ADAPT–MTU HAI at QIAS 2025: Dual-Expert LLM Fine-Tuning and Constrained Decoding for Arabic Islamic Inheritance Reasoning

Shehenaz Hossain
ADAPT Centre HAI
Computer Science Department
Munster Technological University
shehenaz.hossain@mymtu.ie

Haithem Afli
ADAPT Centre HAI
Computer Science Department
Munster Technological University
haithem.afli@mtu.ie

Abstract

We present ADAPT–MTU HAI's submission to Subtask 1 of the QIAS 2025 shared task, which focuses on Arabic multiple-choice question answering (MCQ) for Islamic inheritance law. This domain presents unique challenges, requiring models to navigate precise fractional computations, exclusion rules, and doctrinal nuances under strict format constraints. Our system employs a dual-expert architecture based on ALLaM-7B, integrating a LoRA-fine-tuned model specialised for inheritance reasoning with its generalist base counterpart. A custom constrained decoding mechanism ensures output compliance, while arbitration between the two models enhances answer stability. Our system achieves 60.0% accuracy on the development set and 54.7% on the official blind test set—substantially improving upon the baseline. We analyse common failure modes and discuss implications for structured legal reasoning using large language models.

1 Introduction

Islamic inheritance reasoning (Mohammedi, 2012), or *ʿilm al-mawārīth*, is a formalised branch of Islamic jurisprudence that governs the distribution of a deceased person's estate among legally entitled heirs (Ajani et al., 2013; Chebet et al., 2014). Its rules involve fixed fractional shares (*farāʾiḍ*)[1], eligibility conditions, and precedence mechanisms such as exclusion (*ḥajb*), redistribution (*radd*), and proportional adjustment (*ʿawl*) (Rahman et al., 2017; Samia and Khaled, 2018; Tabassum et al., 2019). These provisions demand precise arithmetic and deep doctrinal understanding—posing significant challenges for large language models (LLMs), especially in Arabic and under strict formatting constraints.

Subtask 1 of QIAS 2025 (Bouchekif et al., 2025a)[2] evaluates LLMs on Modern Standard Arabic multiple-choice inheritance problems. Each question presents a scenario with six options (A–F), of which only one is correct. The dataset spans *Beginner* cases (e.g., eligibility and basic share allocation) and *Advanced* scenarios (e.g., multi-decedent cases and complex fractional reasoning). Final leaderboard rankings are based on accuracy over a 1,000-item hidden test set.

Our system addresses two core challenges: (i) producing legally and numerically grounded responses within a linguistically and culturally faithful framework, and (ii) enforcing strict single-letter output compliance despite inherent generative variability. We implement a *dual-expert* architecture built on the ALLaM-7B model family (Bari et al., 2024)[3], combining a LoRA-fine-tuned model (Hu et al., 2021) specialised for inheritance reasoning with its original base variant. This is paired with deterministic constrained decoding to ensure output validity without compromising reasoning fidelity. Experiments confirm strong performance on development data, laying a foundation for broader application and extension.

2 Related Work

Research on automating *farāʾiḍ* (Muhammad, 2020) has explored expert systems, rule-based reasoning, and ontologies to encode inheritance rules. Forward-chaining approaches[4] have demonstrated how heir eligibility and share allocation can be derived deterministically from case facts, though such systems scale poorly to complex scenarios. Ontological frameworks like AraFamOnto (Zouaoui and Rezeg, 2021) represent kinship relations and con-

[1] https://ir.uitm.edu.my/id/eprint/44401/

[2] https://sites.google.com/view/qias2025/home?authuser=0

[3] https://huggingface.co/ALLaM-AI/ALLaM-7B-Instruct-preview

[4] https://ir.uitm.edu.my/id/eprint/44401/

Proceedings of The Third Arabic Natural Language Processing Conference, pages 923–928
November 8-9, 2025

straints explicitly, enabling more generalisable inference. These symbolic systems offer transparency and correctness but lack flexibility. Mathematical treatments further highlight the difficulty of exact fractional reasoning—an open challenge for purely neural models (Rahman et al., 2017)—which motivates hybrid approaches that combine symbolic logic with LLM outputs.

