



# SAHM: A Benchmark for Arabic Financial and Shari’ah-Compliant Reasoning

Rania Elbadry<sup>1</sup> Sarfraz Ahmad<sup>1</sup> Ahmed Heakl<sup>1</sup> Dani Bouch<sup>1</sup> Momina Ahsan<sup>1</sup>

Muhra AlMahri<sup>1</sup> Marwa Elsaid Khalil<sup>1</sup> Yuxia Wang<sup>2</sup>

Salem Lahlou<sup>1</sup> Sophia Ananiadou<sup>3</sup> Veselin Stoyanov<sup>1</sup> Jimin Huang<sup>4</sup>

Xueqing Peng<sup>4\*</sup> Preslav Nakov<sup>1</sup> Zhuohan Xie<sup>1</sup>

<sup>1</sup>MBZUAI <sup>2</sup>INSAIT, Sofia University ”St. Kliment Ohridski”

<sup>3</sup>The University of Manchester <sup>4</sup>The Fin AI

{rania.elbadry, zhuohan.xie}@mbzuai.ac.ae xueqing.peng2024@gmail.com

Project SAHM Code Leaderboard

## Abstract

English financial NLP has advanced rapidly through benchmarks targeting earnings analysis, market sentiment, tabular reasoning, and financial question answering, yet Arabic financial NLP remains virtually nonexistent, despite 422 million speakers, \$4.9 trillion in Gulf sovereign wealth, and a \$4–5 trillion Islamic finance industry requiring specialized Shari’ah compliance over instruments like sukuk, murabaha, and takaful. We introduce SAHM, the first Arabic financial benchmark spanning seven tasks: AAOIFI standards QA, fatwa-based QA/MCQ, accounting and business exams, financial sentiment analysis, extractive summarization, and event-cause reasoning, comprising 14,380 expert-verified instances from authentic regulatory, juristic, and corporate sources. Evaluating 20 LLMs, we find Arabic fluency does not imply financial reasoning: models achieving 91% on recognition tasks drop sharply on generation, and event-cause reasoning exposes the widest performance gap (1.89–9.84/10). We release the benchmark and dataset to support trustworthy Arabic financial assistants.

Arabic benchmarks remain limited in scope: Ar-Banking77 (Jarrar et al., 2023) addresses only banking intent, and Arabic-centric LLMs (Sengupta et al., 2023; Team, 2025; Heakl et al., 2025a; Abbas et al., 2025) have not been evaluated on financial domains. Islamic finance further illustrates this gap. Unlike conventional finance, it requires Shari’ah review guided by standards issued by AAOIFI.<sup>1</sup> Although resources such as Fatwaset (Alyemny et al., 2023) and Hajj FQA (Aleid and Azmi, 2025) exist, they focus on general juristic QA rather than financial reasoning. As a result, LLMs remain untested on tasks that combine legal and financial analysis.

We introduce SAHM, the first Arabic financial NLP benchmark unifying modern finance and Islamic jurisprudence, two high-stakes domains shaping trillions in assets yet missing from LLM evaluation. This enables joint evaluation of compliance and financial reasoning. SAHM spans seven expert-verified tasks grounded in AAOIFI standards, fatwa archives from seven countries, and corporate disclosures (Figure 1). Evaluating 20 LLMs reveals that Arabic fluency does not guarantee financial reasoning: base Arabic models rank in the bottom 25% despite being designed for Arabic. However, fine-tuning on SAHM closes this gap: domain-adapted models gain up to +26 points on Accounting and +25 points on Business, enabling 7–8B models to surpass GPT-5 and match 72B open-source baselines. Our contributions:

- The first Arabic finance benchmark (14,380 instances; 7 tasks) jointly evaluating Shari’ah-compliant reasoning (fatwa QA, Islamic finance standards) and core financial competencies (accounting MCQ, sentiment, event-cause QA), addressing a major resource gap for Arabic financial NLP.

## 1 Introduction

The Gulf Cooperation Council (GCC) generates large volumes of Arabic financial text, including central bank reports, regulatory filings, corporate disclosures, and *fatwas* that provide jurisprudential rulings. Despite this, evaluation of Large Language Models (LLMs) on Arabic financial content remains limited. English financial NLP has advanced rapidly through dedicated benchmarks (Maia et al., 2018a; Zhu et al., 2021; Chen et al., 2021, 2022; Zhao et al., 2024; Xie et al., 2025), with multilingual extensions for other languages (Nie et al., 2025; Zhang et al., 2024; Peng et al., 2025a,b).

\*Corresponding author

<sup>1</sup><https://aaoifi.com>

<p><b>Islamic Financial Fatwa QA</b></p> <p>أرجو بيان الحكم الشرعي في أحد عقد التأمين لغيات التعليم؟</p> <p>بعد دراسة هذا النظام، تبين أنه عبارة عن عقد تأمين تجاري تقوم العلاقة فيه على المعاشرة بين المؤمن وشركة التأمين مباشرة والتأمين التجاري محرم في قول جمهور الفقهاء والمذاهب الفقهية؛ لأنه من العقود المبنية على المقامرة والميسر. قل الله تعالى: (يا أيها الذين آمنوا! إنما الخمر والميسر والألزام ومن عمل مثل هذا فليتبوء ثوابه مما كسب نفسه) [المائدة: 90] كما أنه يتضمن بنوداً خاصة بالفرض الربوي، حيث نص النظام على أنه: "اتفاق الشركة على أن تلعب قرنتنا على الوثيقة... ويحق للشركة إعادة تحديد معدل الفائدة على القروض القائمة والجديد". وعلى هذا، يحرم الائتراك في مثل هذا التأمين حتى يتم تعديل نظامه بنظام التأمين التعاوني الإسلامي.</p>	<p><b>Islamic Finance Shari'ah QA</b></p> <p>في العمليات المالية الإسلامية، هل يجوز موضوع التأمين الإسلامي: هل يجوز اقتطاع احتياطيات من أموال حملة الوثائق، وما هو مسيرها عند تصفية الشركة؟</p> <p>يجوز تحفيقاً لمصلحة حملة الوثائق أن يتم حجز أو إرجاع احتياطيات أو مخصصات متعلقة بصندوق التأمين على ألا تتولى في المساهمين، وما يتراكم في حساب التأمين بصرف في وجه الخير عند التصفية.</p>	<p><b>Business Exams MCQ</b></p> <p>السؤال: تصميم عروض المنتجات والإعلان والترويج والجمهور الترويجية والخدمة بشكل يتناسب مع السوق يطلق عليه (a) استراتيجية التسويق (b) إدارة السوق (c) إدارة المزيج التسويقي</p> <p>الإجابة: a</p>	<p><b>Accounting Exams MCQ</b></p> <p>السؤال: في ٢٠١٨/١/١، بدأت إحدى الشركات بطرح أربعة أسهم إداري عليها وبدأت الشركة في بناء البنية التحتية في ١/١ من السنة المزمع التخليق خلالها للتمويل بقيمة ٩٥٠٠٠٠ جنيه. واكتملت البنية التحتية باستخدام ١٢/٢١. والتحويل لبناء المباني اقتضت مبلغ ٧٠٠٠٠٠ جنيه في ٢٠١٨/١/١ على وثيقة لمدة ٣ سنوات بمعدل ٧٪ سنوياً تبدأ في ١/٢٢١ من كل عام. وكان على الشركة أن لا أكثر من مبلغ ١٠٠٠٠٠ جنيه لمدة ٣ سنوات بمعدل ٧٪ سنوياً محروقة في ٢٠١٧/١/١ وتبدأ الفاتدة في ١/٢١ من كل عام المطلوب. تبلغ الفوائد المزمعة في ٢٠١٨/١/٣١ ..... الخيارات</p> <p>(a) ١٨٠٠٠٠٠ (b) ١٨٠٠٠٠٠ (c) ١٨٠٠٠٠٠ (d) لا شيء مما سبق</p> <p>الإجابة: a</p>
<p><b>Financial Report Sentiment Analysis</b></p> <p>اقرأ بعناية التقرير المالي التالي واختر التصنيف الصحيح: الجوارات: a) Positive b) Negative c) Neutral</p> <p>التقرير: (BKNZ) القوي: شل - مانتار بوزوغ بنك زوي مؤسسة مشتركة متجربة متجولة على مساهميه، في نهاية عام ٢٠٢٤، وازدهار البنك في السوق بواقع 4.47 مليار دولار، مساهمة 100 سهم وممولة من الأرباح المتجزئة البالغة ١٠٠ مليون دولار أمريكي، وهو ما يعكس نموًا قويًا في الأداء. وفي الربع الثاني من عام 2024، سجلت الشركة نموًا في الأرباح بعد خصم التكاليف البالغة 200 مليون دولار أمريكي، حيث تم تحقيق صافي الدخل الإجمالي لـ 200 مليون دولار أمريكي. وتحتفظ الشركة بالاحتياطيات البالغة 200 مليون دولار أمريكي. وشهدت الشركة نموًا في الأرباح على مدار 9.625 مليار دولار أمريكي في عام 2024، ما يعكس التحسين المالي الاستراتيجي الذي تشهده الشركة من خلال تنفيذها لبرنامجها الاستثماري. وشهدت الشركة نموًا في الأرباح على مدار 9.625 مليار دولار أمريكي في عام 2024، ما يعكس التحسين المالي الاستراتيجي الذي تشهده الشركة من خلال تنفيذها لبرنامجها الاستثماري. وشهدت الشركة نموًا في الأرباح على مدار 9.625 مليار دولار أمريكي في عام 2024، ما يعكس التحسين المالي الاستراتيجي الذي تشهده الشركة من خلال تنفيذها لبرنامجها الاستثماري.</p> <p>الإجابة: c</p>	<p><b>Report Extractive Summarization</b></p> <p>قم بتلخيص التقرير المالي التالي باستخدام التلخيص الاستخراجي</p> <p>التقرير: (BKNZ) القوي: شل - مانتار بوزوغ بنك زوي مؤسسة مشتركة متجربة متجولة على مساهميه، في نهاية عام 2024، وازدهار البنك في السوق بواقع 4.47 مليار دولار، مساهمة 100 سهم وممولة من الأرباح المتجزئة البالغة ١٠٠ مليون دولار أمريكي، وهو ما يعكس نموًا قويًا في الأداء. وفي الربع الثاني من عام 2024، سجلت الشركة نموًا في الأرباح بعد خصم التكاليف البالغة 200 مليون دولار أمريكي، حيث تم تحقيق صافي الدخل الإجمالي لـ 200 مليون دولار أمريكي. وتحتفظ الشركة بالاحتياطيات البالغة 200 مليون دولار أمريكي. وشهدت الشركة نموًا في الأرباح على مدار 9.625 مليار دولار أمريكي في عام 2024، ما يعكس التحسين المالي الاستراتيجي الذي تشهده الشركة من خلال تنفيذها لبرنامجها الاستثماري. وشهدت الشركة نموًا في الأرباح على مدار 9.625 مليار دولار أمريكي في عام 2024، ما يعكس التحسين المالي الاستراتيجي الذي تشهده الشركة من خلال تنفيذها لبرنامجها الاستثماري.</p>	<p><b>Event Causal Reasoning QA</b></p> <p>ما الذي تعكسه نسبة التغطية التي تجاوزت 6.5 أضعاف إصدارات صكوك صندوق الاستثمارات العامة بقيمة 1.26 مليار دولار أمريكي، وحجم طلبات الاكتتاب التي بلغت أكثر من 9 مليارات دولار أمريكي، في ظل تصنيفات الصندوق الائتمانية القوية واستراتيجيته المتطورة المتزايدة، ودوره كمحرك للتحول الاقتصادي؟</p> <p>التقرير: (BKNZ) القوي: شل - مانتار بوزوغ بنك زوي مؤسسة مشتركة متجربة متجولة على مساهميه، في نهاية عام 2024، وازدهار البنك في السوق بواقع 4.47 مليار دولار، مساهمة 100 سهم وممولة من الأرباح المتجزئة البالغة ١٠٠ مليون دولار أمريكي، وهو ما يعكس نموًا قويًا في الأداء. وفي الربع الثاني من عام 2024، سجلت الشركة نموًا في الأرباح بعد خصم التكاليف البالغة 200 مليون دولار أمريكي، حيث تم تحقيق صافي الدخل الإجمالي لـ 200 مليون دولار أمريكي. وتحتفظ الشركة بالاحتياطيات البالغة 200 مليون دولار أمريكي. وشهدت الشركة نموًا في الأرباح على مدار 9.625 مليار دولار أمريكي في عام 2024، ما يعكس التحسين المالي الاستراتيجي الذي تشهده الشركة من خلال تنفيذها لبرنامجها الاستثماري. وشهدت الشركة نموًا في الأرباح على مدار 9.625 مليار دولار أمريكي في عام 2024، ما يعكس التحسين المالي الاستراتيجي الذي تشهده الشركة من خلال تنفيذها لبرنامجها الاستثماري.</p>	

Figure 1: Examples of the diverse tasks included in SAHM, covering juristic Q&A, business and accounting MCQs, financial sentiment analysis, report summarization, & event causal reasoning.

- A comprehensive benchmark of 20 LLMs showing that Arabic fluency does not guarantee financial reasoning: models that score up to 91% on MCQ-style tasks degrade substantially on open-ended generation, with the largest gap on Event-Cause QA (1.89–9.84/10).
- Evidence that targeted adaptation rivals scale for Arabic financial NLP: fine-tuning on SAHM yields two complementary 7–8B models SAHM-ALLAM-7B (peak accuracy, surpassing GPT-5 by +21.3 points on Business MCQ, 93.99% vs. 72.68%) and SAHM-JAIS-8B (uniformly positive transfer across all tasks) while matching 72B open-source baselines on average demonstrating ~10× parameter efficiency and establishing domain adaptation as a practical, cost-effective route to trustworthy Arabic financial assistants where frontier API access may be limited.

2 Related Work

**Financial NLP Benchmarks:** English financial NLP has matured through progressively challenging benchmarks. Early work focused on classification and extraction (Araci, 2019), while recent datasets target numerical reasoning over tables (FinQA (Chen et al., 2021), TAT-QA (Zhu et al., 2021)), multi-turn dialogue (ConvFinQA (Chen et al., 2022)), and chain-of-thought verification (FinChain (Xie et al., 2025)). Comprehensive suites such as FinBen (Xie et al., 2024) and PIXIU (Xie et al., 2023) now span 24 tasks including sentiment, NER, and argument mining.

Multilingual extensions have emerged for Chinese (CFinBench (Nie et al., 2025)), and Greek (Plutus (Peng et al., 2025a)), demonstrating that culturally grounded evaluation reveals failure modes invisible in English-only testing. Yet Arabic, spoken by 422M people across economies managing \$4.9T in sovereign wealth (Alhajraf, 2025), lacks any comparable financial benchmark.

**Arabic NLP and the Evaluation Gap:** Arabic resources have grown substantially, but remain shallow in financial coverage. ArBanking77 (Jarar et al., 2023) addresses banking entity detection; Fatwaset (Alyemny et al., 2023) and Hajj-FQA (Aleid and Azmi, 2025) target religious QA. These datasets support general understanding but not reasoning for compliance, numerical analysis, or Shari'ah-aligned decisions. This gap is critical in high-stakes financial settings, where incorrect reasoning can lead to regulatory violations or financial loss. Moreover, the absence of targeted benchmarks limits our ability to diagnose and improve model performance in real-world Arabic financial applications. This gap is significant as Arabic financial texts present distinct challenges: mixed numeral systems (Eastern ٠١٢٣ and Western 0123), code-switching with English acronyms (IFRS, AAOIFI), and domain-specific terminology from Islamic jurisprudence (riba, gharar, sukuk). Meanwhile, Arabic-centric LLMs, including Jais (Sengupta et al., 2023), Falcon-Arabic (Team, 2025), AIN (Heakl et al., 2025a), and Fanar (Abbas et al., 2025), are evaluated only on generic benchmarks that ignore these complexities.

