Do Arabic LLMs Really Reason? Ceiling Effects, Instruction-Following, and Context Injection on Classical Arabic Puzzles

Mayar Osama
German University in Cairo
mayar.osama@guc.edu.eg

Joody Khattab
German University in Cairo
joody.khattab@student.guc.edu.eg

Mervat AbuElkheir
German University in Cairo
mervat.abuelkheir@guc.edu.eg

Abstract

We evaluate five large language models, three general-purpose multilingual models (GPT-4o mini, Qwen-2.5-7B-Instruct, LLaMA-3.1-8B-Instruct) and two Arabic-native/Arabic-first models (Jais-2-8B-Chat, ALLaM-7B-Instruct-preview), on Arabic logical reasoning using five prompting strategies. While all three multilingual models reach near-ceiling accuracy (95.6–100%) on a standard Arabic reasoning benchmark, the same models score only 0–60% on a harder set of twenty classical Arabic literary puzzles requiring domain expertise. We show that (1) standard-benchmark accuracy substantially overstates genuine reasoning capability; (2) failures of Arabic-specialised models under structured (symbolic) prompting, in which the model labels premises and names the inference rule applied, are attributable to instruction-following limitations rather than reasoning deficits; and (3) injecting domain-specific rule files into the prompt (context injection) helps weaker models but degrades stronger models' accuracy via a context poisoning effect, where injected examples replace answers the model would otherwise get right, with the benefit of context injection inversely related to a model's baseline capability.

1 Introduction and Related Work

Large language models (LLMs) have achieved near-saturation performance on general-purpose reasoning benchmarks in English (Cobbe et al., 2021; Hendrycks et al., 2020), prompting the creation of harder benchmarks such as Humanity's Last Exam (Phan et al., 2025) to keep pace with rising capability. Prompting strategies such as chain-of-thought (CoT) (Wei et al., 2022), zero-shot reasoning cues (Kojima et al., 2022), and self-consistency decoding (Wang et al., 2022) have been shown to substantially improve LLM reasoning in English, and Shi et al. (2022) extended CoT to multilingual settings, showing it can transfer across typologically diverse languages, including underrepresented ones. However, whether these gains hold for Arabic, a morphologically rich language with comparatively few structured-reasoning resources, remains underexplored.

Two further prompting-adjacent findings motivate this work. First, self-reflection, in which a model critiques and revises its own answer, does not reliably improve reasoning: Huang et al. (2023) show that LLMs struggle to self-correct without external feedback and can degrade after attempting to do so, while Liu et al. (2024) find that the outcome of self-reflection is highly sensitive to the wording of the reflection prompt. Second, benchmark accuracy is an increasingly unreliable proxy for genuine reasoning ability as models saturate available test sets (Fodor, 2025), which motivates evaluating models on tasks specifically constructed to resist this ceiling effect.

On the Arabic side, dedicated Arabic-centric LLMs have emerged to close the resource gap, including Jais (Sengupta et al., 2023; Anwar et al., 2026). Yet even large-scale Arabic reasoning benchmarks derived from standardised exams, such as GATmath and GATLc (AlBallaa et al., 2025), report a top accuracy of only 66.9% and 64.3% across the models and scales tested, underscoring the difficulty of Arabic reasoning relative to English. A parallel line of work targets Islamic inheritance law (*'ilm al-fara'id*) as a domain requiring precise multi-step legal-arithmetic reasoning: Bouchekif et al. (2025) find that open-source 7–8B models score below 50% on multiple-choice inheritance questions despite frontier models exceeding 90%, and Almasoud et al. (2026) confirm with the MirathQA dataset that sequential rule application over family-relationship structures is a persistent failure point; Aldahoul et al. (2025) similarly find that even majority-voting ensembles of frontier models leave a substantial gap for smaller open-source models. Finally, injecting external

knowledge into the prompt at inference time is not uniformly beneficial: Yoran et al. (2023) show that retrieval-augmented prompting can reduce accuracy when the injected context is irrelevant to or conflicts with a model's existing knowledge, a finding that anticipates the context poisoning effect we observe in Section 3.

In this paper, symbolic prompting refers to prompts that ask the model to extract the key facts of a problem as labelled premises, name the inference rule it applies (e.g., Modus Ponens, Transitivity), and then derive the conclusion step by step. Context injection refers to prepending domain-specific reference text (here, rule files for inheritance law, grammar, and jurisprudence) to the prompt at inference time, without fine-tuning. Context poisoning refers to the case where injected context lowers accuracy because the model reuses the injected material's illustrative examples instead of giving an answer it would otherwise have produced correctly.

