Sahm (سهم): A Benchmark for Arabic Financial and Shari'ah-Compliant Reasoning

Anonymous ACL submission

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 19 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, evaluation framework, and an instruction-tuned model to support trustworthy Arabic financial assistants.

1 Introduction

Arabic is spoken by over 422 million people across 27 countries (Al-Khalifa et al., 2025). This includes the Gulf Cooperation Council (GCC), where sovereign wealth funds (SWFs) and capital markets shape global financial flows. The six GCC states host 14 SWFs that manage about $4.9 trillion USD (Alhajraf, 2025). Recent reports from the IMF highlight their central role in economic diversification and international investment (Korniyenko and Xin, 2025). The region generates large volumes of financial text, including central bank reports, regulatory filings, corporate disclosures, and *fatwas* that provide jurisprudential rulings. Despite this, systematic evaluation of Large Language Models (LLMs) on Arabic financial content remains limited.

In contrast, English financial NLP has advanced rapidly through dedicated benchmarks for sentiment analysis, QA, and numerical reasoning (Maia et al., 2018a; Zhu et al., 2021; Chen et al., 2021, 2022; Zhao et al., 2024; Xie et al., 2025), with multilingual extensions emerging for other languages (Nie et al., 2025; Zhang et al., 2024; Peng et al., 2025a,b). Arabic benchmarks remain limited in scope, ArBanking77 (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. It represents 4-5 trillion USD globally across banking, capital markets (e.g., *sukuk*), and insurance or *takaful* (LSEG, 2024; IFSB, 2025). Unlike conventional finance, it requires Shari'ah review. Financial products must avoid *riba* (interest), *gharar* (excessive uncertainty), and *maysir* (speculation), and they are guided by standards issued by The Accounting and Auditing Organization for Islamic Financial Institutions (AAOIFI).[1] 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. Sahm spans seven expert-verified tasks grounded in AAOIFI standards, fatwa archives from seven countries, and corporate disclosures (Figure 1). Evaluating 19 LLMs reveals that Arabic fluency does not guarantee financial reason-

[1] https://aaoifi.com

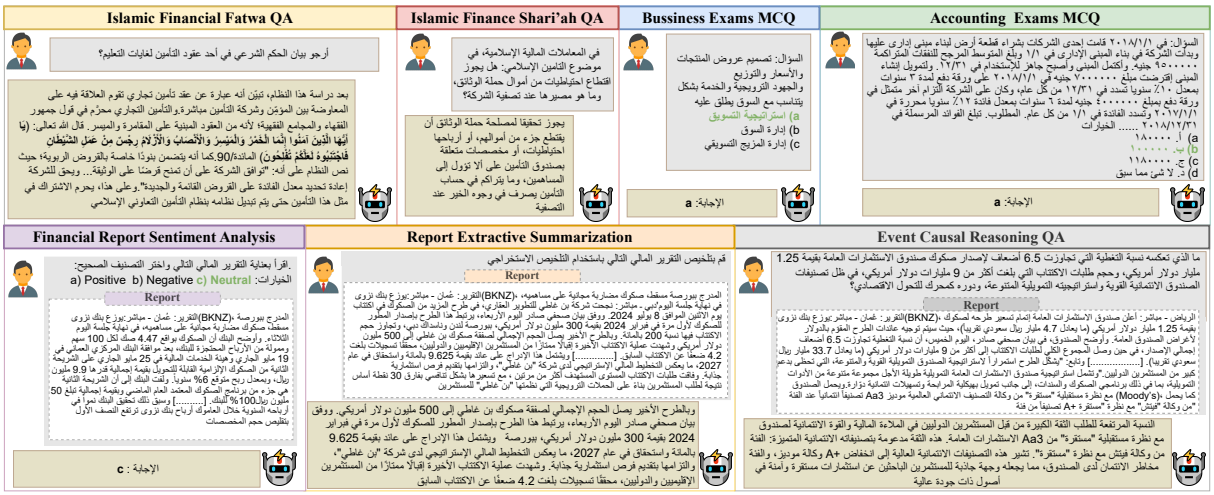

Figure 1: **Examples of the diverse tasks included in Sahm**, covering juristic Q&A, business and accounting MCQs, financial sentiment analysis, report summarization, & event causal reasoning.

ing and targeted instruction tuning can close this gap, with our 7B model surpassing GPT-5 by >20 points. 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.
- A comprehensive benchmark of 19 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 scale alone is insufficient: a targeted 7B instruction-tuned model exceeds GPT-5 by >20 points on accounting and outperforms all open-source baselines (including 72B models), highlighting instruction tuning as a practical route to trustworthy Arabic financial assistants.