Parallel efforts in Arabic question answering (QA) have produced increasingly sophisticated resources. The Qur'an QA shared task (OSACT 2022[5]) advanced reading comprehension over scripture, prompting adaptations of Arabic BERT for retrieval and extraction tasks (Malhas et al., 2022, 2023). Datasets such as HAQA and QUQA support supervised QA for Hadith and Qur'an texts (Alnefaie et al., 2023a), while Hajj-FQA (Aleid and Azmi, 2025) contains over 2,800 QA pairs based on fatwas about the Hajj pilgrimage. Large-scale efforts like Tafsir QA and Hadith QA (Qamar et al., 2024) illustrate the difficulties of long-context reasoning. As noted in surveys (Samia and Khaled, 2018), challenges in Arabic QA persist—including dialect variation, sparse annotations, and domain sensitivity—all of which affect legal-religious domains.

The application of large language models (LLMs) (Team et al., 2025; Sengupta et al., 2023; Huang et al., 2024; Bari et al., 2024) to Islamic inheritance is still emerging. Bouchekif et al. (2025b) and Samia and Khaled (2018) have benchmarked GPT-3.5 and GPT-4 on Sunni inheritance cases involving *ḥajb*, residuary rules, and disqualifications. Their results highlight key limitations: hallucinated logic, vague or ungrounded reasoning, and high sensitivity to prompt phrasing (Mohammed et al., 2025; Alnefaie et al., 2023b). Abbasi (2025) extended this evaluation to Sunni and Shiʿi rules using GPT-4, Gemini, and DeepSeek, finding that domain-aligned prompting and arbitration strategies improve reliability. Broader guidance for building domain-faithful LLMs (Patel et al., 2023) stresses curated data, evaluation rigour, and culturally consistent output constraints. Symbolic approaches, such as the formal rule-based method in (Abdelwahab et al., 2016), remain a useful complement for improving doctrinal accuracy.

Recent work on Arabic cultural and dialectal evaluation (Hossain et al., 2025; de Francony et al., 2019) and benchmarks like CamelEval (Qian et al., 2024) and AraDiCE (Mousi et al., 2024) have highlighted the importance of dialectal robustness, cultural sensitivity, and domain awareness—factors critical to inheritance reasoning. Our work builds on these insights, framing the task as a constrained Arabic MCQ problem and leveraging domain-specific prompting, deterministic decoding, and dual-expert arbitration to ensure both legal validity and output conformity.

3 Dataset

We use the official SubTask 1 dataset, consisting of 20,000 training MCQs, plus 1000 development and 1000 test questions. Each item has six options (A–F) with one correct answer, spanning two difficulty levels (beginner and advanced) and covering diverse inheritance scenarios, including fractional share computation, heir eligibility, and monetary allocation.

4 Methodology

Our SubTask 1 system is designed to maximise accuracy on Arabic multiple-choice inheritance reasoning questions while guaranteeing strict compliance with the required output format. We adopt a *dual-expert inference framework* built on the **ALLAM-7B** family, integrating parameter-efficient fine-tuning, domain-specific prompt design, and deterministic constrained decoding.This section details the architecture, training methodology, and inference workflow.

4.1 System Overview

The core principle of our approach is to leverage the complementary strengths of two model variants: a *domain-specialised fine-tuned model* and its *unmodified base counterpart*. The fine-tuned model (FT-ALLAM-7B) is optimised for Islamic inheritance reasoning, learning task-specific patterns from curated training data. The base model (ALLaM-7B-Instruct-preview) (Bari et al., 2024)[6]preserves the generalisation capacity of the original pre-trained model. By running both in parallel and reconciling their outputs via an arbitration mechanism, we aim to reduce systematic biases while retaining the accuracy benefits of domain adaptation.

