

# Evaluating Metalinguistic Knowledge in Large Language Models across the World’s Languages

Tjaša Arčon<sup>1</sup>, Matej Klemen<sup>1</sup>, Marko Robnik-Šikonja<sup>1</sup>, and Kaja Dobrovoljc<sup>1,2,3</sup>

<sup>1</sup>University of Ljubljana, Faculty of Computer and Information Science, Ljubljana, Slovenia

<sup>2</sup>University of Ljubljana, Faculty of Arts, Ljubljana, Slovenia

<sup>3</sup>Jožef Stefan Institute, Ljubljana, Slovenia

\*Corresponding author: Tjaša Arčon, Email: tjas.a.arcon@fri.uni-lj.si

February 3, 2026

## Abstract

Large language models (LLMs) are routinely evaluated on language use tasks, yet their explicit knowledge about linguistic structure remains poorly understood. Existing linguistic benchmarks typically focus on narrow phenomena, emphasize high-resource languages, and rarely evaluate metalinguistic knowledge—explicit reasoning about language structure rather than language use. Using accuracy and macro  $F_1$ , together with majority-class and chance baselines, we analyse overall performance and examine variation by linguistic domains and language-related factors. Our results show that metalinguistic knowledge in current LLMs is limited: GPT-4o performs best but still achieves only moderate accuracy (0.367), while open-source models lag behind. All models perform above chance but fail to outperform the majority-class baseline, suggesting they capture broad cross-linguistic patterns but lack fine-grained grammatical distinctions. Performance varies across linguistic domains, with lexical features showing the highest accuracy and phonological features among the lowest, partially reflecting differences in online visibility. At the language level, accuracy shows a strong and consistent association with digital language status: languages with higher digital presence and resource availability are evaluated more accurately, while low-resource languages exhibit substantially lower performance. Analyses of predictive factors confirm that resource-related indicators (Wikipedia size, corpus availability) are more informative predictors of model accuracy than geographical, genealogical or sociolinguistic factors. Together, these results suggest that LLMs’ metalinguistic knowledge is fragmented and strongly shaped by data availability, rather than reflecting broadly generalizable grammatical competence across the world’s languages. We release our benchmark as an open-source dataset to support systematic evaluation of metalinguistic knowledge across the world’s languages and to encourage greater global linguistic diversity in future LLMs.

**Keywords:** large language models; metalinguistic knowledge; large-scale multilingual evaluation; low-resource languages; WALS

## 1 Introduction

Large language models (LLMs) are routinely evaluated on tasks ranging from text generation to question answering, but rarely on their explicit knowledge of language structure. In other words, while we know that LLMs can *use* language fluently (Chang et al., 2024), we know far less about what they know about language itself—a gap that is especially pronounced for low-resource languages, where limited training data may result in even more fragmented or unreliable linguistic representations. Explicit linguistic knowledge includes awareness of grammatical properties such as word order, agreement, case marking, or phonological patterns, which underpin linguistic analysis and explanation. Understanding whether LLMs possess such knowledge is crucial, particularly as they are increasingly employed in linguistically informed tasks such as annotation, grammatical analysis, and cross-linguistic comparison (Beguš et al., 2025; Kellert et al., 2025; Ramji and Ramji, 2025; Waldis et al., 2024), as well as in language documentation, where they are used to accelerate transcription, translation, morphological analysis, glossing, and grammatical description, crucial for the preservation of endangered languages (Berez-Kroeker et al., 2023; Spencer and Kongborirak, 2025; Tanzer et al., 2024).

To support these goals, recent research has begun to probe LLMs’ linguistic knowledge through grammatical classification tasks (Ide et al., 2025), feature-specific evaluations and analyses (Beguš et al., 2025), as well as an increasing number of targeted benchmarks testing phenomena such as agreement, acceptability, and metalinguistic reasoning (Jumelet et al., 2025; Zhang et al., 2024; Behzad et al., 2023). While these efforts provide valuable insights into particular aspects of linguistic competence, they remain narrowly scoped, typically focusing on specific phenomena, tasks, or small language subsets, with a strong emphasis on English and other high-resource languages. As a result, current evaluations provide only a fragmented picture of LLMs’ linguistic knowledge and offer little insight into how such knowledge generalizes across the world’s languages. This limitation is particularly problematic for low-resource and under-documented languages, where the lack of systematic evaluation obscures model weaknesses and risks reinforcing existing biases, making models appear linguistically competent while relying primarily on patterns learned from a small number of digitally dominant languages.

To address this gap, we explore the methodological potential of the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013) as a framework for multilingual evaluation of explicit linguistic (metalinguistic) knowledge in LLMs (as illustrated in Figure 1). WALS documents nearly two hundred grammatical features across more than 2,600 languages, spanning linguistic domains from phonology and morphology to lexicon and syntax, and thus provides a unique basis for large-scale, cross-linguistic evaluation. We systematically

convert WALS features into natural-language questions to construct a QA-style benchmark covering all available languages, and use this benchmark to conduct a multidimensional evaluation of several LLMs.

In doing so, we address the following research questions:

**RQ1:** How accurately do LLMs answer metalinguistic questions about linguistic features across a large and diverse set of languages?

**RQ2:** How does LLM performance vary across different linguistic domains?

**RQ3:** How does LLM performance vary across languages, and which factors are associated with this variation?

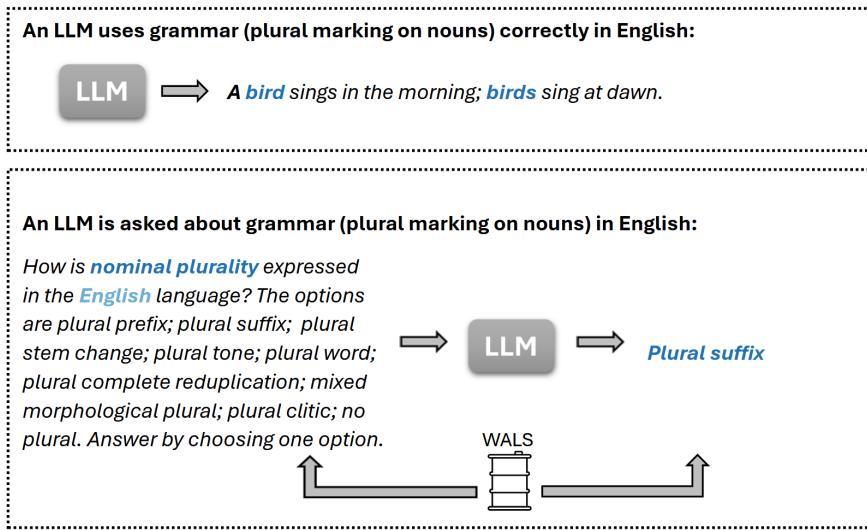


Figure 1: High-level overview of our evaluation setup. While LLMs can use grammatical patterns correctly in language generation (top example), we assess their explicit linguistic knowledge by querying models with WALS-based multiple-choice questions and comparing their responses to the corresponding ground-truth feature values documented in WALS (bottom example).

Our results show that metalinguistic knowledge in current LLMs is limited: even the best-performing model achieves only moderate accuracy, while open-source models lag further behind. Performance varies across linguistic domains, with lexical features showing the highest accuracy and phonological features among the lowest, and across languages, with low-resource languages exhibiting substantially lower performance than digitally well-supported ones. These findings highlight the importance of broad, cross-linguistic evaluation when assessing LLMs’ linguistic competence. To support such evaluation, our main contributions are as follows:

1. **New multilingual benchmark:** We introduce a massively multilingual benchmark for evaluating explicit linguistic (metalinguistic) knowledge in LLMs, grounded in the World Atlas of Language Structures.
2. **Large-scale evaluation:** Using this benchmark, we conduct a large-scale evaluation covering 2,660 languages—including a substantial proportion of low-resource and under-documented languages—and analyse how LLM performance varies across domains and across languages with different levels of digital support.
3. **Methodological insights:** We discuss limitations of using WALS for metalinguistic benchmarking, such as uneven language coverage and categorical feature design, and their implications for future evaluation frameworks.