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 (Jarrar et al., 2023) addresses banking intent detection; Fatwaset (Alyemny et al., 2023) and HajjFQA (Aleid and Azmi, 2025) target religious QA. These datasets support general language understanding, but do not evaluate the reasoning required for regulatory compliance, numerical analysis, or Shari'ah-aligned decision making. 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.

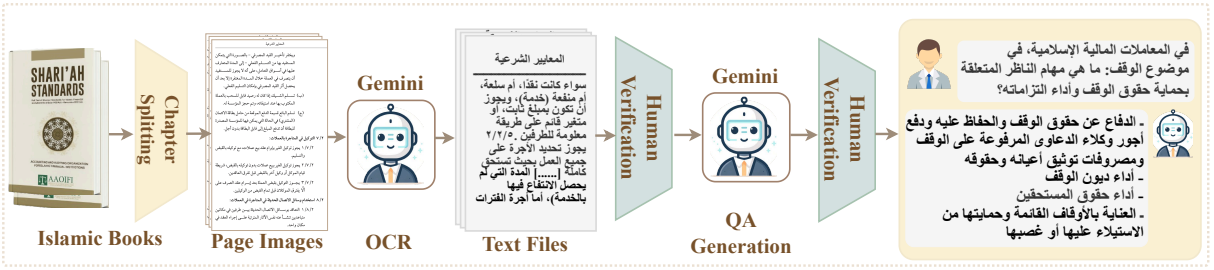

Figure 2: **Pipeline for constructing the Islamic Finance Shari'ah Standards QA dataset.** A hybrid LLM-shuman pipeline converts AAOIFI standards into QA pairs through OCR and generation stages, each followed by expert verification to ensure linguistic accuracy and legal fidelity.

3 Sahm

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[2] issuance (Pomeranz, 1997; Islamic Financial Services Board (IFSB), 2024; Saudi Central Bank, 2024). Yet, most financial benchmarks implicitly assume Western instruments such as interest-bearing loans and conventional bonds, leaving models untested on regionally critical reasoning about contract permissibility, legal constraints, and Shari'ah compliance. To address this gap, we construct the first Islamic Shari'ah Standards QA dataset directly from the official 1,264-page AAOIFI compendium spanning 52 standards chapters, enabling systematic evaluation of LLMs on rule-based Islamic financial reasoning.

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 native Islamic finance experts manually verified the extracted text to preserve diacritics, numerals, and domain-specific terminology.

In a review of a 25% sample, the experts measured a high exact-match rate of 98.7±0.7% with a 95% confidence interval and strong inter-annotator agreement ($\kappa = 0.962$), confirming the reliability of the OCR pipeline. The remaining 1.3% of mismatched characters consisted primarily of minor orthographic or formatting issues (e.g., spacing, punctuation, and occasional diacritics), which annotators corrected in the canonical text used for QA construction; in the audited sample, we did not observe OCR errors that altered the substance of any Shari'ah ruling. After cleanup, we grouped the verified text into thematic clusters, e.g., Murabaḥa (cost-plus sale) and used Gemini-2.5-Pro to draft candidate Arabic question–answer pairs. Domain experts then refined and validated these samples to ensure that each question accurately captured its intended ruling and included all mandatory 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. Examples from the dataset and the full pipeline appear in Figures 1 and 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 4). 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 5), including زكاة (almsgiving), ربا (usury), and مرابحة (cost-plus financing), then transformed these long, formal texts into concise QA pairs via Gemini-2.5-pro while preserving their juristic meaning. 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 pipeline removes greet-

[2] صكوك (sukuk) are Shari'ah-compliant financial certificates representing ownership in underlying assets rather than interest-bearing debt.

Task	Dataset	Processed	Source	Train	Test	Metric	Tested Capabilities
MCQ	Accounting Exams MCQ	416	Exam Questions	249	167	Accuracy	Domain-specific reasoning (Accounting)
	Business Exams MCQ	457	Business Exams	274	183	Accuracy	Business and management reasoning
	Islamic Financial Fatwa MCQ	2,000	Religious QA Bank	–	2,000	Accuracy	Fatwa-based reasoning and comprehension
	Financial Report Sentiment Analysis MCQ	200	Social Media / Reviews	120	80	Accuracy	Sentiment and polarity detection
	Total MCQ	**3,073**	Mixed Sources	**643**	**2,430**	Accuracy	Multi-domain reasoning
Open-Ended	Event–Cause Reasoning QA	200	News / Event Corpora	120	80	Avg. score (/10)	Causal reasoning
	Islamic Fatwa QA	11,953	Religious QA Bank	9,953	2,000	Avg. score (/10)	Faith-based legal reasoning
	Islamic Sharīʿa Standards QA	2,027	Financial Standards	1,216	811	Avg. score (/10)	Standards and legal-domain comprehension
	Report Extractive Summarization	200	Financial Reports / Standards	120	80	ROUGE (R-1/R-2/R-L)	Faithful extraction and content selection
	Total QA	**14,380**	Mixed Sources	**11,409**	**2,971**	Mixed (LLM-judge, ROUGE)	Open-ended Arabic reasoning

