AraLingBench: A Human-Annotated Benchmark for Evaluating Arabic Linguistic Capabilities of Large Language Models

Mohamad Zbib[1,2,*], Hasan Abed Al Kader Hammoud[1,*], Sina Mukalled[2], Nadine Rizk[2]
Fatima Karnib[2], Issam Lakkis[2], Ammar Mohanna[2], Bernard Ghanem[1]

[1]King Abdullah University of Science and Technology (KAUST)
[2]American University of Beirut (AUB)

[*]Equal contribution.

Correspondence: mohamad.zbib@kaust.edu.sa, hasanabedalkader.hammoud@kaust.edu.sa

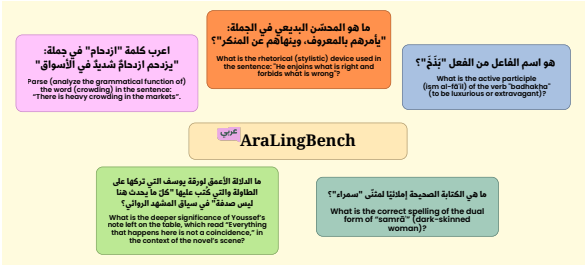

Figure 1: **Sample Questions from AraLingBench.** Example items illustrating the five linguistic categories: grammar, morphology, spelling, reading comprehension, and syntax. Each question targets a distinct aspect of Arabic linguistic competence and is crafted by expert annotators to assess genuine linguistic understanding.

Abstract

We present *AraLingBench*, a fully human annotated benchmark for evaluating the Arabic linguistic competence of large language models (LLMs). The benchmark spans five core categories: grammar, morphology, spelling, reading comprehension, and syntax, through 150 expert-designed multiple choice questions that directly assess structural language understanding. Evaluating 35 Arabic and bilingual LLMs reveals that current models demonstrate strong surface level proficiency but struggle with deeper grammatical and syntactic reasoning. AraLingBench highlights a persistent gap between high scores on knowledge-based benchmarks and true linguistic mastery, showing that many models succeed through memorization or pattern recognition rather than authentic comprehension. By isolating and measuring fundamental linguistic skills, *AraLingBench* provides a diagnostic framework for developing Arabic LLMs. The benchmark and evaluation code are available on Hugging Face and GitHub.

1 Introduction

Natural language processing in Arabic has progressed rapidly through Arabic and bilingual LLMs (Al-Khalifa et al., 2025; Inoue et al., 2021), yet evaluation has lagged. The field still lacks reliable methods to test whether models truly *understand* Arabic at the linguistic level rather than excel only at generation or factual recall.

Existing benchmarks such as BALSAM (Almatham et al., 2025), CamelEval (Qian et al., 2024), and 3LM (Boussaha et al., 2025) emphasize knowledge and problem-solving. EXAMS (Hardalov et al., 2020) , ArabicMMLU, (Koto et al., 2024) and Dialectal ArabicMMLU (Altakrori et al., 2025) rarely probe grammatical agreement, morphological derivation, or orthographic conventions, leaving the linguistic foundations of Arabic understanding largely untested.

Arabic demands mastery of complex morphology, rich inflection and derivation, and flexible syntax. Figure 1 highlights five interconnected skills: grammar (*Nahw*), morphology (*Sarf*), orthography (*Imlaa*), reading comprehension (*Fahm al-logha*), and syntactic structure (*Tarkib Lughawi*). Current evaluations often assume these abilities instead of measuring them directly.

We introduce **AraLingBench**, a human-annotated benchmark built to evaluate Arabic LLMs on core linguistic competence. It contains 150 multiple-choice questions evenly split across the five categories. Questions are authored and reviewed by trained Arabic linguists to ensure linguistic validity, clarity, and reasoning grounded in language rather than factual memory. Figure 1 provides representative items.

We evaluate more than thirty Arabic and bilingual LLMs with AraLingBench. Scores correlate with general benchmarks such as ArabicMMLU, yet many high-performing models rely on surface pattern recognition or retrieval. They excel in spelling and reading comprehension but struggle with grammar, morphology, and syntax, the skills required for authentic proficiency.

AraLingBench separates fluent text production

Proceedings of the 2nd Workshop on NLP for Languages Using Arabic Script (AbjadNLP 2026), pages 385–393
March 28, 2026.

from true linguistic mastery, enabling diagnosis of strengths and weaknesses that knowledge-centric benchmarks miss.

Contributions.

1. Introduce **AraLingBench**, a fully human-annotated benchmark covering grammar, morphology, spelling, reading comprehension, and syntax.

2. Evaluate more than 30 Arabic and bilingual LLMs, revealing persistent deficits in grammatical and morphological reasoning despite strong general scores.

3. Analyze cross-benchmark relationships showing that AraLingBench captures a distinct dimension of ability beyond surface-level or retrieval-based performance.

2 Related Work

The rapid development of Arabic LLMs has transformed NLP for Arabic-speaking communities, creating an urgent need for linguistically grounded evaluation. We summarize (1) the evolution of Arabic language models, (2) the expansion of Arabic evaluation benchmarks, and (3) how AraLingBench fills the persistent gap in linguistic assessment.