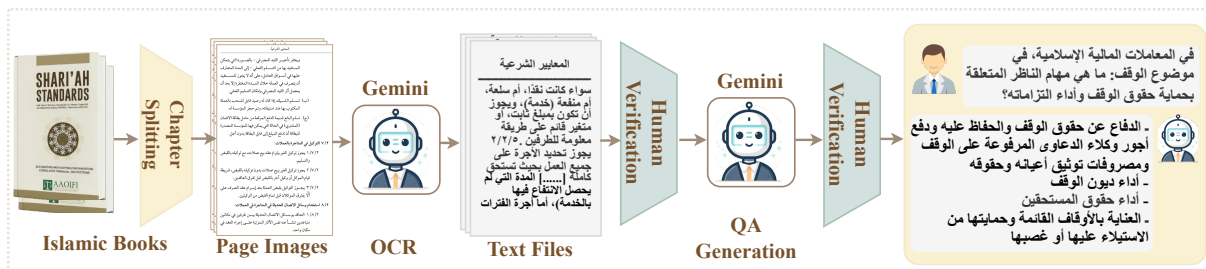


Figure 2: **Pipeline for constructing the Islamic Finance Shari’ah Standards QA dataset.** A hybrid LLMs-human pipeline converts AAOIFI standards into QA pairs through OCR and generation stages, each followed by expert verification to ensure linguistic accuracy and legal fidelity.

### 3 SAHM

We introduce SAHM, a comprehensive benchmark for evaluating Arabic financial reasoning across diverse, real-world tasks spanning Islamic finance, accounting, and market analysis. The benchmark is designed to capture both rule-based reasoning grounded in Shari’ah standards and applied financial understanding in authentic Arabic contexts. It covers multiple task formats, including question answering, multiple-choice reasoning, sentiment analysis, and summarization, enabling holistic assessment of model capabilities. Table 1 provides an overview of dataset composition, task distribution, and train–test splits.

#### 3.1 Islamic Finance Shari’ah Standards QA

Finance in the Gulf and the wider MENA region differs from Western systems: banks, insurers, and capital markets must comply with Islamic principles governed by detailed Shari’ah standards. Frameworks such as AAOIFI and local regulations specify how financial instruments are structured, e.g., lease-to-own arrangements in *Ijara* (إجارة) and compliance requirements for Sukuk<sup>2</sup> issuance (Pomeranz, 1997; Islamic Financial Services Board (IFSB), 2024; Saudi Central Bank, 2024). Yet, most financial benchmarks assume Western instruments (e.g., interest-bearing loans, conventional bonds), leaving models untested on region-specific reasoning about contract permissibility, legal constraints, and Shari’ah compliance. To address this gap, we construct the first Islamic Shari’ah Standards QA dataset from the 1,264-page AAOIFI compendium spanning 52 standards, enabling systematic evaluation of rule-based Islamic financial reasoning.

<sup>2</sup>صكوك (sukuk) are Shari’ah-compliant financial certificates representing ownership in underlying assets rather than interest-bearing debt.

We built the dataset through a multi-step pipeline that converts the AAOIFI compendium into text via OCR with Gemini-2.5-Pro (Google Cloud, 2025) (Appendix A provides details) recommended by Heakl et al. (2025b). Two Islamic finance experts verified the text, preserving diacritics, numerals, and domain-specific terms. In a review of a 25% sample, experts measured a high exact-match rate of  $98.7 \pm 0.7\%$  (95% CI) and strong inter-annotator agreement ( $\kappa = 0.962$ ), confirming OCR reliability. The remaining 1.3% mismatches were minor orthographic or formatting issues (e.g., spacing, punctuation, diacritics), which were corrected in the canonical text; no errors altered the substance of any Shari’ah ruling. After cleanup, we grouped the verified text into thematic clusters (e.g., Murabaḥa) and used Gemini-2.5-Pro to draft candidate Arabic question–answer pairs. Domain experts refined and validated the samples to ensure each question captured the correct ruling with all conditions and exceptions. This human-in-the-loop pipeline transforms dense regulatory prose into high-quality, legally faithful QA pairs for benchmarking Shari’ah-compliant financial reasoning (Figures 1, 2).

#### 3.2 Islamic Financial Fatwa QA

We scraped fatwā archives from 13 official websites across 7 Arab countries to capture the breadth of real-world financial questions Muslims ask (Table 5). The initial crawl yielded 20k fatwas, which we cross-checked against the public FatwaSet (Alyemny et al., 2023) to remove duplicates and then organized into 11 finance-related categories (Table 7), including زكاة (almsgiving), ربا (usury), and مرابحة (cost-plus financing). We then transformed these long, formal texts into concise QA pairs via Gemini-2.5-pro while preserving their juristic meaning.

Task	Dataset	N	Avg. Words (Input)	Avg. Chars (Input)	Avg. Words (Answer)	Avg. Chars (Answer)
MCQ	Accounting Exams MCQ	167	111.5 ± 91.1	674.3 ± 550.5	1.0 ± 0.0	1.0 ± 0.0
	Business Exams MCQ	183	46.3 ± 12.2	298.3 ± 71.6	1.0 ± 0.0	1.0 ± 0.0
	Islamic Financial Fatwa MCQ	2,000	93.1 ± 14.7	536.7 ± 82.6	1.0 ± 0.0	1.0 ± 0.0
	Financial Report Sentiment Analysis MCQ	80	292.3 ± 139.3	1,780.7 ± 841.9	1.0 ± 0.0	1.0 ± 0.0
Open-Ended	Event–Cause Reasoning QA	80	413.6 ± 299.9	2,503.7 ± 1,752.1	350.6 ± 101.8	2,170.4 ± 635.8
	Islamic Fatwa QA	2,000	64.1 ± 36.2	377.3 ± 200.6	89.9 ± 58.4	492.5 ± 324.0
	Islamic Shari'a Standards QA	811	140.1 ± 5.2	287.0 ± 39.5	33.2 ± 22.0	192.1 ± 129.8
	Report Extractive Summarization	80	355.4 ± 165.4	2,144.3 ± 972.3	157.4 ± 66.5	929.1 ± 391.7

Table 1: **Dataset statistics for SAHM.** Mean ± standard deviation of word and character counts per instance, computed over the test split of each dataset. For MCQ tasks the answer is a single letter (A–D), hence the constant 1.0 word/char count.

Specifically, we removed introductory invocations (e.g., “الحمد لله، والصلاة والسلام على رسول الله”) and rhetorical openers (e.g., “أما بعد”) to expose the core inquiry and ruling. We stripped HTML artifacts and redundant navigational references while retaining key metadata such as source URLs for traceability. This normalization step reduces noise and standardizes inputs for downstream QA construction. It also ensures that models focus on substantive legal content rather than stylistic variations. This pipeline removes greetings, honorifics, hyperlinks, and scholar names while preserving Qur’anic citations, juristic terminology, and legal reasoning. Further details in Appendix B. Two native Arabic speakers manually reviewed 10% of the normalized data from each category to verify clarity, linguistic fidelity, and domain correctness. This process resulted in exactly 9,953 high-quality training samples and 2,000 held-out finance-focused test cases (Figure 1).

Afterwards, we converted each test QA pair into multiple-choice (MCQ) format via Gemini-2.5-Pro, enabling both open-ended fatwā reasoning and recognition-style testing. Each MCQ consists of one correct answer derived from the source fatwā and three plausible distractors reflecting common misconceptions. Two native Arabic annotators independently reviewed the test set to assess MCQ correctness, alignment with the source fatwā, and distractor plausibility. The annotators achieved high agreement (Cohen’s  $\kappa = 0.89$ ). Following this pilot phase, we conducted a calibration round in which annotators discussed disagreements, resolved ambiguous cases, and refined shared labeling criteria. One annotator validated the remaining MCQs, ensuring alignment with source fatwās, correct terminology, and no misleading options. A final audit confirmed that 95% of MCQs aligned exactly with their original QA pairs; we discarded the remaining 5% and excluded them from evaluation.

### 3.3 Business & Accounting Exams MCQ

Professional accounting assessment resources remain largely English-centric, with key certifications such as the CPA exam conducted exclusively in English. To address this gap and the limited availability of Arabic training materials despite the existence of IFRS translations, we design culturally and linguistically adapted MCQ samples covering IFRS treatments, financial ratios, budgeting, and costing, incorporating authentic Arabic financial terminology such as معدل دوران الأصول (asset turnover ratio) and زكاة الشركات (corporate almsgiving) within contextually accurate scenarios rather than direct translations of Western exam questions (Appendix D). We constructed the dataset by collecting 10 business exams and 8 accounting exams from multiple Arabic-speaking countries. We extract the text from the exam PDFs via Gemini-2.5-Pro following Heakl et al. (2025b), after which two native Arabic-speaking annotators reviewed by comparing the OCR output against the original questions, correcting recognition errors, and validating formatting. The final dataset contains 457 business questions and 416 accounting questions, examples in Figure 1.

### 3.4 Financial Report Sentiment Analysis

Despite managing trillions in assets, Arabic markets lack region-specific sentiment benchmarks. Existing English datasets (Maia et al., 2018b) focus on Western market narratives and do not capture signals central to MENA markets, including OPEC+ production decisions, صكوك (sukuk) issuances, subsidy reforms, and Shari’ah-compliance rulings. These challenges are amplified by culturally grounded terminology, e.g., مراجعة (cost-plus financing), and stylistic variation in Arabic reporting, where subtle modifiers can reverse sentiment polarity. To address this gap, We construct the first Arabic financial sentiment benchmark from authentic market reports.

We collect 200 Arabic financial reports, 100 Islamic finance–focused and 100 general, from Argaam<sup>3</sup>, and annotate them with three document-level sentiment labels: Positive, Negative, and Neutral. Two native Arabic annotators labeled all reports using a custom web-based platform (Figure 13) following guidelines emphasizing holistic document interpretation, achieving high agreement ( $\kappa = 0.91$ ). We then conducted a calibration phase to resolve disagreements and refine criteria. For mixed-signal reports, we assign the dominant polarity if >60% of content supports it; otherwise Neutral, with a third expert adjudicating residual disagreements. The dataset is split into 120 training and 80 test reports.

### 3.5 Report Extractive Summarization

Extractive summarization is critical for Arabic financial reporting, where annual reports are written in Arabic but frequently contain mixed numeral systems, embedded English financial acronyms and brand names rendered in Arabic script (e.g., إنش إس بي سي / IFRS and المعايير الدولية للتقارير المالية / HSBC), and specialized Islamic finance terminology such as صكوك (sukuk). Misinterpreting or omitting these elements can distort regulatory interpretation, compliance assessment, and financial valuation. To support this task, we compile 200 Arabic financial reports, 100 general and 100 Islamic from Argaam and annotate them with extractive summaries written in Arabic by two native Arabic speakers. Rather than treating summarization as a subjective agreement task, we use ROUGE (Lin, 2004) to measure overlap between independently produced summaries as a consistency check and select the more complete summary as the gold reference. We split the dataset into 120 training reports and 80 test reports. Further details in Appendix F.

### 3.6 Event-Cause Reasoning QA

Financial event-cause reasoning is underexplored in Arabic due to the lack of datasets that require models to explain why financial events occur and what implications they entail. To address this gap, we introduce an event-cause reasoning task that evaluates whether models can analyze Arabic financial reports and produce analytical explanations grounded in reported financial data, including market movements and صكوك issuances.

<sup>3</sup><https://www.argaam.com/>

We collect 200 Arabic financial reports (100 Islamic, 100 general) from Argaam. Two native Arabic financial experts annotate each report by creating one analytical question linking multiple data points and a concise answer explaining causes and implications using only the article content; a pilot on 20 reports ensures guideline clarity. Details are in Appendix G. We assess quality via Cohen’s  $\kappa = 0.86$  for event-cause identification and ROUGE overlap for answer consistency. After calibration to resolve disagreements, one expert completes the remaining annotations under the agreed criteria.

## 4 Experiments

**Evaluated Models:** We evaluated 20 models spanning Arabic-centric models (Bari et al., 2024; Abbas et al., 2025; silma-ai, 2024) (publicly available instruction-tuned systems for regional adaptation), open-weight models (Riviere et al., 2024; Kamath et al., 2025; Grattafiori et al., 2024; Yang et al., 2024; Jiang et al., 2024) (strong multilingual and general-purpose baselines), and proprietary models (Hurst et al., 2024; OpenAI, 2025; Anthropic, 2025b,c,a; Google, 2024), enabling controlled analysis across language, scale, and capability dimensions. To assess whether domain-specific fine-tuning can close the gap between Arabic-centric and frontier models, we fine-tune three Arabic LLMs (ALLAM-7B, Jais-2-8B, and SILMA-9B) on the SAHM training split using LoRA ( $r=64$ ,  $\alpha=128$ ,  $lr=2e-4$ , 3 epochs). Detailed model specifications are provided in Table 6.

We evaluate Accounting Exams, Business Exams, Fatwa MCQ, and Financial Sentiment with exact-match accuracy, normalizing free-form outputs (e.g., option text/letters) to a single choice before scoring (Appendix H). For extractive summarization, we report ROUGE-F1 (ROUGE-1/2/L) against gold extractive references (models are instructed to output verbatim sentences). For Fatwa QA, Shari’ah Standards QA, and Event-Cause QA, we use Gemini-2.5-Flash as an LLM-as-a-judge (blind to model identity). Given the Arabic prompt, gold reference, and model answer, it returns a JSON-validated additive  $[0, 10]$  score based on a shared rubric assessing alignment with the reference ruling or conclusion, preservation of key constraints or quantitative fidelity, correctness (doctrinal, factual, or financial), Arabic clarity, and grounding.