Table 1: **Task-level breakdown of the SAHM.** We report dataset provenance, sizes, train-test splits, evaluation metrics, and the financial, legal, and causal reasoning capabilities evaluated across MCQ and open-ended tasks.

ings, honorifics, hyperlinks, and scholar names while preserving Qurʾānic 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 then validated the remaining MCQs under the calibrated guidelines, ensuring that each item precisely matched its source fatwā, preserved juristic terminology, and avoided 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 of dollars in assets, Arabic financial markets lack sentiment benchmarks tailored to region-specific financial discourse. 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 the use of culturally grounded financial terminology, e.g., مرابحة (cost-plus financing) and stylistic variation in Arabic financial reporting, where subtle modifiers can reverse sentiment polarity. To address this gap, we construct the first Arabic financial sentiment benchmark based on authentic market reports rather than translated proxies. We collect 200 Arabic financial reports: 100 Islamic finance–focused and 100 general from Argaam[3], 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 annotation platform (Figure 8) following shared guidelines that emphasize holistic document interpretation rather

[3] https://www.argaam.com/

than sentence-level cues resolving in a high inter-annotator agreement (Cohen's $\kappa = 0.91$). We then conducted a calibration phase where annotators resolved disagreements and refined decision criteria. For mixed-signal reports, we assign the dominant polarity if >60% of content supports it; otherwise Neutral. A third expert adjudicates residual disagreements. We split the dataset 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., المعايير الدولية للتقارير المالية / IFRS and إتش إس بي سي / 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 or regulatory 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. We collect 200 Arabic financial reports: 100 Islamic finance–focused and 100 general from *Argaam*. Two native Arabic financial experts annotate each report by formulating one analytical question that links multiple reported data points and by writing a concise expert answer explaining the underlying causes and implications using only information in the article. A pilot phase on 20 reports ensures guideline clarity. We provide further details in Appendix G. We assess annotation quality at two levels: Cohen's $\kappa = 0.86$ measures agreement on event-cause identification, while ROUGE overlap serves as a consistency check for independently written answers. After calibration to resolve disagreements and align on edge cases, one expert completes the remaining annotations under the agreed criteria.

4 Experiments

Evaluated Models. We evaluated **19** 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 (OpenAI 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 instruction-tuning can close the gap between Arabic-centric and frontier models, we fine-tune **SAHM-7B-Instruct** on the SAHM training split (11.4k instances) using ALLaM-7B as the base model with LoRA (r=64, α=128, lr=2e-4, 3 epochs). Detailed model specifications are provided in Table 7.

Evaluation Measures. 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 original Arabic prompt, gold reference, and model answer, it returns a JSON-validated, additive (sum-of-components) $[0, 10]$ score under a shared rubric assessing alignment with the reference ruling/conclusion, preservation of key constraints or quantitative fidelity, correctness (doctrinal/factual or financial reasoning), Arabic clarity, and directness/grounding. We validate the judge with two expert Arabic annotators on 200 randomly sampled outputs across the three tasks

