# QU-NLP at QIAS 2026: Multi-Stage QLoRA Fine-Tuning for Arabic Islamic Inheritance Reasoning

**Mohammad AL-Smadi**
Qatar University
Doha, Qatar
malsmadi@qu.edu.qa

**Abstract**

Islamic inheritance law (علم المواريث, *'ilm al-mawārīth*) presents a challenging domain for evaluating large language models' structured reasoning capabilities, requiring multi-step legal analysis, rule-based blocking decisions, and precise fractional calculations. We present QU-NLP's submission to the QIAS 2026 shared task on Arabic Islamic inheritance reasoning. Our approach employs a multi-stage Quantized Low-Rank Adaptation (QLoRA) fine-tuning strategy on Qwen3-4B: (1) domain adaptation on 3,166 Islamic fatwa records to acquire inheritance terminology and jurisprudential reasoning patterns, followed by (2) task-specific training on 12,000 structured inheritance cases to optimize JSON-formatted output generation. Using 4-bit NF4 quantization with rank-128 LoRA adapters, our model achieves 90% MIR-E (Mawarith Inheritance Reasoning Evaluation) score on the test set, demonstrating competitive performance while requiring minimal computational resources. Our results show that domain-specific pre-adaptation combined with structured output training enables small language models to perform complex legal reasoning tasks effectively, matching commercial systems such as Gemini-2.5-flash.

## 1. Introduction

Large language models (LLMs) have demonstrated remarkable capabilities across diverse NLP tasks (OpenAI et al., 2024), yet their ability to perform structured, rule-based reasoning under legal constraints remains insufficiently evaluated. Islamic inheritance law (علم المواريث, *'ilm al-mawārīth*) offers a demanding testbed: solving a case requires identifying eligible heirs, applying blocking rules, assigning Qur'anic shares, detecting adjustment mechanisms عول (*'awl*, proportional reduction when shares exceed unity) or رد (*radd*, redistribution of surplus), and computing the final distribution—errors at any stage propagate deterministically.

Recent work on LLMs for Islamic knowledge tasks (Bhatia et al., 2026; Malhas et al., 2022; Mubarak et al., 2025) reveals systematic failures on structured reasoning (Bouchekif et al., 2025b). Prior inheritance tasks were evaluated using MCQs dataset (Bouchekif et al., 2025a; Elrefai et al., 2025; AL-Smadi, 2025; Almasoud et al., 2026), which prevent assessment of intermediate reasoning of inheritance. The QIAS 2026 shared task addresses this by requiring complete JSON reasoning traces evaluated by the multi-component MIR-E metric (Bouchekif et al., 2026), enabling fine-grained error analysis impossible under answer-selection formats.

We employ multi-stage QLoRA fine-tuning on Qwen3-4B (Qwen Team, 2025), achieving 90% MIR-E—matching Gemini-2.5-flash while using a 4B parameter model, and substantially outperforming larger open-weight baselines (Bouchekif et al., 2026).

## 2. Task and Dataset

### 2.1. Task Complexity

While Islamic inheritance follows deterministic jurisprudential rules, the QIAS 2026 task evaluates *end-to-end neural reasoning from natural language to structured output*. Neural models must simultaneously handle natural language understanding (parsing diverse Arabic expressions for family relationships), conditional logic through learned patterns (share assignments depending on other heirs), hierarchical blocking through pattern recognition (precedence rules such as sons blocking grandsons), conditional algorithm selection (detecting which of three adjustment types applies based on share totals and heir presence), numerical precision in text generation (fractions and percentages as exact strings), and structured output generation (syntactically valid JSON with consistent Arabic terminology).

The task difficulty arises not from the logical complexity of inheritance rules—which are well-defined—but from learning and applying them through neural pattern matching on natural language while generating structured outputs with exact numerical precision. With class imbalances (*radd* appears in only 2.8% of cases, see Section 2.2), models must generalize to unseen heir configurations without access to gold intermediate representations, making *MAWARITH* a demanding testbed for end-to-end legal reasoning.