2.1 Arabic Language Models

Arabic language models progressed through encoder and decoder generations. Early encoders such as AraBERT (Antoun et al., 2020), MARBERT (Abdul-Mageed et al., 2021), ARBERT (Abdul-Mageed et al., 2021), and CAMeLBERT (Inoue et al., 2021) established discriminative baselines.

Generative architectures shifted focus to text production and instruction following. AraGPT2 enabled Arabic generation, followed by bilingual and multilingual systems such as JAIS (Sengupta et al., 2023) (13B to 30B parameters) and ALLaM (Bari et al., 2025) (7B to 13B). AceGPT (Huang et al., 2024) emphasizes cross-lingual transfer on LLaMA-2, and Hala (Hammoud et al., 2025) uses synthetic bilingual data for fine-tuning.

Recent models broaden coverage: Yehia (Navid-AI, 2025) ranks highly across benchmarks, Fanar (Fanar Team, 2025) leverages extensive compute, and Atlas-Chat (Shang et al., 2025) and AraianGPT(Koubaa et al., 2024) specialize in dialectal variants. Table 1 summarizes architectures, scales, and training strategies from scratch pretraining to multilingual adaptation.

2.2 Evaluation Benchmarks for Arabic

The Arabic evaluation ecosystem now includes more than forty benchmarks (Alzubaidi et al., 2025), most centered on factual reasoning rather than linguistic competence. ArabicMMLU (Koto et al., 2024) and EXAMS (Hardalov et al., 2020) dominate academic and professional topics, while 3LM (Boussaha et al., 2025) spans Arabic, STEM, and coding.

Multi-task platforms such as ORCA (Elmadany et al., 2023), AlGhafa (Almazrouei et al., 2023), and BALSAM (Almatham et al., 2025) enable broader comparisons. Domain-specific suites target legal (Hijazi et al., 2024), medical (Daoud et al., 2025), financial, cultural (Alwajih et al., 2025; Mousi et al., 2025; Sadallah et al., 2025), and dialectal settings. Below an overview of these resources and shows that none explicitly evaluate core linguistic competence, defining AraLingBench's contribution.

2.3 Positioning AraLingBench

AraLingBench directly targets linguistic understanding rather than assuming it. It assesses grammar, morphology, spelling and orthography, reading comprehension, and syntax through carefully constructed questions that isolate specific phenomena. Expert authorship and review provide an interpretable resource that complements existing benchmarks while re-centering linguistic competence in Arabic NLP evaluation.

3 AraLingBench Construction

3.1 Data Collection Process

Five Arabic linguistics experts at the American University of Beirut designed AraLingBench through four stages:

Phase 1: Question Generation. Experts authored original question-answer pairs for the five categories, using references only for inspiration.

Phase 2: Difficulty and Diversity Filtering. Native Arabic speakers reviewed clarity and perceived difficulty; items were kept only if challenging and diverse in phenomena and format.

Phase 3: Expert Quality Control. A senior linguist refined accuracy, phrasing, and category alignment to ensure one unambiguous correct answer.

Table 1: Comparison of major Arabic language models showing architecture types, parameter scales, and training approaches.

Model	Size	Architecture	Training Approach
AraBERT (Antoun et al., 2020)	110M / 335M	Encoder (BERT)	Pretrained from scratch
MARBERT (Abdul-Mageed et al., 2021)	∼163M	Encoder (BERT)	Pretrained from scratch
CAMeLBERT (Inoue et al., 2021)	110M	Encoder (BERT)	Pretrained from scratch
JABER (Ghaddar et al., 2021)	∼125M	Encoder (BERT)	Pretrained from scratch
AraGPT2 (Antoun et al., 2021)	125M to 1.5B	Decoder (GPT-2)	Pretrained from scratch
JAIS (Sengupta et al., 2023)	13B/30B/70B	Decoder-only Transformer	Pretrained from scratch; +Chat SFT
ALLaM (Bari et al., 2025)	7B / 13B / 34B / 70B	Decoder-only Transformer	Continued pretraining + SFT
AceGPT(Huang et al., 2024)	7B / 13B	Decoder	Continued pretraining (Llama-2) + SFT
Hala (Hammoud et al., 2025)	350M/700M/1.2B/9B	Decoder	SFT on synthetic bilingual supervision
Atlas-Chat (Shang et al., 2025)	9B	Decoder	Dialect-focused SFT
Yehia (Navid-AI, 2025)	7B	Decoder	Instruction tuning (SFT/DPO)
Fanar (Fanar Team, 2025)	9B	Decoder	Continued pretraining (Gemma-2-9B) + SFT
SUHAIL (ZeroOne AI, 2025)	14B	Decoder	Inst. tuning / LoRA on multilingual base
ArabianGPT (Koubaa et al., 2024)	1.5B	Decoder	Continued pretraining + SFT
Jais-Adapted (Inception, 2024)	13B	Decoder	Instruction tuning (from Llama-2)

Notes:
· *Pretrained from scratch:* Trained on Arabic data from initialization.
· *Continued pretraining:* Further trained from a multilingual base model.
· *SFT:* Supervised fine-tuning.
· *LoRA:* Low-rank adaptation for efficient fine-tuning.

