ALBA: A European Portuguese Benchmark for Evaluating Language and Linguistic Dimensions in Generative LLMs

Inês Vieira[1], **Inês Calvo**[1], **Iago Paulo**[1,2], **James Furtado**[1,2], **Rafael Ferreira**[1,2], **Diogo Tavares**[1,2], **Diogo Glória-Silva**[1,2], **David Semedo**[1,2], **João Magalhães**[1,2],
[1]NOVA University of Lisbon, Portugal, [2]NOVA LINCS
{im.vieira,i.calvo,df.semedo,jmag}@fct.unl.pt
{im.paulo,jh.furtado,rah.ferreira,dc.tavares,dmgc.silva}@campus.fct.unl.pt

Abstract

As Large Language Models (LLMs) expand across multilingual domains, evaluating their performance in under-represented languages becomes increasingly important. European Portuguese (pt-PT) is particularly affected, as existing training data and benchmarks are mainly in Brazilian Portuguese (pt-BR). To address this, we introduce ALBA, a linguistically grounded benchmark designed from the ground up to assess LLM proficiency in linguistic-related tasks in pt-PT across eight linguistic dimensions, including Language Variety, Culture-bound Semantics, Discourse Analysis, Word Plays, Syntax, Morphology, Lexicology, and Phonetics and Phonology. ALBA is manually constructed by language experts and paired with an LLM-as-a-judge framework for scalable evaluation of pt-PT generated language. Experiments on a diverse set of models reveal performance variability across linguistic dimensions, highlighting the need for comprehensive, variety-sensitive benchmarks that support further development of tools in pt-PT[1].

1 Introduction

While Large Language Models (LLMs) have generally progressed remarkably, their progress in lower-resource languages has been less marked (Zhang et al., 2023; Kim et al., 2024). High-resource languages form the prime focus, frequently relegating low-resource languages benchmarks to machine translation (MT) datasets. European Portuguese (pt-PT) exemplifies this issue, as data is overwhelmingly dominated by Brazilian Portuguese (pt-BR), leading to systematic biases in which pt-PT is frequently conflated with pt-BR during both training and evaluation. As a result, various assessments provide only a partial, and often misleading, picture of LLM capabilities for the pt-PT variety.

Figure 1: ALBA allows the assessment of LLM generative capabilities across eight linguistic dimensions.

Existing evaluation frameworks for pt-PT suffer from two major limitations. First, most benchmarks are English-centric or rely on machine translation from English (Lai et al., 2023; Thellmann et al., 2024). While MT offers a scalable and convenient solution, it introduces a systemic bias that obscures language-specific phenomena such as wordplay, rhyme, or idiomatic expressions, making it unsuitable for fine-grained linguistic evaluation. Second, due to the under-representation of pt-PT data, most models frequently default to pt-BR lexical, morphological, or syntactic patterns, producing outputs that lack pt-PT linguistic authenticity (Piot et al., 2025; Lopes et al., 2024; Riley et al., 2023; González et al., 2026).

Facing the need to assess LLMs' generative quality in pt-PT, we introduce ALBA (**A**utomated **L**inguistics **B**enchmark for baseline **A**ssessment), a benchmark manually created by domain experts that departs from standard binary/multiple-choice evaluation in favor of text generation in order to

[1]https://github.com/AMALIA-LLM/alba-benchmark

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 1, pages 697–707

engage with language on multiple levels. In this work, inspired by the efforts done in other low-resource languages facing similar challenges (Kim et al., 2024; Norman et al., 2025), we propose a selection of eight linguistic dimensions to evaluate the linguistic quality of LLMs for European Portuguese (pt-PT): Language Variety, Culture-bound Semantics, Discourse Analysis, Word Plays, Syntax, Morphology, Lexicology, and Phonetics and Phonology (Figure 1).

To support scalable and reliable evaluation, we further propose a rigorously validated LLM-as-a-judge framework (Gu et al., 2024) for scoring open-ended responses. This judge was calibrated against expert annotations, ensuring alignment with native-speaker intuition while enabling systematic comparison across models. Our contributions are threefold:

- ALBA, a novel benchmark, created by language experts, specifically designed to evaluate the pt-PT linguistic capabilities of generative language models (Section 3);
- An LLM-as-a-judge framework that leverages ALBA to assess the pt-PT generation quality of LLMs (Section 4);
- An extensive evaluation study of LLMs, revealing fine-grained strengths and weaknesses across ALBA's eight linguistic dimensions (Section 5).

By moving beyond machine translation-based datasets, ALBA offers a linguistically faithful and variety-aware framework that advances the assessment of LLM proficiency in linguistic-related tasks in pt-PT.

2 Related Work

Language Evaluation Benchmarks. Although benchmarks for evaluating the quality of LLM language output have been previously developed, research efforts predominantly focus on high-resource languages, while low-resource languages are often confined to machine translation benchmarks. This limitation is not unique to pt-PT, as many languages similarly lack evaluation benchmarks developed for native-language assessment.

In the case of Korean, Kim et al. (2024) created CLIck, a benchmark that harnesses QA pairs from examinations and textbooks to address the lack of non-translated resources and cultural nuance in existing benchmarks. As for Danish, the Danish NLU Benchmark (Norman et al., 2025), created with the purpose of diagnosing and potentially remedying language and cultural biases that LLMs have in low-resource languages, has tasks involving synonymy, semantic similarity, word sense disambiguation, sentiment of words in context, entailment, and idiom interpretation.