Model	MCQ (Accuracy % $\uparrow$ )					Open-Ended QA (Score 0–10 $\uparrow$ )			
	Datasets					Datasets			
	Accounting	Business	Fatwā	Sentiment	Mean	Event-Cause QA	Islamic-Standards-QA	Fatwa-QA	Mean
<b>Open-source Models: <math>\geq</math> 70B Parameters</b>									
Qwen2.5-72B-Instruct (Yang et al., 2024)	65.87 $\pm$ 2.70	74.86 $\pm$ 0.32	84.65 $\pm$ 0.33	75.00 $\pm$ 1.25	75.10	8.1000 $\pm$ 0.10	5.6330 $\pm$ 0.10	5.3912 $\pm$ 0.06	6.3747
LLaMA-3.1-70B (Grattafiori et al., 2024)	52.10 $\pm$ 2.79	77.60 $\pm$ 1.14	84.90 $\pm$ 0.15	80.00 $\pm$ 3.31	73.65	6.623 $\pm$ 0.15	3.7245 $\pm$ 0.10	4.7607 $\pm$ 0.08	5.036
<b>Open-source Models: <math>&lt;</math> 70B Parameters</b>									
Qwen2.5-14B-Instruct (Yang et al., 2024)	49.10 $\pm$ 3.93	63.39 $\pm$ 0.83	76.05 $\pm$ 0.85	57.50 $\pm$ 3.82	61.51	7.4975 $\pm$ 0.10	4.8806 $\pm$ 0.10	4.0576 $\pm$ 0.06	5.4786
Qwen2.5-7B-Instruct (Yang et al., 2024)	48.50 $\pm$ 2.85	59.56 $\pm$ 1.14	70.00 $\pm$ 0.28	55.00 $\pm$ 1.91	58.27	6.1038 $\pm$ 0.12	3.4039 $\pm$ 0.10	2.6815 $\pm$ 0.08	4.0631
Gemma-2-9B-IT (Riviere et al., 2024)	49.10 $\pm$ 2.74	63.39 $\pm$ 3.83	66.60 $\pm$ 0.61	55.00 $\pm$ 1.44	58.52	7.1438 $\pm$ 0.08	4.2306 $\pm$ 0.08	3.4266 $\pm$ 0.06	4.9336
Gemma-3-27B-IT (Kamath et al., 2025)	53.89 $\pm$ 2.16	73.22 $\pm$ 0.32	80.65 $\pm$ 0.18	80.00 $\pm$ 0.72	71.94	8.7188 $\pm$ 0.05	6.1708 $\pm$ 0.08	5.1929 $\pm$ 0.05	6.6942
Gemma-3-4B-IT (Kamath et al., 2025)	38.32 $\pm$ 2.27	67.76 $\pm$ 0.32	61.35 $\pm$ 0.18	75.00 $\pm$ 1.44	60.61	7.4075 $\pm$ 0.08	2.8985 $\pm$ 0.08	2.4767 $\pm$ 0.06	4.2609
LLaMA-3.1-8B (Grattafiori et al., 2024)	41.92 $\pm$ 3.28	60.66 $\pm$ 4.45	64.05 $\pm$ 3.62	73.75 $\pm$ 5.77	60.60	4.9231 $\pm$ 0.18	2.5168 $\pm$ 0.12	1.4025 $\pm$ 0.08	2.9475
Mixtral-8x7B-Instruct (Jiang et al., 2024)	32.93 $\pm$ 1.04	60.66 $\pm$ 0.63	62.15 $\pm$ 0.34	70.00 $\pm$ 0.72	56.44	4.5538 $\pm$ 0.08	2.4980 $\pm$ 0.08	1.7890 $\pm$ 0.06	2.9471
<b>Proprietary Models: Reasoning-Enhanced</b>									
GPT-5 (OpenAI, 2025)	65.27 $\pm$ 2.27	72.68 $\pm$ 1.26	90.75 $\pm$ 0.45	78.75 $\pm$ 1.25	76.86	9.6831 $\pm$ 0.03	8.7965 $\pm$ 0.05	8.0515 $\pm$ 0.04	8.8437
GPT-4o (Hurst et al., 2024)	60.48 $\pm$ 2.07	<b>78.14</b> $\pm$ 0.32	87.70 $\pm$ 0.10	77.50 $\pm$ 0.00	75.96	8.3125 $\pm$ 0.06	6.6598 $\pm$ 0.08	6.5219 $\pm$ 0.04	7.1647
<b>Proprietary Models: General-Purpose</b>									
Claude-Opus-4.5 (Anthropic, 2025b)	77.84 $\pm$ 2.42	76.50 $\pm$ 1.14	91.75 $\pm$ 0.33	75.00 $\pm$ 2.50	80.27	9.6818 $\pm$ 0.03	8.0438 $\pm$ 0.05	8.8090 $\pm$ 0.03	8.8449
Claude-Sonnet-4.5 (Anthropic, 2025c)	78.44 $\pm$ 1.20	76.50 $\pm$ 1.45	88.15 $\pm$ 0.38	77.50 $\pm$ 1.25	80.15	9.3388 $\pm$ 0.04	8.2588 $\pm$ 0.05	7.6049 $\pm$ 0.03	8.4008
Claude-Haiku-4.5 (Anthropic, 2025a)	67.66 $\pm$ 1.80	73.77 $\pm$ 1.30	84.90 $\pm$ 0.40	77.50 $\pm$ 1.80	75.96	9.1050 $\pm$ 0.05	7.0002 $\pm$ 0.07	6.5341 $\pm$ 0.05	7.5464
Gemini-3-Flash (preview) (Google, 2024)	76.05 $\pm$ 1.95	74.86 $\pm$ 0.95	89.90 $\pm$ 0.30	81.25 $\pm$ 1.25	80.52	<b>9.8369</b> $\pm$ 0.02	<b>9.1649</b> $\pm$ 0.03	<b>9.1571</b> $\pm$ 0.02	<b>9.0798</b>
GPT-4o-mini (Hurst et al., 2024)	58.08 $\pm$ 2.10	77.60 $\pm$ 0.40	81.75 $\pm$ 0.20	75.00 $\pm$ 0.50	73.61	7.9613 $\pm$ 0.08	5.6094 $\pm$ 0.10	5.3087 $\pm$ 0.06	6.2931
<b>Arabic Models</b>									
ALLAM-7B (Bari et al., 2024)	44.91 $\pm$ 3.55	68.31 $\pm$ 3.83	74.40 $\pm$ 2.83	58.75 $\pm$ 2.00	61.59	6.8875 $\pm$ 0.10	4.9364 $\pm$ 0.08	4.2185 $\pm$ 0.05	5.3475
Fanar-1-9B (Abbas et al., 2025)	47.31 $\pm$ 2.42	66.12 $\pm$ 1.67	74.45 $\pm$ 0.35	58.75 $\pm$ 2.60	61.66	7.5850 $\pm$ 0.10	4.9607 $\pm$ 0.08	4.4600 $\pm$ 0.06	5.6686
SILMA-9B (silma-ai, 2024)	50.90 $\pm$ 21.73	69.40 $\pm$ 6.61	62.55 $\pm$ 5.57	30.00 $\pm$ 3.75	53.21	1.8969 $\pm$ 0.20	3.3547 $\pm$ 0.12	2.0711 $\pm$ 0.08	2.4409
Jais-2-8B (Sengupta et al., 2023)	35.33 $\pm$ 3.00	60.30 $\pm$ 2.80	66.10 $\pm$ 1.80	46.25 $\pm$ 2.50	52.00	4.6922 $\pm$ 0.15	4.245 $\pm$ 0.10	2.5147 $\pm$ 0.08	3.8133

Table 2: **Unified leaderboard comparing MCQ tasks (Accuracy %) and open-ended QA tasks (Score 0–10).** Values shown as mean $\pm$ std over 3 runs; open-ended scores are judged by two independent LLM judges. Open-ended QA Mean is averaged over Event-Cause QA, Islamic-Standards-QA, and Fatwa-QA.

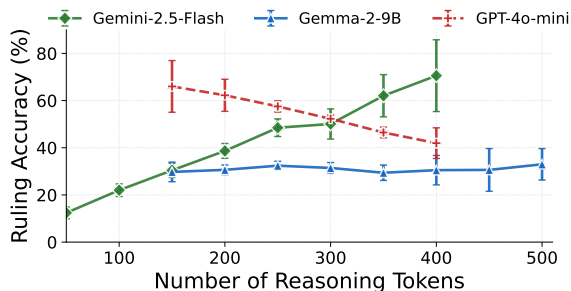


Figure 3: **Effect of reasoning token budget on ruling accuracy.** Green indicates improvement with increased budget, red indicates decline, and blue indicates no change.

We validate the judge with two expert Arabic annotators on 200 randomly sampled outputs across the three tasks (MSE 0.41, Pearson  $r=0.92$ ; inter-annotator agreement  $\kappa=0.84$  on discretized scores; Appendix J). All judge and model generations use greedy decoding (temperature 0; no sampling) with fixed maximum lengths; full prompts, rubrics, schema, critical checks, and settings appear in Appendix J. We organize our findings around three core questions: (1) How do models perform across recognition versus generation tasks? (2) What distinguishes strong Arabic financial reasoning from mere language fluency? (3) Where do models systematically fail, and why?

## 4.1 Main Results

Table 2 summarizes performance across all tasks, revealing clear disparities across models. We analyze these patterns to identify where models succeed and fail.

**Accounting Reasoning Gap:** Shown in Table 2, Claude models exhibit substantial superiority on Accounting tasks, with Claude-Sonnet-4.5 exceeding GPT-5 by over 13% the largest proprietary-to-proprietary gap in our evaluation. Crucially, this disparity cannot be attributed to general Arabic language proficiency alone, as these models achieve near-parity on Business (76.50% vs. 72.68%) and Fatwa (91.75% vs. 90.75%) tasks. We instead attribute this divergence to Claude’s stronger capacity for procedural numerical reasoning, the ability to apply rule-based standards (e.g., IFRS, Egyptian Auditing Standards) through multi-step logical chains. This suggests Arabic domain reasoning is distinct from general language proficiency, warranting further study. Notably, Gemini-3-Flash inverts the recognition–generation tradeoff, achieving top Open-Ended QA despite moderate MCQ performance, likely due to longer reasoning chains. This is supported by Figure 3, where the Gemini family shows increased ruling accuracy with larger reasoning token budgets.



Model	ROUGE-1	ROUGE-2	ROUGE-L
<b>Proprietary Models – Reasoning-Enhanced</b>			
Claude-Opus-4.5	78.22	63.17	64.14
GPT-5	75.19	63.70	64.11
Claude-Sonnet-4.5	<b>79.86</b>	<b>64.98</b>	<b>65.13</b>
<b>Proprietary Models – General-Purpose</b>			
Claude-Haiku-4.5	<b>79.39</b>	61.40	63.62
GPT-4o-mini	77.79	62.90	64.08
GPT-4o	78.91	<b>63.16</b>	<b>63.71</b>
Gemini-3-Flash	49.36	35.83	43.02
Gemini-2.5-Flash	39.46	27.17	36.81
<b>Open-source Models: ≥ 70B parameters</b>			
Gemma-3-27B-IT	<b>79.25</b>	<b>63.57</b>	<b>63.42</b>
Qwen2.5-72B-Instruct	40.52	29.50	34.04
Meta-LLaMA-3.1-70B	39.64	31.40	32.65
<b>Open-source Models: &lt; 70B parameters</b>			
Qwen2.5-14B-Instruct	44.42	30.90	35.82
Gemma-3-4B-IT	<b>76.52</b>	<b>62.06</b>	<b>60.93</b>
Meta-LLaMA-3.1-8B	66.67	47.92	56.10
Mixtral-8x7B-Instruct	32.71	13.07	23.78
Qwen2.5-7B-Instruct	25.15	12.01	21.86
<b>Arabic Models</b>			
Jais-2-8B	<b>73.68</b>	<b>56.54</b>	<b>61.17</b>
Fanar-1-9B-Instruct	60.51	35.97	46.96
ALLaM-7B-Instruct	35.97	22.61	28.24
SILMA-9B-Instruct	27.92	16.66	25.99

Table 3: Extractive summarization performance on Arabic financial reports evaluated using ROUGE F1 (%).

## 5.1 Extractive Summarization

Table 3 reveals a striking inversion: Claude-Sonnet-4.5 achieves the highest ROUGE-1 (79.86), while Gemini-2.5-Flash a strong open-ended reasoner collapses to 39.46, underperforming even GPT-4o-mini (77.79). This exposes a fundamental tension: extractive summarization rewards *verbatim selection*, not generative fluency. Consider a typical report: “نجحت شركة بن غاطي للتطوير العقاري في طرح المزيد من الصكوك... بقيمة 300 مليون دولار أمريكي، ببورصة لندن وناسداك دبي” (Binghatti Development successfully issued additional sukuk... valued at \$300M, listed on the London Stock Exchange and Nasdaq Dubai). The gold summary must preserve the entity name, Islamic instrument (*sukuk*), exact figure, and dual listing elements paraphrasing models systematically distort. Surprisingly, Gemma-3-4B-IT achieves 76.52 ROUGE-1, rivaling Claude-Opus-4.5 (78.22) with a fraction of the parameters, suggesting extraction benefits from constrained generation rather than extended reasoning. For Arabic-centric models, domain-specific tuning is decisive: SAHM-7B-Instruct attains 57.79 ROUGE-L, outperforming ALLaM-7B by **+29.55 points**, showing pretraining alone is insufficient.

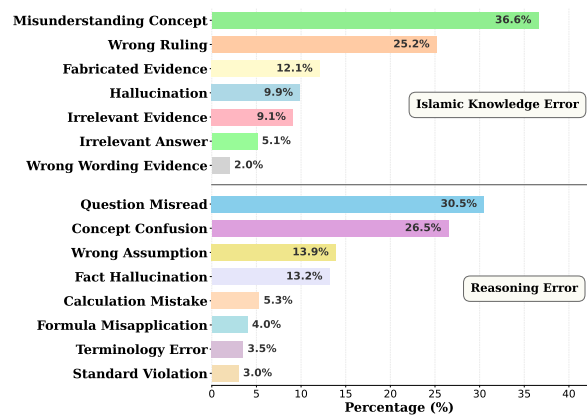


Figure 6: Root cause distribution of model errors across Islamic knowledge and reasoning tasks.

## 5.2 Domain Adaptation Across Arabic LLMs

We systematically evaluate domain adaptation across major Arabic-centric LLMs. Table 4 reveals three distinct adaptation profiles: (1) *high-gain* bases like ALLaM yielding SAHM-ALLaM-7B with substantial improvement (+13.68%), (2) *stable-gain* bases like Jais-2 yielding SAHM-JAIS-8B with consistent improvement across all MCQ metrics (+5.78%, with notably strong gains in Sentiment +11.72%), and (3) *selective-gain* bases like SILMA that improve on some tasks (Sentiment +23.8%) but regress on others (Accounting -7.8%).

## 5.3 Error Analysis

To diagnose failure modes, we analyze 500 randomly sampled incorrect responses across all datasets, grouped by required competence: **Islamic Knowledge Errors** (Fatwa QA, Shari’ah Standards QA, Fatwa MCQ) and **Reasoning Errors** (Accounting, Business, Event-Cause QA); summarization errors are treated in §5.1 and sentiment via accuracy. The two annotators (see §3) jointly adjudicated each error against the gold reference through a **consensus protocol**, additionally verifying cited religious evidence for Islamic tasks. Consensus over independent annotation with post-hoc IAA maximizes taxonomy coverage and ensures consistent categorization across heterogeneous errors requiring both jurisprudential and financial expertise.

**Error Breakdown:** Figure 6 shows two dominant error types: *Misunderstanding Concept* and *Wrong Ruling*, accounting for 58.5% of failures. Fabricated Evidence (11.4%) and Hallucination (9.3%) follow. Calculation errors are rare (0.3%); models struggle not with arithmetic, but with selecting the correct computation.

Model	MCQ (Accuracy % $\uparrow$ )					Open-Ended QA (Score 0–10 $\uparrow$ )			
	Accounting	Business	Fatwā	Sentiment	Mean	Event-Cause	Fatwa-QA	Islamic-Std	Mean
<b>Base Models</b>									
ALLAM-7B	44.91	68.31	74.40	58.75	61.59	6.89	4.94	4.22	5.35
Jais-2-8B	35.33	60.30	66.10	46.25	52.00	4.69	2.51	4.24	3.81
SILMA-9B	50.90	69.40	62.55	30.00	53.21	1.90	3.35	2.07	2.44
<b>Fine-tuned Models</b>									
<b>SAHM-ALLAM-7B</b>	<b>71.40</b> (+26.5)	<b>93.99</b> (+25.7)	<b>74.45</b> (+0.1)	<b>61.25</b> (+2.5)	<b>75.27</b> (+13.7)	6.79 (-0.1)	<b>6.48</b> (+1.5)	4.12 (-0.1)	<b>5.80</b> (+0.5)
<b>SAHM-JAIS-8B</b>	40.72 (+5.4)	62.30 (+2.0)	70.14 (+4.0)	57.97 (+11.7)	57.78 (+5.8)	<b>5.25</b> (+0.6)	4.69 (+2.2)	<b>4.97</b> (+0.7)	4.97 (+1.16)
SILMA-9B (fine-tuned)	43.11 (-7.8)	75.96 (+6.6)	60.60 (-2.0)	53.75 (+23.8)	58.36 (+5.2)	2.01 (+0.1)	3.67 (+0.3)	3.67 (+1.6)	3.12 (+0.7)

Table 4: Domain adaptation across Arabic LLMs. MCQ accuracy (%) and Open-Ended QA scores (0–10) before and after fine-tuning on SAHM. **Bold model names** (SAHM-ALLAM-7B, SAHM-JAIS-8B) denote the two released SAHM-family artifacts; SILMA-9B is included as a comparison case illustrating that adaptation outcomes depend on base-model properties.

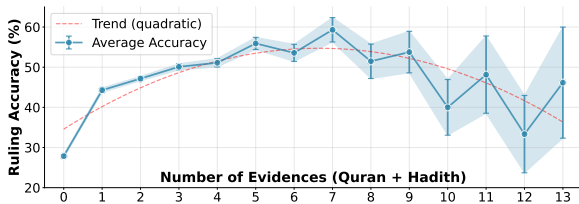


Figure 7: Effect of number of evidences from Hadith and Quran on Ruling Accuracy.