Benchmark	Year	Primary Focus	Type	Source	Ling.
ArabicMMLU (Koto et al., 2024)	2024	Knowledge (MMLU)	MC	Native	No
EXAMS (Hardalov et al., 2020)	2020	Knowledge (Exams)	MC	Native	No
AlGhafa (Almazrouei et al., 2023)	2023	Multi-task NLP	MC	Mixed	No
ORCA (Elmadany et al., 2023)	2023	Multi-task NLP	Mixed	Mixed	No
BALSAM (Almatham et al., 2025)	2025	Platform	Mixed	Mixed	No
3LM (Boussaha et al., 2025)	2025	STEM + Code	Mixed	Mixed	No
ArabLegalEval (Hijazi et al., 2024)	2024	Legal	Mixed	Mixed	No
MedArabiQ(Daoud et al., 2025)	2025	Medical	Mixed	Mixed	No
CamelEval (Qian et al., 2024)	2024	Instruction	Open	Mixed	No
AraDiCE (Mousi et al., 2025)	2025	Dialectal + Cultural	Mixed	Mixed	No
PalmX (Alwajih et al., 2025)	2025	Cultural	MC	Mixed	No
ACVA (Huang et al., 2024)	2024	Cultural Values	MC	Mixed	No
AraLingBench (Ours)	2025	Linguistic	MC	Native	Yes

Notes:
· *MC:* Multiple-choice.
· *Open:* Open-ended.
· *Mixed:* Various formats.
· *Ling.:* Focuses on core linguistic capabilities.

Phase 4: Difficulty Annotation. Three annotators rated difficulty on {1, 2, 3} (Easy, Medium, Hard) with majority voting.

The final benchmark contains 150 human-authored questions, evenly split across five linguistic categories, explicitly targeting core Arabic skills for LLM evaluation.

3.2 Benchmark Statistics

Figure 2 summarizes balance and difficulty. Each category has 30 questions. Difficulty skews toward Medium to maximize discriminative power: 50 Easy (33.3%), 74 Medium (49.3%), and 26 Hard (17.3%).

Most items use four choices (125, 83.3%), with 25 using three options. Correct answers vary across positions (A: 34.0%, B: 27.3%, C: 26.0%, D: 12.7%) without systematic positional bias.

4 Experimental Evaluation

We evaluate over 30 Arabic and bilingual large language models on AraLingBench to assess their linguistic competence. Our analysis is structured around four key research questions, each probing a distinct dimension of model performance and benchmark validity.

4.1 Evaluation Setup

Model Selection. We evaluated 35 leaderboard models from 350M to 70B parameters, covering Arabic-specific systems (Hala, Fanar, Yehia), bilingual models (JAIS, ALLaM), and multilingual bases adapted for Arabic (Qwen2.5, Phi-4).

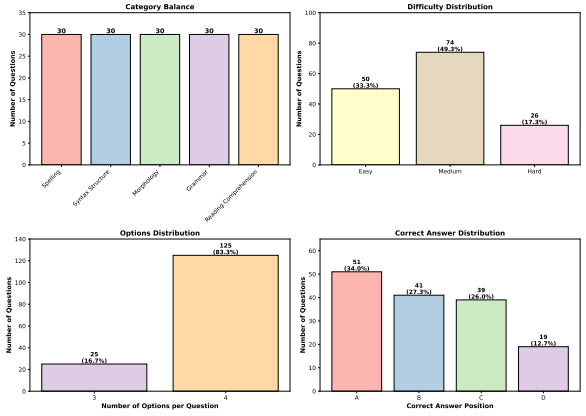

Figure 2: **Overview of AraLingBench.** Category balance, difficulty distribution, formats, and answer position frequencies, illustrating even coverage across skills and difficulty levels.

Evaluation Protocol. Models answered Arabic multiple-choice questions in a uniform zero-shot format with options A to D and a single-letter response. We did not use few-shot examples or chain-of-thought prompting.

Metrics. We report accuracy per category and overall, enabling direct comparison with existing Arabic benchmarks.

4.2 RQ1: Do Models Exhibit Balanced Linguistic Competence?

Motivation. True competence requires balanced mastery of spelling, syntax, morphology, grammar, and reading comprehension. We test whether models remain balanced or specialize.

Results. Table 2 shows three tiers. Top models (Yehia-7B, ALLaM-7B) reach about 72 to 74% average accuracy, mid-tier models (Fanar, Qwen2.5-14B variants) score 55 to 62%, and smaller or less specialized models fall below 50%. Spelling and Reading Comprehension are easiest (median about 58 to 60%), while Syntax is hardest (median about 48%). Even top models show wide gaps: Yehia-7B scores 86.7% in Spelling but 53.3% in Syntax, revealing asymmetric skill development. Morphology also lags despite Arabic's rich structure.

Figure 3 highlights broad interquartile ranges, indicating heterogeneous effects of architecture and training across linguistic dimensions.