In the remainder of this paper, Section 2 provides an overview of existing benchmarks for evaluating linguistic knowledge in LLMs; Section 3 introduces the WALS database and describes the construction of our benchmark; Section 4 details the models and evaluation protocol; Section 5 presents the results, and Section 6 discusses their implications and outlines directions for future work.

## 2 Background and related work

LLMs are becoming increasingly important in the scientific methodology (Lu et al., 2024) across a range of fields, including linguistics (Klemen et al., 2025; Spencer and Kongborirak, 2025; Singh et al., 2023). As a result, their evaluation has become increasingly important, as it determines how effectively different models handle specific tasks. In practice, such evaluation is typically carried out through benchmarks that allow for systematic comparison between models. This section examines benchmarks that specifically assess linguistic competence and organizes them according to different dimensions of linguistic knowledge.

Recent evaluation work on LLMs examines multiple aspects of linguistic knowledge, indicating that LLMs’ linguistic ability is better understood as a set of distinct layers rather than a single unified competence. Some available benchmarks focus on explicit grammatical knowledge, testing whether models apply specific grammatical rules and distinguish between correct and incorrect grammatical usage (Section 2.1). Other benchmarks scrutinize metalinguistic competence, assessing the ability of LLMs to act as linguists by reasoning explicitly about language, identifying linguistic structures, or performing linguistic analyses in different languages (Section 2.2). A third group of benchmarks focuses on pedagogical linguistic knowledge, treating LLMs as potential language teachers that can explain a wide variety of grammatical rules in different languages (Section 2.3). Finally, certain benchmarks, such as Holmes (Waldis et al., 2024), investigate the linguistic performance of LLMs at the level of internal embeddings using probing instead of evaluating observable outputs through prompting. As the present work is concerned with linguistic knowledge that is

accessible through direct interaction with LLMs, we restrict our survey below to benchmarks based on prompting rather than internal probing techniques. Accordingly, the following overview is organized from benchmarks that assess surface grammatical competence to those targeting metalinguistic and pedagogical evaluations.

## 2.1 Evaluation of grammatical competence

Several benchmarks investigate the surface linguistic capability. Jumelet et al. (2025) compile a massively multilingual benchmark, MultiBLiMP 1.0, consisting of minimal pairs that test formal grammatical knowledge, evaluating morphosyntactic subject-verb and subject-participle agreement for number, person, and gender across 101 languages. They evaluate 42 language models on grammatical preference using probability-based differences between minimal pairs. Their results show strong performance for high-resource languages, which drops sharply for low-resource languages, even for larger models that consistently outperform smaller ones. Accuracy correlates strongly with language frequency in Common Crawl, suggesting that grammatical competence is mainly data-driven and may deteriorate during post-training.

Similarly, the MELA benchmark (Zhang et al., 2024) assesses whether a model can distinguish between grammatical and ungrammatical sentences, encompassing morphology and syntax features such as word order, agreement, and relative clauses. It measures the linguistic acceptability of presented sentences in ten typologically diverse languages. The findings demonstrate that modern LLMs can perform human-like acceptability judgments across multiple languages, but open-source models lag significantly behind closed models. The benchmark does not include any low-resource language.

PhonologyBench (Suvarna et al., 2024) is an English-only benchmark that tests how well LLMs understand phonology through the grapheme-to-phoneme task, syllable counting, and rhyme judgement. It occupies an intermediate position between surface linguistic competence and explicit metalinguistic reasoning since rhyme judgment reflects surface phonological behaviour, while the other two tasks require implicit phonological analyses. The benchmark is used to evaluate six major LLMs. The study demonstrates that LLM competence remains below human performance, especially for tasks that require abstract phonological reasoning such as syllable counting. Performance varies widely across models and tasks, although LLMs exhibit some phonological awareness despite being trained on texts, indicating that some phonological structure is indirectly learned from orthography.

LINGGYM (Yang et al., 2025) is another benchmark that bridges surface grammatical knowledge and metalinguistic reasoning. The benchmark tests whether models can infer a multiple-choice masked word or word-gloss pair in a sentence based on provided linguistic information, so the models need to apply grammatical descriptions to reconstruct linguistic structure. The benchmark is multi-lingual and spans across eighteen low-resource languages, many of them severely underrepresented. Without grammatical cues models perform

only slightly above chance, but with structured linguistic information accuracy improves across all models. However, even strong LLMs show poor performance, especially on unseen languages, complex morphological paradigms, and abstract grammatical rules.

## 2.2 Evaluation of metalinguistic competence

As having the structure is not the same as talking about the structure, some benchmarks test how well LLMs answer metalinguistic questions about different languages. The first publicly available corpus of metalinguistic questions and answers was ELQA (Behzad et al., 2023), with over 70,000 metalinguistic questions from English learners, collected from two online Stack Exchange forums, covering topics such as grammar, meaning, fluency, and etymology. In contrast to benchmarks evaluating surface grammaticality, ELQA does not test preference or acceptability, but instead assesses whether models can generate accurate and informative linguistic explanations. The results suggest that, although the LLM outputs are fluent, their linguistic validity and correctness are below human performance. Explanations are often partially incorrect or misleading, with models performing better on meaning-related questions than on explicit grammatical analysis.

A dataset that evaluates how well models deal with metalinguistic self-reference was developed by Thrush et al. (2024). The dataset consists of two subtasks: i) generation, where models continue statements with truth-preserving completions, and ii) verification, where they judge the truth of completed statements. To assess whether models can handle metalinguistic language in general, minimally different metalinguistic control tasks without self-reference are included. The study concludes that models struggle with metalinguistic self-reference and perform at or near chance in all domains. Although GPT-4 shows improvement, it remains well below human performance.

IOLBENCH (Goyal and Dan, 2025) evaluates a different kind of metalinguistic knowledge by focusing on linguistic reasoning based on puzzles that are derived from the International Linguistics Olympiad (ILO). The benchmark tests whether models can infer grammatical systems from linguistic data and comes to the conclusion that current LLMs struggle with linguistic tasks that require explicit rule induction, especially without prior knowledge. Moreover, LingBench++ (Lian et al., 2025) is also derived from ILO problems targeting inductive linguistic reasoning across over 90 low-resource and typologically diverse languages. Additionally, it measures LLM reasoning quality and analyses how reasoning unfolds. The results indicate that even strong LLMs struggle with abstract grammatical rule induction. They perform worst on the phonological rule system and multi-rule grammatical systems, but slightly better on lexical and morphological pattern matching.

### 2.3 Evaluation of pedagogical competence

The third group of benchmarks tests pedagogical linguistic knowledge. CPG-EVAL (Wang, 2025) is the first benchmark designed to measure pedagogical grammar competence of LLM in teaching Chinese as a second language. It checks whether models can correctly recognize and discriminate teaching-oriented grammar rules for Chinese. It emerges that LLMs perform strongly on simple grammar recognition, but their performance drops sharply with increasing task complexity.

Similarly, a part of the CLTE benchmark (Xu et al., 2025) addresses linguistic knowledge as part of a broader benchmark that evaluates the pedagogical competence of LLMs functioning as language teachers for Chinese as a second language. This study also confirms that models struggle with pedagogical competence related to linguistic and grammar explanation as performance is below human teacher standards.

### 2.4 Our contribution

Across the benchmarks reviewed above, a number of recurring observations have been reported, including better performance of larger models compared to smaller ones, an advantage for high-resource languages over low-resource ones, and higher accuracy on surface-level grammatical tasks than on tasks requiring abstract, metalinguistic, or pedagogical reasoning. However, these observations emerge from heterogeneous benchmarks that differ substantially in task design, linguistic scope, and language coverage, making it difficult to assess the extent to which such patterns generalize across languages and types of linguistic knowledge, making it difficult to assess how robust they are across the world’s languages and across different types of linguistic knowledge—particularly for typologically diverse and low-resource languages, which remain largely underrepresented.