### 2.2. MAWARITH Dataset

The *MAWARITH* dataset (Bouchekif et al., 2026) comprises 12,500 Arabic inheritance cases follow-



*Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks*, pages 199–203
May 11, 2026. 

ing the majority opinion (الجمهور), split into 12,000 training and 500 test instances covering 36 distinct heir categories. The majority (92.3%) are simple cases; 4.9% involve عول (*'awl*) and 2.8% involve رد (*radd*). For Stage 1 domain adaptation we additionally use 3,166 Islamic fatwa records from Islamweb[1] introducing jurisprudential terminology (الورثة *al-waratha*, الحجب *al-ḥajb*, العصبة *al-'aṣaba*) and reasoning patterns.

| Split | Simple | 'awl | radd | Total |
|---|---|---|---|---|
| Training | 11,079 | 577 | 344 | 12,000 |
| Test | 456 | 39 | 5 | 500 |
| Total | 11,535 | 616 | 349 | 12,500 |

Table 1: MAWARITH dataset composition.

## 3. Methodology

We use Qwen3-4B (Qwen Team, 2025) fine-tuned in two stages via QLoRA (Dettmers et al., 2023) with 4-bit NF4 quantization, rank $r = 128$, $\alpha = 256$, applied to all projection layers. Stage 1 performs causal language modeling on 3,166 fatwa records (lr $210^{-4}$, 2 epochs) to acquire inheritance terminology. Stage 2 trains on 12,000 structured inheritance cases (lr $310^{-5}$, 6 epochs) with a system prompt enforcing JSON-only output with keys `heirs`, `blocked`, `shares`, `awl_or_radd`, `post_tasil`. Inference uses greedy decoding (temperature 0, max 1024 tokens).

Post-processing applies: (1) typographic normalization (باقى→باقي, tatweel removal); (2) structural deduplication, removing heirs appearing in both `heirs` and `blocked`; (3) label normalization, replacing unrecognised `awl_or_radd` strings with a fraction-sum inference while preserving all valid labels. A fourth *PostTasil* variant additionally recalculates `post_tasil` when identical to unadjusted `shares`.

## 4. MIR-E Evaluation Metric

MIR-E (Bouchekif et al., 2026) decomposes inheritance reasoning into four weighted components: Heirs & Blocking ($S_h$), Share Assignment ($S_s$), Adjustment ($S_a$), and Final Allocation ($S_f$):

$$\text{MIR-E} = \alpha_h S_h + \alpha_s S_s + \alpha_a S_a + \alpha_f S_f \quad (1)$$

where $\alpha_h = \alpha_s = \alpha_f = 0.30$ and $\alpha_a = 0.10$. $S_a$ is evaluated conditionally, receiving a non-zero score only when $S_h = S_s = 1$.

[1] https://www.islamweb.net/

## 5. Results and Discussion

### 5.1. Overall Performance

Table 2 shows component scores on the 500-case test set. Our 4B model achieves 90.0% MIR-E, matching Gemini-2.5-flash (90.1%) while outperforming all open-weight baselines by a wide margin—including models with 8–30 more parameters.

| Model | $S_h$ | $S_s$ | $S_a$ | MIR-E |
|---|---|---|---|---|
| Gemini-2.5-flash | 94.5% | 92.9% | 89.4% | **90.1%** |
| **QU-NLP (Ours)** | **97.1%** | **94.3%** | 84.6% | **90.0%** |
| Qwen3-32B | 69.0% | 44.6% | 26.5% | 43.7% |
| GPT-OSS-120B | 69.3% | 32.7% | 27.1% | 39.1% |
| LLaMA-3.3-70B | 64.8% | 40.3% | 21.5% | 39.0% |
| Fanar-Sadiq | 62.1% | 36.7% | 20.4% | 36.8% |
| Fanar-C-2-27B | 58.4% | 31.4% | 17.8% | 32.8% |

Table 2: Component-wise MIR-E on the 500-case test set (Basic post-processing). Baselines from (Bouchekif et al., 2026).

### 5.2. Component-Wise and Per-Category Analysis

**Heir Identification ($S_h = 97.1\%$):** Highest among all systems, exceeding Gemini by 2.6pp and all open-weight baselines (58.4%–69.3%).

**Share Assignment ($S_s = 94.3\%$):** Exceeds Gemini by 1.4pp; multi-stage training allows Stage 1 to acquire fractional notation before Stage 2 practices assignment.