Interpretation. Current Arabic LLMs lack balanced competence. Most prioritize surface tasks such as spelling or lexical retrieval over structural understanding of syntax and morphology, likely mirroring training corpora rich in orthography but sparse in explicit grammatical signals.

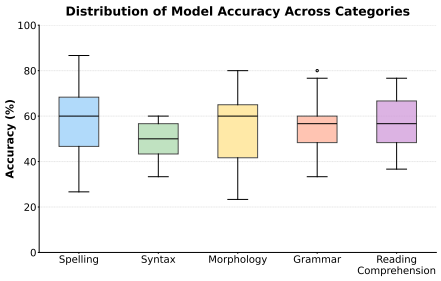

Figure 3: **Category-level accuracy distribution.** Models perform best on Spelling and Reading Comprehension, with Syntax remaining the most difficult category.

4.3 RQ2: How Do Linguistic Skills Correlate?

Motivation. Correlation across linguistic categories reveals whether competence develops holistically or through independent skills.

Results. Figure 4 shows strong links between Grammar and Morphology ($r = 0.83$) and between Spelling and Grammar ($r = 0.86$), reflecting shared reliance on word structure and agreement. Spelling and Reading Comprehension correlate moderately ($r \approx 0.51$). Syntax remains comparatively independent with correlations near 0.13 to 0.43, implying distinct representational needs.

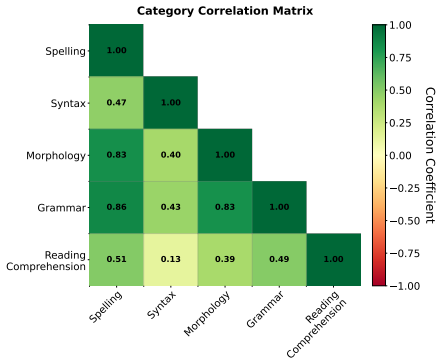

Figure 4: **Inter-category correlations.** Grammar and Morphology show the strongest relationship ($r = 0.83$), while Syntax remains comparatively independent, suggesting distinct representational mechanisms.

Interpretation. Arabic linguistic competence emerges as overlapping skill clusters. Morphology and grammar form a coupled subsystem, while syntax stands apart and may require targeted induc-

Model	Spelling	Syntax	Morphology	Grammar	Reading Comp.	Average
Yehia-7B-preview	86.7	53.3	80.0	80.0	70.0	74.0
ALLaM-7B-Instruct-preview	86.7	60.0	73.3	73.3	76.7	74.0
Yehia-7B-Reasoning-preview	80.0	50.0	80.0	76.7	73.3	72.0
Yehia-7B-DPO-Reasoning-preview	80.0	50.0	80.0	76.7	73.3	72.0
Yehia-7B-SFT-Reasoning-preview	76.7	36.7	66.7	76.7	73.3	66.0
tempmotacilla-cinerea-0308	63.3	60.0	60.0	60.0	70.0	62.7
Qwen2.5-Lumen-14B	70.0	56.7	63.3	60.0	60.0	62.0
Saka-14B	66.7	56.7	63.3	60.0	60.0	61.3
SUHAIL-14B-preview	60.0	60.0	70.0	63.3	53.3	61.3
lambda-qwen2.5-14b-dpo-test	70.0	60.0	60.0	60.0	56.7	61.3
Qwen2.5-14B	60.0	46.7	60.0	70.0	66.7	60.7
Qwen2.5-14B-Gutenberg-1e-Delta	70.0	56.7	60.0	60.0	56.7	60.7
Rombos-LLM-V2.6-Qwen-14b	70.0	60.0	60.0	56.7	56.7	60.7
Fanar-1-9B-Instruct	60.0	43.3	73.3	63.3	60.0	60.0
Qwen2.5-14B-Instruct	66.7	56.7	60.0	56.7	53.3	58.7
Qwen3-8B-Base	60.0	60.0	60.0	60.0	43.3	56.7
Hala-9B	63.3	40.0	63.3	46.7	60.0	54.7
emirati-14b-v2	63.3	46.7	60.0	56.7	46.7	54.7
Qwen2.5-7B-Instruct-abliterated-v2	53.3	53.3	43.3	56.7	60.0	53.3
SILMA-9B-Instruct-v1.0	53.3	53.3	66.7	56.7	36.7	53.3
T.E-8.1	43.3	43.3	60.0	50.0	66.7	52.7
Marco-LLM-AR-V2	53.3	46.7	66.7	56.7	36.7	52.0
recoilme-gemma-2-9B-v0.4	56.7	40.0	63.3	53.3	46.7	52.0
Qwen2.5-7B-Instruct	50.0	50.0	40.0	50.0	66.7	51.3
Qwen2.5-7B-Instruct-Uncensored	46.7	40.0	46.7	50.0	66.7	50.0
Josiefied-Qwen2.5-7B	43.3	50.0	40.0	46.7	60.0	48.0
Qwen2.5-7B-Instruct-abliterated	46.7	40.0	40.0	46.7	66.7	48.0
SauerkrautLM-Nemo-12b-Instruct	46.7	40.0	46.7	43.3	56.7	46.7
Phi-4-mini-instruct	46.7	43.3	53.3	43.3	43.3	46.0
Qwen2-7B-Instruct	50.0	46.7	36.7	43.3	50.0	45.3
Marco-LLM-AR-V4	53.3	33.3	40.0	50.0	36.7	42.7
Qwen2.5-3B-Instruct	36.7	50.0	26.7	50.0	36.7	40.0
Hala-1.2B	26.7	43.3	30.0	36.7	56.7	38.7
Hala-350M	36.7	43.3	30.0	46.7	36.7	38.7
Hala-700M	43.3	36.7	23.3	33.3	53.3	38.0