In this work, we advance the state of the art by introducing a new large-scale multilingual benchmark for evaluating metalinguistic knowledge in LLMs and using it for a systematic, multi-dimensional analysis. Our framework enables evaluation across a broad range of languages and linguistic domains, supports principled comparisons across language groups, and allows us to examine how performance varies with linguistic domain, language characteristics, and resource-related factors. This provides a more comprehensive and fine-grained view of LLMs’ metalinguistic capabilities than prior evaluations focused on smaller language samples or isolated phenomena.

## 3 Benchmark construction from WALS

The World Atlas of Language Structures (WALS) is a large typological database documenting structural properties of the world’s languages (Dryer and Haspelmath, 2013). It covers 192 features across 2,660 languages, with each language annotated for a subset of features based on available descriptive sources. Figure

2 illustrates a typical WALS feature: for each feature, WALS defines a set of possible values and documents which value is attested in which languages.

We chose WALS as the basis for our benchmark because it provides human-verified ground-truth labels across a broad set of languages, including many low-resource ones, and its feature-value structure translates naturally into a multiple-choice QA format. In the following subsections, we describe the language inventory and feature structure in more detail (Sections 3.1-3.2), then explain how we construct the benchmark (Section 3.3).

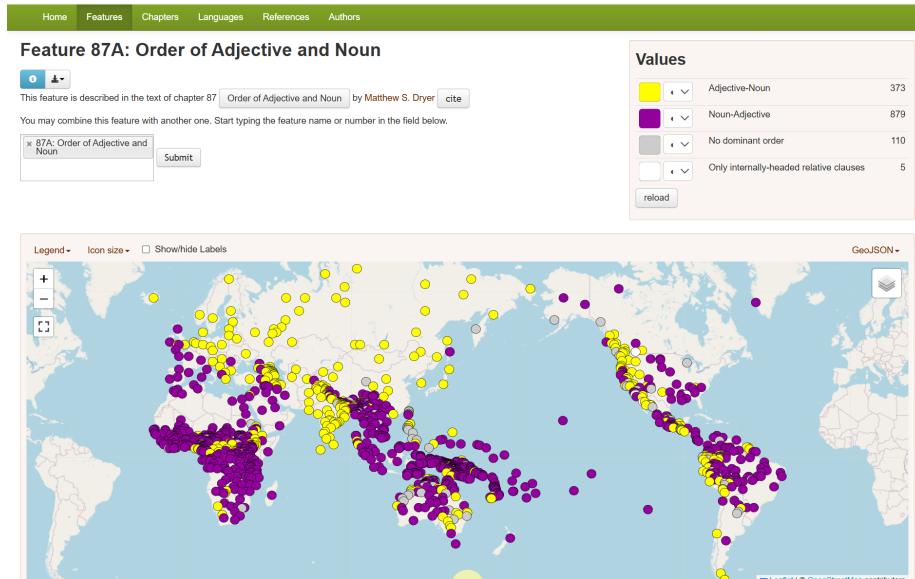


Figure 2: A feature page from WALS Online illustrating how each feature defines a set of possible values (right panel) and maps their distribution across languages (bottom panel).

### 3.1 Languages and samples in WALS

WALS contains data on 2,660 languages, each annotated with metadata such as genus, family, and ISO 639-3 code, enabling genealogical and geographical analyses. The languages are distributed across six major macro-areas: Africa, Eurasia, Papua and Oceania, North America, South America, and Australia. In addition to the full language inventory, the WALS authors also define a curated 100-language sample designed to maximize genealogical and areal diversity and mitigate biases arising from the over-representation of well-documented language families and regions. Because this sample exhibits substantially higher feature coverage (95–159 features per language), we use it as a complementary dataset in our language-level analyses (Section 5.3) to disentangle effects of annotation sparsity from genuine cross-linguistic differences in model perfor-

mance. Table 1 summarizes the distribution of languages across macro-areas for both the full WALS database and the WALS 100-language sample.

Table 1: Distribution of languages across macro-geographical areas in the full WALS database and the WALS 100-language sample.

Macroarea	WALS	WALS-100
Africa	606	17
Eurasia	659	28
Papua and Oceania	560	17
North America	396	18
South America	258	13
Australia	183	7

### 3.2 Linguistic features and domains in WALS

WALS documents 192 structural features that capture different aspects of grammatical organization across languages. Each feature is defined by a fixed set of discrete values representing alternative structural options. Languages are annotated with a single value per feature where data are available; for example, for *Feature 33: Coding of Nominal Plurality*, the value attested for English is *plural suffix* (see Figure 1), and for *Feature 87: Order of Adjective and Noun* it is *adjective-noun* (Figure 2).

Table 2: Distribution of WALS features across linguistic domains, showing the number of features per domain, the number of possible values per feature, and the number of languages for which each feature is attested (reported as minimum–maximum range with mean  $\mu$ ).

Linguistic domain	Num. features	Num. values	Num. lang. per feat.
Word order	56	2 - 28 ( $\mu = 7.61$ )	5 - 1518
Nominal categories	29	2 - 21 ( $\mu = 4.83$ )	71 - 1066
Simple clauses	26	2 - 23 ( $\mu = 5.92$ )	118 - 1157
Phonology	20	2 - 8 ( $\mu = 7.61$ )	40 - 567
Verbal categories	17	2 - 28 ( $\mu = 3.95$ )	193 - 1131
Lexicon	13	2 - 21 ( $\mu = 7.15$ )	72 - 617
Morphology	12	2 - 8 ( $\mu = 5.17$ )	145 - 969
Nominal syntax	8	3 - 8 ( $\mu = 6$ )	124 - 301
Complex sentences	7	2 - 7 ( $\mu = 4.86$ )	112 - 283
Sign languages	2	3 - 6 ( $\mu = 4.50$ )	35 - 38
Clicks (Other)	1	4 - 4 ( $\mu = 4$ )	143 - 143
Writing systems (Other)	1	5 - 5 ( $\mu = 5$ )	6 - 6

The features are grouped into 12 linguistic domains, ranging from phonology

and morphology to clause structure and word order (Table 2). Although the resulting language–feature matrix is sparse (i.e. not every feature is documented for every language), it contains as many as 76,475 manually verified data points. Because all annotations are curated by domain experts based on descriptive linguistic sources, WALS provides a reliable ground-truth resource for evaluating knowledge of specific structural properties across individual languages.

### 3.3 Benchmark construction

To construct a multilingual benchmark from WALS, we transform its structured representation of linguistic features and annotations into a set of explicit metalinguistic benchmark items. Each WALS feature is mapped to a single grammatical question accompanied by a fixed set of answer options, and each item corresponds to a documented grammatical property of a specific language. The annotated WALS value for a given language–feature pair serves as the ground-truth label.

The resulting benchmark comprises 192 distinct question types, one for each WALS feature. These question types function as reusable templates and are instantiated across all languages for which WALS provides annotations, yielding a large set of language-specific question–answer pairs. Questions are derived from WALS feature descriptions, while answer options reflect the corresponding feature value categories. Below is an example for feature 129A (*Hand and Arm*):

**Question:** *How are the concepts of ‘hand’ and ‘arm’ expressed in the X language?*

**Answer options:**

- *Identity – a single word denotes both ‘hand’ and ‘arm’*
- *Differentiation – separate words denote ‘hand’ and ‘arm’*

For more complex features, WALS encodes a large number of highly compressed and terminology-heavy value labels that combine multiple grammatical properties. For instance, Feature 144L (*The Position of Negative Morphemes in SOV Languages*) distinguishes various patterns of negation placement using symbolic shorthand (e.g. *NegSOV* as one of the possible values). To handle such cases, we systematically rephrased both feature names and value labels into clearer formulations that spell out the relevant grammatical configurations (e.g. *What is the position of negative words in subject-object–verb clauses in the X language?* as the question, and *Negative word before the subject, object, and verb* as one of its possible answers). This makes questions more interpretable for both models and readers, and provides some control over potential surface-level memorization effects—a point we return to in Section 6.