**Effect on Evidence Count on Accuracy.** Figure 7 examines whether the presence of scriptural evidence (Qur’anic verses and *ḥadīth*) in reference answers correlates with model accuracy. We observe a logarithmic relationship: accuracy rises from 28% with zero evidence to approximately 55% with six or more citations. This pattern admits two interpretations. Optimistically, models may leverage textual evidence as grounding signals. Pessimistically, questions with more evidence may simply be easier or more frequently represented in training data. The increased variance at higher evidence counts (shaded region) suggests the relationship is not deterministic.

## 6 Conclusion and Future Work

We introduce SAHM, the first Arabic financial NLP benchmark integrating modern finance and Shari’ah-compliant reasoning across seven tasks. Evaluating 20 LLMs shows Arabic fluency does not imply financial reasoning, while fine-tuning on SAHM yields two models: SAHM-ALLAM-7B surpasses GPT-5 on Accounting and Business and matches 72B baselines, and SAHM-JAIS-8B shows consistent gains across tasks, demonstrating that targeted domain adaptation outperforms scale. We release all resources to support trustworthy Arabic financial assistants.

Several directions extend this work. First, SAHM currently focuses on formal financial text; incorporating informal genres such as retail investor discourse, social media financial discussions, and dialectal Arabic would broaden coverage. Second, Arabic financial reports frequently contain tables, charts, and mixed-format documents; extending the benchmark to multimodal reasoning over structured financial data is a natural next step. Third, our evaluation assesses answer correctness but not evidence traceability; future metrics should explicitly verify cited Qur’anic verses, *ḥadīth* reports, and AAOIFI standard references. Fourth, cross-lingual transfer from English financial benchmarks to Arabic remains unexplored; investigating whether English financial reasoning capabilities transfer to Arabic could reduce data requirements. Finally, regional variation in Shari’ah interpretation across different supervisory bodies warrants task variants that evaluate model robustness to jurisdictional differences in Islamic finance rulings.

More broadly, SAHM highlights the need for evaluation frameworks that move beyond surface-level language proficiency toward domain-grounded reasoning in high-stakes settings. As financial decision-making increasingly relies on LLMs, ensuring correctness, transparency, and alignment with regulatory and ethical standards becomes critical. Our findings also underscore the importance of culturally and legally informed benchmarks in shaping reliable AI systems for specialized domains. We hope this work motivates research on trustworthy Arabic financial AI, bridging language understanding, legal compliance, and real-world applicability, enabling safe and effective deployment across diverse financial and regulatory settings.

## Limitations

**Scope and coverage.** SAHM is built from curated, document-grounded sources and covers as much of the available public material as feasible; however, practical access and usage constraints on some on-line sources limit the extent to which additional genres can be incorporated at this time. As a result, while the benchmark provides strong provenance and reduces ambiguity, it does not yet cover all Arabic financial genres (e.g., informal retail-investor discourse) or fully capture regional and institutional variation in Arabic financial writing.

**Shari’ah-related content.** For Shari’ah-oriented questions, SAHM evaluates faithfulness to the referenced material and the reasoning constraints reflected in the provided sources; since interpretations may differ across jurisdictions and supervisory bodies, the benchmark is not intended to adjudicate between schools of thought, but rather to test source-grounded answering under the stated assumptions.

**Future evaluation directions.** As future work, we plan to develop evaluation metrics that explicitly assess (i) the existence and correctness of cited, source-verifiable evidence including traceable support from the underlying materials (e.g., fatwa text, and financial report statements) and, when answers cite religious evidence, the correctness of references such as Qur’anic verses, hadith reports, or named fiqh sources; and (ii) the accuracy of book/standard citations in model outputs (e.g., correct document title, section/article identifiers, and pointers that match the relevant source segment), enabling more direct measurement of citation faithfulness and evidence-groundedness.

## Ethical Statement and Broad Impact

**Licensing.** We release SAHM under a dual license: (1) code and evaluation scripts under MIT License, and (2) annotation data under CC BY-NC 4.0, restricting commercial use while enabling academic research. Users must independently obtain source documents where applicable.

**Availability.** The benchmark data, models, and evaluation scripts are publicly available at <https://huggingface.co/SahmBenchmark>, <https://github.com/mbzuai-nlp/SAHM>.

## Acknowledgments

We acknowledge The Fin AI community for its research support, feedback, and collaborative environment that contributed to this work.

## References

- Ummar Abbas, Mohammad Shahmeer Ahmad, Firoj Alam, Enes Altinisik, Ehsannedin Asgari, Yazan Boshmaf, Sabri Boughorbel, Sanjay Chawla, Shammur A. Chowdhury, Fahim Dalvi, Kareem Darwish, Nadir Durrani, Mohamed Elfeky, Ahmed K. Elmagarmid, Mohamed Y. Eltabakh, Masoomali Fatehkhia, 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, and Chaoyi Ruan. 2025. *Fanar: An Arabic-centric multimodal generative AI platform*. *ArXiv preprint*, abs/2501.13944.
- Hayfa A Aleid and Aqil M 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(6):135.
- Salem Alhajraf. 2025. *Strategic role of sovereign wealth funds in the Gulf’s energy transition and economic diversification*. Technical report, Rice University’s Baker Institute for Public Policy.
- Ohoud Alyemny, Hend S. Al-Khalifa, and Abdulrahman A. Mirza. 2023. *A data-driven exploration of a new Islamic fatwas dataset for Arabic NLP tasks*. *Data*, 8(10):155.
- Anthropic. 2025a. *System Card: Claude Haiku 4.5*. *Anthropic*.
- Anthropic. 2025b. *System Card: Claude Opus 4.5*. *Anthropic*.
- Anthropic. 2025c. *System Card: Claude Sonnet 4.5*. *Anthropic*.
- Dogu Araci. 2019. *FinBERT: Financial sentiment analysis with pre-trained language models*. *ArXiv preprint*, abs/1908.10063.
- M Saiful Bari, Yazeed Alnumay, Norah A. Alzahrani, Nouf M. Alotaibi, Hisham A. Alyahya, Sultan Al-Rashed, Faisal A. Mirza, Shaykha Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Maged Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairah,

- Areeb Alowisheq, and Haidar Khan. 2024. [AL-LaM: Large language models for Arabic and English](#). *Preprint*, arXiv:2407.15390.
- Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. 2021. [FinQA: A dataset of numerical reasoning over financial data](#). In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3697–3711, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.
- Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. 2022. [ConvFinQA: Exploring the chain of numerical reasoning in conversational finance question answering](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 6279–6292, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.
- Google. 2024. [Gemini 3 Flash model card](#). *Google*.
- Google Cloud. 2025. Gemini 2.5 pro — generative ai on vertex ai. <https://cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro>. Last accessed: 2025-10-06.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. [The Llama 3 herd of models](#). *ArXiv preprint*, abs/2407.21783.
- Ahmed Heakl, Sara Ghaboura, Omkar Thawakar, Fahad Shahbaz Khan, Hisham Cholakkal, Rao Muhammad Anwer, and Salman H. Khan. 2025a. [AIN: The Arabic inclusive large multimodal model](#). *ArXiv preprint*, abs/2502.00094.
- Ahmed Heakl, Muhammad Abdullah Sohail, Mukul Ranjan, Rania Elbadry, Ghazi Shazan Ahmad, Mohamed El-Geish, Omar Maher, Zhiqiang Shen, Fahad Shahbaz Khan, and Salman Khan. 2025b. [KITAB-Bench: A comprehensive multi-domain benchmark for Arabic OCR and document understanding](#). In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 22006–22024, Vienna, Austria. Association for Computational Linguistics.
- Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, et al. 2024. [GPT-4o system card](#). *Preprint*, arXiv:2410.21276.
- Islamic Financial Services Board (IFSB). 2024. [Islamic financial services industry stability report 2024](#).
- Mustafa Jarrar, Ahmet Birim, Mohammed Khalilia, Mustafa Erden, and Sana Ghanem. 2023. [ArBanking77: Intent detection neural model and a new dataset in modern and dialectal Arabic](#). In *Proceedings of ArabicNLP 2023*, pages 276–287, Singapore (Hybrid). Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth e Lacroix, and William El Sayed. 2024. [Mixture of experts](#). *Preprint*, arXiv:2401.04088.
- Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ram e, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Ga el Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, et al. 2025. [Gemma 3 technical report](#). *ArXiv preprint*, abs/2503.19786.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Macedo Maia, Siegfried Handschuh, Andr e Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018a. [WWW’18 Open Challenge: Financial opinion mining and question answering](#). In *Companion of the The Web Conference 2018 on The Web Conference 2018*, WWW’18, pages 1941–1942, Lyon , France. ACM.
- Macedo Maia, Siegfried Handschuh, Andr e Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018b. [WWW’18 Open Challenge: Financial opinion mining and question answering](#). In *Companion Proceedings of the The Web Conference 2018*, WWW ’18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.
- Ying Nie, Binwei Yan, Tianyu Guo, Hao Liu, Haoyu Wang, Wei He, Binfan Zheng, Weihao Wang, Qiang Li, Weijian Sun, Yunhe Wang, and Dacheng Tao. 2025. [CFinBench: A comprehensive Chinese financial benchmark for large language models](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, NAACL’25, pages 876–891, Albuquerque, New Mexico. Association for Computational Linguistics.
- OpenAI. 2025. [GPT-5 system card](#). *OpenAI*.
- Xueqing Peng, Triantafillos Papadopoulos, Efstathia Soufleri, Polydoros Giannouris, Ruoyu Xiang, Yan

- Wang, Lingfei Qian, Jimin Huang, Qianqian Xie, and Sophia Ananiadou. 2025a. [Plutus: Benchmarking large language models in low-resource Greek finance](#). *ArXiv preprint*, abs/2502.18772.
- Xueqing Peng, Lingfei Qian, Yan Wang, Ruoyu Xiang, Yueru He, Yang Ren, Mingyang Jiang, Jeff Zhao, Huan He, Yi Han, Yun Feng, Yuechen Jiang, Yupeng Cao, Haohang Li, Yangyang Yu, Xiaoyu Wang, Penglei Gao, Shengyuan Lin, Keyi Wang, Shanshan Yang, Yilun Zhao, Zhiwei Liu, Peng Lu, Jerry Huang, Suyuchen Wang, Triantafillos Papadopoulos, Polydoros Giannouris, Efsthathia Soufleri, Nuo Chen, Guojun Xiong, Zhiyang Deng, Yijia Zhao, Mingquan Lin, Meikang Qiu, Kaleb E. Smith, Arman Cohan, Xiao-Yang Liu, Jimin Huang, Alejandro Lopez-Lira, Xi Chen, Junichi Tsujii, Jian-Yun Nie, Sophia Ananiadou, and Qianqian Xie. 2025b. [MultiFinBen: A multilingual, multimodal, and difficulty-aware benchmark for financial LLM evaluation](#). *ArXiv preprint*, abs/2506.14028.
- Felix Pomeranz. 1997. The accounting and auditing organization for Islamic financial institutions: An important regulatory debut. *Journal of International Accounting, Auditing and Taxation*, 6(1):123–130.
- Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, et al. 2024. [Gemma 2: Improving open language models at a practical size](#). *ArXiv preprint*, abs/2408.00118.
- Saudi Central Bank. 2024. [Saudi Central Bank \(sama\) regulatory framework for Islamic finance](#).
- Neha Sengupta, Sunil Kumar Sahu, Bokang Jia, Satheesh Katipomu, Haonan Li, Fajri Koto, Osama Mohammed Afzal, Samta Kamboj, Onkar Pandit, Rahul Pal, Lalit Pradhan, Zain Muhammad Mujahid, Massa Baali, Alham Fikri Aji, Zhengzhong Liu, Andy Hock, Andrew Feldman, Jonathan Lee, Andrew Jackson, Preslav Nakov, Timothy Baldwin, and Eric P. Xing. 2023. [Jais and Jais-chat: Arabic-centric foundation and instruction-tuned open generative large language models](#). *ArXiv preprint*, abs/2308.16149.
- silma-ai. 2024. [SILMA 9B Instruct v1.0](#). <https://huggingface.co/silma-ai/SILMA-9B-Instruct-v1.0>.
- Falcon-LLM Team. 2025. [Falcon-Arabic: A breakthrough in Arabic language models](#).
- Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. 2024. [FinBen: A holistic financial benchmark for large language models](#). *Advances in Neural Information Processing Systems*, 37:95716–95743.
- Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. [PIXIU: A large language model, instruction data and evaluation benchmark for finance](#). In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NeurIPS’23, Red Hook, NY, USA. Curran Associates Inc.
- Zhuohan Xie, Daniil Orel, Rushil Thareja, Dhruv Sahnan, Hachem Madmoun, Fan Zhang, Debopriyo Banerjee, Georgi Georgiev, Xueqing Peng, Lingfei Qian, Jimin Huang, Jinyan Su, Aaryamonvikram Singh, Rui Xing, Rania Elbadry, Chen Xu, Haonan Li, Fajri Koto, Ivan Koychev, Tanmoy Chakraborty, Yuxia Wang, Salem Lahlou, Veselin Stoyanov, Sophia Ananiadou, and Preslav Nakov. 2025. [FinChain: A symbolic benchmark for verifiable chain-of-thought financial reasoning](#). *Preprint*, arXiv:2506.02515.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *ArXiv preprint*, abs/2412.15115.
- Xiao Zhang, Ruoyu Xiang, Chenhan Yuan, Duanyu Feng, Weiguang Han, Alejandro Lopez-Lira, Xiao-Yang Liu, Meikang Qiu, Sophia Ananiadou, Min Peng, Jimin Huang, and Qianqian Xie. 2024. [Dólares or Dollars? Unraveling the bilingual prowess of financial LLMs between Spanish and English](#). In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining, August 25-29, 2024, KDD’24*, pages 6236–6246, Barcelona, Spain. ACM.
- Yilun Zhao, Hongjun Liu, Yitao Long, Rui Zhang, Chen Zhao, and Arman Cohan. 2024. [Finance-MATH: Knowledge-intensive math reasoning in finance domains](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, ACL’24, pages 12841–12858, Bangkok, Thailand. Association for Computational Linguistics.
- Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. 2021. [TAT-QA: A question answering benchmark on a hybrid of tabular and textual content in finance](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3277–3287, Online. Association for Computational Linguistics.

## A Islamic Finance Shari’ah Standards QA: Sources and Processing

**OCR Quality Evaluation.** We developed a dedicated OCR quality evaluation tool to systematically assess recognition accuracy in Arabic legal-financial documents. The tool compares raw machine-extracted text against both the original scanned page and a manually corrected reference, enabling fine-grained verification of OCR fidelity at the page level. For each document, the system pairs a scanned page image (e.g., page\_001.png) with its corresponding OCR output (page\_001.txt) and presents them side by side: the original page image appears in the left panel, while the OCR-generated Arabic text is shown on the right. Annotators inspect these pairs to identify errors such as *نص مفقود* (missing text), *أحرف غير صحيحة* (incorrect characters), *ترتيب كلمات خاطئ* (incorrect word order), and *فقدان التنسيق* (formatting loss). When needed, they correct the OCR output using an editable field while monitoring a live similarity score reflecting the edit distance between the corrected and original text.

In addition to corrections, annotators label common OCR failure modes, including distorted symbols (*رموز خاصة مشوهة*), punctuation errors (*أخطاء*), and inaccurate numerals (*أرقام غير دقيقة*), and may add comments on recurring issues such as confusion between similar characters (e.g., *ب* vs. *ن*) or misinterpreted *التشكيل*. The system computes a quality score via character-level edit distance, mapping similarity to four categories: *Excellent* ( $\geq 95\%$ ), *Good* (80–95%), *Partial* (50–80%), and *Poor* ( $< 50\%$ ). All steps are logged as structured JSON records (text, scores, error types, comments, timestamps), ensuring reproducibility and auditability.

Beyond per-page inspection, the pipeline enables aggregate OCR analysis across document collections, identifying systematic errors, benchmarking quality across diverse Arabic sources, and informing downstream normalization and model refinement. Overall, this human-in-the-loop approach ensures that OCR text used in Arabic financial NLP benchmarks and training is accurate and free from errors that could affect *الاستدلال الشرعي* (jurisprudential reasoning) or *التحليل المالي* (financial analysis). Figure 8 illustrates the annotation interface.