Table 2: **Model performance across AraLingBench categories.** Top-performing models reach about 72 to 74% accuracy but show large intra-category variance; top models are highlighted in red.

tive biases such as hierarchical modeling beyond general pretraining.

4.4 RQ3: Does General Benchmark Performance Predict Linguistic Competence?

Motivation. Rankings on general benchmarks such as ArabicMMLU or EXAMS may not reflect linguistic understanding. We test that assumption.

Results. Table 3 compares AraLingBench to seven benchmarks. High ArabicMMLU scores do not guarantee strong linguistic competence (e.g., Hala-9B scores 65.6% on ArabicMMLU but 54.7% on AraLingBench). Models tuned heavily on synthetic instruction data perform well on knowledge tasks yet lag on linguistic ones, while instruction tuning with real data (Yehia, ALLaM) aligns general and linguistic competence more closely. Domain specialists can excel in narrow tasks but still falter on basic linguistic skills.

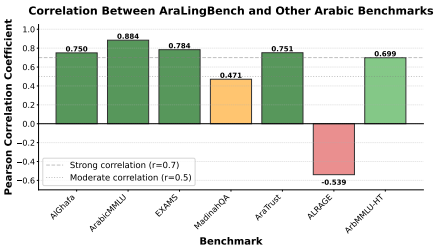

Figure 5: **Cross-benchmark correlations.** Pearson coefficients between AraLingBench and seven major Arabic benchmarks reveal strong alignment with language understanding tasks but weak or negative correlation with retrieval-augmented systems.

Interpretation. Figure 5 shows strong correlations with language understanding benchmarks such as ArabicMMLU ($r = 0.884$), EXAMS ($r = 0.784$), and AraTrust ($r = 0.751$), supporting the link between linguistic competence and general reasoning. Negative correlation with retrieval-augmented evaluation (ALRAGE, $r = -0.539$)

Table 3: **Cross-benchmark comparison.** Performance of major Arabic LLMs across eight benchmarks shows limited predictive power between knowledge-based and linguistic evaluations.

Model	AlGhafa	ArabicMMLU	EXAMS	MadinahQA	AraTrust	ALRAGE	ArbMMLU-HT	AraLingBench
Navid-AI/Yehia-7B-preview	70.8	64.9	52.1	54.4	87.5	76.6	53.4	74.0
ALLaM-7B-Instruct-preview	69.5	64.9	51.6	54.2	86.9	76.8	52.8	74.0
Yehia-7B-Reasoning-preview	75.2	66.3	52.7	55.0	80.8	73.3	55.3	72.0
Hala-9B	78.3	65.6	53.8	70.4	89.6	–	61.4	54.7
Fanar-1-9B-Instruct	76.4	65.8	52.7	73.4	88.3	77.0	58.6	60.0
Qwen2.5-14B-Instruct	72.3	60.0	53.6	35.6	86.1	78.9	55.7	58.7
Qwen2.5-7B-Instruct	65.6	52.3	39.7	62.7	80.7	77.4	40.3	51.3
Hala-1.2B	59.2	48.6	43.4	41.6	71.7	–	44.2	38.7
Hala-700M	55.5	45.9	40.6	34.7	65.2	–	39.4	38.0
Hala-350M	51.4	41.2	36.9	34.5	52.1	–	35.4	38.7

indicates that heavy reliance on retrieval can inflate scores without genuine understanding.

4.5 RQ4: Does Question Difficulty Align with Model Performance?

Motivation. We test whether human difficulty labels align with model performance across categories.

Results. Figure 6 shows non-monotonic trends: median accuracy is 58% on Easy, 50% on Medium, and 54% on Hard items. Some models (e.g., Qwen3-8B-Base) perform better on Hard than Medium questions, indicating mismatched difficulty perception. Effects vary by category, with Syntax showing the clearest downward trend. Leading models such as Yehia and ALLaM degrade only slightly across levels (about 76% to 69%).

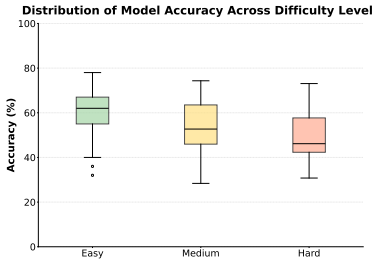

Figure 6: **Performance by difficulty level.** Model accuracy does not decrease monotonically with annotated difficulty; Hard questions occasionally yield higher accuracy than Medium ones.