Each benchmark entry corresponds to a single linguistic feature and includes a feature identifier and name, a task question, a fixed set of possible answers, and a language-keyed map of ground-truth answers derived from WALS. Ground-truth annotations are provided only for languages attested for each feature. The

dataset is split by feature rather than by language, with features stratified by linguistic domain and assigned to training, validation, and test splits to ensure balanced domain coverage and prevent feature leakage. The datasets are stored in JSON Lines (JSONL) format. The prompt is stored separately from the question content. The benchmark is released as an open-source dataset under the CC-BY-4.0 licence<sup>1</sup>.

## 4 Model setup and evaluation procedure

This section first introduces the LLMs and prompting strategy used (Section 4.1), and then presents the evaluation framework and metrics (Section 4.2). We then describe a set of language-level properties used in our analyses (Section 4.3), such as measures of digital presence and language proximity, which are later examined as potential predictors of model performance.

### 4.1 LLM models and prompting strategy

We tested the benchmark on three models: one large proprietary (GPT-4o) and two large open-source models (Llama-3.3-70B; Gemma-3-27B). The GPT-4o model is chosen as a representative of current state-of-the-art models, while open models are evaluated to examine potential differences in behavior and support reproducibility of our experiments. Smaller models were not systematically evaluated, as they showed poor performance in initial tests, a finding that is consistent with prior research on linguistic knowledge in LLMs (Jumelet et al., 2025).

We prompted GPT-4o through API requests, while the open-source models were run on a high-performance computing cluster for inference. We used a zero-shot prompting strategy, prompting the model with the feature question for each of the languages listed under each linguistic feature. For each WALS feature, models were prompted with a single feature-derived question instantiated for each language annotated for that feature, together with a fixed set of predefined answer options. This setup constrains the task to explicit selection among alternatives rather than free-form generation.

We set the temperature parameter to 0.2 to reduce randomness in model outputs and encourage consistent and deterministic behaviour across runs. This value was selected based on preliminary experiments. An example prompt is shown below:

*“How large is the consonant inventory in the English language? The options are Small; Moderately small; Average; Moderately large; Large. Answer with one of the options only. Do not explain.”*

---

<sup>1</sup>A preliminary version is available at: [https://github.com/Oranzna/metalinguistic\\_benchmark](https://github.com/Oranzna/metalinguistic_benchmark). The final version will be archived on CLARIN.SI and Hugging Face.

## 4.2 Evaluation framework and metrics

This section presents the evaluation metrics and how they are applied at three levels of analysis: overall model performance, performance across linguistic domains, and performance across individual languages.

### 4.2.1 Overall performance evaluation

We evaluated model performance using **accuracy** and **macro  $F_1$**  metrics. Accuracy was calculated separately for each feature as the proportion of correctly generated feature values across all languages for which that feature is documented. Since feature-value classes are frequently imbalanced, with some values occurring much more frequently than others, we also reported macro  $F_1$  for each feature, which assigns equal weight to all classes and therefore provides a more balanced assessment of performance across both frequent and rare values.

To contextualize model performance, we compare accuracy against two simple baselines:

- **Random chance baseline.** The expected accuracy from random selection among the available options. This varies by feature depending on the number of possible values (e.g., 50% for binary features).
- **Majority class baseline.** The accuracy achieved by always predicting the most frequent value for a given feature. Performance above this baseline indicates that a model captures more than just the dominant pattern.

### 4.2.2 Evaluation by linguistic domain

To evaluate performance across linguistic domains, we computed weighted accuracy for each domain. Since features vary considerably in how many languages they cover, we weighted each feature’s accuracy by the proportion of languages it represents within its domain. This means that broadly attested features—those documented across many languages—contribute more to domain-level scores, providing more robust estimates of model performance than features with sparse coverage.

To enable fair comparison across domains with different baseline difficulties, we also computed relative accuracy gain over the majority-class baseline at the feature level, then aggregated using the same weighting procedure:

$$\text{Relative Accuracy Gain}(f) = \frac{\text{Accuracy}_f - \text{Baseline}_f}{\text{Baseline}_f} \quad (1)$$

where  $f$  denotes a feature. This normalises for the fact that some domains have higher majority-class baselines than others (i.e. are inherently easier to predict due to more skewed value distributions).

#### 4.2.3 Evaluation by language

Language performance was measured as the proportion of linguistic features that a model answered correctly out of the total number of features present for that language in WALS.

Direct comparison of model performance at the level of individual languages is challenging due to highly uneven feature coverage in WALS: many languages are annotated for only a small number of features, making per-language accuracy estimates unstable and difficult to interpret. Consequently, ranking all languages would conflate model performance with annotation sparsity. To address this, we adopted a two-stage approach.

First, we perform a coarse-grained analysis by grouping languages according to digital status following the six-class taxonomy of Joshi et al. (2020), which categorises languages based on the availability of labelled and unlabelled resources, ranging from class 0 (very low digital presence, no unlabelled data) to class 5 (dominant digital presence, significant resource investment). This allows us to assess how metalinguistic performance varies with digital support at the group level, aggregating accuracy within each status category rather than comparing individual languages. We perform this analysis on both the full WALS dataset and the WALS 100-language sample.

Second, for the WALS 100-language sample, which provides substantially denser and more uniform annotation (95–159 features per language), we additionally report the top- and bottom-performing languages per model as illustrative examples of language-level variation.

### 4.3 Identifying external factors associated with model performance

To investigate which factors are associated with variation in model performance, we examine external variables at two levels of analysis. At the domain level (Section 4.3.1), we analyse the online visibility of individual linguistic features. At the language level (Section 4.3.2), we examine a set of language-level predictors spanning linguistic, sociolinguistic, and resource-related dimensions.

#### 4.3.1 Domain-level predictor

We check whether model performance across different linguistic domains is related to the online footprint of each of the 192 WALS features. The WALS feature name serves as the search keyword; if the name is too general (e.g. *tone*), it is refined to ensure linguistic relevance (e.g. *tone in language*). Although Google search engine result counts are approximations, they serve as a reasonable proxy for the presence of linguistic features in online texts. We obtain approximate hit counts using the Google Search API, which provides reproducible result estimates.

We compute the Pearson correlation coefficient ( $r$ ) between domain-level accuracy and the average number of hits for features within each domain. We

apply a  $\log_{10}$  transformation to search result counts to account for their wide range and to enable meaningful comparison across domains with very different levels of online prevalence.

#### 4.3.2 Language-level predictors

To examine which factors are associated with language-level performance, we consider eight predictors spanning various dimensions related to digital presence, sociolinguistic status, and linguistic relatedness. Several of these predictors capture partially overlapping aspects of language use and visibility; accordingly, our analysis focuses on their relative importance rather than treating them as independent causal factors. To ensure comparability across languages, we restrict this analysis to the WALS 100-language sample. We consider the following language-level predictors:

- **Resource availability.** We use the aforementioned digital status taxonomy proposed by Joshi et al. (2020), which classifies languages into six categories based on the availability of labelled and unlabelled resources, ranging from very low digital presence (e.g. Bora) to dominant digital presence (e.g. Spanish)
- **Digital language support.** Ethnologue’s global digital language support scale<sup>2</sup> classifies languages into five levels, from *still* (no digital support) to *thriving* (supported by advanced tools, including AI).
- **Language vitality.** Ethnologue’s vitality scale classifies languages into four levels based on intergenerational transmission and institutional use, ranging from *institutional* (the language is used in institutions outside of home and community) to *extinct* (the language is no longer used).
- **Wikipedia size.** Wikipedia<sup>3</sup> size is used as an indicator of a language’s digital presence. Languages with more articles are usually better represented in digital environments and have a more active digital community. We choose the number of articles as an indicator of digital language presence.
- **UD corpus size.** The size of a language’s Universal Dependencies (UD) treebanks (de Marneffe et al., 2021) indicates the availability of curated, grammatically annotated resources and reflects the degree of attention the language has received in computational linguistics research. We use the number of tokens available per language in the latest UD release (Zeman et al., 2024).
- **Geographical macroregion.** The broad geographic area where a language is primarily spoken, following the six macroareas defined in WALS (Africa, Eurasia, Papunesia, Australia, North America, South America).