[5] https://sites.google.com/view/quran-qa-2022

[6] https://huggingface.co/ALLaM-AI/ALLaM-7B-Instruct-preview

4.2 Prompt Engineering

Each instance is wrapped in a fixed template aligned with our fine-tuning setup. We prepend *four* few-shot exemplars drawn from the official training set, curated to cover eligibility determination (*ḥajb*), fixed-share allocation, residual (*ʿaṣaba*) distribution, and least-common-multiple (LCM) normalization for fractional shares. Exemplars follow a *reason-then-answer* pattern to provide procedural signals while preserving a concise, single-letter target format. The test item is then presented with six options (A–F) and an explicit `Answer:` cue, which constrains the model to a one-letter, format-compliant output.

```
INSTRUCTION_EN:
  "You are an expert in Islamic inheritance law.
   Think step-by-step; output ONE uppercase letter (A--
F)."

SHOTS (k=4; TRAIN-only; fixed order; reason->answer)
  SHOT 1
    Question_AR: {EX1_Q_AR}
    Options_AR:  A){...} B){...} C){...} D){...} E){...} F){...}
    Steps_AR:    {EX1_STEPS_AR}
    Answer:      {EX1_LABEL}    # A--F
  SHOT 2 ... SHOT 4 (same fields)

TARGET
  Question_AR: {Q_AR}
  Options_AR:  A){...} B){...} C){...} D){...} E){...} F){...}
  Steps_AR:    {REASONING_CUE_AR}   # no gold label
  Answer:      # model outputs ONE letter only
```

Figure 1: Few-shot prompt schema used at inference.

4.3 Fine-Tuning Procedure

FT-ALLAM-7B is trained using Low-Rank Adaptation (LoRA) applied to the attention projection layers (`q_proj`, `k_proj`, `v_proj`, `o_proj`) with rank $r = 16$, scaling factor $\alpha = 32$, and dropout of 0.05. The model is fine-tuned on 20k MCQs from the official dataset using a causal language modelling (CLM) objective, concatenating the prompt and the gold answer letter. Training runs for 5 epochs with an effective batch size of 16 (via gradient accumulation), learning rate 3×10^{-5} with cosine decay, weight decay of 0.01, and `bf16` precision. The sequence length is capped at 512 tokens, and gradient checkpointing is enabled to manage memory. The best checkpoint is selected based on accuracy over the 1k-item development set.

4.4 Constrained Decoding

To enforce the requirement of single-letter predictions (A–F), we implement a custom logits processor that masks all vocabulary tokens except the six valid options at each decoding step. We fix `max_new_tokens=1`, `temperature=0.0`, and `do_sample=False` to ensure deterministic, format-compliant outputs. This replicates `logit_bias`-style behaviour in a Hugging Face–compatible framework.

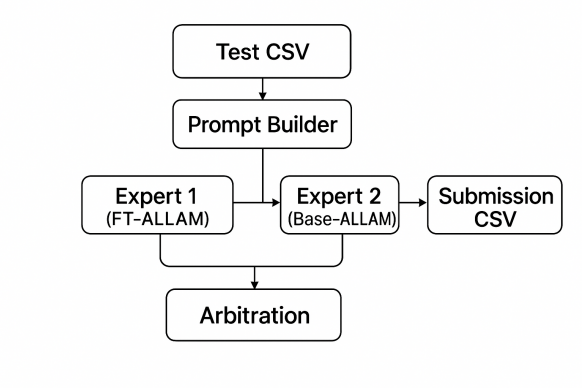

Figure 2: Dual-expert inference pipeline for SubTask 1. Prompts are constructed from the test set, processed independently by FT-ALLAM-7B and Base-ALLAM-7B under constrained decoding, and reconciled via arbitration to produce the final submission.

4.5 Dual-Expert Arbitration

At inference, both FT-ALLAM-7B and Base-ALLAM-7B process each question using identical prompts and decoding constraints. If both models agree, their answer is accepted; in the case of disagreement, the FT-ALLAM-7B prediction is chosen, as it consistently outperforms the base model on the development set. This arbitration strategy preserves the fine-tuned model's domain-specific precision while allowing the base model to act as a corrective filter.

4.6 Pipeline Summary

As shown in Figure 2, the pipeline operates in five steps: (1) parse the test CSV to extract question–option pairs; (2) format each instance using the few-shot template; (3) generate predictions from FT-ALLAM-7B and Base-ALLAM-7B under constrained decoding; (4) apply arbitration to resolve disagreements; (5) export final answers in the submission format. This compact, modular setup ensures reproducibility and allows easy integration of additional experts or symbolic verifiers.

5 Experimental Setup

We retain the original Arabic text with minimal cleaning (e.g., removing option prefixes) and embed each instance in the fixed few-shot template. FT-ALLAM-7B, derived from ALLAM-7B via LoRA

($r = 16$, $\alpha = 32$, dropout 0.05), is fine-tuned on ∼20k MCQs for 5 epochs (batch size 16, LR 3×10^{-5}, cosine schedule, weight decay 0.01, `bf16`, max length 512, gradient checkpointing). Inference is constrained to A–F via a custom logits processor, with tie-breaks favouring the fine-tuned model. All experiments run on NVIDIA A100 (80GB) using Hugging Face Transformers and PEFT.