4.6 Detailed Performance Visualization

Two heatmaps summarize model behavior. Figure 7 shows category accuracy: top models perform well overall with Spelling up to 86.7% but still face a Syntax ceiling near 60.0%. Mid-range and lower models display uneven strengths, often struggling with Morphology and Syntax. Figure 8 summarizes Easy, Medium, and Hard performance, where top models drop only modestly and mid-range models show irregular patterns, reinforcing the mismatch between human labels and model difficulty.

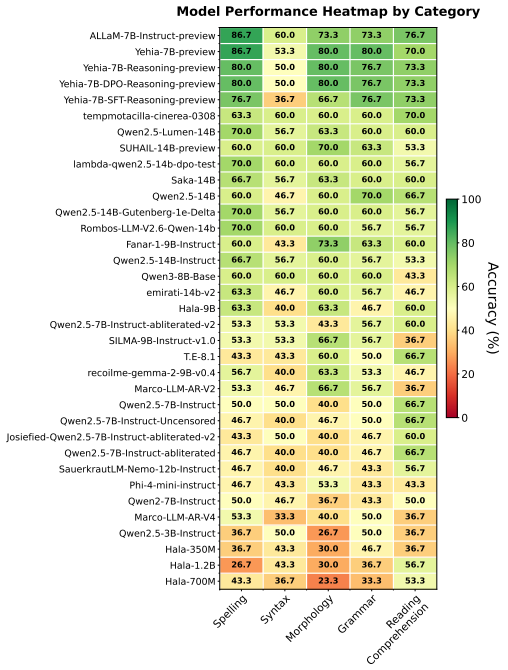

Figure 7: Model performance heatmap across the five AraLingBench linguistic categories. Accuracy values are shown for 35 evaluated models, sorted by weighted average performance. Color intensity ranges from red (low) through yellow (moderate) to green (high).

Interpretation. Non-monotonic scaling shows that model challenge diverges from human-perceived complexity. Hard items may include constructions frequent in pretraining, while Medium items can demand integrative reasoning. Calibrating difficulty benefits from both human annotation and pilot testing on representative models.

4.7 Summary of Experimental Findings

Our evaluation yields four main insights:

1. Arabic LLMs display highly uneven linguistic competence, excelling in surface-level abili-

ties (spelling, comprehension) but struggling with deeper structural understanding (syntax, morphology).

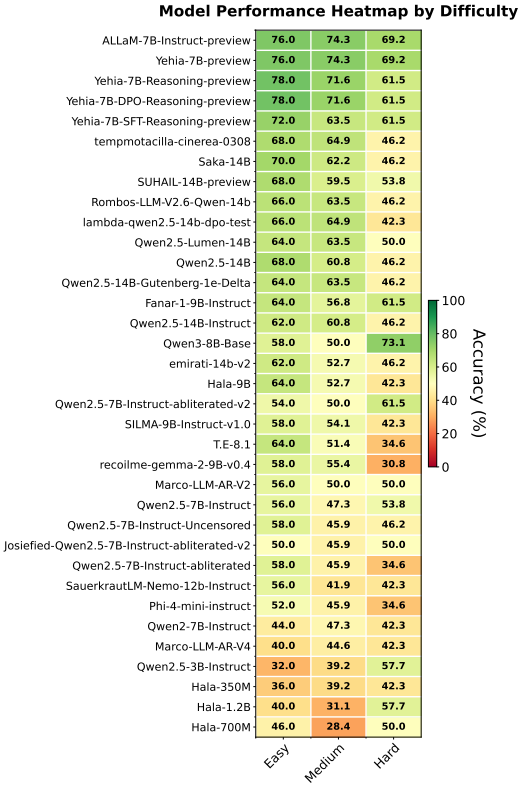

Figure 8: Model performance heatmap across AraLingBench difficulty levels (Easy, Medium, Hard) for 35 evaluated models, sorted by weighted average performance. Color intensity ranges from red (low accuracy) through yellow (moderate) to green (high).

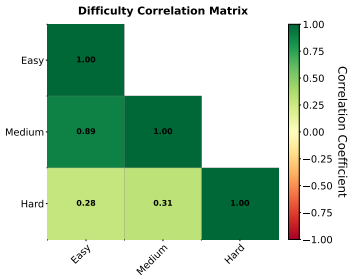

Figure 9: **Difficulty-level correlations.** Strong positive relationships ($r > 0.65$) indicate consistent model ranking despite non-monotonic accuracy patterns.

2. Linguistic skills correlate moderately but not uniformly: grammar and morphology form a tightly coupled subsystem, while syntax remains largely independent.

3. General benchmark success does not guarantee linguistic competence. Although overall correlations are strong ($r > 0.75$), certain training regimes especially retrieval-heavy or synthetic setups inflate benchmark scores without improving true linguistic understanding.

4. Human-assigned difficulty labels only partially align with model performance, highlighting the need to jointly consider cognitive and data-driven measures of challenge.