Building on this literature, we evaluate five LLMs spanning general-purpose multilingual and Arabic-specialised categories on both a standard Arabic reasoning benchmark and a harder classical Arabic puzzle set, using the five prompting strategies and context-injection design described in Section 2, to test whether high accuracy on the former reflects genuine reasoning capability. Three research questions guide this investigation: **RQ1**: does benchmark accuracy reflect genuine reasoning capability? **RQ2**: are Arabic-model failures under symbolic prompting reasoning failures or instruction-following failures? **RQ3**: does context injection help or harm reasoning, and does this depend on a model's baseline capability?

2 Methodology

This section describes the datasets, models, prompting strategies, evaluation metric, and context-injection design used to test the research questions above. A prompting strategy (Section 2.3) is one of five ways of presenting a task to the model, and a condition (Section 2.5) is one of three prompt settings used only on the puzzle set: C1 (generic prompts), C2 (puzzle-specific prompts without context), and C3 (puzzle-specific prompts with injected rule files). The main dataset uses all five strategies, with 100 questions for each model and strategy. The puzzle set uses four strategies (all except Self-Reflection) under C1 to C3, except that Jais-2-8B could not be run under C3 (Section 3).

2.1 Dataset

The main evaluation uses the Arabic Reasoning Dataset (Logic), containing approximately 1,000 tasks spanning deductive, inductive, and abductive reasoning. After deduplication, 553 unique tasks remained (297 deduction, 256 induction); abduction items were held out and evaluated separately via LLM-as-a-judge. A fixed random seed (`seed=42`) was used for sampling across all model-strategy combinations to ensure comparability, with 100 questions evaluated per combination.

A secondary, harder evaluation set was constructed from twenty classical Arabic literary puzzles extracted from *Al-Ahaji wa-al-Algaz al-Adabiyya* (Riddles and Literary Puzzles) by Abdul-Hay ibn Hasan Kamal.[1] These puzzles are organised into five domain categories: grammar ($n = 3$, abduction), inheritance ($n = 4$, deduction), arithmetic ($n = 8$, deduction), jurisprudence ($n = 4$, deduction and abduction), and astronomy ($n = 1$, abduction). They require specialised domain knowledge (classical Arabic case-marking, Islamic inheritance law, fiqh) absent from the standard benchmark.

2.2 Models to be Compared

Five models were evaluated (Table 1), selected to vary along three dimensions: scale, training paradigm, and language specialisation. Three are general-purpose multilingual models (GPT-4o mini, accessed through the OpenAI API, and the open-weight Qwen-2.5-7B-Instruct and LLaMA-3.1-8B-Instruct); two are Arabic-specialised (Jais-2-8B-Chat, trained from scratch on a 150B+ Arabic-dominant corpus; ALLaM-7B-Instruct-preview, Arabic-first with bilingual instruction tuning). All models used greedy decoding (temperature = 0) with `max_new_tokens=1000`.

2.3 Prompting Strategies

Five prompting strategies were applied across all models, in order of increasing structural complexity (the puzzle set used all but Self-Reflection): (1) **Direct**: zero-shot, no reasoning scaffold; (2) **Chain-of-Thought (CoT)**: one few-shot worked example demonstrating step-by-step reasoning; (3)

[1] Al-Ahaji wa-al-Algaz al-Adabiyya (Riddles and Literary Puzzles), by Abdul-Hay ibn Hasan Kamal, 2nd edition, Nadi al-Ta'if al-Adabi (Al-Ta'if Literary Club), 1401 AH (~1981 CE).

Model	Params	Category
GPT-4o mini	∼8B	Proprietary, multilingual
Qwen-2.5-7B-Instruct	7B	Open, multilingual
LLaMA-3.1-8B-Instruct	8B	Open, multilingual
Jais-2-8B-Chat	8B	Arabic-native
ALLaM-7B-Instruct-preview	7B	Arabic-first

Table 1: Models evaluated in this study.

Model	Main best	Main Symb.	Puzzle best
GPT-4o mini	1.000	1.000	0.60
Qwen-2.5-7B	1.000	1.000	0.55
LLaMA-3.1-8B	1.000	0.956	0.05
Jais-2-8B	1.000	0.840	0.18
ALLaM-7B	0.970	0.440	0.08

Table 2: Accuracy per model (Arabic instructions): highest across prompting strategies (best) and under Symbolic prompting on the main dataset, and highest across the four strategies on the 20-puzzle set (C2).

Symbolic: the model formalises premises as labelled facts and names the applied inference rule (e.g., Modus Ponens, Transitivity) before deriving the conclusion step by step; (4) **CoT+Symbolic (Hybrid)**: combines a worked CoT example with rule naming; (5) **Self-Reflection**: a two-pass strategy in which the model produces a CoT answer (Pass-1), then reviews and revises it (Pass-2).