---

<sup>2</sup><https://www.ethnologue.com/>

<sup>3</sup><https://wikistats.wmcloud.org/display.php?t=wp>

- **Language family.** We use the top-level genealogical family assignments provided by WALS, which cover all languages in the 100-language sample. WALS distinguishes major language families (e.g. Indo-European, Niger–Congo, Austronesian), treating language isolates as single-language families. For example, English is classified as Indo-European.
- **Proximity to English.** We include typological distance measures based on lang2vec representations (Littell et al., 2017), following the distance-based analysis framework of Van Der Goot et al. (2025). Lang2vec encodes languages as vectors of typological features, allowing us to quantify structural similarity to English and assess whether such similarity is associated with higher model performance.

To assess the relative importance of these predictors, we divide languages into three accuracy groups (high, middle, low) and train a random forest classifier to predict group membership using 10-fold cross-validation. Cross-validated performance was assessed with the Matthews correlation coefficient (MCC). Random forests provide interpretable feature-importance scores, allowing us to determine which language-level factors are most strongly associated with model performance. For selected predictors, we additionally report Spearman’s ( $\rho$ ) to quantify monotonic relationships with accuracy.

## 5 Results and analysis

In this section, we present the results of evaluating three LLMs using the newly constructed WALS-based benchmark (Section 3) and experimental setup (Section 4). We begin by examining overall performance (Section 5.1), then analyse performance across linguistic domains and its relationship with online feature visibility (Section 5.2), and finally examine performance across languages and which factors predict language-level variation (Section 5.3).

### 5.1 Overall LLM performance

We first examined LLM performance across individual features by computing accuracy and macro  $F_1$  for each feature. Overall model performance is reported as the unweighted mean of these feature scores.

Overall performance is low across all models (Table 3). GPT-4o achieved the highest accuracy (0.367), followed by Llama-3.3-70B (0.265) and Gemma-3-27B (0.246), with the same ranking for macro  $F_1$ . All models perform well above the chance baseline (0.234), indicating that they capture some systematic regularities rather than guessing at random. However, none outperform the majority-class baseline (0.539), meaning their predictions fail to improve upon simply selecting the most frequent feature value.

Together, these results show that metalinguistic question answering remains a challenging task for current LLMs. While models capture broad grammatical

regularities, their knowledge reflects dominant cross-linguistic patterns rather than fine-grained, language-specific distinctions.

Table 3: Overall LLM performance on the WALS-based metalinguistic benchmark. We report unweighted mean accuracy and macro  $F_1$  across all 192 grammatical features.

LLM model	Accuracy	Macro $F_1$
Chance baseline	0.234	
Majority-class baseline	0.539	
GPT-4o	0.367	0.228
Llama-3.3-70B	0.265	0.157
Gemma-3-27B	0.246	0.129

## 5.2 LLM performance across linguistic domains

We next examined how model performance varies across linguistic domains (Figure 3). Accuracy was highest for questions related to lexicon and verbal categories, and lowest for phonology, nominal syntax, and sign languages. This pattern was consistent across all three models, though GPT-4o additionally showed moderately strong performance on nominal categories.

Relative accuracy gains over the majority-class baseline - which account for differing baseline difficulties across domains - confirm this pattern. All domains show negative gains, but the magnitude varies substantially. For GPT-4o, the smallest deficits appear for nominal categories (-0.03) and morphology (-0.14), while the largest deficits appear for sign languages (-0.59) and nominal syntax (-0.47). The pattern is similar for the other models, with sign languages, nominal syntax, and phonology consistently showing the weakest performance across all three LLMs.

To investigate whether this variation reflects differences in how well linguistic phenomena are represented online, we computed the correlation between domain-level accuracy and mean Google search hit counts for features within each domain (see Section 4.3.1). We excluded three domains containing only one or two features (Clicks, Writing Systems, Sign Languages), as their estimates are less stable. For GPT-4o, accuracy correlates strongly with online visibility ( $r = 0.715$ ; Figure 4), with a similar pattern for Gemma-3-27B ( $r = 0.571$ ). No such relationship was observed for Llama-3.3-70B ( $r = 0.045$ ), which may be attributed to differences in training data composition and curation strategies.

Together, these results show that metalinguistic knowledge in LLMs is unevenly distributed across linguistic domains, with some domains substantially easier than others. Our analysis suggests that the online visibility of linguistic phenomena may partially account for this variation, although the relationship is not consistent across models and warrants further investigation.

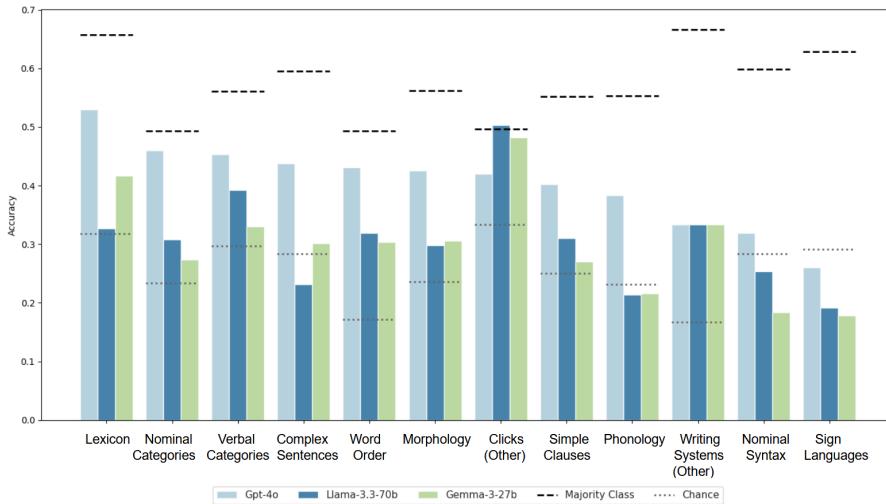


Figure 3: Normalized LLM accuracy across linguistic domains relative to majority-class and chance baselines, ranked by GPT-4o performance.

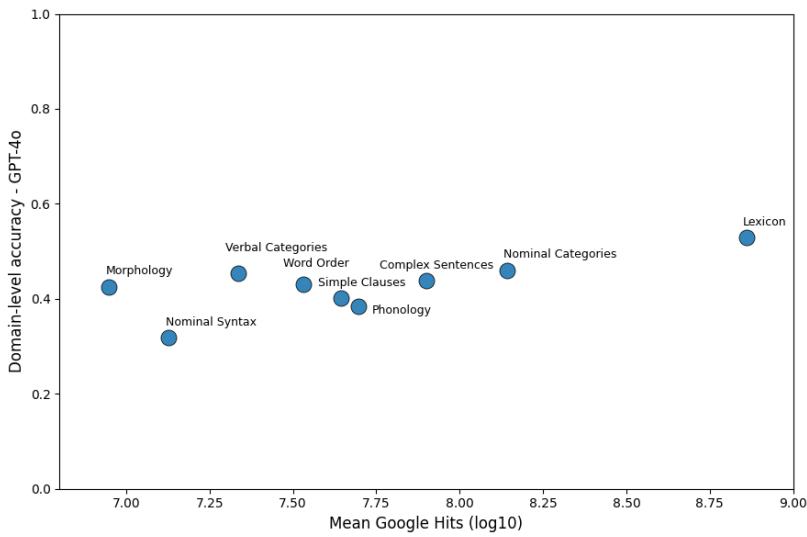


Figure 4: Correlation between domain-level accuracy and online visibility (mean Google hits per domain) for GPT-4o ( $r = 0.715$ ).

## 5.3 LLM performance across languages

Finally, we examine how LLM performance varies across languages, focusing on performance by digital language status (Section 5.3.1), illustrative top- and bottom-performing languages (Section 5.3.2), and language-level predictors of model accuracy (Section 5.3.3).