Model	MCQ (Accuracy % ↑)					Open-Ended QA (Score 0–10 ↑)			
	Datasets					Datasets			
	Accounting	Business	Fatwā	Sentiment	Mean	Event-Cause QA	Islamic-Standards-QA	Fatwa-QA	Mean
Open-source Models: ≥ 70B Parameters									
Qwen2.5-72B-Instruct (Yang et al., 2024)	65.87	74.86	84.65	75.00	75.10	8.1000	5.6330	5.3912	6.3747
LLaMA-3.1-70B (Grattafiori et al., 2024)	52.10	77.60	84.90	80.00	73.65	6.623	3.7245	4.7607	5.036
Open-source Models: < 70B Parameters									
Qwen2.5-14B-Instruct (Yang et al., 2024)	49.10	63.39	76.05	57.50	61.51	7.4975	4.8806	4.0576	5.4786
Qwen2.5-7B-Instruct (Yang et al., 2024)	48.50	59.56	70.00	55.00	58.27	6.1038	3.4039	2.6815	4.0631
Gemma-2-9B-IT (Riviere et al., 2024)	49.10	63.39	66.60	55.00	58.52	7.1438	4.2306	3.4266	4.9336
Gemma-3-27B-IT (Kamath et al., 2025)	53.89	73.22	80.65	80.00	71.94	8.7188	6.1708	5.1929	6.6942
Gemma-3-4B-IT (Kamath et al., 2025)	38.32	67.76	61.35	75.00	60.61	7.4075	2.8985	2.4767	4.2609
LLaMA-3.1-8B (Grattafiori et al., 2024)	41.92	60.66	64.05	73.75	60.60	4.9231	2.5168	1.4025	2.9475
Mixtral-8x7B-Instruct (Jiang et al., 2024)	32.93	60.66	62.15	70.00	56.44	4.5538	2.4980	1.7896	2.9471
Proprietary Models: Reasoning-Enhanced									
GPT-5 (OpenAI, 2025)	65.27	72.68	90.75	78.75	76.86	9.6831	8.7965	8.0515	8.8437
GPT-4o (OpenAI et al., 2024)	60.48	**78.14**	87.70	77.50	75.96	8.3125	6.6598	6.5219	7.1647
Proprietary Models: General-Purpose									
Claude-Opus-4.5 (Anthropic, 2025b)	77.84	76.50	91.75	75.00	80.27	9.6818	8.0438	8.80906	8.8449
Claude-Sonnet-4.5 (Anthropic, 2025c)	78.44	76.50	88.15	77.50	80.15	9.3388	8.2588	7.6049	8.4008
Claude-Haiku-4.5 (Anthropic, 2025a)	67.66	73.77	84.90	77.50	75.96	9.1050	7.0002	6.5341	7.5464
Gemini-3-Flash (preview) (Google, 2024)	76.05	74.86	89.90	81.25	80.52	**9.8369**	**9.1649**	**9.1571**	**9.0798**
GPT-4o-mini (OpenAI et al., 2024)	58.08	77.60	81.75	75.00	73.61	7.9613	5.6094	5.3087	6.2931
Arabic Models									
ALLAM-7B (Bari et al., 2024)	44.91	68.31	74.40	58.75	61.59	6.8875	4.9364	4.2185	5.3475
Fanar-1-9B (Abbas et al., 2025)	47.31	66.12	74.45	58.75	61.66	7.5850	4.9607	4.4600	5.6686
SILMA-9B (silma-ai, 2024)	50.90	69.40	62.55	30.00	53.21	1.8969	3.3547	2.0711	2.4409
SAHM-7B-Instruct (ours)	**71.40**	**93.99**	**74.45**	**61.25**	**75.27**	**6.785**	**6.482**	**4.124**	**5.7970**

Table 2: **Unified leaderboard comparing MCQ tasks (Accuracy %) and open-ended QA tasks (Score 0-10).** Open-ended QA scores are averaged over Event-Cause QA, Islamic-Standards-QA, and Fatwa-QA.

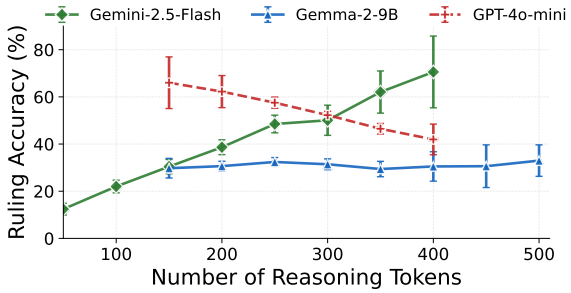

Figure 3: **Effect of reasoning token budget on ruling accuracy.** Green indicates improvement with increased budget, red indicates decline, and blue indicates no change.

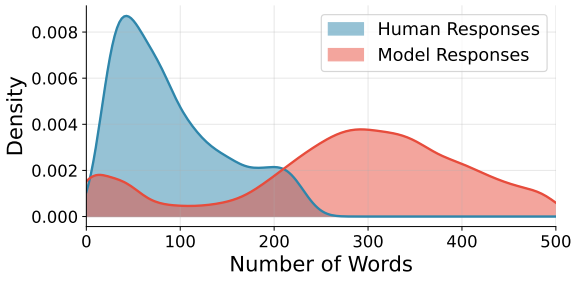

Figure 4: **Models Talk More, Not Better.** Despite models generating 4-6× more fatwas text than human, models do not achieve proportionally higher accuracy, indicating that verbosity serves as proxy for uncertainty rather than expertise.

(MSE 0.41, Pearson r=0.92; inter-annotator agreement κ=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.

5 Results

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?