Linguistics Benchmarks. Linguistically grounded evaluation has an important role in the assessment of language performance. Various benchmarks evaluate LLMs through linguistically motivated tasks across multiple languages, offering structured insights into specific aspects of model behavior. For example, IOLBENCH (Goyal and Dan, 2025), derived from the International Linguistics Olympiad, is primarily focused on linguistics-oriented reasoning.

Others concentrate on a single linguistic subfield or on narrowly defined task types. This is the case for PhonologyBench (Suvarna et al., 2024), which evaluates phonological awareness in LLMs, and TACOMORE (Li and Wang, 2024), a prompting framework tailored for corpus-based discourse analysis with tasks centered on keyword, collocate, and concordance analysis. In addition, other Portuguese linguistics benchmarks are unavailable in pt-PT, such as BRoverbs (Almeida et al., 2025), which evaluates the comprehension of pt-BR proverbs.

Other works have explored the evaluation of specific linguistic competencies, including morphological generalization through compositionality in Turkish and Finnish (Ismayilzada et al., 2025), as well as wordplay detection for authorship attribution in French (Cafiero and Puren, 2025). These approaches highlight the diversity and depth of linguistically informed evaluation, while also underscoring the language-specific nature of many linguistic phenomena.

European Portuguese Benchmarks. Evaluation benchmarks for pt-PT have been developed using a variety of approaches. Some benchmarks are derived via MT from English resources; for example, PORTULAN ExtraGLUE (Osório et al., 2024) builds on the GLUE (Wang et al., 2019) and SuperGLUE (Wang et al., 2020), enabling Portuguese models to be evaluated on tasks originally designed for English. Other benchmarks are manually translated to preserve subtle linguistic distinc-

tions. BATS-PT (Oliveira et al., 2024) is a manual translation of the lexicographic portion of the Bigger Analogy Test Set (BATS) (Gladkova et al., 2016), supporting analogical reasoning while retaining language-specific nuances. In parallel, several resources are conceived directly in Portuguese, either by adapting native texts or focusing on specific tasks. CALAME-PT (Lopes et al., 2024) evaluates text completion, and ASSIN 2 (Real et al., 2020) provides manually annotated sentence pairs for semantic similarity and textual entailment.

While translation enables rapid expansion of benchmarks across languages, many linguistic phenomena do not transfer cleanly, potentially compromising evaluation validity. This limitation has motivated the creation of resources designed natively in Portuguese. ALBA addresses this challenge by covering multiple linguistic dimensions and being developed entirely in pt-PT from the outset.

3 ALBA Benchmark Dataset

When designing a benchmark for evaluating pt-PT, it is essential to capture the subtle nuances of the language. This creates an opportunity to tailor the benchmark specifically to pt-PT, considering both linguistic and cultural dimensions. ALBA was developed with the goal of covering a wide range of linguistic aspects, organizing questions into different branches of linguistics to grasp the finer points of language from a macroscopic perspective, so as to obtain a baseline assessment of language and linguistic performance in generative LLMs.

In specific, ALBA departs from standard binary/multiple-choice evaluation and emphasizes text generation to engage with language on multiple levels. It evaluates not only understanding and grammar, but also the ability to construct and deconstruct language creatively, as in poetry and word plays. Additionally, it incorporates culturally embedded knowledge, such as proverbs, tongue twisters, and riddles, and accounts for language variations across regions.

The methodology used to create the questions and reference answers, as well as the reasoning behind each linguistic dimension, is detailed below.

3.1 Methodology

Given the breadth of linguistics, priority was given to domains most relevant to LLM evaluation in pt-PT. In particular, ALBA is structured around eight core linguistic dimensions, carefully selected by language experts to capture essential aspects of the language. These dimensions encompass a broad range of linguistic capabilities: Language Variety, Culture-bound Semantics, Discourse Analysis, Word Plays, Syntax, Morphology, Lexicology, and Phonetics and Phonology.

Dimension	Example
Language Variety	*Transforma esta frase que está em português do Brasil e coloca-a em português Europeu. Respeita o registo: se houver termos informais ou gíria na frase, tenta encontrar equivalentes na língua de chegada. "Valeu, cara. A gente leva ele junto."*
Culture-bound Semantics	*Dá-me exemplos de expressões portuguesas que mencionem chuva.*
Discourse Analysis	*Podes criar uma frase com dois pleonasmos?*
Word Plays	*Dá-me um exemplo de um pangrama ou pantograma em português. Inclui todas as letras e acentos do alfabeto português europeu.*
Syntax	*Coloca esta frase na forma passiva: "Ele vendeu a mota."*
Morphology	*Cria uma lista de dez palavras com oito letras que acabem em "ão".*
Lexicology	*Dá-me 5 antónimos para "suave".*
Phonetics & Phonology	*Escreve um poema de três versos com esquema rímico aaa (em que todos os versos rimam) com uma aliteração com "e".*

Table 1: Example questions of all 8 ALBA dimensions.

Expert-based Reference Questions. For each dimension, diverse and curated questions were created, resulting in a total of 800 questions, with 100 per dimension, created in two rounds. In the initial round, 30 questions per dimension were created along with corresponding answers for LLM-as-a-Judge validation. In the second round, 70 further questions were created. All questions were created by two language experts (co-authors) holding Master's degrees in linguistics-related fields and with expertise in European Portuguese. Each question was authored by one expert and reviewed by the other to ensure accuracy and alignment with the intended dimension. Example questions for each dimension are provided in Table 1.