### 5.3.1 Performance by language status

To examine the relationship between model performance and language resourcedness, we grouped languages by digital status according to the six-class taxonomy of Joshi et al. (2020): 0 = very low digital presence, no unlabelled data (2,191 languages, e.g., Bora); 1 = low, some unlabelled data (222 languages, e.g., Navajo); 2 = low, some labelled data (19 languages, e.g., Zulu); 3 = moderate, insufficient labelled data (28 languages, e.g., Hebrew); 4 = strong, large unlabelled but less labelled data (18 languages, e.g., Hungarian); 5 = dominant, significant resource investment (7 languages, e.g., Spanish). Figures 5–7 present the distribution of mean accuracy by digital status for all three models, shown for both the full WALS dataset and the 100-language sample. A clear pattern emerges: languages with higher digital status achieve higher accuracy across all models. Because digital status is ordinal, we report Spearman’s correlation ( $\rho$ ). Correlations are relatively weak for the full dataset (GPT-4o:  $\rho = 0.227$ ; Llama-3.3-70B:  $\rho = 0.23$ ; Gemma-3-27B:  $\rho = 0.182$ ), but substantially stronger for the 100-language sample (GPT-4o:  $\rho = 0.734$ , Llama-3.3-70B:  $\rho = 0.710$  and Gemma-3-27B:  $\rho = 0.598$ ).

These results reveal three key patterns. First, metalinguistic performance varies substantially across languages: models consistently struggle more with some languages than others. Second, resource availability emerges as a strong predictor of this variation across all three models — languages with limited digital presence perform worse regardless of model architecture or size. Third, GPT-4o, the largest model in our comparison, achieves the highest accuracy overall, with the advantage most pronounced for well-resourced languages. This suggests that increased model capacity amplifies the benefit of abundant training data, but does not compensate for the lack of it.

### 5.3.2 Top- and bottom-performing languages

To complement the population-level analysis, we next examine performance differences between individual languages. Since direct per-language comparison is unreliable for the full WALS dataset—where some languages are annotated for only a handful of features—we restrict this analysis to the WALS 100-language sample, which provides a more uniform feature coverage.

Tables 4 and 5 list the ten highest- and lowest-performing languages for each model. Across all models, the top-performing languages are predominantly high-resource, digitally well-supported languages such as English, German, French, Spanish, and Mandarin. In contrast, the bottom-performing languages are largely low-resource languages with limited digital presence, including Barasano,

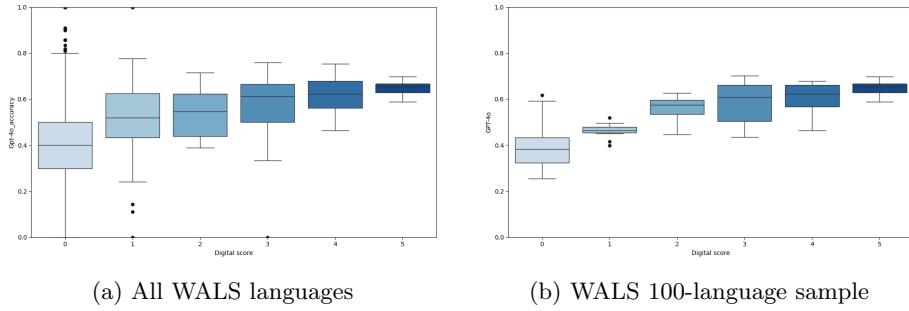


Figure 5: Distribution of accuracy by digital status (0 = very low to 5 = dominant) for GPT-4o.

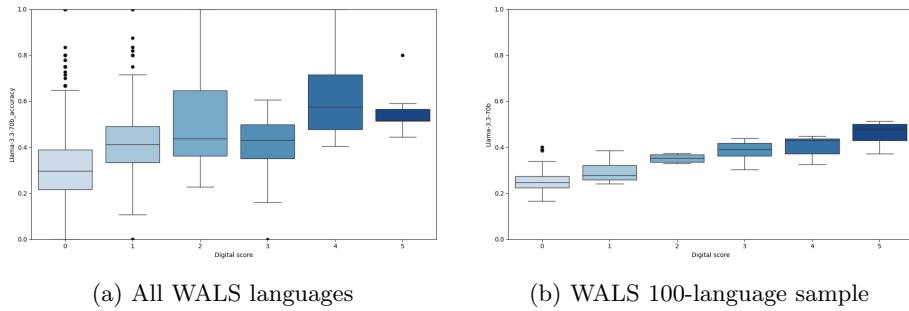


Figure 6: Distribution of accuracy by digital status (0 = very low to 5 = dominant) for Llama-3.3-70b.

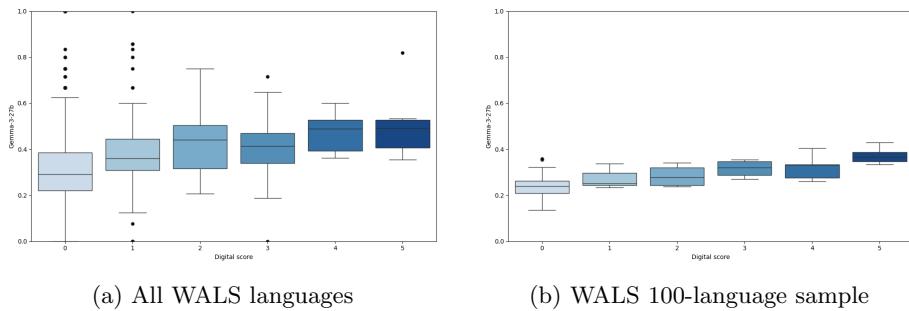


Figure 7: Distribution of accuracy by digital status (0 = very low to 5 = dominant) for Gemma-3-27b.

Imonda, Wichí, and Kutenai — a pattern consistent with the status analysis above (Figures 5–7).

Table 4: Top ten languages by accuracy for each model (WALS 100-language sample).

<b>GPT-4o Accuracy</b>	<b>Llama-3.3-70b Accuracy</b>	<b>Gemma-3-27b Accuracy</b>
Hebrew (M) 0.702	French 0.513	Spanish 0.429
English 0.698	English 0.503	Turkish 0.405
Russian 0.680	Mandarin 0.497	English 0.392
German 0.675	German 0.478	French 0.382
Thai 0.669	Vietnamese 0.448	German 0.365
Finnish 0.665	Spanish 0.445	Korean 0.358
Spanish 0.658	Turkish 0.442	Indonesian 0.355
Vietnamese 0.657	Hebrew (M) 0.440	Mandarin 0.354
Mandarin 0.654	Finnish 0.432	Kannada 0.354
French 0.653	Russian 0.430	Greek (M) 0.351

Table 5: Bottom ten languages by accuracy for each model (WALS 100-language sample).

<b>GPT-4o Accuracy</b>	<b>Llama-3.3-70b Accuracy</b>	<b>Gemma-3-27b Accuracy</b>
Barasano 0.313	Canela 0.199	Mixtec 0.176
Lavukaleve 0.311	Otomí 0.193	Maybrat 0.173
Rama 0.307	Apurinã 0.192	Maricopa 0.165
Nama 0.307	Otomí 0.193	Karok 0.162
Alamblak 0.306	Paiwan 0.190	Canela 0.162
Wari’ 0.306	Kutenai 0.188	Imonda 0.160
Wichí 0.302	Wichí 0.186	Slave 0.155
Sanuma 0.298	Mixtec 0.182	Apurinã 0.154
Imonda 0.278	Maybrat 0.179	Lakhota 0.149
Maricopa 0.254	Kayardild 0.167	Kutenai 0.136

### 5.3.3 Predictors of language-level performance

To examine which factors best predict language-level accuracy, we trained a random forest classifier to predict performance group (high, middle, low) based on the eight predictors described in Section 4.3.2.<sup>4</sup> Model training was performed using 10-fold cross validation. Cross-validated performance, evaluated using the MCC, was 0.581 for GPT-4o, 0.589 for Llama-3.3-70b, and 0.403 for

<sup>4</sup>We measure association, which may not necessarily translate into causation.