6 Results and Analysis

6.1 Overall Performance

Our Dual-Expert **ALLaM-7B** system attains **60.0%** accuracy on the official SubTask 1 development set and **54.7%** on the test set. For reference, Bouchekif et al. (2025b) evaluate **ALLaM-7B** as a base, zero-shot model and report an **overall accuracy of 42.9%** (aggregate over Beginner+Advanced); they do not report separate dev/test splits.

System	Dev Acc (%)	Test/Overall (%)
Dual-Expert ALLaM-7B (Ours)	**60.0**	**54.7**
ALLaM-7B(Base, Zero-Shot)	–	42.9

Table 1: QIAS SubTask 1 accuracy. Baseline score (42.9%) is reported by Bouchekif et al. (2025b) as an overall aggregate; dev/test splits are not provided.

6.2 Error Analysis

A qualitative examination of the system's incorrect predictions reveals several recurring error types:

- **Eligibility errors:** In some cases, the model fails to correctly determine heir eligibility, particularly when multiple residuaries are present and certain heirs should be excluded under *ḥajb* rules.
- **Fractional calculation errors:** The model occasionally miscomputes aggregated shares, especially in scenarios involving *awl* adjustments where the expansion of denominators is required.
- **Redistribution errors:** In instances requiring *radd*, residual shares are sometimes redistributed incorrectly, resulting in deviations from proportional allocation.
- **Numerical confusion:** The model is occasionally misled by distractor options that are numerically close to the correct answer, often due to minor inaccuracies in intermediate computations.

Among these, fractional calculation errors and numerical confusion were the most prevalent, accounting for the majority of observed mistakes. These error types were particularly impactful in *Advanced* questions, where multiple layers of arithmetic reasoning and legal constraints interact, amplifying the effect of even minor computational deviations.

Although the absolute accuracy of our system (60.0% dev, 54.7% test) is below the current leaderboard peak, the results validate the robustness of our dual-expert architecture in a challenging reasoning domain. The approach achieves a substantial gain over the random baseline, maintains consistent performance across evaluation splits, and guarantees strict output-format compliance. Moreover, the modular design offers a clear path toward further enhancements, such as the integration of symbolic share calculators or retrieval-augmented prompts, which are expected to address the advanced fractional reasoning errors identified in our analysis.

7 Conclusion and Future Work

We introduced a dual-expert large language model system for structured Islamic inheritance reasoning in Arabic, combining parameter-efficient fine-tuning with deterministic output control. Our architecture, based on ALLaM-7B, achieved competitive performance in QIAS Subtask 1 and demonstrated strong generalisability across question types. The system's strengths include strict output compliance, modular design, and reproducibility, while limitations remain in handling complex fractional arithmetic and legal exclusions. In future work, we plan to incorporate rule-based verifiers, enrich training with curated and synthetic edge cases, and explore retrieval-augmented and multi-agent frameworks to further enhance reasoning accuracy and robustness in domain-specific applications.

Acknowledgments

This research was supported by the ADAPT Research Centre at Munster Technological University. ADAPT is funded by Taighde Éireann – Research Ireland through the Research Centres Programme and co-funded under the European Regional Development Fund (ERDF) via Grant 13/RC/2106_P2.

We also thank the QIAS 2025 shared task organisers for their efforts in preparing the datasets and evaluation platform, and the anonymous reviewers for their valuable feedback and constructive suggestions, which have helped to improve the quality and clarity of this work.

References

Zubair Abbasi. 2025. Augmented learning: Generative artificial intelligence and islamic inheritance law. Islamic Law Blog Roundtable Essay.

Elnaserledinellah Mahmood Abdelwahab, Karim Daghbouche, and Nadra Ahmad Shannan. 2016. The algorithm of islamic jurisprudence (fiqh) with validation of an entscheidungsproblem. *Preprint*, arXiv:1604.00266.

Salako Taofiki Ajani, Bhasah Abu Bakar, and Mikail Ibrahim. 2013. The value of islamic inheritance in consolidation of the family financial stability. *IOSR Journal of Humanities and Social Science*, 8(3).