This approach departs from previous works that were either focused on specific linguistic reasoning capabilities (Norman et al., 2025) or took inspiration from purely test-based formats (Kim et al., 2024) by merging the two into a dataset that aims to offer a high quality dataset that breaks complex linguistic reasoning tasks into focused components, allowing for targeted evaluation of model capabili-

Input Question	Correct Output (5)	Moderate Output (4)	Incorrect Output (1)
Qual é o significado da expressão "lágrimas de crocodilo"?	A expressão "lágrimas de crocodilo" representa lágrimas que são fingidas ou exageradas e é usada para descrever um sofrimento que não é genuíno. Exemplo: "Antes de chegar a polícia, aquela mulher esteve aqui meia hora aos berros a dizer que o acidente era culpa nossa. Agora está ali com lágrimas de crocodilo para ver se não perde a carta de condução."	A expressão "lágrimas de crocodilo" representa lágrimas que são fingidas ou exageradas e é usada para descrever um sofrimento que não é genuíno. Exemplo: "Antes da polícia chegar, aquela mulher esteve aqui meia hora gritando, dizendo que o acidente era culpa nossa. Agora tá ali com lágrimas de crocodilo, tentando não perder a carteira de motorista."	A expressão "lágrimas de crocodilo" refere-se a lágrimas provocadas por medo de crocodilos.

Table 2: Input-Output pair example from Culture-bound Semantics with accompanying expert created answers at various correctness levels.

ties. These tasks include adjusting register or tone, proofreading, solving riddles, and composing poetry, all aligned with the eight dimensions and the pt-PT linguistic context. The following section presents each dimension in detail.

Expert-based Reference Answers. To measure LLM responses' quality and to calibrate assessment methods, the same set of language experts created a set of reference responses. For each of the eight linguistic dimensions, the first 30 questions from round one were used, and for each one, the experts produced three distinct responses corresponding to different quality tiers, as illustrated in Table 2. All responses were independently rated on a 1–5 Likert scale, where 1 corresponds to an incorrect, substantially flawed and/or low language quality response, and 5 corresponds to a fully correct, complete and/or high language quality response.

This process resulted in a total of 720 expert-rated responses, which serve as ground truth for evaluating and optimizing LLM judges.

3.2 Linguistic Dimensions

In this section, we present ALBA's linguistic dimensions, along with a detailed explanation of what each dimension is designed to target.

3.2.1 Language Variety

This dimension evaluates LLMs' ability to distinguish between pt-PT and pt-BR. It further extends this distinction to dialectal variation, defined as "a way of talking that belongs to a particular part of a country" (Crystal, 2010, p. 72), by targeting regionalisms. The existence of multiple varieties of Portuguese, coupled with the resource disparity between pt-BR and pt-PT, makes it challenging to obtain an output from an LLM without the influence of pt-BR, even when using a pt-PT input.

Taking this into account, this dimension is composed of questions targeting tasks such as identifying the language variety, adapting one variety into another, recognizing the terms that flag a text or sentence as pertaining to one variety or the other, as well as identifying commonly used differing terms and expressions with the same meaning or use in the distinct varieties. Moreover, we intend to evaluate LLMs' ability to discern the richness of pt-PT through regional variations, i.e., local words and phrases, that are specific to various parts of a country (Crystal, 2010, p. 72). In our case, the focus are terms and expressions used in different parts of mainland Portugal (Center, North, Trás-os-Montes, Alentejo, Algarve) and in the islands (archipelagos of Madeira and Azores).

3.2.2 Culture-bound Semantics

In linguistics, semantics is the study of meaning in language (Crystal, 2010). Instead of a traditional semantical approach, ours is bound to pt-PT culture, aiming to evaluate LLMs' capacity of recognizing cultural-focused aspects, i.e., idiomatic expressions and sayings used frequently in oral communication. In addition, with ALBA, the intention is for LLMs to identify the meaning behind proverbs and idiomatic expressions in other languages and find their equivalent or parallel in pt-PT.

The aim is to evaluate the model's ability to go beyond the literal meaning and identify the underlying message behind popular expressions, idiomatic expressions and proverbs that are intrinsic to Portuguese culture and language, as well as testing its overall knowledge of these expressions.

3.2.3 Discourse Analysis

"Discourse analysis is the study of what we humans do with language and how we do it." (Gee, 2018). This linguistic dimension was implemented due to the need to evaluate LLMs' capability to infer meaning and interpret longer texts. At a practical level, it refers, for example, to discursive communication and interaction, text style and text typology.

The aim is to evaluate the model's ability to ana-

lyze, adapt and extract information from a text in tasks such as proofreading, summarization, data extraction, subject-specific text generation, text completion, and text adaptation, which are then broken down into more targeted tasks, such as the recognition, change, creation or generation of register, text typology, keywords, figures of speech, paraphrase, direct and indirect speech, divergences in content and chronological order.

3.2.4 Word Plays

In this dimension, language is twisted, manipulated, reconstructed and made to fit a mold through the use of word plays (such as palindromes, pangrams, isograms, marsupial words, anagrams, acrostics, calligrams, alphabet soups), with an additional focus on word and letter count, letter recognition, and rule-based sentence generation.

Overall, the aim is to test the limits of language manipulation in tasks that usually prove too challenging for LLMs (Pawar et al., 2025; Shin and Kaneko, 2024), regardless of the language being used, with the ultimate objective of evaluating how well a model can complete these tricky, convoluted tasks, if at all, without compromising language quality and fluency.