**Adjustment Detection ($S_a = 84.6\%$):** Trails Gemini by 4.8pp. Both adjustment types are rare (*'awl* 4.9%, *radd* 2.8%), limiting training exposure. Per-category analysis (Table 3) reveals that *'awl* cases (79.2%) underperform *radd* cases (83.0%) despite nearly twice as many training examples (577 vs. 344 training, 39 vs. 5 test)—the bottleneck is arithmetic complexity, not data frequency. Notably, *radd* cases achieve 100% on both $S_h$ and $S_s$, with failures confined to detection and final arithmetic, arguing directly against pattern memorization during model training.

### 5.3. Effect of Post-Processing

Post-processing yields a net gain of $+0.2$pp (89.8%→90.0%), concentrated entirely in $S_h$ ($+0.7$pp) from structural deduplication—removing heirs that appeared in both `heirs` and `blocked`. Share fractions and `awl_or_radd` labels are taken directly from model output without correction. Valid labels ('awl عول, radd رد, none لا) for `awl_or_radd` are never overridden: a model may write a wrong share fraction for a residuary

| Type | N | MIR-E | $S_h$ | $S_s$ | $S_a$ | $S_f$ |
|---|---|---|---|---|---|---|
| Simple | 456 | 90.8% | 96.4% | 94.1% | 86.2% | 83.5% |
| *'awl* | 39 | 79.2% | 96.2% | 95.1% | 66.7% | 50.4% |
| *radd* | 5 | 83.0% | 100% | 100% | 80.0% | 50.0% |

Table 3: Per-category MIR-E. *'Awl* underperforms *radd* despite nearly twice as many training examples.

heir, causing the fraction sum to exceed 1, yet its `awl_or_radd: none`, label is still correct—correcting the label based on the sum of wrong fractions would turn a right answer into a wrong one.

The PostTasil variant produces results identical to Basic across all metrics, confirming that the model's `post_tasil` arithmetic is correct in every case it correctly classifies the adjustment type. The 29 calculation errors (5.8%) are therefore confined to the final percentage generation step, not addressable by any post-hoc symbolic layer (see Section 5.4).

## 5.4. Error Analysis

To understand our model's limitations, we conduct detailed error analysis on all 500 test cases, comparing model predictions against gold standard references across all reasoning components.

### 5.4.1. Error Distribution

Table 4 shows the error distribution. The three discrete failure categories (29+19+11 = 59 cases) are non-semantic: the model understands inheritance law but fails in arithmetic text generation (calculation), label precision (heir ID), or rare-event detection (radd). Residue label avoidance is reported separately as a systematic representational phenomenon rather than a discrete failure.

| Error Type | Cases | % | Impact |
|---|---|---|---|
| Calculation | 29 | 5.8% | $-1.7$pp |
| Residue label avoidance† | 314 | 62.9% | $-0.85$pp |
| Heir Identification | 19 | 3.8% | $-0.4$pp |
| Radd Detection | 11 | 2.2% | $-0.2$pp |

Table 4: Error distribution with impact on MIR-E. Discrete failures total 59 cases (11.8%); impacts are estimated. †Among the 417 cases requiring باقى التركة, the model substitutes an explicit fraction in 314 (75.3%). Because 83.1% of those 314 write the correct fraction, MIR-E's tolerance absorbs most of the penalty: shares score gap = 0.045, giving 314/500 x 0.045 x 0.30 = $-0.85$pp (0.97 with label vs. 0.92 without). Not summed with discrete failures.

### 5.4.2. Error Patterns and Implications

**Arithmetic vs. Semantic Errors:** The most impactful errors (calculation, 29 cases, see Table 4) are *non-semantic*—the model understands inheritance law correctly but fails in final arithmetic. This suggests errors occur in the text-generation step of the final output field rather than in core reasoning, making them addressable through constrained decoding without retraining.

**Complexity Degradation:** Performance varies with case complexity. For simple cases involving 2–4 heirs, the model achieves approximately 91.7% MIR-E. For medium-complexity cases with 5–7 heirs, performance decreases slightly to around 88.9%, while for complex cases involving $\geq$8 mentioned heirs, it further declines to approximately 87.0%. This gradual degradation indicates a moderate impact of complexity on performance, with the model maintaining strong accuracy even in more complex scenarios.