5.1 Main Results

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 that Arabic domain rea-

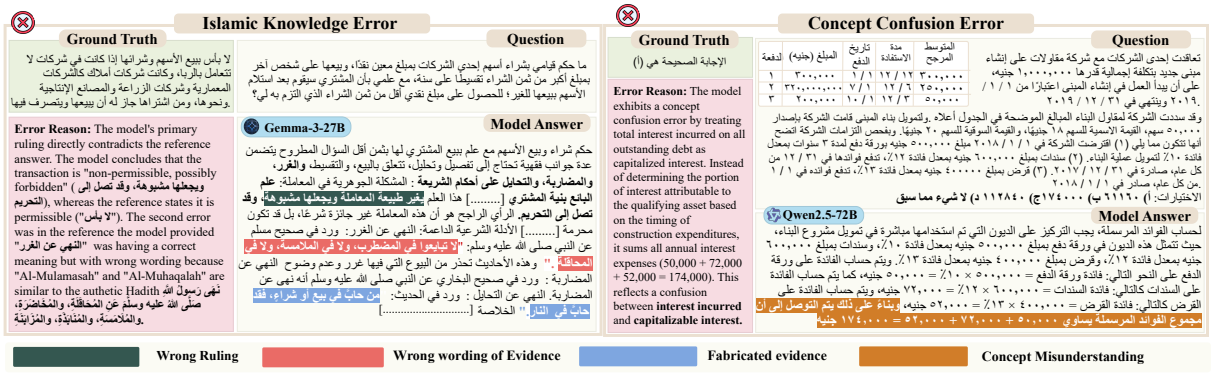

Figure 5: **Qualitative error analysis showing representative failure modes.** *Left:* Islamic knowledge error where Gemma-3-27B incorrectly rules a permissible transaction as forbidden, citing fabricated evidence with wrong wording of authentic Hadith. *Right:* Concept confusion error where Qwen2.5-72B conflates total interest incurred with capitalizable interest in a construction loan scenario.

soning capabilities may constitute an independent axis from general language proficiency, warranting architectural investigation in future work. Notably, Gemini-3-Flash inverts the recognition-generation tradeoff, achieving the highest Open-Ended QA score despite moderate MCQ performance, likely because generative tasks afford extended reasoning chains. This is supported by Figure 3, where the Gemini family shows increased ruling accuracy with larger reasoning token budgets.

Arabic Fluency ≠ Domain Reasoning: Event-Cause QA Exposes the Gap Arabic-centric pretraining provides strong foundations for Islamic jurisprudence tasks, but fails to transfer to financial reasoning (Accounting, Business). Domain-specific instruction-tuning on Sahm closes this gap, improving Accounting accuracy by 26% over the ALLAM-7B base model beating all open source models. Event-Cause QA emerges as the "true IQ test" for Arabic financial reasoning. The spread (1.89-9.84) is the widest in the table, nearly the full scale. Proprietary models cluster tightly at the top (9.1-9.8), then a cliff drops everyone else below 8.7. This is where Arabic-specific models expose their limits as the task requires causal reasoning over Arabic financial text. Language fluency does not imply domain reasoning. The task demands compositional causal inference that neither Arabic pretraining nor raw scale can approximate. Qualitative analysis of failure cases (Figure 5) reveals two distinct error patterns: models exhibit surface-level familiarity with Islamic terminology without grounding in authoritative sources, for instance, Gemma-3-27B incorrectly rules a permissible transaction as forbidden while citing fabricated *ḥadīth* evidence, and they conflate related but distinct financial concepts, as when Qwen2.5-72B confuses total interest incurred with capitalizable interest by summing all expenses rather than computing weighted-average expenditures.

The Recognition-Generation Gap. A model that can identify correct Islamic rulings when presented as options should, in principle, generate coherent fatwās from scratch. Our results challenge this assumption. On Fatwa MCQ, Claude-Opus-4.5 and GPT-5 achieve 91.75% and 90.75% accuracy, respectively. However, their Fatwa QA scores drop to 8.81 and 8.05 out of 10, a gap suggesting that recognition and generation tap fundamentally different competencies. Figure 4 illuminates one mechanism behind this gap. Human fatwās peak at approximately 50 words; model responses peak at 300 words, a 4-6× inflation. Despite this verbosity, models do not achieve proportionally higher scores. We interpret this pattern as *verbosity as uncertainty*: when models lack confident knowledge, they hedge with additional text rather than committing to precise rulings. This finding has practical implications for deployment, response length may serve as a useful signal for answer confidence in Arabic financial QA systems.

5.2 Extractive Summarization

Table 3 reveals a striking inversion: Claude-Sonnet-4.5 achieves the highest ROUGE-L (65.13), while GPT-5 our strongest model on open-ended reasoning collapses to 33.37, underperforming even GPT-4o-mini (64.08). 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...