3.2.5 Syntax

In linguistics, syntax is the subject that studies and describes the system of rules that are followed when building sentences (Koeneman and Zeijlstra, 2017). From a practical standpoint, it refers to sentence categorization and correlation between words. This ALBA dimension aims to check to which extent models are capable of understanding the language foundations, i.e., order, dependence, and hierarchy of words in sentences.

Our examples convey the following evaluation spectra: subject, verb, types of objects, other sentence constituents, types of clauses, active and passive voices, types of sentences. The overall aim of these targeted tasks is to evaluate the model's ability to complete more overreaching complex syntax-related exercises, such as proofreading, which makes use of textual fine-tuning and overall rephrasing for clarity and textual enhancement, as well as text and sentence analysis.

3.2.6 Morphology

In linguistics, morphology "is the study of word formation, including the ways new words are coined in the languages of the world, and the way forms of words are varied depending on how they're used in sentences." (Lieber, 2009, p. 9) and stands for the description and analysis of words' internal structure. In practice, it refers, for example, to word inflection according to number, gender, and tense.

This dimension aims to check to which extent LLMs are capable of understanding and replicating word formation and inflection in pt-PT, so as to enable higher-complexity related tasks (e.g. changing the register of a text by changing the verb tenses from first person singular to third person singular in pt-PT, and vice-versa). Our examples convey the following evaluation spectra: verb tenses, adjective degrees, morphological constituents (i.e., affixes, interfixes, suffixes), prepositions, thematic vowels.

3.2.7 Lexicology

Lexicology stands for the linguistic dimension that studies "lexis, understood as the stock of words in a given language, i.e., its vocabulary or lexicon" (Jackson and Amvela, 2000, p. 1). Thus, it analyses words, including their origin and morphological, syntactical and semantic features, as well as describing processes at the basis of new words' formation. This dimension aims to check to which extent LLMs are capable of understanding relations between words, particularly their lexical mastery range, hierarchy and lexical network.

In order to better perceive how a model might perform in tasks, such as error correction, proofreading, changes in clarity and conciseness, as well as lexical richness, these tasks are dissected into their most fundamental parts so as to extract the knowledge and skills which they rely on. Therefore, this dimension focuses on the following evaluation spectra: synonymy, antonymy, homophony, homography, hyponymy, hypernymy, lexical field, semantic field, word family, neologism, archaism.

3.2.8 Phonetics and Phonology

Phonetics and Phonology were grouped in one dimension, since, from distinct perspectives, both subjects study language sounds and, as a result, they are complementary. Phonetics is the study of speech sounds (Crystal, 2010), meaning the subject that studies and describes the physical, perceptive, articulatory, and acoustic features of speech sounds. On the other hand, phonology is the study of sound structure in language (Odden, 2005), meaning the subject that studies languages' speech sounds and sound patterns.

This ALBA dimension aims to check to which

extent models are capable of understanding structure and internal constraints of language, while assessing their sensitivity to abstract representation. Our dataset conveys the following evaluation spectra: vowels, consonants, diphthongs, hiatus, syllables classification, alliterations, tongue twisters, rhyme types, rhyming words, scansion, phonetic transcription, sounds repetition, and speech sounds classification.

The aim is to evaluate the model's understanding of sound as it relates to the language being evaluated and its ability to perform creativity and language related tasks, such as the creation of poems and tongue twisters, as well as literature and language analysis related tasks (e.g. scansion), while breaking down the necessary components for completing these kinds of tasks, such as rhyme, accentuation, metrics, and structure.

The idea behind this is that if a model is not able to successfully complete the Phonetics and Phonology related tasks that have been broken down into individual exercises (e.g. rhyme, metrics, structure), it will not be able to do more complex tasks (e.g. create an original sonnet or analyze the rhyme scheme of a poem).

4 ALBA's LLM-as-a-Judge Framework

Given the complexity of evaluating linguistic competence across multiple dimensions and the subjective nature of assessing open-ended responses, we adopted an LLM-as-a-judge approach for evaluating model outputs (Gu et al., 2024). This methodology allows for scalable evaluation while maintaining the nuanced understanding required for linguistic assessment.

Our judge selection methodology consists of three stages:

1. Creation of expert-annotated reference responses (Section 3.1);
2. Systematic exploration of prompt configurations and judge models;
3. Selection of the most reliable judge and configuration for benchmark evaluation.

4.1 LLM Judge Configuration

We systematically explored the LLM-as-a-judge configuration space to identify the setup that best aligns automated evaluations with human judgments. In particular, we examined the effects of prompt language, few-shot example selection, and the choice of judge model.

Prompt Design. The judge prompt assigns the model the role of a professional text evaluator and receives a detailed expert-defined rubric with a 1 (Very Bad) to 5 (Very good) scoring scale, covering precision, linguistic quality, and completeness. When few-shot examples are used, they are appended after the rubric to guide scoring. The judge is instructed to first produce an explicit reasoning trace using a chain-of-thought (Wei et al., 2022) before providing the final numeric score, a process shown to improve evaluation reliability and consistency (Liu et al., 2023; Wang et al., 2025).

Prompt Language. We tested prompts written in pt-PT and English (EN) to evaluate whether native-language prompting improves judgment accuracy for pt-PT content.