2.4 Evaluation Metric

Correctness on the main dataset was determined by a hybrid metric applying, in priority order: numeric strict match, exact match after normalisation, gender normalisation match, short-answer containment, token-level F1, and semantic similarity. Abduction items and the classical puzzle set, whose answers are open-ended, were instead evaluated using an LLM-as-a-judge framework (Claude Sonnet 4.5 as judge) and human annotation respectively, classifying responses as correct, partially correct, or incorrect. Puzzle accuracy was computed as $(N_{correct} + 0.5 \times N_{partial})/N_{total}$, with $N_{total} = 20$.

2.5 Context Injection

For the puzzle set, three experimental conditions were run: C1, generic prompts from the main dataset (baseline); C2, puzzle-specific prompts framing the task as a classical Arabic riddle, without context; and C3, puzzle-specific prompts with domain-specific rule files (inheritance, grammar, fiqh) prepended to the prompt (context injection). C2 and C3 are the no-context and with-context conditions compared in Section 3. Rule files were identical across all models (1,032–1,226 tokens). The arithmetic and astronomy categories received no injected context and served as within-experiment controls.

3 Experiments and Results

This section addresses RQ1–RQ3 in turn, using the main dataset and 20-puzzle results described in Section 2.

3.1 Benchmark Accuracy and Reasoning Capability

On the main dataset, GPT-4o mini, Qwen-2.5-7B, and LLaMA-3.1-8B all saturate at 95.6–100% accuracy across most prompting strategies (Table 2), making it impossible to differentiate their reasoning quality on this benchmark. The same three models score dramatically lower on the twenty classical Arabic puzzles under puzzle-specific prompting without context (C2): 60%, 55%, and 5% respectively (best strategy per model). Under the generic prompts of C1, all models scored 0/20 under every strategy. This gap indicates that near-perfect main-dataset accuracy reflects the benchmark's low difficulty ceiling rather than genuine reasoning capability.

3.2 Instruction-Following versus Reasoning Failures

Symbolic prompting, in which the model labels premises, names the inference rule applied, and derives the conclusion step by step, reveals the sharpest divergence between model categories on the main dataset (Table 2). GPT-4o mini and Qwen-2.5-7B maintain 1.000 accuracy; LLaMA-3.1-8B maintains 0.956; but Jais-2-8B drops to 0.840 and ALLaM-7B collapses to 0.440. Qualitative inspection of outputs indicates this drop is not a reasoning failure: Jais-2-8B's outputs contain correct reasoning steps that fail to satisfy the template's formatting constraints, and token-count analysis confirmed prompts remained well within context limits. ALLaM-7B instead responds to the symbolic instruction by generating LaTeX-formatted academic notation and exhaustive enumerations, a training

Model	No-ctx	With-ctx	Δ
GPT-4o mini	60%	55%	−5pp
Qwen-2.5-7B	55%	55%	0pp
LLaMA-3.1-8B	5%	18%	+13pp
Jais-2-8B	18%	N/A	n/a
ALLaM-7B	8%	20%	+12pp

Table 3: Best puzzle-set accuracy (across the four strategies) without (C2) vs. with (C3) context injection.

artefact rather than a reasoning error.

3.3 Context Injection and Baseline Capability

Injecting domain-specific rule files into the puzzle prompts produces opposite effects depending on the model's unaided (no-context, C2) performance (Table 3). The two weakest models on the puzzle set improve substantially: LLaMA-3.1-8B gains +13 percentage points (5%→18%), largely because the rule files signal that Islamic-content questions are legitimate academic material, reducing safety-filter refusals; ALLaM-7B gains +12 points (8%→20%), as the rule text anchors generation and prevents runaway enumeration under Symbolic prompting. Conversely, GPT-4o mini loses 5 points (60%→55%) and Qwen-2.5-7B is unchanged, because injected rule-file examples overwrite correct parametric answers with the file's own illustrative examples, a context poisoning effect. In output inspection, GPT-4o mini answered two jurisprudence puzzles correctly under Symbolic prompting without context, but with the rule file its answers mirrored the file's own examples, and Qwen-2.5-7B showed the same pattern on both puzzles. Because Qwen-2.5-7B's best score is unchanged, the effect is visible at the level of individual puzzles rather than in the best-strategy figure, and for GPT-4o mini under Symbolic prompting alone accuracy falls from 60% to 48%. Jais-2-8B could not be evaluated under context injection due to an architectural constraint: the model generates invalid (BOS) tokens for prompts exceeding approximately 600 tokens, a threshold every rule file exceeds. Across all models and conditions, the inheritance category remained unsolved (0/4 correct in all 144 attempts: four strategies and four puzzles for five models in C2 and four in C3), indicating the persistent gap is a relational-reasoning limitation that rule knowledge alone cannot resolve.