Model	ROUGE-1	ROUGE-2	ROUGE-L
Proprietary Models – Reasoning-Enhanced			
Claude-Opus-4.5	78.22	63.17	64.14
GPT-5	41.19	33.37	33.37
Claude-Sonnet-4.5	79.86	64.98	65.13
Proprietary Models – General-Purpose			
Claude-Haiku-4.5	79.39	61.40	63.62
GPT-4o-mini	77.79	62.90	64.08
GPT-4o	78.91	63.16	63.71
Gemini-3-Flash	49.36	35.83	43.02
Gemini-2.5-Flash	39.46	27.17	36.81
Open-source Models: $\geq$ 70B parameters			
Gemma-3-27B-IT	79.25	63.57	63.42
Qwen2.5-72B-Instruct	40.52	29.50	34.04
Meta-LLaMA-3.1-70B	39.64	31.40	32.65
Open-source Models: $<$ 70B parameters			
Qwen2.5-14B-Instruct	44.42	30.90	35.82
Gemma-3-4B-IT	76.52	62.06	60.93
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
Arabic Models			
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
SAHM-7B-Instruct (ours)	67.48	53.21	57.79

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

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 60.93, rivaling Claude-Opus-4.5 (64.14) with a fraction of the parameters, suggesting extraction benefits from constrained generation rather than extended reasoning. For Arabic-centric models, domain-specific tuning proves decisive: SAHM-7B-Instruct attains 57.79, outperforming ALLaM-7B by **+29.55 points**, demonstrating that Arabic pretraining alone does not confer financial extraction competence targeted domain adaptation does.

5.3 Error Analysis

To understand *why* models fail, we analyze 500 randomly sampled incorrect responses, categorizing failures by root cause.

Error Breakdown. Figure 6 reveals that two error types dominate: *Misunderstanding Concept* and *Wrong Ruling*, together accounting for 58.5% of all failures. Fabricated Evidence (11.4%) and Hallucination (9.3%) follow. Notably, calculation mistakes contribute only 0.3%, models rarely fail at arithmetic but frequently fail at *knowing which arithmetic to perform*.

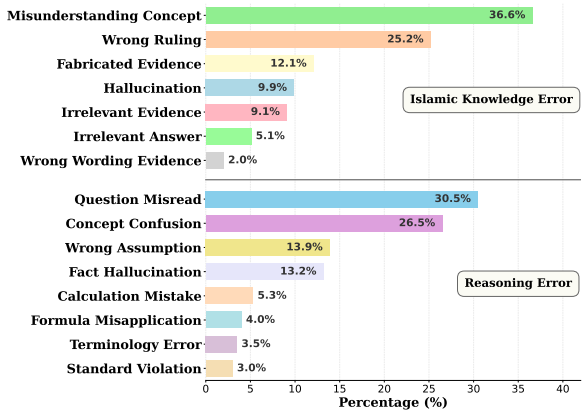

Figure 6: Root cause distribution of model errors across Islamic knowledge and reasoning tasks.

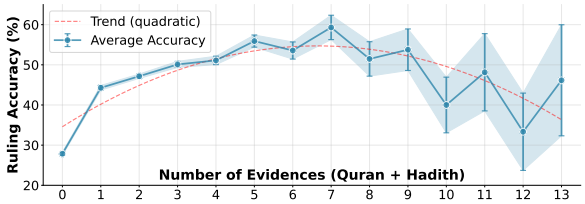

Figure 7: Effect of number of evidences from Hadith and Quran on Ruling Accuracy.

Effect on Evidence Count on Accuracy. Figure 7 examines whether the presence of scriptural evidence (Qur'ānic 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

We introduced SAHM, the first Arabic financial NLP benchmark integrating modern finance and Shari'ah-compliant reasoning across seven tasks. Evaluating 19 LLMs reveals Arabic fluency does not imply financial reasoning, and our 7B model surpasses GPT-5 by $>$20 points—demonstrating targeted domain adaptation outperforms scale. We release all resources to support trustworthy Arabic financial assistants.

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 online 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.

Future Work

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.

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 Fatehkia, Anastasios Fragkopoulos, Maram Hasanain, Majd Hawasly, Mus'ab Husaini, Soon-Gyo Jung, Ji Kim Lucas, Walid Magdy, Safa Messaoud, Abubakr Mohamed, Tasnim Mohiuddin, Basel Mousi, Hamdy Mubarak, Ahmad Musleh, Zan Naeem, Mourad Ouzzani, Dorde Popovic, Amin Sadeghi, Husrev Taha Sencar, Mohammed Shinoy, Omar Sinan, Yifan Zhang, Ahmed Ali, Yassine El Kheir, Xiaosong Ma, and Chaoyi Ruan. 2025. Fanar: An Arabic-centric multimodal generative AI platform. *ArXiv preprint*, abs/2501.13944.