Few-Shot Examples. We evaluate the effect of few-shot prompting on evaluation quality by varying both the number of examples (2–5 samples per prompt) and the example selection strategy (Brown et al., 2020). Each example comprises three candidate responses reflecting distinct quality levels: correct, moderate, and incorrect (Section 3.1). For example selection, we consider the following strategies:

- **Random**: samples chosen at random;
- **Similarity**: selecting semantically dissimilar samples to maximize diversity;
- **Length-Diverse**: samples with maximally different response lengths (e.g., shortest, median, longest) are selected to increase diversity in structure.

Candidate Judge Models. In addition to varying the prompt, we evaluated multiple LLMs as candidate judges. The selected models, Gemini-2.5 (Team, 2025a), DeepSeek (DeepSeek-AI, 2025, 2024), and GPT-5 (OpenAI, 2025), were chosen for their strong reasoning capabilities, robustness in zero- and few-shot settings, and multilingual understanding.

4.2 Experimental Protocol

From the expert-rated inputs (Section 3.1), we created two disjoint subsets. In particular, we used two-thirds of the data as a validation set to optimize

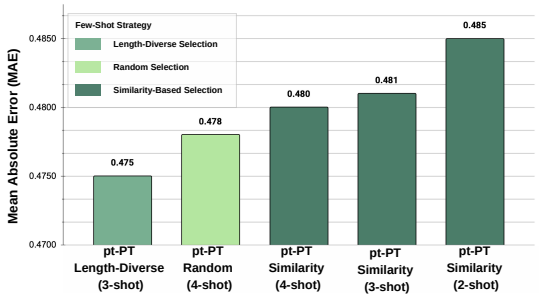

Figure 2: Top five judge configurations ranked by MAE on the validation set.

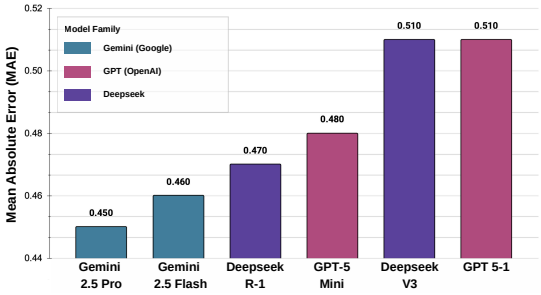

Figure 3: MAE of different LLM judges on the evaluation set using the optimal configuration.

the judge prompt (language, and few-shot configuration), while the remaining served as a held-out evaluation set for selecting the LLM judge model.

Judge performance was measured by the Mean Absolute Error (MAE) between the scores given by the LLM judge and the expert ratings, where a lower MAE indicates closer alignment with human judgment.

4.3 Results

Prompt and Few-Shot Configuration. Figure 2 presents the five best-performing configurations on the validation set. Across all cases, the results consistently favor the pt-PT prompt. In particular, the configuration using three few-shot samples selected with the length-diverse strategy achieved the lowest MAE (0.475), demonstrating a good alignment with expert ratings.

Model Selection. Using the optimal configuration identified in the previous section, Figure 3 reports the results for each candidate model. The results show that Gemini-2.5-Pro achieves the lowest MAE, emerging as the most reliable judge.

Based on these results, the final LLM judge setup employs Gemini-2.5-Pro with a pt-PT prompt and three few-shot samples. Generation is performed using greedy decoding..

5 Experiments

In this section, we analyze the performance of multiple LLM models in terms of language and linguistic generation capabilities according to ALBA's dimensions and the LLM-as-a-judge framework.

5.1 Baseline Language Models

We evaluate multilingual instruction-tuned LLMs with 7B–12B parameters, including fully open-source and open-weight models from major families (e.g., OLMo, Mistral, Gemma, Qwen, and LLaMA). Models were selected based on strong reported Portuguese and multilingual performance, public availability, and architectural diversity, while controlling for model scale to enable fair comparison. We additionally include GPT-5 and Gemini-2.5-Pro[2] as frontier baselines, contextualizing open-weight results against current proprietary state-of-the-art systems.

5.2 Overall Results

Table 3 presents the performance of instruction-tuned models on the ALBA benchmark.

Overall, fully open models achieve lower scores. OLMo-2 performs poorly due to its monolingual (English-only) training. Multilingual models such as Euro-LLM and Apertus-8B perform better, but still lag behind the strongest open-weight models. AMALIA, tailored for European Portuguese (pt-PT), achieves the strongest results among fully open models, even outperforming the larger Gemma 3-12B in culturally bound semantics and lexicology, highlighting the benefits of language-specific specialization. The Gemma models show good results, with Gemma 3-12B achieving the highest score (51.1) among open-weight models, showing a good understanding of pt-PT linguistics. Qwen 3-8B follows closely (49.8), delivering competitive performance despite its smaller size, likely due to its explicit reasoning capabilities. However, all open models exhibit clear limitations in more fine-grained dimensions. In Phonetics and Phonology, errors frequently involve metric inconsistencies, rhyme misclassifications, tonic stress misplacement, and phonetic hallucinations. In Word Plays, models often produce word hallucinations or fail to correctly manipulate characters.

[2]Using the same model as the judge may introduce potential bias (Ye et al., 2025), however, our judge validation showed good alignment with human judgments.