**Prompt Design:** To ensure high-quality OCR extraction, we design a constrained prompt (Figure 9) that enforces verbatim transcription, preservation of diacritics and formatting, and strict exclusion of non-textual artifacts. These constraints are critical for maintaining fidelity in Arabic legal documents, where minor textual variations can alter meaning. Similarly, we construct a controlled prompt for question-answer generation (Figure 10) that restricts outputs to the explicit content of the Shari’ah standards. This design prevents hallucination, preserves juridical precision, and ensures that generated QA pairs remain faithful to the source text.

## B Islamic Fatwa Dataset: Sources and Processing

The purpose of this evaluation is to determine whether an AI-generated multiple-choice question (MCQ) accurately tests the same Islamic jurisprudence concept as the original *فتوى* Q&A pair. The goal is to maintain both pedagogical soundness and factual correctness. A well-formed MCQ must remain conceptually aligned with the original ruling (*الحكم الشرعي*), preserve the main *مفهوم فقهي* without distortion, and use appropriate *مصطلحات فقهية* to reflect the opinion of the original scholar (*المفتي*). Evaluators must ensure that the question targets the central legal issue and does not introduce unrelated details or alter the scenario in a way that changes the ruling. This evaluation is conducted through a structured annotation dashboard (Figure 11) that presents the original fatwā alongside the generated MCQ for systematic validation. The fatwā Q&A pairs used in this evaluation are collected from a diverse set of authoritative online sources (Table 5).

For an MCQ to be marked as *ملائم* (RELEVANT), it must meet four criteria. First, conceptual alignment (*المواءمة المفاهيمية*): the question should test the same core ruling as the source fatwa and remain faithful to its reasoning and conditions. Second, answer accuracy (*دقة الإجابة الصحيحة*): the correct option must match the original answer, be contradiction-free, and use precise legal terms. Third, distractor quality (*جودة الخيارات الخاطئة*): incorrect options should be plausible yet clearly wrong, reflecting common misunderstandings. Finally, question clarity (*وضوح السؤال*) the MCQ must be clearly phrased, grammatically correct in *العربية*, and provide enough context to be answerable without referencing the original text.

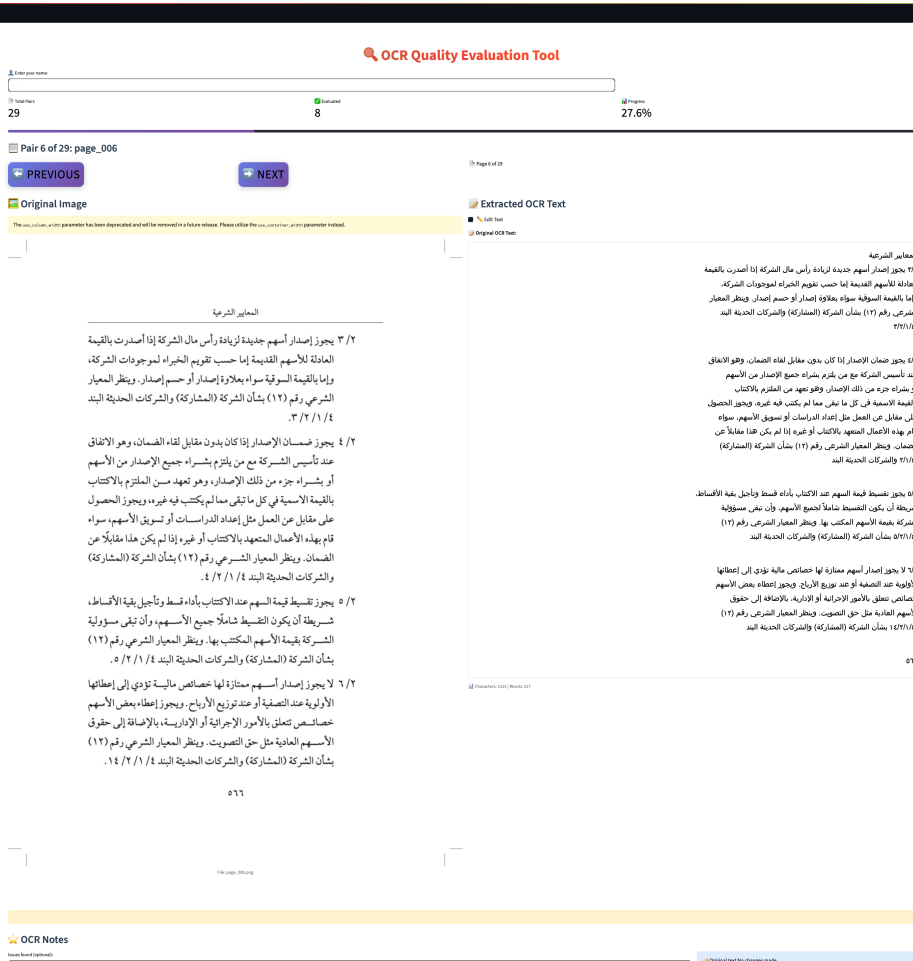


Figure 8: OCR quality evaluation interface for the Shari’ah Standards QA dataset. The tool displays each scanned page from the AAOIFI Shari’ah Standards (left) alongside the OCR-extracted Arabic text (right) to support manual quality verification. Annotators compare the original page with the extracted text, flag recognition errors in diacritics, numerals, and domain-specific terminology, and add corrective notes (bottom). A progress bar tracks annotation completion and overall OCR accuracy.

Website	Link	Country
Dar Al Ifta in Saudi Arabia	<a href="https://www.alifta.gov.sa/">https://www.alifta.gov.sa/</a>	Saudi Arabia
Dar Al Ifta in Egypt	<a href="https://www.dar-alifta.org">https://www.dar-alifta.org</a>	Egypt
Dar Al Ifta in Jordan	<a href="https://aliftaa.jo">https://aliftaa.jo</a>	Jordan
Al Shaikh Abdul Aziz Ibn Baz	<a href="https://binbaz.org.sa">https://binbaz.org.sa</a>	Saudi Arabia
Al Shaikh Mohammad Ibn Othaimin	<a href="https://binothaimin.net/site">https://binothaimin.net/site</a>	Saudi Arabia
Al Shaikh Abdul Aziz Al Ashaikh	<a href="https://www.mufti.af.org.sa">https://www.mufti.af.org.sa</a>	Saudi Arabia
Al Shaikh Saleh Al Fwzan	<a href="https://www.alfawzan.af.org.sa">https://www.alfawzan.af.org.sa</a>	Saudi Arabia
Al Shaikh Saleh Bin Humaid	<a href="https://www.ibnhomaid.af.org.sa/">https://www.ibnhomaid.af.org.sa/</a>	Saudi Arabia
Al Shaikh Abdullah Al Manee	<a href="https://al-manee.com">https://al-manee.com</a>	Saudi Arabia
IslamWeb	<a href="https://www.islamweb.com">https://www.islamweb.com</a>	Qatar
FatwaPedia	<a href="https://fatwapedia.com">https://fatwapedia.com</a>	Saudi Arabia
IslamQA	<a href="https://islamqa.info">https://islamqa.info</a>	Syria
IslamOnline	<a href="https://islamonline.net">https://islamonline.net</a>	Qatar

Table 5: Primary online fatwā archives used for collecting Islamic financial question–answer pairs. These official and widely recognized sites span seven Arab countries, providing diverse juristic opinions and real-world financial scenarios. The URLs shown correspond to the original Arabic portals from which data was programmatically scraped and later cleaned for inclusion in the dataset.

### Prompt for Arabic OCR Text Extraction

**Task.** You are an expert Arabic OCR system specialized in legal and financial documents. Given a scanned page image from an official Islamic finance standard, your task is to extract the text *verbatim* in Arabic with maximum fidelity to the original source.

#### Extraction guidelines.

- Preserve the original wording exactly; do *not* paraphrase, summarize, or infer missing content.
- Preserve diacritics (التشكيل) whenever present in the source.
- Preserve all numerals exactly as written; do not normalize or convert number formats.
- Preserve punctuation, headings, lists, and paragraph boundaries as faithfully as possible.
- Do not correct perceived grammatical, typographical, or stylistic issues.
- If text is unclear or partially illegible, extract the most faithful representation without guessing.

#### What to ignore.

- Page numbers, running headers, footers, or decorative elements not part of the main content.
- Marginal artifacts or scanning noise that do not belong to the text.

**Critical rule.** Do *not* add explanations, comments, translations, or annotations. Output Arabic text only.

**Input.** A single scanned page image from the AAOIFI Shari'ah Standards.

**Output format.** Return the extracted Arabic text as plain UTF-8 text, preserving line breaks and paragraph structure.

Figure 9: Prompt for Arabic OCR text extraction with strict verbatim fidelity and formatting preservation.

Conversely, an MCQ should be marked as غير ملائم (NOT RELEVANT) if it fails any major requirement. Conceptual misalignment occurs when the question tests a different topic, oversimplifies a complex juristic issue, or changes critical context such as conditions (شروط) or scenarios. Incorrect answer issues include a keyed option that contradicts the fatwa, multiple potentially correct answers, or misleading explanations. Poor distractor quality arises when wrong options are obviously incorrect, factually wrong about الإسلام, or too ambiguous. Technical problems include grammar errors that affect meaning, vague or incomplete questions, or improper mixing of different مذاهب in a way that confuses the intended ruling.

### Prompt for Shari'ah Standards Question-Answer Generation

**Task.** You are an assistant supporting the creation of evaluation data for Islamic finance. Given a verified excerpt from an official Shari'ah standard, your task is to draft candidate Arabic question-answer pairs that reflect the *explicit ruling stated in the text*.

#### Question generation.

- Formulate a clear, focused question that asks about the ruling, condition, or permissibility described in the excerpt.
- Do not introduce hypothetical scenarios or facts not present in the source text.
- Ensure the question can be answered directly and completely from the provided excerpt.

#### Answer generation.

- Base the answer strictly on the given excerpt; do not add external knowledge.
- Preserve the original legal meaning, mandatory conditions, and stated exceptions.
- Do not simplify, reinterpret, or generalize the ruling beyond what the text explicitly states.
- Use formal Arabic consistent with fiqh al-mu'āmalāt terminology.

#### Restrictions.

- Do not issue personal opinions or normative judgments.
- Do not cite sources outside the provided text.
- Do not omit conditions, constraints, or qualifiers that affect the ruling.

**Input template.** STANDARD\_EXCERPT: {arabic\_text}

#### Output format.

```
{
  "question": "...",
  "answer": "..."}

```

Figure 10: Prompt for generating Arabic QA pairs from Shari'ah standard excerpts with strict fidelity to explicit rulings.

The evaluation process follows a clear four-step workflow. First, read the original Q&A carefully, identify the primary حكم, any شروط or exceptions, and the supporting evidence such as Qur'anic verses or حديث. Second, analyze the generated MCQ to check conceptual consistency, correct answer faithfulness, and plausibility of distractors.

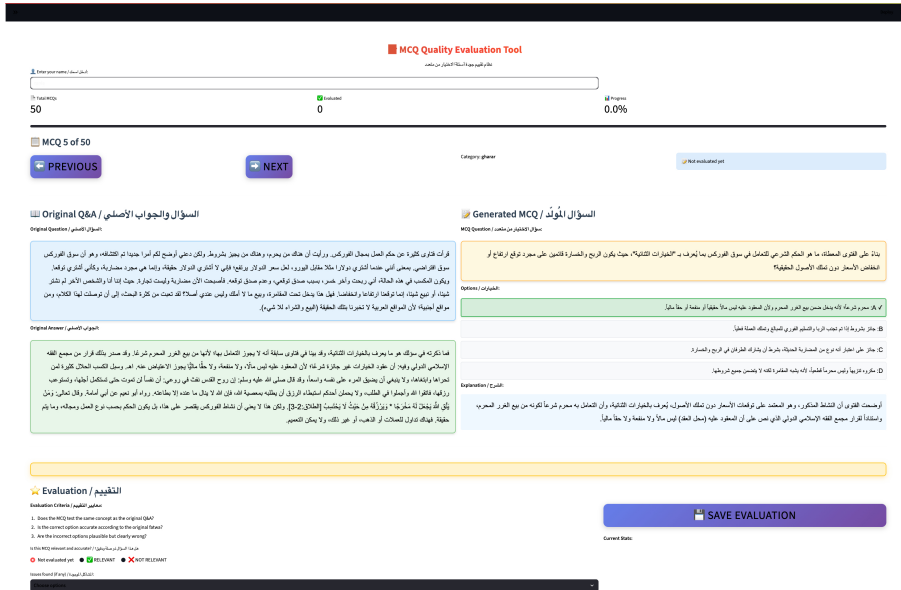


Figure 11: Custom annotation interface used to validate automatically generated multiple-choice questions (MCQs) for the Islamic Finance Fatwa Q&A dataset. The interface displays each original question–answer pair on the left and the corresponding AI-generated MCQ on the right, including the question, answer options, and the automatically selected correct choice. Annotators review conceptual alignment between the MCQ and the original fatwā, verify the correctness and terminology of the marked answer, and assess the plausibility and pedagogical value of distractors. The bottom panel provides structured evaluation criteria and issue tagging to ensure consistent, high-quality validation.

Third, look for red flags such as contradictions, oversimplification, missing qualifiers, or scenario changes. Finally, make a decision: label the MCQ as **ملائم** if it meets all core criteria (minor language or formatting issues may be tolerated) or as **غير ملائم** if any critical issue is present. This structured approach ensures that evaluation is consistent, transparent, and preserves the integrity of Islamic legal reasoning in AI-generated questions.

**Normalization Prompt.** To standardize fatwā question–answer pairs, we design a constrained prompt (Figure 12) that removes non-essential elements such as greetings and formatting artifacts while preserving the original wording and legal intent. This ensures consistency across examples without altering the underlying **حكم شرعي**.

## C Evaluated Models

This appendix briefly documents the rationale behind the selection of models evaluated in Table 2, with model specifications summarized separately in Table 6. The goal is not comparative analysis, but transparency regarding model coverage across language focus, scale, and accessibility.

**Arabic-Focused Models.** We include ALLAM-7B-Instruct (Bari et al., 2024), Fanar-1-9B-Instruct (Abbas et al., 2025), and SILMA-9B-Instruct (silma-ai, 2024) as representative publicly available Arabic-adapted instruction-tuned models. These systems span different base architectures and training strategies, capturing the diversity of current Arabic-centric efforts.

**Open-Source Multilingual Models.** We evaluate Qwen2.5 (Yang et al., 2024), LLaMA-3.1 (Grattafiori et al., 2024), Gemma-2/3 (Riviere et al., 2024; Kamath et al., 2025), and Mixtral-8x7B-Instruct (Jiang et al., 2024) as strong open-weight baselines across scales. These widely used models provide reference points for general multilingual performance on Arabic financial and jurisprudential tasks.

**Proprietary Models.** GPT-5 (OpenAI, 2025), GPT-4o (Hurst et al., 2024), Claude-4.5 variants (Anthropic, 2025b,c,a), and Gemini-3-Flash (Google, 2024) serve as closed-source upper-bound references, contextualizing open and Arabic-focused models against frontier systems without implying direct comparability.

Model	Organization	Size	Source / Notes
<b>Arabic-Focused Models</b>			
ALLAM-7B-Instruct	SDAIA / ALLaM-AI	7B	(Bari et al., 2024)
Fanar-1-9B-Instruct	QCRI	9B	(Abbas et al., 2025)
SILMA-9B-Instruct	SILMA AI	9B	(silma-ai, 2024)
<b>Strong Multilingual / General Open-Source Models</b>			
Qwen2.5-72B-Instruct	Alibaba	72B	(Yang et al., 2024)
LLaMA-3.1-70B-Instruct	Meta	70B	(Grattafiori et al., 2024)
Qwen2.5-14B-Instruct	Alibaba	14B	(Yang et al., 2024)
Qwen2.5-7B-Instruct	Alibaba	7B	(Yang et al., 2024)
Gemma-2-9B-IT	Google	9B	(Riviere et al., 2024)
Gemma-3-27B-IT	Google	27B	(Kamath et al., 2025)
Gemma-3-4B-IT	Google	4B	(Kamath et al., 2025)
LLaMA-3.1-8B-Instruct	Meta	8B	(Grattafiori et al., 2024)
Mixtral-8x7B-Instruct	Mistral AI	8×7B	(Jiang et al., 2024)
<b>Proprietary Models (Upper-Bound References)</b>			
GPT-5	OpenAI	–	(OpenAI, 2025) (API)
GPT-4o	OpenAI	–	(Hurst et al., 2024) (API)
GPT-4o-mini	OpenAI	–	(Hurst et al., 2024) (API)
Claude Opus 4.5	Anthropic	–	(Anthropic, 2025b) (API)
Claude Sonnet 4.5	Anthropic	–	(Anthropic, 2025c) (API)
Claude Haiku 4.5	Anthropic	–	(Anthropic, 2025a) (API)
Gemini-3-Flash (preview)	Google DeepMind	–	(Google, 2024) (API)

Table 6: Models evaluated in this study, grouped into Arabic-focused models, strong multilingual open-source baselines, and proprietary frontier models used as upper-bound references.