4 Discussion and Conclusion

For RQ1, the near-perfect main-dataset accuracy achieved by GPT-4o mini, Qwen-2.5-7B, and LLaMA-3.1-8B reflects the benchmark's low difficulty ceiling rather than genuine reasoning capability: the same models fall to 0–60% on puzzles requiring classical Arabic domain knowledge. A direct consequence is that prompting-strategy comparisons are uninformative once a model saturates a benchmark, as the five strategies cannot be meaningfully differentiated for GPT-4o mini, Qwen, or LLaMA on the main dataset, and only the harder puzzle set exposes real differences between them.

For RQ2, for models that do not saturate the main dataset (Jais-2-8B, ALLaM-7B), the dominant factor under Symbolic prompting is instruction-following capacity rather than reasoning depth. This has a direct implication for evaluation design: a single accuracy number conflates two distinct failure modes, inability to reason and inability to comply with an output format, and future evaluations of Arabic-specialised models should separate these axes, for example by extracting reasoning content independently of tag- or template-compliance.

For RQ3, context injection functions as a corrective mechanism only for models with a genuine capability deficit. Structural failures such as ALLaM's runaway generation and LLaMA's safety-filter misfires are the kinds of problems context injection can plausibly fix, since injected text disambiguates the domain and anchors output length. In contrast, for models that already carry the relevant knowledge parametrically (GPT-4o mini, Qwen-2.5-7B), injected rule-file examples overwrite that knowledge with the file's own worked cases, a poisoning effect that argues against applying retrieval- or rule-augmentation uniformly without first establishing whether the target model needs the injected knowledge at all.

Across all three findings, the inheritance category is the one persistent failure: no model, prompting strategy, or context condition produces a correct answer on any inheritance puzzle (Section 3). This is consistent with dedicated inheritance benchmarks reporting sub-50% accuracy for open-source 7–8B models (Bouchekif et al., 2025; Almasoud et al., 2026), and indicates the gap is a relational-reasoning limitation (identifying that two described family members are the same person) rather than a knowledge gap that rule injection can close.

Limitations. The puzzle set is small ($n = 20$), so category-level findings such as the 0/4 inheritance result should be read as suggestive rather than statistically robust. The models evaluated span 7–8B parameters, so findings may not generalise to substantially larger Arabic or multilingual models. Correctness on the puzzle set relies on human annotation rather than an automated metric, introducing potential subjectivity into partial-credit scoring.

Conclusion. Near-ceiling accuracy on the standard benchmark (95.6–100% for three multilingual models) substantially overstates genuine reasoning capability, as the same models score only 0–60% on puzzles requiring classical Arabic domain expertise; the sharp accuracy drop of Arabic-specialised models under Symbolic prompting is attributable to instruction-following limitations rather than reasoning deficits; and injecting domain-specific rule files helps models with a genuine capability gap but degrades stronger models via a context poisoning effect. Together, these results argue for evaluating Arabic LLM reasoning on tasks that resist benchmark saturation, and for treating prompting-strategy and context-injection choices as model-specific rather than universally beneficial. Future work should extend this evaluation to larger Arabic-native models and test whether fine-tuning, rather than inference-time context injection, can close the Islamic inheritance reasoning gap identified here.

References

Safa AlBallaa et al. 2025. GATmath and GATLc: Comprehensive benchmarks for evaluating Arabic large language models. *PLOS One*.

Nouar Aldahoul et al. 2025. Benchmarking the legal reasoning of LLMs in Arabic Islamic inheritance cases. *arXiv preprint arXiv:2508.15796*.

A. Almasoud et al. 2026. MirathQA: A dataset for evaluating large language models on Hanbali Islamic inheritance reasoning tasks. *Data in Brief*.

Mohamed Anwar et al. 2026. Jais 2: A family of Arabic-centric open large language models. *arXiv preprint arXiv:2608.13580*.

Abdessalam Bouchekif et al. 2025. Assessing large language models on Islamic legal reasoning: Evidence from inheritance law evaluation. *arXiv preprint arXiv:2509.01081*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

James Fodor. 2025. Line goes up? inherent limitations of benchmarks for evaluating large language models. *arXiv preprint arXiv:2502.14318*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *arXiv preprint arXiv:2205.11916*.

Fengyuan Michael Liu et al. 2024. Self-reflection makes large language models safer, less biased, and ideologically neutral. *arXiv preprint arXiv:2406.10400*.

Long Phan et al. 2025. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*.

Neha Sengupta et al. 2023. Jais and jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models. *arXiv preprint arXiv:2308.16149*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2023. Making retrieval-augmented language models robust to irrelevant context. *arXiv preprint arXiv:2310.01558*.