Model	ALBA	Language Variety	Culture-bound Semantics	Discourse Analysis	Word Plays	Syntax	Morphology	Lexicology	Phonetics & Phonology
Fully open models									
OLMo 2-7B (OLMo et al., 2025)	16.9	9.5	9.5	43.3	4.0	24.5	26.0	14.0	4.8
Salamandra-7B (Gonzalez-Agirre et al., 2025)	27.4	27.8	30.5	52.8	4.8	26.5	36.3	29.3	11.5
EuroLLM-9B (Martins et al., 2024)	38.5	41.0	40.0	67.0	11.3	43.3	44.0	47.5	13.8
Apertus-8B (Apertus, 2025)	38.7	37.8	38.3	70.3	13.8	47.5	45.0	45.8	11.5
AMALIA-9B (Simplício et al., 2026)	43.6	48.3	47.8	73.0	14.8	42.3	49.8	53.8	19.0
Open weight models									
Mistral-7B (Jiang et al., 2023)	21.7	15.5	17.5	50.0	3.5	32.5	26.3	26.8	1.8
Ministral-8B (Team, 2024b)	35.6	32.0	37.5	61.0	12.0	45.8	50.8	34.0	11.8
LLaMA 3.1-8B (Team, 2024a)	31.3	27.8	23.5	60.5	19.3	35.0	38.5	29.3	17.0
Gervasio-8B (Santos et al., 2024)	31.1	29.0	22.8	61.8	17.0	38.3	39.5	25.8	15.0
Qwen 2.5-7B (Yang et al., 2024)	31.0	26.3	24.0	63.0	11.0	42.8	38.3	32.3	10.3
Qwen 3-8B (Team, 2025c)	49.8	44.8	35.8	77.5	31.0	70.3	53.8	44.5	41.0
Gemma 2-9B (Rivière et al., 2024)	41.1	40.8	35.5	78.5	22.8	44.8	48.5	41.3	16.8
Gemma 3-12B (Team, 2025b)	51.1	52.8	41.5	85.5	34.3	58.0	56.8	50.8	29.5
Close source models									
GPT-5 (OpenAI, 2025)	**91.0**	**98.3**	**88.3**	**98.5**	**91.5**	78.3	**98.5**	88.8	**86.3**
Gemini 2.5-Pro (Team, 2025a)	90.1	96.5	85.3	98.0	85.3	**85.8**	95.8	**89.5**	84.8

Table 3: Performance of models on ALBA across linguistic dimensions, based on LLM-as-a-judge evaluation with original 1–5 ratings rescaled to a 0–100 range. Bold indicates the best overall model, while underlining denotes the best-performing non-closed-source model.

Even in the dimensions where open models performed better, there were still recurring errors. In Culture-bound Semantics, models often failed to recognize or generate culturally intrinsic elements and humor. In Language Variety, they hallucinated slang or regional terms and confused language varieties. In Discourse Analysis, common issues included misidentification of rhetorical devices and difficulty detecting irony. Terminological inconsistencies and unintended language mixing were also frequent across dimensions. In Figure 4, we show the model and judge outputs for three different examples.

From a scaling perspective, a clear gap remains between open-weight and closed-source models. Closed-source models reach consistently high performance across all dimensions, including the more complex ones, benefiting from larger scale and broader training data.

5.3 Results Analysis

The results obtained on ALBA are consistent with previously established trends in the evaluation of LLMs on linguistic tasks. As observed in other benchmarks (Li and Wang, 2024; Norman et al., 2025; Waldis et al., 2024), LLMs tend to perform well on syntactic, lexical, and discourse-level tasks, while exhibiting substantially lower performance in other linguistic dimensions.

As previously established (Suvarna et al., 2024; Ismayilzada et al., 2025; Shin and Kaneko, 2024), LLMs exhibit limitations in tasks involving phonology, morphology, and wordplay, particularly in tasks involving rhyme, scansion, syllable segmentation, morphological composition, and character-level manipulations such as reordering or counting.

This contrast in performance may be attributed to the way LLMs process natural language, specifically the segmentation of words into tokens. Although tokenization can enhance performance on tasks involving syntax and grammatical structures (Choudhary et al., 2025; Waldis et al., 2024; Warstadt et al., 2023), it can adversely affect others, such as word insertion and retrieval, as well as character counting, insertion, and deletion (Shin and Kaneko, 2024).

The results in ALBA further substantiate previously reported discrepancies in the linguistic performance of LLMs, extending prior findings from linguistic benchmarks to the context of pt-PT.

6 Conclusion

In this work, we introduced ALBA, a linguistically grounded benchmark for pt-PT that encompasses structural, semantic, cultural, and variety-sensitive aspects of the language. Unlike previous benchmarks, ALBA broadens the scope of tasks to provide a more comprehensive evaluation of language generation and linguistic competence. ALBA includes eight linguistic dimensions and 800 expert-crafted questions, supported by a validated LLM-as-Judge framework.

Our results show that current open models perform better on straightforward tasks such as Discourse Analysis and Syntax but struggle with more intricate areas, including Phonetics & Phonology