Gemma-3-27b, indicating a moderate-to-strong association between the predictors and performance group. Figure 8 shows the feature importance rankings for all three models.

Resource-related factors emerge as the strongest predictors across all models. Wikipedia size ranks highest for all three (GPT-4o: 0.148; Llama-3.3-70B: 0.151; Gemma-3-27B: 0.136), followed by resource availability (GPT-4o: 0.125; Llama-3.3-70B: 0.124; Gemma-3-27B: 0.103). This is consistent with the digital-status analysis above: languages with larger digital footprints are easier for models to answer questions about, likely because more descriptive and metalinguistic content about these languages is available in training data.

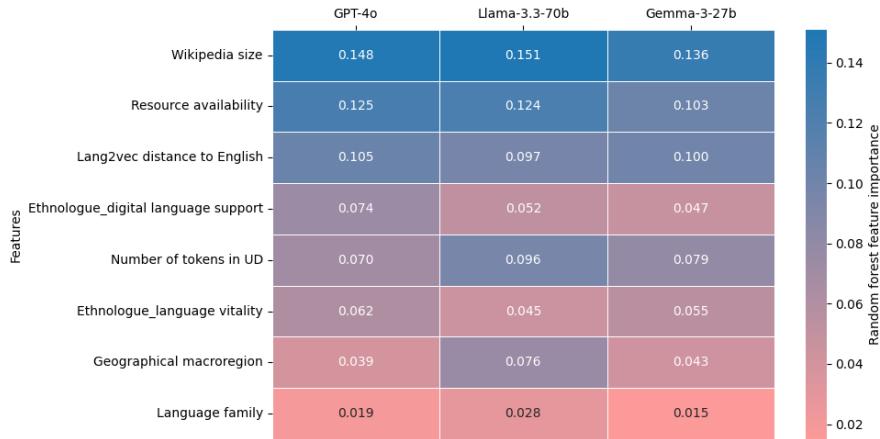


Figure 8: Random forest feature importance for predicting language-level accuracy group (high, middle, low) across three models.

Notably, Lang2vec distance to English ranks third across all models (GPT-4o: 0.105; Llama-3.3-70B: 0.097; Gemma-3-27B: 0.100), ahead of sociolinguistic factors such as language vitality and geographical macroregion. This suggests that typological similarity to English — the dominant language in LLM training data — provides an additional advantage beyond mere resource availability. We explain this by noting that English is likely the main source of training data for the tested LLMs.

In contrast, language family is consistently the weakest predictor (GPT-4o: 0.019; Llama-3.3-70B: 0.028; Gemma-3-27B: 0.015), indicating that genealogical relatedness contributes little to model accuracy when resource-related factors are taken into account. Sociolinguistic factors such as Ethnologue language vitality and geographical macroregion fall in the middle range, suggesting that while these factors play some role, they are less informative than direct measures of digital presence.

Together, these results indicate that LLM metalinguistic performance is

shaped primarily by data availability, with typological proximity to English as a secondary factor. Genealogical and sociolinguistic properties of languages, by contrast, explain relatively little of the variation.

## 6 Discussion

In this work, we introduced a massively multilingual benchmark for evaluating metalinguistic knowledge in LLMs, derived from grammatical features documented in the World Atlas of Language Structures. Using this benchmark, we evaluated three contemporary LLMs across nearly two hundred linguistic features and more than 2,600 languages. Our results confirm earlier findings from smaller-scale studies that LLMs exhibit limited explicit grammatical knowledge but extend them to a global scale and a much broader range of linguistic domains. We show that metalinguistic performance varies substantially across domains and languages, with particularly low accuracy for phonological and syntactically complex features, and systematic disadvantages for low-resource languages across all models. Beyond these findings, the benchmark opens many opportunities for further exploration leveraging the rich human-curated knowledge encoded in WALS: analyses targeting specific domains, geographical regions, or language families; correlations with external factors beyond those examined here; and systematic comparison across a wider range of models and prompting strategies.

At the same time, using WALS as a benchmarking resource entails several methodological limitations. As a typological database, WALS was designed to capture broad structural distinctions relevant for cross-linguistic comparison, rather than to provide exhaustive grammatical descriptions. Its feature inventory reflects particular descriptive traditions and theoretical choices, which could be expanded or complemented in future benchmarks. Second, feature coverage is sparse and uneven, making generalizations difficult; we address this by using the WALS 100-language sample, though future benchmarks could target feature subsets with more uniform attestation across languages (or draw on similar resources like Grambank (Skirgård et al., 2023), which offers more systematic per-language coverage, though without the phonological and lexical domains examined here). Third, WALS represents grammatical properties as discrete feature values, which do not necessarily reflect the gradient, context-dependent patterns of language use observed in corpora (Yan and Liu, 2023; Levshina et al., 2023; Baylor et al., 2023). While discrete categories provide a clear evaluation target, future benchmarks could complement them with corpus-based representations that capture gradience and variation in actual language use (e.g., Klemen et al., 2025; Baylor et al., 2024).

The use of WALS also introduces some experimental considerations. Because WALS is publicly available, models may have encountered its content during training. While low overall accuracy and systematic variation across domains and languages suggest that direct retrieval is not a significant factor, future work could introduce additional controls such as paraphrased questions

and answers across the full set of features (see Section 3.3) or newly added features. A further limitation is the multiple-choice prompt format. Recent evidence suggests that models may exploit option-level artifacts, such as elimination heuristics or surface cues in the answer choices, leading them to select correct options without fully solving the underlying task Raman et al. (2025). However, the low accuracy observed here suggests that such prompt-induced effects were not primary drivers of model performance.

Despite these limitations, our benchmark provides an unprecedented new resource for evaluating LLMs across low-resource and underdocumented languages. Our findings reveal that limited digital presence affects not only performance on standard NLP tasks, but also models’ explicit knowledge about language itself—a critical gap given the growing interest in using LLMs to support language documentation and preservation. By making such disparities visible at a global scale, our work underscores the importance of inclusive evaluation frameworks and demonstrates the value of leveraging human-curated linguistic knowledge to probe model behaviour beyond surface language use. We release the benchmark as an open-source dataset to support this goal.

## 7 Conclusion

We introduced a massively multilingual benchmark for evaluating metalinguistic knowledge in LLMs, drawing on the linguistic features and languages documented in WALS. Our evaluation of three contemporary models reveals that metalinguistic knowledge in current LLMs is limited and fragmentary: accuracy is low overall, varies systematically across linguistic domains, and correlates strongly with resource availability. These findings suggest that current LLMs mainly reflect the distribution of digitally available data rather than exhibiting generalizable grammatical knowledge across the world’s languages.

By scaling metalinguistic evaluation to over 2,600 languages, this study provides the most comprehensive assessment of LLMs’ explicit linguistic knowledge to date, revealing that the languages most in need of computational support are precisely those about which models know the least. We release the benchmark as an open-source resource to support further research on metalinguistic evaluation across the world’s languages.

## Acknowledgments

The work was primarily supported by the Large Language Models for Digital Humanities project (GC-0002), funded by the Slovene Research and Innovation Agency (ARIS), and the ARIS core research programme P6-0411. Additional support was provided by the EU ERA Chair grant no. 101186647 (AI4DH).

Competing interests: None.

Generative AI tools were used to assist with language editing.