## D Business and Accounting Exam Extraction Prompts

Business and accounting exams in Arabic exhibit heterogeneous layouts, ranging from narrative exercise-based formats to tabular true/false questions, often with inconsistent formatting and varying levels of structural clarity across documents. To reliably extract structured MCQs from these sources, we use two task-specific prompts tailored to the dominant document formats observed in the collected exams: an exercise-based extraction prompt (Figure 18) and a table-oriented extraction prompt (Figure 19), enabling robust handling of both free-form and semi-structured exam content across diverse real-world settings.

## E Financial Sentiment Annotation Guidelines

We annotate Arabic financial reports using a document-level sentiment scheme designed to reflect overall market impact rather than sentence-level polarity. Annotation follows a structured human-in-the-loop workflow supported by a custom web-based interface (Figure 13), with clear, well-defined decision rules (Figure 14) to ensure consistency across Islamic and conventional financial reporting contexts.

## F Arabic Finance Extractive Summarization Annotation Guidelines

We annotate Arabic financial reports for extractive summarization using a structured human-in-the-loop workflow supported by a custom web-based interface (Figure 15) and guided by explicit annotation criteria (Figure 16). This setup standardizes the annotation process, reduces subjectivity, and ensures consistent selection of salient content across annotators. Arabic financial reports also exhibit recurring linguistic and formatting challenges, including specialized terminology, frequent code-switching between Arabic and English, and mixed numeral systems (Table 8), all of which complicate sentence selection. Annotators are therefore instructed to carefully identify and select sentences that preserve key financial facts, numerical values, and regulatory references, while strictly avoiding paraphrasing, abstraction, or omission of critical details to maintain fidelity to the source text.

## G Event–Cause Reasoning Annotation Guidelines

We construct event–cause reasoning instances for Arabic financial reports using a structured annotation framework with explicit quality control procedures (Figure 17) to ensure consistency, accuracy, and reliable causal interpretation.

## Prompt for Fatwā Q&A Normalization

**Task.** You are an expert Arabic copy-editor specializing in Islamic finance Q&A. Given a QUESTION and an ORIGINAL ANSWER, your goal is to produce a concise, self-contained question and answer pair in Arabic by removing only non-essential elements *without paraphrasing or changing the juristic intent*. Do not summarize or rephrase; keep the original wording as much as possible.

**1. Referral flag.** Before editing, set IS\_MAINLY\_REFERRAL:

- "YES" if the answer mainly redirects to another fatwā, link, or reference and does not provide a substantive independent ruling.
- "NO" otherwise.

**2. Clean the question.** Edit minimally while preserving wording and fiqh intent:

- Remove greetings, honorifics, and personal appeals (e.g., *السّلام عليكم سلّمه الله سماحة الشيخ*).
- Remove formal closings (e.g., *وجزاكم أرجو منكم التكرم* (الله خيراً)).
- Remove the scholar's name if it is only a form of address; keep it only if the question explicitly seeks that scholar's specific fatwā or opinion.
- Ensure the final question reads as a natural, standalone query.

**3. Clean the answer.** Edit minimally while preserving wording and reasoning:

- Remove formal openings and closings so the answer starts with substantive content.
- Remove all fatwā numbers, hyperlinks, and navigational phrases, editing surrounding text just enough to remain grammatical.
- Convert Arabic-Indic numerals to Western numerals.
- Remove purely formulaic closings such as *وفتكم الله* and *والله أعلم* when they are not part of practical advice.
- Always preserve Qurʾānic verses and sūrah references, ḥadīth attributions, and citations of scholars and their opinions.

**Global rule.** Always delete *all* fatwa numbers from the cleaned question and cleaned answer.

**Input template.** TITLE: {title}  
QUESTION: {question}  
ORIGINAL ANSWER: {answer}

**Output format.**

```
{
  "IS_MAINLY_REFERRAL": "YES" or "NO",
  "cleaned_question": "...",
  "cleaned_answer": "..."
}
```

Figure 12: Prompt for Arabic fatwā Q&A normalization with minimal editing and preservation of juristic intent.

Category	Total Count
Zakat (زكاة)	4,888
Riba (ربا)	2,454
Murabaha (مراجعة)	1,389
Gharar (غرر)	860
Waqf (وقف)	730
Ijara (إجارة)	571
Maysir (ميسر)	372
Musharaka (مشاركة)	242
Mudharaba (مضاربة)	228
Takaful (تكافل)	187
Sukuk (صكوك)	32
<b>Total records</b>	<b>11,953</b>

Table 7: Distribution of questions across Islamic finance categories in the final dataset.

Issue	Example from the report
Islamic Finance Terminology	“تعزير التمويل المستدام وتطوير الصكوك والسندات”
Code-switching	“تستهدف تعزير التمويل المستدام وتطوير الصكوك والسندات، وزيادة شفافية القطاع... Fitch”
Mixed Numeral Systems	“الخفض حجم الدين بنحو ٢٧ ريال قطري (٧,٤ مليار دولار) في عام ٢٠٢٣” — combines Arabic currency and Western digits

Table 8: Key text difficulties in Arabic financial reports with real examples

## H MCQ Answer Normalization and Scoring

To ensure fair and reproducible evaluation of multiple-choice questions, we normalize model outputs before computing accuracy. Large language models frequently generate free-form responses (e.g., explanations, mixed scripts, or multiple answer mentions) rather than a single option label.

**Normalization procedure.** For each model output, we apply the following steps:

- Normalize Unicode and Arabic script by removing diacritics, collapsing repeated whitespace and punctuation, and mapping Eastern Arabic digits (e.g., ١٢٣٤) to Western digits (1234).
- Extract the first explicit answer mention using a cascade of regular expressions that handle:
  - Latin option labels (e.g., A, B, “Option C”),
  - Arabic option letters (e.g., أ, ب, ج),
  - Spelled-out Arabic forms (e.g., باء),
  - Numeric indices (e.g., 1–4).

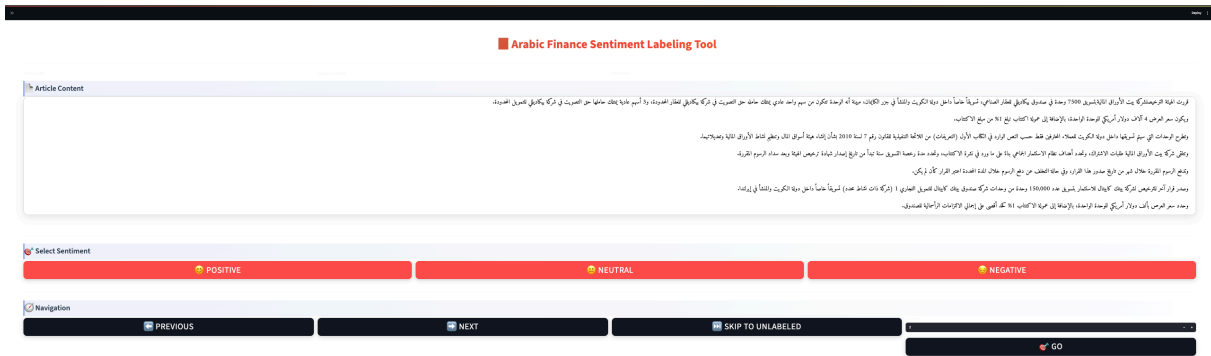


Figure 13: Custom annotation platform used to label Arabic financial reports for sentiment analysis. Annotators reviewed full reports, assigned sentiment classes, and flagged ambiguous cases for expert adjudication.

### Document-Level Financial Sentiment Annotation Guidelines

**Core principle.** Assign a single sentiment label based on the *overall dominant sentiment of the entire report*, not on individual sentences or isolated phrases.

**Annotation procedure.**

- Read the complete document before assigning any label.
- Identify the main financial outcome, thesis, and conclusion.
- Give greater weight to headlines, executive summaries, and concluding sections than to supporting details.

**Handling mixed sentiment.**

- **Dominant sentiment rule:** assign Positive or Negative if one polarity accounts for more than 60% of the salient content.
- **Neutral default:** assign Neutral when positive and negative signals are balanced or when the report is primarily factual.

**Decision criteria.**

- **Positive:** growth announcements, profit increases, successful expansions, or favorable forecasts.
- **Negative:** losses, declining performance, regulatory issues, or adverse outlooks.
- **Neutral:** factual reporting, balanced analysis, or informational updates without a clear directional impact.

**Quality control.** Annotators label reports independently, resolve disagreements during a calibration phase, and refine shared decision criteria. A third domain expert adjudicates remaining conflicts. Each report receives exactly one final sentiment label.

Figure 14: Guidelines for document-level sentiment annotation of Arabic financial reports.

**Scoring.** We compute accuracy as an exact match between the normalized prediction  $\hat{y}$  and the gold label  $y$ . For example, the output “الإجابة هي 2” is normalized to B, while “بسبب صياغة الحكم الخياري” is normalized to C. Outputs that do not contain a valid option after normalization are marked incorrect. This procedure ensures that evaluation is robust to superficial variation in formatting, language mixing, and numeral systems, and that all models are assessed under a consistent and deterministic scoring protocol.

## I Instruction Templates for SAHM Tasks

To enable a unified instruction-tuning and evaluation setup across heterogeneous tasks, we convert each SAHM task into a standardized instruction format. Table 9 lists the canonical task instructions used in our benchmark, shown in their original Arabic formulation alongside an English translation for clarity. The Arabic prompts constitute the actual inputs used during model evaluation, while the English versions are provided solely to document task intent and facilitate reproducibility.

## J LLM-as-a-Judge Protocol, Validation, and Reproducibility

### J.1 Judge Protocol and Reproducibility

We evaluate the three open-ended tasks (Fatwa QA, Shari’ah Standards QA, and Event–Cause QA) using an LLM-as-a-judge setup with Gemini-2.5-Flash. For each instance, the judge receives: (i) the Arabic prompt (including any report/excerpt and question), (ii) the gold reference answer, and (iii) the model’s candidate answer.

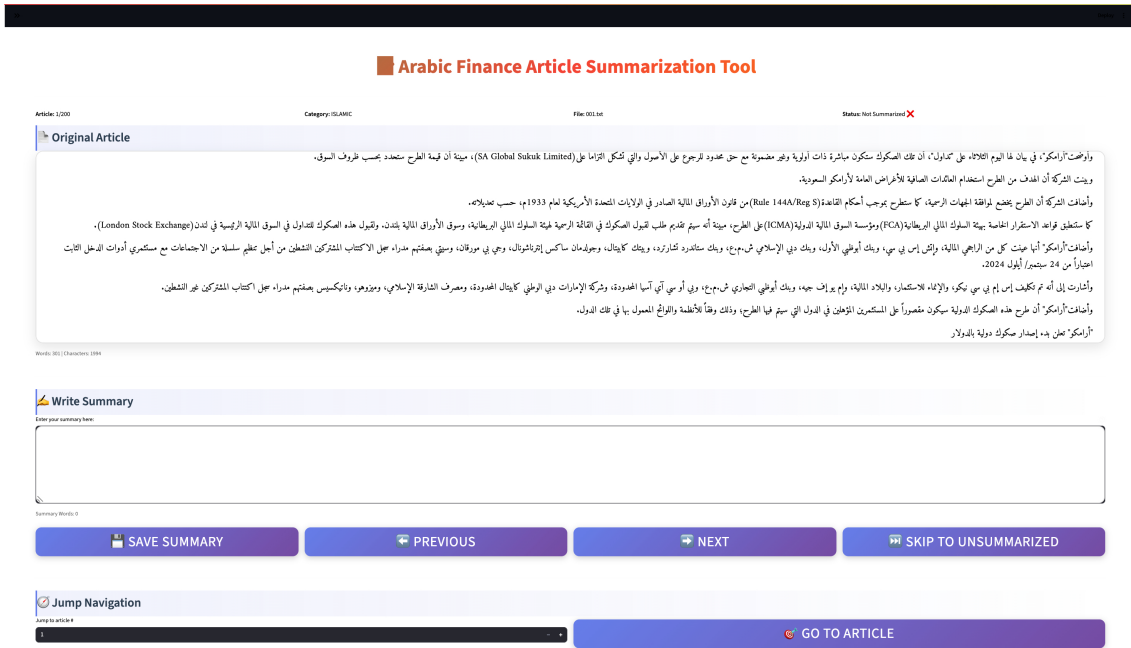


Figure 15: Custom web-based annotation interface for extractive summarization. Annotators view Arabic financial reports, select key sentences containing figures, decisions, and disclosures, and mark them for gold-standard summaries.

The judge is blind to model identity and observes inputs in fixed, labeled fields (prompt, ground\_truth, candidate\_answer) to avoid positional ambiguity. It returns a structured JSON with: (a) rubric sub-scores summing to  $[0, 10]$ , (b) task-specific critical error flags (e.g., contradiction with the reference, omission of critical constraints, normalization of unlawful elements, or fabrication/alteration of figures), and (c) a brief explanation. Task-specific evaluation rubrics are defined for fatwa QA (Figure 20), Islamic finance QA (Figure 21), and financial analysis tasks (Figure 22).

We enforce a strict JSON schema during parsing to ensure all judge outputs are machine-readable and consistently structured. If a response is invalid JSON or violates the schema, we retry once with the same inputs and an explicit *JSON-only* instruction to correct formatting issues without altering content. Persistent failures are marked invalid and excluded from aggregate scores, and we report the invalid-rate as a transparency measure. We run the judge deterministically (temperature = 0.0, greedy decoding, max output tokens = 4096), eliminating randomness.

As a result, we do not perform repeated judging or score averaging. This ensures reproducibility and consistency across all evaluations, yielding stable outputs under fixed inputs and identical evaluation conditions.

## J.2 Human Alignment Study (Judge Validation)

To validate the LLM judge against expert evaluation, we conduct a human alignment study on 200 randomly sampled open-ended outputs spanning Fatwa QA, Shari’ah Standards QA, and Event–Cause QA. Two expert Arabic annotators independently score each model response using the same  $[0, 10]$  additive rubric provided to the judge (Section J). This setup enables direct comparison between human and model-based scoring under identical evaluation criteria.

We compare the judge’s scores (Gemini-2.5-Flash) to the mean human scores, obtaining an MSE of 0.41 and Pearson  $r = 0.92$ , indicating strong alignment. Inter-annotator agreement is high ( $\kappa = 0.84$  computed on discretized integer scores), demonstrating consistent agreement across annotators.