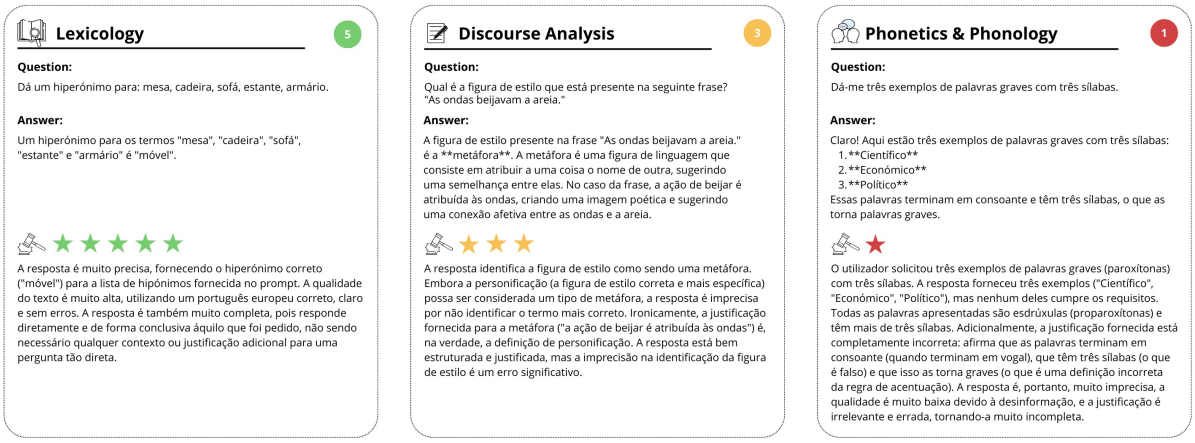

Figure 4: Illustrative answers and judge evaluations from Ministral on various ALBA dimensions.

and Word Plays, underscoring the importance of diverse, linguistically grounded data.

In summary, ALBA provides a language attuned framework to measure proficiency in linguistic-related tasks in pt-PT. Future work should expand these capabilities to additional under-represented languages and linguistic phenomena, enhancing its coverage, relevance, and utility.

Acknowledgments

This work was supported by the AMALIA project under Measure RE-C05-i08 of the Portuguese national Programa de Recuperação e Resiliência. We also acknowledge the support of Fundação para a Ciência e Tecnologia (FCT) and the NOVA LINCS project (UID/04516/2025). Finally, we thank the Barcelona Supercomputing Center (BSC) for providing the computational resources that made this work possible.

References

Thales Sales Almeida, Giovana Kerche Bonás, and João Guilherme Alves Santos. 2025. Broverbs – measuring how much llms understand portuguese proverbs. *Preprint*, arXiv:2509.08960.

Project Apertus. 2025. Apertus: Democratizing open and compliant llms for global language environments. *CoRR*, abs/2509.14233.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, and 26 others. 2020. Language models are few-shot learners. In *NeurIPS 2020, December 6-12, 2020, virtual*.

Florian Cafiero and Marie Puren. 2025. A Riddle in a Haystack: LLM Detection of Intricate Wordplays in Colette and Willy's Novels for Authorship Attribution. In *Digital Humanities 2025*, Lisbonne, Portugal. Universidade Nova de Lisboa and Universidade de Coimbra and Universidade de Aveiro.

Mukund Choudhary, KV Aditya Srivatsa, Gaurja Aeron, Antara Raaghavi Bhattacharya, Dang Khoa Dang Dinh, and 4 others. 2025. Unveiling: What makes linguistics olympiad puzzles tricky for llms? *Preprint*, arXiv:2508.11260.

David Crystal. 2010. *A Little Book of Language*. Yale University Press, Cornwall.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *CoRR*, abs/2501.12948.

James Paul Gee. 2018. *Introducing Discourse Analysis - From Grammar to Society*. Routledge, Oxford.

Anna Gladkova, Aleksandr Drozd, and Satoshi Matsuoka. 2016. Analogy-based detection of morphological and semantic relations with word embeddings: what works and what doesn't. In *Proceedings of the Student Research Workshop, SRW@HLT-NAACL 2016, NAACL: Human Language Technologies, USA*, pages 8–15. The ACL.

José Ángel González, Ian Borrego-Obrador, Álvaro Romo Herrero, Areg Mikael Sarvazyan, Mara Chinea-Rios, and 2 others. 2026. Iberbench: LLM evaluation on iberian languages. *Comput. Speech Lang.*, 96:101899.

Aitor Gonzalez-Agirre, Marc Pàmies, Joan Llop, Irene Baucells, Severino Da Dalt, and 19 others. 2025. Salamandra technical report. *CoRR*, abs/2502.08489.

Satyam Goyal and Soham Dan. 2025. Iolbench: Benchmarking llms on linguistic reasoning. *Preprint*, arXiv:2501.04249.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, and 7 others. 2024. A survey on llm-as-a-judge. *CoRR*, abs/2411.15594.

Mete Ismayilzada, Defne Circi, Jonne Sälevä, Hale Sirin, Abdullatif Köksal, and 4 others. 2025. Evaluating morphological compositional generalization in large language models. *Preprint*, arXiv:2410.12656.

Howard Jackson and Etienne Zé Amvela. 2000. *Words, Meaning and Vocabulary: An Introduction to Modern English Lexicology*. Continuum, London.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, and 13 others. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. Click: A benchmark dataset of cultural and linguistic intelligence in korean. *Preprint*, arXiv:2403.06412.

Olaf Koeneman and Hedde Zeijlstra. 2017. *Introducing Syntax*. Cambridge University Press, Cambridge.

Viet Dac Lai, Chien Van Nguyen, Nghia Trung Ngo, Thuat Nguyen, Franck Dernoncourt, and 2 others. 2023. Okapi: Instruction-tuned large language models in multiple languages with reinforcement learning from human feedback. In *Proceedings of EMNLP 2023 - System Demonstrations, Singapore*, pages 318–327. ACL.

Bingru Li and Han Wang. 2024. Tacomore: Leveraging the potential of llms in corpus-based discourse analysis with prompt engineering. *Preprint*, arXiv:2412.10139.