## References

- Baylor, E., Ploeger, E., and Bjerva, J.** (2023). The past, present, and future of typological databases in NLP. In **Bouamor, H., Pino, J., and Bali, K.**, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 1163–1169, Singapore. Association for Computational Linguistics.
- Baylor, E., Ploeger, E., and Bjerva, J.** (2024). Multilingual gradient word-order typology from Universal Dependencies. In **Graham, Y. and Purver, M.**, editors, *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 42–49.
- Beguš, G., Dabkowski, M., and Rhodes, R.** (2025). Large linguistic models: investigating LLMs' metalinguistic abilities. *IEEE Transactions on Artificial Intelligence*.
- Behzad, S., Sakaguchi, K., Schneider, N., and Zeldes, A.** (2023). ELQA: A corpus of metalinguistic questions and answers about English. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2031–2047.
- Berez-Kroeker, A. L., Gabber, S., and Slayton, A.** (2023). Recent advances in technologies for resource creation and mobilization in language documentation. *Annual Review of Linguistics*, 9(Volume 9, 2023):195–214.
- Chang, Y., Wang, X., Wang, J., Wu, Y., Yang, L., Zhu, K., Chen, H., Yi, X., Wang, C., Wang, Y., and others** (2024). A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.
- de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D.** (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.
- Dryer, M. S. and Haspelmath, M.** (2013). WALS online (v2020.4).
- Goyal, S. and Dan, S.** (2025). IOLBENCH: Benchmarking LLMs on linguistic reasoning. *arXiv preprint arXiv:2501.04249*.
- Ide, Y., Nishida, Y., Vasselli, J., Oba, M., Sakai, Y., Kamigaito, H., and Watanabe, T.** (2025). How to make the most of LLMs' grammatical knowledge for acceptability judgments. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7416–7432.
- Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M.** (2020). The state and fate of linguistic diversity and inclusion in the NLP world. In **Jurafsky, D., Chai, J., Schluter, N., and Tetreault, J.**, editors,

*Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 6282–6293.

**Jumelet, J., Weissweiler, L., Nivre, J., and Bisazza, A.** (2025). Multi-BLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *arXiv preprint arXiv:2504.02768*.

**Kellert, O., Tyagi, N., Imran, M., Licona-Guevara, N., and Gómez-Rodríguez, C.** (2025). Parsing the switch: LLM-based UD annotation for complex code-switched and low-resource languages. *arXiv preprint arXiv:2506.07274*.

**Klemen, M., Arčon, T., Terčon, L., Robnik-Šikonja, M., and Dobrovoljc, K.** (2025). Towards corpus-grounded agentic LLMs for multilingual grammatical analysis. *arXiv preprint arXiv:2512.00214*.

**Levshina, N., Namboodiripad, S., Allasonnière-Tang, M., Kramer, M., Talamo, L., Verkerk, A., Wilmoth, S., Rodriguez, G. G., Gutton, T. M., Kidd, E., Liu, Z., Naccarato, C., Nordlinger, R., Panova, A., and Stoynova, N.** (2023). Why we need a gradient approach to word order. *Linguistics*, 61(4):825–883.

**Lian, D.-C., Huang, R.-S., Chen, P.-E., Lim, C., Lin, Y.-K., Tseng, G.-Y., Yang, Z.-C., Lin, Z.-Y., Chen, P.-C., and Hsieh, S.-K.** (2025). LingBench++: A linguistically-informed benchmark and reasoning framework for multi-step and cross-cultural inference with LLMs. *arXiv preprint arXiv:2507.16809*.

**Littell, P., Mortensen, D. R., Lin, K., Kairis, K., Turner, C., and Levin, L.** (2017). URIEL and lang2vec: Representing languages as typological, geographical, and phylogenetic vectors. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 8–14, Valencia, Spain. Association for Computational Linguistics.

**Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D.** (2024). The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

**Raman, N., Lundy, T., and Leyton-Brown, K.** (2025). Reasoning models are test exploiters: Rethinking multiple-choice. *arXiv preprint arXiv:2507.15337*.

**Ramji, R. and Ramji, K.** (2025). Inductive linguistic reasoning with large language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 22783–22810.

- Singh, C., Morris, J. X., Aneja, J., Rush, A., and Gao, J.** (2023). Explaining data patterns in natural language with language models. In **Belinkov, Y., Hao, S., Jumelet, J., Kim, N., McCarthy, A., and Mohedbi, H.**, editors, *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pp. 31–55.
- Skirgård, H., Haynie, H. J., Blasi, D. E., Hammarström, H., Collins, J., Latarche, J. J., Lesage, J., Weber, T., Witzlack-Makarevich, A., Passmore, S., Chira, A., Maurits, L., Dinnage, R., Dunn, M., Reesink, G., Singer, R., Bowern, C., Epps, P., Hill, J., Vesakoski, O., Robbeets, M., Abbas, N. K., Auer, D., Bakker, N. A., Barbos, G., Borges, R. D., Danielsen, S., Dorenbusch, L., Dorn, E., Elliott, J., Falcone, G., Fischer, J., Ate, Y. G., Gibson, H., Göbel, H.-P., Goodall, J. A., Gruner, V., Harvey, A., Hayes, R., Heer, L., Miranda, R. E. H., Hübler, N., Huntington-Rainey, B., Ivani, J. K., Johns, M., Just, E., Kashima, E., Kipf, C., Klingenberg, J. V., König, N., Koti, A., Kowalik, R. G. A., Krasnoukhova, O., Lindvall, N. L., Lorenzen, M., Lutzenberger, H., Martins, T. R., German, C. M., van der Meer, S., Samamé, J. M., Müller, M., Muradoglu, S., Neely, K., Nickel, J., Norvik, M., Oluoch, C. A., Peacock, J., Pearcey, I. O., Peck, N., Petit, S., Pieper, S., Poblete, M., Prestipino, D., Raabe, L., Raja, A., Reimringer, J., Rey, S. C., Rizaew, J., Ruppert, E., Salmon, K. K., Sammet, J., Schembri, R., Schlabbach, L., Schmidt, F. W., Skilton, A., Smith, W. D., de Sousa, H., Sverredal, K., Valle, D., Vera, J., Voß, J., Witte, T., Wu, H., Ye, J., Yong, M., Yuditha, T., Zariquiey, R., Forkel, R., Evans, N., Levinson, S. C., Haspelmath, M., Greenhill, S. J., Atkinson, Q. D., and Gray, R. D.** (2023). Grambank reveals the importance of genealogical constraints on linguistic diversity and highlights the impact of language loss. *Science Advances*, 9(16).
- Spencer, P. T. and Kongborirak, N.** (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*, pp. 10214–10227.
- Suvarna, A., Khandelwal, H., and Peng, N.** (2024). PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pp. 1–14.
- Tanzer, G., Suzgun, M., Visser, E., Jurafsky, D., and Melas-Kyriazi, L.** (2024). A benchmark for learning to translate a new language from one grammar book.
- Thrush, T., Moore, J., Monares, M., Potts, C., and Kiela, D.** (2024). I am a strange dataset: Metalinguistic tests for language models. *arXiv preprint arXiv:2401.05300*.

- Van Der Goot, R., Ploeger, E., Blaschke, V., and Samardzic, T.** (2025). DistalLs: A comprehensive collection of language distance measures. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 307–318.
- Waldis, A., Perlitz, Y., Choshen, L., Hou, Y., and Gurevych, I.** (2024). Holmes: A benchmark to assess the linguistic competence of language models. *Transactions of the Association for Computational Linguistics*, 12:1616–1647.
- Wang, D.** (2025). CPG-EVAL: A multi-tiered benchmark for evaluating the Chinese pedagogical grammar competence of large language models. *arXiv preprint arXiv:2504.13261*.
- Xu, L., Li, Q., Peng, T., Li, Z., Zhao, H., and Wang, P.** (2025). Can large language models be good language teachers? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 23968–23982.
- Yan, J. and Liu, H.** (2023). Basic word order typology revisited: A crosslinguistic quantitative study based on UD and WALS. *Linguistics Vanguard*, 9(1):73–85.
- Yang, C., Ma, F., Shi, F., and Zhu, J.** (2025). LingGym: How far are LLMs from thinking like field linguists? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 1314–1340.
- Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., and ...** (2024). Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.
- Zhang, Z., Liu, Y., Huang, W., Mao, J., Wang, R., and Hu, H.** (2024). MELA: Multilingual evaluation of linguistic acceptability. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2658–2674.