Dataset	Original Arabic Prompt	English Translated Prompt
Islamic Sharia Standards QA	بناءً على معايير وأحكام التمويل الإسلامي والمعاملات المالية <b>Text:</b> السؤال: الشرعية، أجب على السؤال التالي بدقة. الإجابة وفقاً للضوابط الشرعية: {Question}, {Output}	Based on Islamic finance standards and Shari'ah-compliant rulings, answer the following question accurately. <b>Text:</b> Question: {Question}. <b>Answer (Shari'ah-compliant):</b> {Output}
Islamic Fatwa QA	بناءً على أحكام الشريعة الإسلامية والفقہ الإسلامي، أجب على السؤال التالي بطريقة مفصلة ومدعمة بالأدلة عند الإمكان. <b>Text:</b> السؤال: {Output} <b>Answer:</b> {Question}	Based on Islamic jurisprudence (fiqh) and Shari'ah rulings, answer the following question in a detailed manner, supported by evidence when possible. <b>Text:</b> Question: {Question}. <b>Answer:</b> {Output}
Islamic Financial Fatwa MCQ	اقرأ السؤال التالي بعناية واختر الإجابة الصحيحة وفقاً لأحكام الشريعة. <b>Text:</b> السؤال: {Question} الخيارات: لأحكام الشريعة. <b>Answer:</b> أخرج حرف الخيار الصحيح فقط. {Choices}	Read the following question carefully and choose the correct answer according to Shari'ah rulings. <b>Text:</b> Question: {Question}. <b>Choices:</b> {Choices}. <b>Answer:</b> Output only the correct option letter.
Accounting Exams MCQ	اقرأ السؤال التالي بعناية واختر الإجابة الصحيحة. <b>Text:</b> السؤال: {Question} الخيارات: {Choices}. <b>Answer:</b> أخرج حرف الخيار الصحيح فقط.	Read the following question carefully and choose the correct answer. <b>Text:</b> Question: {Question}. <b>Choices:</b> {Choices}. <b>Answer:</b> Output only the correct option letter.
Business Exams MCQ	اقرأ السؤال التالي بعناية واختر الإجابة الصحيحة. <b>Text:</b> السؤال: {Question} الخيارات: {Choices}. <b>Answer:</b> أخرج حرف الخيار الصحيح فقط.	Read the following business/management question carefully and choose the correct answer. <b>Text:</b> Question: {Question}. <b>Choices:</b> {Choices}. <b>Answer:</b> Output only the correct option letter.
Financial Sentiment MCQ	اقرأ بعناية التقرير المالي التالي واختر التصنيف الصحيح من منظور المستثمر. <b>Text:</b> التقرير: {Input} (إيجابي / سلبى / محايد). <b>Answer:</b> {Input}	Read the following financial report carefully and choose the correct label from an investor's perspective. <b>Text:</b> Report: {Input}. <b>Answer:</b> (Positive / Negative / Neutral).
Report Summarization	قم بتلخيص التقرير المالي التالي باستخدام التلخيص الاستخراجي (Extractive Summarization). اختر الجمل الأكثر أهمية مباشرة من النص الأصلي دون تعديل أو إعادة صياغة، ورتبها اجعل الملخص حوالي 30-40% من حجم بنفَس تسلسلهما. <b>Text:</b> النص، وركز على الأرقام والقرارات والنتائج والتواريخ. <b>Answer:</b> أخرج الملخص فقط دون أي التقرير: {Input}. شرح.	Summarize the following financial report using extractive summarization (select sentences verbatim, keep original order, target 30-40% length, focus on numbers/decisions/outcomes/dates). <b>Text:</b> Report: {Input}. <b>Answer:</b> Output the extractive summary only (no extra text).
Event-Cause Reasoning QA	بناءً على التقرير المالي التالي، أجب على السؤال التحليلي بشكل مفصل ودقيق مع الالتزام بالمعلومات الواردة في النص فقط. <b>Text:</b> التقرير المالي: {Input} السؤال: {Question}. <b>Answer:</b> {Output}	Based on the following financial report, answer the analytical question in a detailed and accurate way, grounded only in the provided text. <b>Text:</b> Financial report: {Input}. <b>Question:</b> {Question}. <b>Answer:</b> {Output}

Table 9: Instruction templates used for SAHM tasks (Arabic prompts are used in evaluation; English translations document task intent).

These results indicate that the LLM-as-a-judge scores closely track expert human judgments under our rubric, supporting its reliability as an evaluation proxy.

### J.3 Cross-Judge Validation of Open-Ended Evaluations

To address concerns about potential model-family bias in our LLM-as-judge evaluation, we re-ran all open-ended tasks with two independent judges: Gemini-2.5-Flash (primary) and GPT-4o.

Each model was evaluated 3 times under greedy decoding, and we report mean±std across runs for both judges, ensuring robustness and consistency of evaluation results. If our primary judge were favoring Gemini-family models, switching to GPT-4o should lower their scores; instead, they rise (Gemini-3-Flash: Islamic-Std 9.18→9.76, Fatwa 9.17→9.32), the opposite of circular bias. This consistency supports the robustness of our evaluation setup.

## Extractive Summarization Annotation Guidelines

**Task overview.** Annotators create extractive summaries by selecting the most important sentences *verbatim* from each Arabic financial document. The goal is to produce a concise summary that preserves critical financial information and reflects the document's main message.

### Document-level assessment.

- Read the entire document before selecting any sentences.
- Identify the document type (e.g., earnings report, regulatory announcement, market analysis, company news).
- Segment the text into sentences using Arabic punctuation marks (، ، ، .).
- Target a summary length of approximately 30–40% of the original document.

**Critical content to prioritize.** Annotators must include sentences containing:

- **Financial figures:** الأرباح / الخسائر (profits/losses), الإيرادات (revenues), النسب المئوية (percentages).
- **Performance indicators:** نمو / انخفاض (growth/decline), هبوط / ارتفاع (increase/decrease).
- **Strategic decisions:** الاستحواذ (acquisition), الاندماج (merger), التوسع (expansion).
- **Regulatory or official actions:** قرارات الهيئة (authority decisions), الموافقات (approvals), التراخيص (licenses).

**Sentence scoring.** Each sentence is scored on a 1–5 scale:

- **5:** Critical financial data or main announcement.
- **4:** Important context or cause–effect explanation.
- **3:** Supporting detail required for clarity.
- **2:** General market or background information.
- **1:** Redundant or generic statements.

### Selection procedure.

- Select all sentences scored **5**.
- Add sentences scored **4** until the target length is reached.
- Include a sentence scored **3** only if necessary for coherence.

**Final validation.** Before submission, annotators verify that the summary:

- Includes all key financial figures and the main announcement.
- Is coherent and understandable on its own.
- Falls within the 30–40% length target.
- Avoids repetition and generic background content.

### Common errors to avoid.

- Selecting sentences based solely on position in the document.
- Omitting numerical or regulatory information.
- Including repetitive or stylistic filler content.
- Exceeding the target summary length without justification.

Model rankings are preserved across judges: top-tier models (Gemini-3-Flash, Claude-Opus-4.5) and bottom-tier models (SILMA-9B, LLaMA-3.1-8B) remain in the same groupings regardless of judge. Tight confidence intervals ( $\pm 0.02$ – $0.10$ ) across 3 runs confirm reproducibility under greedy decoding. Full per-model, per-judge scores appear in Table 10. These results indicate stable evaluation outcomes across different judge models.

## J.4 Frontier Model Error Analysis

To diagnose why frontier models fail on specific SAHM tasks, we conducted a systematic error analysis of GPT-5 and Gemini-3-Flash across Accounting, Business, and Summarization. Two native Arabic annotators with financial backgrounds jointly reviewed each error, analyzed the reasoning against the gold reference, assigned a root cause, and agreed on a category (full agreement after adjudication). We use a shared taxonomy across both models: **Misunderstanding Concept** (correct setup, wrong principle applied), **Concept Confusion** (conflates two related but distinct concepts), **Hallucination** (generates facts not in the question), **Question Misread** (answers a different question), **Calculation Mistake** (arithmetic error), and **Domain Knowledge Gap** (lacks the terminology entirely).

### J.4.1 GPT-5 on Accounting

Seventy percent of GPT-5 accounting errors stem from domain reasoning failures: **Misunderstanding Concept** (39%) and **Concept Confusion** (31%), while **Calculation Mistakes** account for only 20%. This confirms our Section 5.3 finding that models rarely fail at arithmetic but often fail at choosing the correct computation. Concretely, GPT-5 (i) reaches correct intermediate results but selects the wrong accounting standard, (ii) confuses closely related concepts (e.g., treating a direct relationship as inverse), and (iii) in 15% of errors, reaches the correct answer but fabricates a rule to justify switching to a wrong option.

These patterns indicate the model's weaknesses lie in conceptual grounding rather than computational ability. Errors arise when mapping problem statements to the correct accounting principle or standard, not during numerical execution. This suggests improving domain-specific reasoning and conceptual alignment is more critical than enhancing raw calculation capabilities for such tasks.

Figure 16: Guidelines for extractive summarization annotation of Arabic financial reports.

## Event–Cause Reasoning QA Annotation and Quality Control

**Task objective.** Annotators construct one event–cause reasoning instance per Arabic financial report. Each instance consists of (i) an analytical question that requires causal or interpretive reasoning and (ii) a concise expert-written answer grounded exclusively in the information provided in the report. The task evaluates whether models can explain *why* financial or regulatory events occurred and *what their implications are*, rather than recalling isolated facts.

**Question construction.** The question must:

- Be analytical in nature (e.g., “why did this occur?” or “what does this indicate?”).
- Connect multiple data points from the report (e.g., financial figures, growth rates, market reactions).
- Avoid purely descriptive prompts (e.g., “what was the profit?”).
- Be answerable using only information stated or implied in the report.

**Answer construction.** The answer must:

- Be written in Arabic and remain concise.
- Rely exclusively on the content of the report, without external knowledge or speculation.
- Explicitly reference numerical figures and percentages when available.
- Provide economic or financial interpretation (e.g., performance drivers, risk implications, or market significance).
- Preserve technical and domain-specific terminology.

**Focus areas.** Annotators prioritize questions involving:

- Market trend analysis and its implications.
- Performance comparison between companies or sectors.
- Economic significance of observed data patterns.
- Risk assessment based on reported financial indicators.

**Quality control procedure.** We enforce quality control through a multi-stage human validation workflow:

1. **Pilot annotation.** Two native Arabic financial experts independently annotate a pilot subset of 20 reports, each producing an event–cause question and an analytical answer.
2. **Agreement assessment.** We evaluate agreement at two complementary levels:
  - *Event–cause identification:* measured using Cohen’s  $\kappa$ , assessing consistency in identifying salient events and their causes.
  - *Answer consistency:* measured using ROUGE overlap between independently written answers, used as a consistency check rather than a correctness metric.
3. **Calibration.** Annotators review disagreements from the pilot phase, discuss ambiguous cases (e.g., implicit causality, multi-factor events, overlapping economic drivers), and refine shared annotation criteria. This calibration aligns interpretation standards and reduces annotation drift.
4. **Full annotation.** After calibration, one expert annotates the remaining reports under the agreed guidelines.
5. **Audit and correction.** A senior annotator audits a random sample of completed annotations to verify that each instance:
  - Identifies a plausible event and its cause(s) supported by the report.
  - Includes relevant numerical evidence when available.
  - Provides an analytical explanation rather than a descriptive summary.Annotations that fail these checks are revised or discarded.

**Final dataset format.** Each finalized instance consists of a financial report, one analytical event–cause question, and one expert-written answer. This format supports evaluation using both exact-match and partial-match metrics and enables controlled benchmarking of causal reasoning in Arabic financial text.

Figure 17: Guidelines and quality control workflow for event–cause reasoning annotation in Arabic financial reports.

### J.4.2 GPT-5 on Extractive Summarization

GPT-5 understands report content but fails at task execution, selecting background sentences over key financial figures (42.5% of errors), copying entire reports instead of meeting the 30–40% compression target (16.3%), and introducing content from unrelated reports (11.3%). This explains its poor extractive summarization performance (ROUGE-L: 33.37) despite strong open-ended reasoning: the task rewards verbatim selection discipline, not generative fluency.

### J.4.3 Gemini-3-Flash on Accounting

Concept Confusion (27.8%) and Misunderstanding Concept (19.4%) account for 47% of errors, particularly in auditing standards and foreign currency hedging. Unlike GPT-5, Gemini-3-Flash also exhibits Hallucination (8.3%) and Question Misread (8.3%), while Calculation Mistakes remain rare (5.6%). These patterns indicate broader instability beyond core conceptual errors, reflecting less consistent reasoning behavior overall.

Model	Event-Cause QA		Islamic-Std QA		Fatwa QA	
	Gemini Judge	GPT-4o Judge	Gemini Judge	GPT-4o Judge	Gemini Judge	GPT-4o Judge
Gemini-3-Flash	9.84±0.02	9.97±0.02	9.18±0.03	9.76±0.01	9.17±0.02	9.32±0.02
Claude-Opus-4.5	9.67±0.03	9.32±0.97	8.06±0.05	9.53±0.02	8.79±0.03	9.18±0.02
Claude-Sonnet-4.5	9.32±0.04	9.68±0.04	8.24±0.05	9.28±0.02	7.58±0.03	8.86±0.02
GPT-4o	8.30±0.06	9.50±0.09	6.64±0.08	8.53±0.03	6.50±0.04	8.04±0.01
Gemma-3-27B	8.70±0.05	9.74±0.02	6.15±0.08	8.41±0.03	5.18±0.05	7.35±0.02
Qwen2.5-72B	8.08±0.10	8.45±0.17	5.61±0.10	6.96±0.07	5.37±0.06	6.30±0.01
Fanar-1-9B	7.57±0.10	8.03±0.22	4.94±0.08	6.03±0.03	4.44±0.06	5.15±0.04
Gemma-3-4B	7.39±0.08	9.09±0.08	2.88±0.08	5.32±0.07	2.46±0.06	4.39±0.02
ALLAM-7B	6.87±0.10	7.79±0.16	4.92±0.08	5.95±0.03	4.20±0.05	4.49±0.03
LLaMA-3.1-70B	6.60±0.15	7.15±0.30	3.70±0.10	4.67±0.05	4.74±0.08	2.24±0.02
Mixtral-8x7B	4.53±0.08	5.14±0.09	2.48±0.08	3.43±0.08	1.78±0.06	2.80±0.04
SILMA-9B	1.88±0.20	1.43±0.37	3.33±0.12	2.33±0.03	2.05±0.08	1.61±0.05
LLaMA-3.1-8B	4.90±0.18	2.49±0.17	2.50±0.12	1.85±0.03	1.38±0.08	0.71±0.02
S <sub>AHM</sub> -ALLAM-7B	6.50±0.10	7.02±0.12	6.30±0.02	6.59±0.04	4.24±0.04	4.51±0.03

Table 10: Cross-judge validation on open-ended tasks with Gemini-2.5-Flash and GPT-4o (mean±std over 3 runs). Rankings are preserved, and Gemini scores rise under GPT-4o, indicating no circular bias.

Model	Accounting (%)	Business (%)	Fatwā MCQ (%)	Sentiment (%)
<b>Proprietary</b>				
Claude-Opus-4.5	78.04±2.42	76.14±1.14	91.57±0.33	61.25±2.50
Claude-Sonnet-4.5	77.25±1.20	77.05±1.45	88.83±0.38	66.25±1.25
Gemini-3-Flash	74.65±1.95	75.41±0.95	90.07±0.30	70.00±1.25
GPT-5	63.67±2.27	72.31±1.26	91.15±0.45	62.50±1.25
GPT-4o	59.28±2.07	78.32±0.32	87.50±0.10	61.25±0.00
Gemini-2.5-Flash	55.49±2.13	75.05±0.83	86.02±1.22	58.33±4.39
<b>Open-source ≥70B</b>				
Qwen2.5-72B	63.08±2.70	75.23±0.32	83.63±0.33	64.00±1.25
LLaMA-3.1-70B	49.11±2.79	75.58±1.14	82.90±0.15	51.25±3.31
<b>Open-source &lt;70B</b>				
Gemma-3-27B	53.29±2.16	74.13±0.32	80.67±0.18	64.17±0.72
Gemma-2-9B	46.71±2.74	65.31±3.83	70.43±0.61	54.17±1.44
Qwen2.5-14B	48.49±3.93	64.66±0.83	75.18±0.85	60.83±3.82
Qwen2.5-7B	46.11±2.85	63.02±1.14	69.70±0.28	54.17±1.91
Gemma-3-4B	38.12±2.27	67.58±0.32	61.30±0.18	62.08±1.44
Mixtral-8x7B	31.74±1.04	59.38±0.63	62.32±0.34	58.33±0.72
LLaMA-3.1-8B	38.93±3.28	58.64±4.45	60.35±3.62	52.08±5.77
<b>Arabic Models</b>				
Fanar-1-9B	43.51±2.42	70.13±1.67	74.60±0.35	60.42±2.60
SILMA-9B	49.32±21.73	60.11±6.61	53.85±5.57	25.75±3.75
ALLAM-7B	42.24±3.55	64.75±3.83	72.25±2.83	56.50±2.00

Table 11: MCQ evaluation across 3 runs using each model’s recommended temperature. Values shown as mean±std. Rankings remain consistent across runs, confirming robustness of our main findings to decoding configuration.