**Rare Event Underfitting:** Per-category analysis (Table 3) shows that *'awl* cases underperform *radd* cases (79.2% vs. 83.0%) despite having nearly twice as many more training examples. The primary bottleneck is arithmetic complexity: *'awl* requires multi-fraction common-denominator computation, while *radd* requires only proportional redistribution once the type is identified. Both benefit from oversampling and explicit rule-based fallbacks as future work.

**Residue Label Recall:** The gold standard requires the residuary label باقى التركة in 417 of 500 evaluated cases (83.4%), reflecting the prevalence of male agnate heirs (عصبة) across the test set. The model provides this label in only 103 of those cases (24.7% recall), substituting an explicit fraction in the remaining 314 cases (75.3% avoidance rate). Table 5 summarises the breakdown.

Despite the low recall, the MIR-E cost is only $-0.85$pp because MIR-E's tolerance absorbs 83.1% of avoidance cases: the model computes the numerically correct residue fraction and writes it as an explicit value—for instance, writing `"7/12"` when باقى التركة is expected, which falls within the evaluation tolerance. This reveals a *representational* rather than *computational* failure. Among the 314 avoidance cases, 261 (83.1%) produce the exact fraction a symbolic calculator would derive by subtracting fixed shares from unity—the model has learned to perform residue arithmetic

| Residue label behaviour | Cases | Rate |
|---|---|---|
| Gold requires residue label | 417 | 83.4% of test |
| Model provides label (recall) | 103 | 24.7% of required |
| Model avoids label | 314 | 75.3% of required |
| of which: correct fraction | 261 | 83.1% of avoided |
| of which: wrong fraction | 53 | 16.9% of avoided |
| Global MIR-E cost | | $-0.85$pp |

Table 5: Residue label recall analysis. Despite a 75.3% avoidance rate, the global MIR-E cost is only $-0.85$pp because 83.1% of avoidance cases compute the numerically correct fraction within the evaluation tolerance.

correctly but defaults to explicit fraction notation due to training bias toward fixed-share cases. The failure is therefore in the final token selection, not in the underlying calculation. Constrained decoding enforcing باقى التركة whenever an عصبة heir is present and fixed shares sum to less than unity would recover the correct label in the majority of affected cases without any change to the model weights.

### 5.5. Pipeline Success Rate

Table 6 reports cumulative success rates across the reasoning pipeline, where each row shows the percentage of cases in which all stages up to and including that point score perfectly.

| Reasoning Stage | Success Rate |
|---|---|
| $S_h = 1$ (Heirs correct) | 84.0% |
| $S_h = 1,\ S_s = 1$ (+ Shares correct) | 81.2% |
| $S_h = 1,\ S_s = 1,\ S_a = 1$ (+ Adjustment) | 79.4% |
| All stages correct | 65.5% |

Table 6: Cumulative pipeline success rates. Each row shows the percentage of cases where all stages up to and including that point score perfectly ($S = 1$).

While 65.5% of cases are solved perfectly across all components, the overall MIR-E reaches 90.0%. This gap is explained by the partial-credit design of MIR-E (Bouchekif et al., 2026), which assigns weighted scores to each intermediate stage ($\alpha_h = \alpha_s = \alpha_f = 0.30,\ \alpha_a = 0.10$). A case with correct heirs, shares, and adjustment but wrong final percentages, for instance, still earns $0.30{+}0.30{+}0.10{+}0.00 = 0.70$ MIR-E. It is worth noting that the adjustment score $S_a$ is evaluated conditionally: it receives a non-zero value only when both $S_h = 1$ and $S_s = 1$, reflecting the sequential dependency of the reasoning pipeline. The 34.5% of imperfect cases therefore contribute meaningful partial credit, and a back-of-envelope check confirms the arithmetic: if the 65.5% perfect cases score 1.0 and the remaining 34.5% score on average 0.71, the weighted average yields $(0.655$ x $1.0) + (0.345$ x $0.71) \approx 0.90$, consistent with our reported MIR-E.

The stage-by-stage breakdown reveals where errors enter the pipeline. The drop from heirs-correct (84.0%) to all-correct (65.5%) accumulates across three transitions: heir identification to share assignment ($-2.8$pp), share assignment to adjustment ($-1.8$pp), and adjustment to final distribution ($-13.9$pp). The largest single drop is the last, indicating that arithmetic computation in `post_tasil` is the dominant bottleneck for cases that pass all upstream reasoning stages—consistent with calculation errors being the most impactful error category (5.8% of cases, $-1.7$pp, Table 4).