Taken together, these findings establish AraLingBench as a crucial complement to existing Arabic evaluation suites. It isolates fundamental linguistic understanding, exposing competence gaps that remain invisible in knowledge-oriented benchmarks.

5 Conclusion

We introduced *AraLingBench*, a fully human annotated benchmark targeting grammar, morphology, spelling, reading comprehension, and syntax, isolating linguistic foundations that knowledge-based benchmarks overlook.

Evaluating more than 30 models shows that strong general scores do not guarantee genuine linguistic understanding; many systems still struggle with grammatical and morphological reasoning. AraLingBench offers a diagnostic view that separates superficial fluency from true competence.

We release the benchmark to support Arabic LLMs that generate fluent text and reflect authentic mastery of the language's structure and logic.

Limitations. AraLingBench is intentionally designed as a compact, expert-authored diagnostic benchmark, but this design introduces several limitations. First, the benchmark contains 150 multiple-choice questions (30 per category), which constrains statistical power and the granularity of linguistic phenomenon coverage; small score differences may reflect only a few items and should not be over-interpreted. Second, exclusive reliance on the MCQ format may favor elimination strategies and surface pattern recognition, and it does not fully measure productive linguistic competence in open-ended generation. Third, while we provide human difficulty labels, our results show that difficulty does not always align monotonically with model performance, indicating that perceived difficulty and model difficulty can diverge and motivating future calibration using larger pilots and alternative labeling protocols. Fourth, our category definitions are deliberately high-level to keep the benchmark broadly usable, but boundaries between

grammar, morphology, and syntax can overlap, and deeper alignment with formal linguistic taxonomies remains future work.

References

Muhammad Abdul-Mageed, AbdelRahim Elmadany, and El Moatez Billah Nagoudi. 2021. ARBERT & MARBERT: Deep bidirectional transformers for Arabic. In *ACL*, pages 7088–7105, Online. ACL.

Shahad Al-Khalifa, Nadir Durrani, Hend Al-Khalifa, and Firoj Alam. 2025. The landscape of arabic large language models. *Commun. ACM*, 68(10):54–61.

Rawan Nasser Almatham, Kareem Mohamed Darwish, Raghad Al-Rasheed, Waad Thuwaini Alshammari, Muneera Alhoshan, Amal Almazrua, Asma Al Wazrah, Mais Alheraki, Firoj Alam, Preslav Nakov, and 1 others. 2025. Balsam: A platform for benchmarking arabic large language models. In *ArabicNLP*, pages 258–277.

Ebtesam Almazrouei, Ruxandra Cojocaru, Michele Baldo, Quentin Malartic, Hamza Alobeidli, Daniele Mazzotta, Guilherme Penedo, Giulia Campesan, Mugariya Farooq, Maitha Alhammadi, Julien Launay, and Badreddine Noune. 2023. AlGhafa evaluation benchmark for Arabic language models. In *ArabicNLP*, pages 244–275, Singapore (Hybrid). ACL.

Malik H. Altakrori, Nizar Habash, Abdelhakim Freihat, Younes Samih, Kirill Chirkunov, Muhammed AbuOdeh, Radu Florian, Teresa Lynn, Preslav Nakov, and Alham Fikri Aji. 2025. Dialectalarabicmmlu: Benchmarking dialectal capabilities in arabic and multilingual language models. *arXiv preprint arXiv:2510.27543v1*.

Fakhraddin Alwajih, Abdellah El Mekki, Samar Mohamed Magdy, AbdelRahim A. Elmadany, Omer Nacar, El Moatez Billah Nagoudi, Reem Abdel-Salam, Hanin Atwany, Youssef Nafea, Abdulfattah Mohammed Yahya, Rahaf Alhamouri, Hamzah A. Alsayadi, Hiba Zayed, Sara Shatnawi, Serry Sibaee, Yasir Ech-chammakhy, Walid Al-Dhabyani, Marwa Mohamed Ali, Imen Jarraya, and 25 others. 2025. Palm: A culturally inclusive and linguistically diverse dataset for Arabic LLMs. In *ACL*, pages 32871–32894, Vienna, Austria. ACL.

Ahmed Alzubaidi, Shaikha Alsuwaidi, Basma El Amel Boussaha, Leen AlQadi, Omar Alkaabi, Mohammed Alyafeai, Hamza Alobeidli, and Hakim Hacid. 2025. Evaluating arabic large language models: A survey of benchmarks, methods, and gaps. *arXiv preprint arXiv:2510.13430v2*.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2020. AraBERT: Transformer-based model for Arabic language understanding. In *OSACT*, pages 9–15, Marseille, France. European Language Resource Association.

Wissam Antoun, Fady Baly, and Hazem Hajj. 2021. AraGPT2: Pre-trained transformer for Arabic language generation. In *WANLP*, pages 196–207, Kyiv, Ukraine (Virtual). ACL.

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham Abdullah Alyahya, Sultan AlRashed, Faisal Abdulrahman Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Saad Amin Hassan, Dr. Majed Alrubaian, Ali Alammari, Zaki Alawami, and 7 others. 2025. ALLam: Large language models for arabic and english. In *ICLR*.