Hayfa Aleid and Aqil Azmi. 2025. Hajj-fqa: A benchmark arabic dataset for developing question-answering systems on hajj fatwas. *Journal of King Saud University Computer and Information Sciences*, 37:1–28.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023a. HAQA and QUQA: Constructing two Arabic question-answering corpora for the Quran and Hadith. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 90–97, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Sarah Alnefaie, Eric Atwell, and Mohammad Ammar Alsalka. 2023b. Is GPT-4 a good islamic expert for answering Quran questions? In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 124–133, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, and 6 others. 2024. Allam: Large language models for arabic and english. *Preprint*, arXiv:2407.15390.

Abdessalam Bouchekif, Samer Rashwani, Mohammed Ghaly, Mutaz Al-Khatib, Emad Mohamed, Wajdi Zaghouani, Heba Sbahi, Shahd Gaben, and Aiman Erbad. 2025a. Qias 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In *Proceedings of The Second Arabic Natural Language Processing Conference, ArabicNLP 2025, Suzhou, China, November 5–9, 2025*. 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 Second Arabic Natural Language Processing Conference (ArabicNLP 2025)*, Suzhou, China. Association for Computational Linguistics.

Mahmoud Halima Chebet, Joseph Orero, and Anthony Luvanda. 2014. A knowledgebase model for islamic inheritance.

Gael de Francony, Victor Guichard, Praveen Joshi, Haithem Afli, and Abdessalam Bouchekif. 2019. Hierarchical deep learning for Arabic dialect identification. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 249–253, Florence, Italy. Association for Computational Linguistics.

Shehenaz Hossain, Fouad Shammary, Bahaulddin Shammary, and Haithem Afli. 2025. Enhancing dialectal Arabic intent detection through cross-dialect multilingual input augmentation. In *Proceedings of the 4th Workshop on Arabic Corpus Linguistics (WACL-4)*, pages 44–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

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.

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.

Rana Malhas, Watheq Mansour, and Tamer Elsayed. 2023. Qur'an QA 2023 shared task: Overview of passage retrieval and reading comprehension tasks

over the holy qur'an. In *Proceedings of ArabicNLP 2023*, pages 690–701, Singapore (Hybrid). Association for Computational Linguistics.

Marryam Mohammed, Sama Ali, Salma Khaled, Ayad Majeed, and Ensaf Mohamed. 2025. Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and rag. *Neural Computing and Applications*, pages 1–26.

Omar T Mohammedi. 2012. Sharia-complaint wills; principles, recognition, and enforcement. *NYL Sch. L. Rev.*, 57:259.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2024. Aradice: Benchmarks for dialectal and cultural capabilities in llms. *Preprint*, arXiv:2409.11404.

Busari Muhammad. 2020. *The Islamic Law of Inheritance: Introduction and Theories*.

Shabaz Patel, Hassan Kane, and Rayhan Patel. 2023. Building domain-specific llms faithful to the islamic worldview: Mirage or technical possibility? *Preprint*, arXiv:2312.06652.

Faiza Qamar, Seemab Latif, and Rabia Latif. 2024. A benchmark dataset with larger context for non-factoid question answering over islamic text. *Preprint*, arXiv:2409.09844.

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

Siti Fatimah Abdul Rahman, Abdul Malek Yaakob, Ahmad Adnan Fadzil, and MS Shaban. 2017. Asset distribution among the qualified heirs based on islamic inheritance law. *Contemporary Issues on Zakat Waqf and Islamic Philanthropy*, pages 465–474.

Zouaoui Samia and Rezeg Khaled. 2018. Islamic inheritance calculation system based on arabic ontology (arafamonto). *Journal of King Saud University - Computer and Information Sciences*, 33.

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. *Preprint*, arXiv:2308.16149.

Sadia Tabassum, A. Hoque, Sharaban Twahura, and Mohammad Rahman. 2019. Developing an islamic farayez system applying software engineering. *Jurnal Kejuruteraan*, 31:25–38.

Fanar Team, 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, and 23 others. 2025. Fanar: An arabic-centric multimodal generative ai platform. *Preprint*, arXiv:2501.13944.

Samia Zouaoui and Khaled Rezeg. 2021. Islamic inheritance calculation system based on arabic ontology (arafamonto). *Journal of King Saud University - Computer and Information Sciences*, 33(1):68–76.