Continued on next page

Language	Chapter	Subfield
Mehweb	Assertive copula in Mehweb	Other
Ik	Adverbs	Syntax
Ik	Case	Syntax
Ik	Demonstratives	Semantics
Ik	Morphology	Morphology
Ik	Verbs	Syntax

G A Concrete Prompt Example

Full Prompt Example

You are a linguist specializing in Palula. You are given a sentence along with its morpheme breakdown, gloss, and translation. Words are separated by spaces, and morphemes are separated by hyphens. However, a word and its gloss are missing and represented by an underscore. Based on your understanding, please choose the most appropriate option.

Sentence (with missing item): panǰ phuṭ-í _____ phuṭ-í kir dít-u síinta.

Gloss (with missing item): five foot-PL _____ foot-PL snow fall.PFV=MSG CONDH

The English translation of this sentence is: 'When five or XXX feet snow had fallen...'

Here is a relevant knowledge point for this example, with the related morphemes and glosses masked: Another strategy for quantification, is by means of a partitive noun phrase. It specifies the quantity of the head noun, often itself preceded by or modified by a cardinal numeral. Typically, but not exclusively, the nouns used in such partitive phrases denote containers or measuring terms of various kinds. In many ways it would make sense to describe higher numerals (such as 20, 100, 1000) as heads of partitive phrases, modified by the cardinal numerals 1–19 to express the numbers 21–39, etc.

Options:
A: word: dubhiš=ee=ṣoṛíiš gloss: two.twenty=and=sixteen
B: word: xálak-a gloss: people-PL
C: word: ṣo gloss: six
D: word: xálaka gloss: people

Please only return the letter (A–D). Do not output anything else.

DeepSeek-R1-7B result: C

Correct Answer: C

Figure 8: A full prompt example of Palula and its prediction under the S+G+KP+T setting.