Basma El Amel Boussaha, Leen Al Qadi, Mugariya Farooq, Shaikha Alsuwaidi, Giulia Campesan, Ahmed Alzubaidi, Mohammed Alyafeai, and Hakim Hacid. 2025. 3LM: Bridging Arabic, STEM, and code through benchmarking. In *ArabicNLP*, pages 42–63, Suzhou, China. ACL.

Mouath Abu Daoud, Chaimae Abouzahir, Leen Kharouf, Walid Al-Eisawi, Nizar Habash, and Farah E. Shamout. 2025. Medarabiq: Benchmarking large language models on arabic medical tasks. *arXiv preprint arXiv:2505.03427v2*.

AbdelRahim Elmadany, ElMoatez Billah Nagoudi, and Muhammad Abdul-Mageed. 2023. ORCA: A challenging benchmark for Arabic language understanding. In *Findings of ACL*, pages 9559–9586, Toronto, Canada. ACL.

Fanar Team. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944. Authors: Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed Elmagarmid, Mohamed Eltabakh, Masoomali Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, Chaoyi Ruan.

Abbas Ghaddar, Yimeng Wu, Ahmad Rashid, Khalil Bibi, Mehdi Rezagholizadeh, Chao Xing, Yasheng Wang, Duan Xinyu, Zhefeng Wang, Baoxing Huai, Xin Jiang, Qun Liu, and Philippe Langlais. 2021. Jaber: Junior arabic bert. *ArXiv*, abs/2112.04329.

Hasan Abed Al Kader Hammoud, Mohammad Zbeeb, and Bernard Ghanem. 2025. Hala technical report: Building arabic-centric instruction & translation models at scale. *arXiv preprint arXiv:2509.14008v1*.

Momchil Hardalov, Todor Mihaylov, Dimitrina Zlatkova, Yoan Dinkov, Ivan Koychev, and Preslav

Nakov. 2020. EXAMS: A multi-subject high school examinations dataset for cross-lingual and multilingual question answering. In *EMNLP*, pages 5427–5444, Online. ACL.

Faris Hijazi, Somayah Alharbi, Abdulaziz AlHussein, Harethah Shairah, Reem Alzahrani, Hebah Alshamlan, George Turkiyyah, and Omar Knio. 2024. ArabLegalEval: A multitask benchmark for assessing Arabic legal knowledge in large language models. In *ArabicNLP*, pages 225–249, Bangkok, Thailand. ACL.

Huang Huang, Fei Yu, Jianqing Zhu, Xuening Sun, Hao Cheng, Song Dingjie, Zhihong Chen, Mosen Alharthi, Bang An, Juncai He, Ziche Liu, Junying Chen, Jianquan Li, Benyou Wang, Lian Zhang, Ruoyu Sun, Xiang Wan, Haizhou Li, and Jinchao Xu. 2024. AceGPT, localizing large language models in Arabic. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8139–8163, Mexico City, Mexico. Association for Computational Linguistics.

Inception. 2024. Jais family model card.

Go Inoue, Bashar Alhafni, Nurpeiis Baimukan, Houda Bouamor, and Nizar Habash. 2021. The interplay of variant, size, and task type in Arabic pre-trained language models. In *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, pages 92–104, Kyiv, Ukraine (Virtual). Association for Computational Linguistics.

Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. ArabicMMLU: Assessing massive multitask language understanding in Arabic. In *Findings of ACL 2024*, pages 5622–5640, Bangkok, Thailand. ACL.

Anis Koubaa, Adel Ammar, Lahouari Ghouti, Omar Najar, and Serry Sibaee. 2024. Arabiangpt: Native arabic gpt-based large language model. *Preprint*, arXiv:2402.15313.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *COLING*, pages 4186–4218, Abu Dhabi, UAE. ACL.

Navid-AI. 2025. Yehia 7b preview.

Zhaozhi Qian, Faroq Altam, Muhammad Alqurishi, and Riad Souissi. 2024. Cameleval: Advancing culturally aligned arabic language models and benchmarks. *Preprint*, arXiv:2409.12623.

Abdelrahman Sadallah, Junior Cedric Tonga, Khalid Almubarak, Saeed Almheiri, Farah Atif, Chatrine Qwaider, Karima Kadaoui, Sara Shatnawi, Yaser Alesh, and Fajri Koto. 2025. Commonsense reasoning in Arab culture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7695–7710, Vienna, Austria. Association for Computational Linguistics.

Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, William Marshall, Gurpreet Gosal, Cynthia Liu, Zhiming Chen, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Xudong Han, Sondos Mahmoud Bsharat, and 13 others. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149v2*.

Guokan Shang, Hadi Abdine, Yousef Khoubrane, Amr Mohamed, Yassine Abbahaddou, Sofiane Ennadir, Imane Momayiz, Xuguang Ren, Eric Moulines, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. Atlas-chat: Adapting large language models for low-resource Moroccan Arabic dialect. In *LoResLM*, pages 9–30, Abu Dhabi, United Arab Emirates. ACL.

ZeroOne AI. 2025. Suhail-14b-preview.