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

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 AlRashed, Faisal A. Mirza, Shaykhah Z. Alsubaie, Hassan A. Alahmed, Ghadah Alabduljabbar, Raghad Alkhathran, Yousef Almushayqih, Raneem Alnajim, Salman Alsubaihi, Maryam Al Mansour, Majed Alrubaian, Ali Alammari, Zaki Alawami, Abdulmohsen Al-Thubaity, Ahmed Abdelali, Jeril Kuriakose, Abdalghani Abujabal, Nora Al-Twairesh, Areeb Alowisheq, and Haidar Khan. 2024. ALLaM: 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.

IFSB. 2025. Islamic Financial Services Industry Stability Report 2025: Navigating shallow waters: Addressing structural vulnerabilities and shoring up resilience to global shocks. Technical report, Islamic Financial Services Board, Kuala Lumpur, Malaysia.

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 dialectical 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, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean-bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, Gaël Liu, Francesco Visin, Kathleen Kenealy, Lucas Beyer, Xiaohai Zhai, Anton Tsitsulin, et al. 2025. Gemma 3 technical report. *ArXiv preprint*, abs/2503.19786.

Yevgeniya Korniyenko and Weining Xin. 2025. Gulf Cooperation Council Diversification: The role of foreign investments and sovereign wealth funds. *IMF Working Papers*, 2025(174).

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.

LSEG. 2024. ICD – LSEG Islamic finance development report 2024: From niche to norm. Technical report, Islamic Corporation for the Development of the Private Sector (ICD) and LSEG Data & Analytics. Accessed: 2025-09-14.

Macedo Maia, Siegfried Handschuh, André 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, André 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, Wei Wang, Haoyu and= 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, :, 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.

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, Efstathia 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*, NIPS '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, Jiaxi 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. FinanceMATH: 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 direct corrections, annotators label common OCR failure modes, including distorted symbols (رموز خاصة مشوهة), punctuation errors (أخطاء علامات الترقيم), and inaccurate numerals (أرقام غير دقيقة), and may add targeted comments on recurring issues such as confusion between visually similar Arabic characters (e.g., ب vs. ن) or misinterpretation of التشكيل (diacritics).

The system automatically computes a quantitative quality score using character-level edit distance and maps similarity percentages to four interpretable categories: *Excellent* ($\geq$95%), *Good* (80–95%), *Partial* (50–80%), and *Poor* ($<$50%). All evaluation steps are logged as structured JSON records, including original and corrected text, similarity scores, identified error types, annotator comments, and timestamps, supporting reproducibility and auditability.

Beyond per-page inspection, the pipeline enables aggregate analysis of OCR performance across document collections, allowing researchers to identify systematic error patterns, benchmark OCR quality across heterogeneous Arabic sources, and inform downstream normalization and model refinement. Overall, this human-in-the-loop methodology ensures that OCR text used downstream such as in Arabic financial NLP bench-

marks and model training is verifiably accurate and free from recognition errors that could affect الاستدلال الشرعي (jurisprudential reasoning) or التحليل المالي (financial analysis). Figure 10 illustrates the OCR quality evaluation interface used during annotation.

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. Follow these rules strictly:

- 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.

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-muᶜāmalāt 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. Return a JSON object:

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

B Islamic Fatwa Dataset: Sources and Processing

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., السلام عليكم سلمه الله سماحة الشيخ).

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

Table 4: 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.

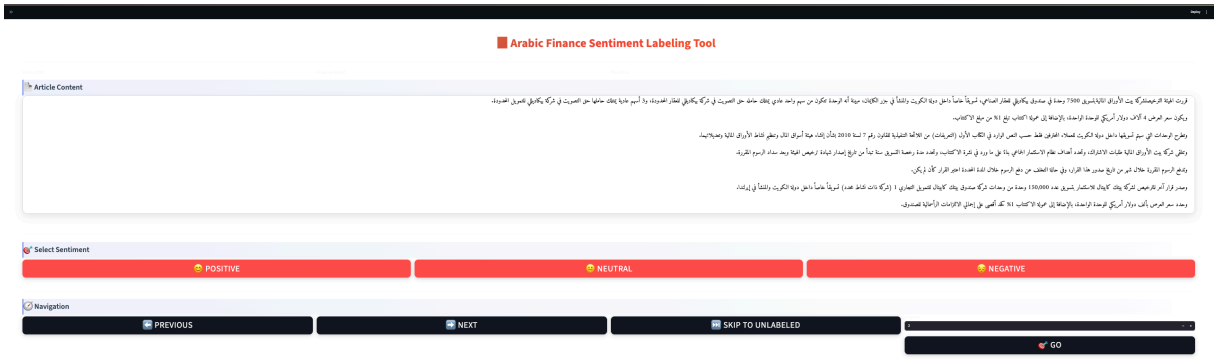

Figure 8: Custom annotation platform used to label Arabic financial reports for sentiment analysis. Annotators reviewed full reports, assigned sentiment classes, and flagged ambiguous cases for expert adjudication.