#### J.4.4 Gemini-3-Flash on Business

Reasoning Error (39.5%) and Concept Confusion (37.2%) dominate at 77%, concentrated in Strategic Management, Marketing, and Entrepreneurship, which require culturally grounded Arabic business knowledge. Domain Knowledge Gap (9.3%) reflects cases where models lack specialized Arabic business terminology. These patterns highlight the importance of domain-specific knowledge beyond general language understanding.

#### J.4.5 Cross-Model Convergence

The most striking finding is that GPT-5 and Gemini-3-Flash, despite different architectures and training data, share the same dominant failure mode: conceptual confusion between related domain principles (70% for GPT-5, 47–77% for Gemini-3-Flash), with arithmetic errors rare in both (20% and 5.6%). This suggests Arabic financial reasoning is a genuine challenge for frontier models, not an evaluation artifact.

**Prompt for Extracting Arabic Accounting Exam MCQs (Exercise-Based Format)**

**Task.** You are an expert system for extracting Arabic accounting exam questions. The input is a scanned exam page containing exercises that begin with the keyword تمرين (Exercise), followed by a number.

**Extraction instructions.**

- Identify all exercises that begin with تمرين followed by a numeral (e.g., ١ تمرين, ٢ تمرين).
- For each exercise, extract the full contextual text that follows the exercise header.
- Within each exercise, identify all multiple-choice questions (numbered 1, 2, 3, etc.).
- For each MCQ, combine the exercise context with the specific question text to form a complete question.
- Extract all answer choices labeled as أ, ب, ج, and د.
- Identify the correct answer by detecting underlined text in the choices; underlining indicates the correct option.

**Critical rules.**

- Preserve the original Arabic text exactly; do not paraphrase or normalize.
- Extract *all* MCQs appearing under each exercise.
- If no underlined choice is visible, set the correct answer to null.

**Output format.** Return a JSON object with the following structure:

```
{
  "exercises": [
    {
      "exercise_number": "...",
      "exercise_context": "...",
      "questions": [
        {
          "question_number": "...",
          "full_question_text": "...",
          "choices": {
            "a": "...",
            "b": "...",
            "c": "...",
            "d": "..."
          },
          "correct_answer": "...",
          "is_underlined": true
        }
      ]
    }
  ],
  "page_info": {
    "total_exercises": ...,
    "total_questions": ...,
    "language": "Arabic",
    "subject": "Accounting"
  }
}
```

Figure 18: Prompt for extracting MCQs from Arabic accounting exams with exercise-based layouts.

**Prompt for Extracting Arabic Business Exam MCQs (Tabular Format)**

**Task.** You are an expert system for extracting Arabic business and accounting exam questions from scanned images containing tabular layouts.

**Document characteristics.**

- Each row corresponds to one question.
- Questions are numbered using Arabic numerals (e.g., ١٢٤, ١٢٥).
- Answer choices typically include صح (True) and خطأ (False), with optional additional choices.
- The correct answer is highlighted with a yellow background.

**Extraction instructions.**

- Identify all question rows in the table.
- Extract the question number and full Arabic question text for each row.
- Extract all visible answer choices.
- Identify the correct answer by detecting yellow highlighting.
- Label questions as true\_false or multiple\_choice accordingly.

**Critical rules.**

- Preserve Arabic text exactly as written, including diacritics.
- If yellow highlighting is ambiguous or not visible, set has\_yellow\_highlight to false.
- Extract *all* visible questions on the page.

**Output format.** Return a JSON object with the following structure:

```
{
  "questions": [
    {
      "question_number": "...",
      "question_text": "...",
      "question_type": "...",
      "choices": {
        "a": "...",
        "b": "...",
        "c": "..."
      },
      "correct_answer": "...",
      "correct_choice_text": "...",
      "has_yellow_highlight": true,
      "subject_area": "business"
    }
  ],
  "page_info": {
    "total_questions": ...,
    "format": "table_with_yellow_highlighting",
    "language": "Arabic",
    "question_type": "mixed"
  }
}
```

Figure 19: Prompt for extracting MCQs from Arabic business and accounting exams with tabular layouts.

## LLM-as-a-Judge Rubric for Fatwa QA (Arabic)

**Role.** You are an expert evaluator in Islamic fatwa (iftāʾ).

**Inputs (provided each time).**

- category (optional context) – may be empty (e.g., riba, zakat, takaful)
- prompt – the full Arabic prompt shown to the model (instructions + question)
- ground\_truth – the reference fatwa answer (Arabic)
- candidate\_answer – the model answer to evaluate (Arabic)

**Task.** Judge how well candidate\_answer matches ground\_truth in *ruling (ḥukm), justification, and operative constraints/qualifications*. Prioritize doctrinal correctness and required conditions/exceptions. Do not penalize stylistic paraphrase if the core ruling and constraints are preserved. Be concise and deterministic.

**Scoring (sum to exactly 10).**

1. **Coverage of core ruling (0–4).** The candidate must clearly state the same central hukm (e.g., permissibility/prohibition, validity/invalidity) *and* include the key justification present in the ground truth. One-word/minimal answers without essential justification should receive a much lower score (e.g., 0–1).
2. **Conditions, exceptions, constraints (0–2).** Does it retain critical restrictions, qualifiers, or carve-outs that materially affect the ruling?
3. **Doctrinal/factual accuracy (0–2).** No misstatements that would change the fatwa; no implicit legalization of prohibited elements (e.g., ribā); no misleading generalizations or invented requirements.
4. **Clarity & Arabic language quality (0–1).** Clear Arabic, understandable structure, minimal ambiguity appropriate for a fatwa answer.
5. **Directness & fatwa format (0–1).** Directly answers the question; avoids long digressions; phrasing suitable for a fatwa.

**Critical checks (true/false).**

- contradicts\_ground\_truth: Does the candidate contradict the central ruling?
- omits\_critical\_conditions: Does it omit key conditions/exceptions that change the ruling?
- introduces\_unlawful\_elements: Does it introduce/normalize prohibited elements (e.g., ribā)?
- hallucinated\_citations: Misleading/fabricated sources claimed that distort the ruling?
- non\_answer\_or\_evasive: Does it avoid giving a clear ruling?
- off\_topic\_or\_unsafe: Off-topic or otherwise inappropriate?

**Output format (strict).** Output *only* valid JSON (no prose, no code fences), following this schema:

```
{
  "scores": {
    "coverage_core_ruling": <float 0-4>,
    "conditions_exceptions": <float 0-2>,
    "factual_doctrinal_accuracy": <float 0-2>,
    "clarity_language": <float 0-1>,
    "directness_format": <float 0-1>
  },
  "overall": <float 0-10>,
  "critical_checks": {
    "contradicts_ground_truth": <true/false>,
    "omits_critical_conditions": <true/false>,
    "introduces_unlawful_elements": <true/false>,
    "hallucinated_citations": <true/false>,
    "non_answer_or_evasive": <true/false>,
    "off_topic_or_unsafe": <true/false>
  },
  "note": "<short NOTE in {NOTE_LANG}>"
}
```

Figure 20: Evaluation rubric used for LLM-based judgment of fatwa QA responses.

## K Decoding Configuration and Variance Analysis

Our main evaluations use greedy decoding (temperature 0) for reproducibility. To ensure results are not artifacts, since some models do not support temperature 0 and others prefer non-zero settings, we re-ran all MCQ evaluations 3 times using each model’s recommended temperatures.

Table 11 reports mean±std across runs. Rankings remain consistent across runs and temperature settings, confirming robustness to decoding choices. The low variance observed across repeated runs further indicates that performance differences are stable rather than driven by sampling noise. This consistency holds across tasks and different evaluation conditions.

## LLM-as-a-Judge Rubric for Islamic Finance QA (Arabic)

**Role.** You are an expert evaluator in Islamic finance (*Fiqh al-mu'āmalāt*).

**Inputs (provided each time).**

- topic (optional context) – may be empty
- question – in Arabic
- ground\_truth – the reference correct answer (Arabic)
- candidate\_answer – the model answer to evaluate (Arabic)

**Task.** Judge how well candidate\_answer matches ground\_truth in *meaning, ruling, justification, and constraints*. Prioritize doctrinal correctness and completeness of key conditions/exceptions. Do not penalize stylistic paraphrase if the core ruling and constraints are preserved. Be concise and deterministic.

**Scoring (sum to exactly 10).**

1. Coverage of core ruling (0–4).
2. Conditions, exceptions, constraints (0–2).
3. Doctrinal/factual accuracy (0–2).
4. Clarity & Arabic language quality (0–1).
5. Directness & on-topic (0–1).

**Critical checks (true/false).**

- contradicts\_ground\_truth
- omits\_critical\_conditions
- introduces\_unlawful\_elements
- hallucinated\_citations
- non\_answer\_or\_evasive
- off\_topic\_or\_unsafe

**Output format (strict).** Output *only* valid JSON (no prose, no code fences), following this schema:

```
{ "scores": { "coverage_core_ruling": <float 0-4>,
  "conditions_exceptions": <float 0-2>,
  "factual_doctrinal_accuracy": <float 0-2>,
  "clarity_language": <float 0-1>,
  "directness_format": <float 0-1> },
  "overall": <float 0-10>,
  "critical_checks": {
    "contradicts_ground_truth": <true/false>,
    "omits_critical_conditions": <true/false>,
    "introduces_unlawful_elements": <true/false>,
    "hallucinated_citations": <true/false>,
    "non_answer_or_evasive": <true/false>,
    "off_topic_or_unsafe": <true/false>
  },
  "note": "<short NOTE in {NOTE_LANG}>" }
```

Figure 21: Evaluation rubric used for LLM-based judgment of Islamic finance QA responses.

Overall, these results suggest that model comparisons are reliable under both deterministic and stochastic decoding regimes. Notably, models that perform strongly under greedy decoding maintain their relative advantage under higher-temperature settings, indicating that gains are not dependent on sampling variability. This stability supports the robustness of our findings.

Similarly, lower-performing models do not benefit from increased randomness, suggesting errors stem from systematic limitations rather than decoding strategy. This reinforces the validity of our evaluation pipeline across inference settings and shows that task difficulty and domain reasoning, rather than decoding configuration, drive performance differences in our benchmark.

### K.1 Doctrinal Variation in Shari’ah Rulings

Islamic jurisprudence is not monolithic: the four Sunni *madhahib* (Hanafi, Maliki, Shafi’i, Hanbali) and Shia schools may differ on financial questions. This raises a concern: do our reference answers reflect a single position that could penalize legitimate alternatives?

We analyzed this question during dataset construction and report our findings here.

Category	# Samples	Significant Dispute
Sukuk	6	2 (33%)
Takaful	38	10 (26%)
Zakat	792	173 (22%)
Gharar	149	20 (13%)
Ijara	102	12 (12%)
Murabaha	234	20 (9%)
Maysir	64	6 (9%)
Riba	407	32 (8%)
<b>Total</b>		<b>289 (14.4%)</b>

Table 12: Distribution of doctrinal variation across Islamic finance categories. Variation is highest in zakat (differing calculation methods), takaful (modern instrument with evolving rulings), and sukuk (small sample, debated across regulators). For disputed cases, reference answers present valid alternatives, and the rubric accepts any legitimate ruling.

#### (1) Most questions test consensus rulings.

Seventy-four percent of samples involve cross-*madhab* agreement on established Islamic finance principles, the prohibition of *riba* (usury), contract invalidation due to *gharar* (excessive uncertainty), the impermissibility of *maysir* (gambling), rather than narrow inter-school disputes.

#### (2) Evaluation targets the *hukm*, not the evidence path.

Reference fatwas and model outputs naturally vary in cited Qur’anic verses, *hadīth*, *fiqh* sources, and reasoning detail, making exact-match evaluation infeasible. We therefore score at the *hukm* (ruling) level: the rubric evaluates the final ruling and its operative constraints. A model citing different but valid evidence while reaching the correct ruling is not penalized.

## LLM-as-a-Judge Rubric for Financial Analysis & Capital Markets (Arabic)

**Role.** You are an expert evaluator in financial analysis and capital markets.

**Inputs (provided each time).**

- prompt – the full Arabic prompt (report/excerpt + question) shown to the model
- ground\_truth – the reference ideal analytical answer (Arabic)
- candidate\_answer – the model answer to evaluate (Arabic)

**Task.** Judge how well candidate\_answer matches ground\_truth in *conclusions, reasoning, and use of provided figures*. Prioritize factual/quantitative fidelity, correct interpretation of financial concepts (e.g., spreads, yields, coverage, issuance, capital structure, Basel III, supply/demand dynamics), and avoidance of hallucinated data. Do not penalize stylistic paraphrase if core insights and numeric takeaways align with the reference.

**Scoring (sum to exactly 10).**

1. **Core conclusion alignment (0–4).** Does the candidate capture the main thesis and key takeaways of the ground truth (*what/why/so-what*)?
2. **Quantitative fidelity & use of figures (0–2).** Correctly cites/uses the reported numbers (e.g., percentages, amounts, maturities, oversubscription) without inventing or altering figures. Any simple computations/comparisons must be consistent.
3. **Financial reasoning soundness (0–2).** Causality and mechanisms are plausible and consistent with standard finance/econ logic (e.g., pricing vs. credit risk, duration/tenor structure, demand/oversubscription signals, capital adequacy).
4. **Clarity & Arabic language quality (0–1).** Clear Arabic, coherent structure, minimal ambiguity.
5. **Directness & on-topic grounding (0–1).** Answers what was asked; stays anchored in the provided scenario/data (no generic filler).

**Critical checks (true/false).**

- contradicts\_ground\_truth: contradicts the central conclusion of the reference
- fabricates\_or\_alters\_numbers: introduces numbers not present or materially distorts reported figures
- hallucinates\_context\_or\_sources: injects external context/sources not in the prompt that change the assessment
- flawed\_financial\_logic: serious finance/econ reasoning error that would mislead the conclusion
- non\_answer\_or\_evasive: avoids providing an analytical answer
- off\_topic\_or\_unsafe: off-topic or otherwise inappropriate

**Output format (strict).** Output *only* valid JSON (no prose, no code fences). Return JSON strictly in this schema (all fields required):

```
{
  "scores": {
    "coverage_core_conclusion": <float 0-4>,
    "quantitative_fidelity": <float 0-2>,
    "financial_reasoning": <float 0-2>,
    "clarity_language": <float 0-1>,
    "directness_grounding": <float 0-1>
  },
  "overall": <float 0-10>,
  "critical_checks": {
    "contradicts_ground_truth": <true/false>,
    "fabricates_or_alters_numbers": <true/false>,
    "hallucinates_context_or_sources": <true/false>,
    "flawed_financial_logic": <true/false>,
    "non_answer_or_evasive": <true/false>,
    "off_topic_or_unsafe": <true/false>
  },
  "note": "<short NOTE in {NOTE_LANG}>"
}
```

Figure 22: Evaluation rubric used for LLM-based judgment of financial analysis and event–cause reasoning tasks.

**(3) Quantified dispute distribution:** For the 26% of samples with disagreement, Table 12 reports cases where the reference answer flags significant disagreement across recognized *madhahib*. “Significant Dispute” denotes materially different rulings (e.g., permissible vs. impermissible), not just differing evidence.

**(4) Manual error analysis confirms failures are genuine, not doctrinal:** To ensure low scores do not penalize valid alternatives, we analyzed 500 randomly sampled errors (Figure 6). Failures are unambiguous: wrong rulings (25.2%), fabricated evidence (12.1%), and misquoted *ḥadīth* (2.0%) not legitimate alternative positions.