## 6. Conclusion

We presented QU-NLP's submission to the QIAS 2026 shared task, achieving 90.0% MIR-E through multi-stage QLoRA fine-tuning of Qwen3-4B—matching Gemini-2.5-flash (90.1%) with a 4B parameter model on consumer hardware. Error analysis identifies failures as predominantly arithmetic text-generation errors rather than reasoning deficiencies: *radd* cases reach 100% heir and share accuracy with failures confined to the final output field; PostTasil equals Basic confirming correct adjustment arithmetic in all classified cases; and 83.1% of residue-avoidance cases compute the correct fraction in explicit form—a representational failure ($-0.85$pp) addressable through constrained decoding without retraining. Domain adaptation on Islamic fatwa records provides jurisprudential grounding absent from general pre-training data.

Future work will focus on: (1) constrained decoding to enforce باقى التركة when عصبة heirs are present and fixed shares sum to less than unity, directly addressing the 75.3% residue avoidance rate; (2) oversampling rare adjustment cases (*radd* 2.8%, *'awl* 4.9%) to reduce underfitting; and (3) extension to other Islamic legal domains and hybrid neural-symbolic architectures for safety-critical deployment.

## References


Mohammad AL-Smadi. 2025. QU-NLP at QIAS 2025 shared task: A two-phase LLM fine-tuning and retrieval-augmented generation approach for islamic inheritance reasoning. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 892–898, Suzhou, China. Association for Computational Linguistics.

Ameera Almasoud, Sharefah Al-Ghamdi, Reem Alqifari, Noof Alfear, and Hend Al-Khalifa. 2026. Mirathqa: A dataset for evaluating large language models on hanbali islamic inheritance reasoning tasks. *Data in Brief*, 65:112589.

Gagan Bhatia, Hamdy Mubarak, Mustafa Jarrar, George Mikros, Fadi Zaraket, Mahmoud Alhirthani, Mutaz Al-Khatib, Logan Cochrane, Kareem Darwish, Rashid Yahiaoui, and Firoj Alam. 2026. From RAG to agentic RAG for faithful islamic question answering.

Abdessalam Bouchekif, Shahd Gaben, Samer Rashwani, Somaya Eltanbouly, Mutaz Al-Khatib, Heba Sbahi, Mohammed Ghaly, and Emad Mohamed. 2026. MAWARITH: A dataset and benchmark for legal inheritance reasoning with llms.

Abdessalam Bouchekif, Samer Rashwani, Emad Soliman Ali Mohamed, Mutaz Alkhatib, Heba Sbahi, Shahd Gaben, Wajdi Zaghouani, Aiman Erbad, and Mohammed Ghaly. 2025a. QIAS 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 851–860, Suzhou, China. Association for Computational Linguistics.

Abdessalam Bouchekif, Samer Rashwani, Heba Sbahi, Shahd Gaben, Mutaz Al Khatib, and Mohammed Ghaly. 2025b. Assessing large language models on islamic legal reasoning: Evidence from inheritance law evaluation. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 246–257, Suzhou, China. Association for Computational Linguistics.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. QLORA: efficient finetuning of quantized llms. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Eman Elrefai, Mohamed Lotfy Elrefai, and Aml Hassan Esmail. 2025. Gumball at QIAS 2025: Arabic LLM automated reasoning in islamic inheritance. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 953–959, Suzhou, China. Association for Computational Linguistics.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2022. Qur'an QA 2022: Overview of the first shared task on question answering over the holy qur'an. In *Proceedinsg of the 5th Workshop on Open-Source Arabic Corpora and Processing Tools with Shared Tasks on Qur'an QA and Fine-Grained Hate Speech Detection*, pages 79–87, Marseille, France. European Language Resources Association.

Hamdy Mubarak, Rana Malhas, Watheq Mansour, Abubakr Mohamed, Mahmoud Fawzi, Majd Hawasly, Tamer Elsayed, Kareem Mohamed Darwish, and Walid Magdy. 2025. Islamiceval 2025: The first shared task of capturing llms hallucination in islamic content. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pages 480–493, Suzhou, China. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, et al. 2024. Gpt-4 technical report.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.