Rochelle Lieber. 2009. *Introducing Morphology*. Cambridge University Press, Cambridge.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *EMNLP 2023, Singapore*, pages 2511–2522. ACL.

Ricardo Lopes, João Magalhães, and David Semedo. 2024. Glória: A generative and open large language model for portuguese. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese, PROPOR 2024, Santiago de Compostela, Galicia/Spain, March 12-15, 2024, Volume 1*, pages 441–453. ACL.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno Miguel Guerreiro, Ricardo Rei, and 10 others. 2024. Eurollm: Multilingual language models for europe. *CoRR*, abs/2409.16235.

Nathalie Norman, Sanni Nimb, Sussi Olsen, Nina Schneidermann, and Bolette S. Pedersen. 2025. *Electronic lexicography in the 21st century (eLex)*.

David Odden. 2005. *Introducing Phonology*. Cambridge University Press, Cambridge.

Hugo Gonçalo Oliveira, Ricardo Rodrigues, Bruno Ferreira, Purificação Silvano, and Sara Carvalho. 2024. BATS-PT: Assessing Portuguese masked language models in lexico-semantic analogy solving and relation completion. In *PROPOR*, pages 207–217, Santiago de Compostela, Galicia/Spain. ACL.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, and 35 others. 2025. 2 olmo 2 furious. *CoRR*, abs/2501.00656.

OpenAI. 2025. Gpt-5 system card.

Tomás Freitas Osório, Bernardo Leite, Henrique Lopes Cardoso, Luís Gomes, João Rodrigues, and 2 others. 2024. PORTULAN ExtraGLUE datasets and models: Kick-starting a benchmark for the neural processing of Portuguese. In *Workshop on Building and Using Comparable Corpora (BUCC) @ LREC-COLING 2024*, pages 24–34. ELRA and ICCL.

Sachin Pawar, Manoj Apte, Kshitij Jadhav, Girish Keshav Palshikar, and Nitin Ramrakhiyani. 2025. Broken words, broken performance: Effect of tokenization on performance of llms. *Preprint*, arXiv:2512.21933.

Paloma Piot, José Ramom Pichel Campos, and Javier Parapar. 2025. Bridging gaps in hate speech detection: Meta-collections and benchmarks for low-resource iberian languages. *CoRR*, abs/2510.11167.

Livy Real, Erick Fonseca, and Hugo Goncalo Oliveira. 2020. The assin 2 shared task: a quick overview. In *PROPOR*, pages 406–412. Springer.

Parker Riley, Timothy Dozat, Jan A. Botha, Xavier Garcia, Dan Garrette, and 3 others. 2023. FRMT: A benchmark for few-shot region-aware machine translation. *Trans. Assoc. Comput. Linguistics*, 11:671–685.

Morgane Rivière, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, and 94 others. 2024. Gemma 2: Improving open language models at a practical size. *CoRR*, abs/2408.00118.

Rodrigo Santos, João Silva, Luís Gomes, João Rodrigues, and António Branco. 2024. Advancing generative AI for portuguese with open decoder gervásio PT. *CoRR*, abs/2402.18766.

Andrew Shin and Kunitake Kaneko. 2024. Large language models lack understanding of character composition of words. *Preprint*, arXiv:2405.11357.

Afonso Simplício, Gonçalo Vinagre, Miguel Ramos, Diogo Tavares, Rafael Ferreira, and 17 others. 2026. AMALIA: An open source large language model for european portuguese. In *PROPOR*, Salvador, Bahia, Brazil.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. Phonologybench: Evaluating phonological skills of large language models. *Preprint*, arXiv:2404.02456.

Gemini Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Gemma Team. 2025b. Gemma 3 technical report. *CoRR*, abs/2503.19786.

Llama Team. 2024a. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Mistral AI Team. 2024b. Ministral 8b instruct 2410. `https://huggingface.co/mistralai/Ministral-8B-Instruct-2410`. Released October 2024.

Qwen Team. 2025c. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, and 6 others. 2024. Towards multilingual LLM evaluation for european languages. *CoRR*, abs/2410.08928.

Andreas Waldis, Yotam Perlitz, Leshem Choshen, Yufang Hou, and Iryna Gurevych. 2024. Holmes: A benchmark to assess the linguistic competence of language models. *Preprint*, arXiv:2404.18923.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, and 3 others. 2020. Superglue: A stickier benchmark for general-purpose language understanding systems. *Preprint*, arXiv:1905.00537.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Glue: A multi-task benchmark and analysis platform for natural language understanding. *Preprint*, arXiv:1804.07461.

Yicheng Wang, Jiayi Yuan, Yu-Neng Chuang, Zhuoer Wang, Yingchi Liu, and 5 others. 2025. DHP benchmark: Are llms good NLG evaluators? In *Findings of the ACL: NAACL 2025, Albuquerque USA, April 29 - May 4, 2025*, pages 8079–8094. ACL.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, and 2 others. 2023. Blimp: The benchmark of linguistic minimal pairs for english. *Preprint*, arXiv:1912.00582.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, and 4 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS 2022*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, and 36 others. 2024. Qwen2.5 technical report. *CoRR*, abs/2412.15115.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, and 7 others. 2025. Justice or prejudice? quantifying biases in llm-as-a-judge. In *ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Xiang Zhang, Senyu Li, Bradley Hauer, Ning Shi, and Grzegorz Kondrak. 2023. Don't trust chatgpt when your question is not in english: A study of multilingual abilities and types of llms. *Preprint*, arXiv:2305.16339.