- Remove formal closings (e.g., أرجو منكم التكرم وجزاكم 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 وفقكم 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. Return a JSON object:

{

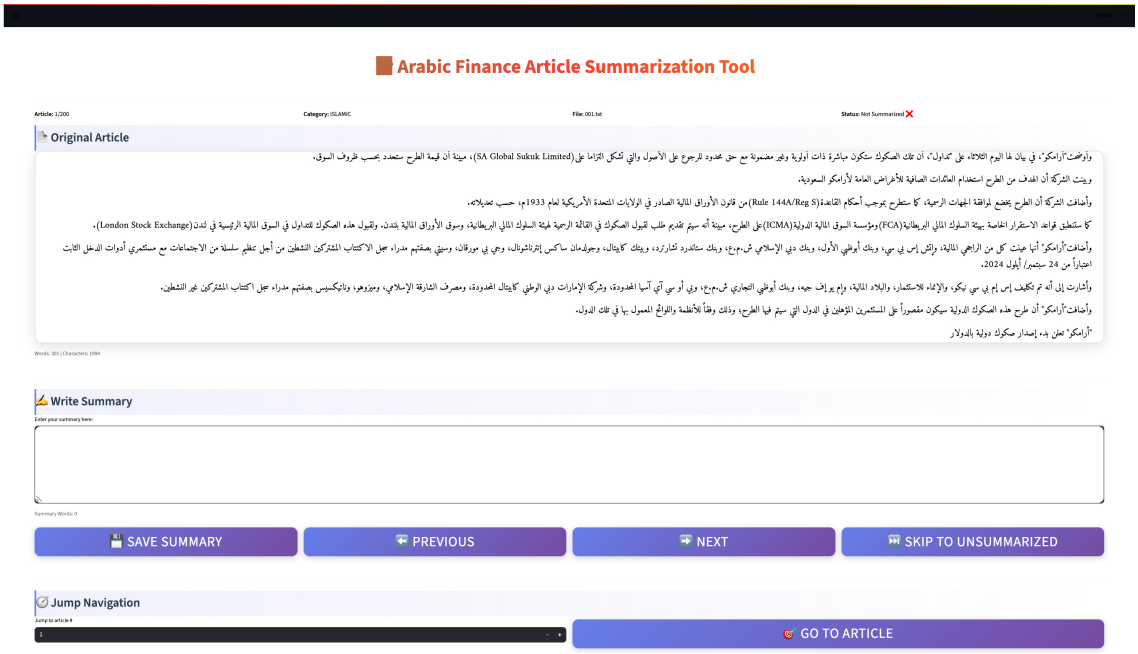

Figure 9: Custom web-based annotation interface for extractive summarization. Annotators view Arabic financial reports, select key sentences containing figures, decisions, and disclosures, and mark them for gold-standard summaries.

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

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 (المُفتي). 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.

For an MCQ to be marked as ملائم (RELEVANT), it must meet four main criteria. First, conceptual alignment (المواءمة المفاهيمية) the question should test the same core ruling as the source fatwa and stay faithful to its reasoning and conditions. Second, correct answer accuracy (دقة الإجابة الصحيحة) the indicated correct option must exactly match the original answer, remain free of contradictions, and use precise Islamic legal terms. Third, distractor quality (جودة الخيارات الخاطئة) incorrect options should be plausible but clearly wrong according to the fatwa, reflecting common misunderstandings rather than random or nonsensical answers. 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.

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

Category	Total Count
Zakat (زكاة)	4,888
Riba (ربا)	2,454
Murabaha (مرابحة)	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
Total records	**11,953**

Table 5: Distribution of questions across Islamic finance categories in the final dataset.

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

Table 6: Key text difficulties in Arabic financial reports with real examples

way that confuses the intended ruling.

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, faithfulness of the correct answer, and plausibility of distractors. 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.

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 7. The goal is not comparative analysis, but transparency regarding model coverage across language focus, scale, and accessibility.

Arabic Focused Models: ALLAM-7B-Instruct (Bari et al., 2024), Fanar-1-9B-Instruct (Abbas et al., 2025), and SILMA-9B-Instruct (silma-ai, 2024) were selected to represent publicly available instruction-tuned models explicitly adapted for Arabic. These systems reflect different training strategies and base model lineages, providing coverage of current Arabic-centric development efforts.

Open-Source Multilingual Models. Qwen2.5 models (Yang et al., 2024), LLaMA-3.1 models (Grattafiori et al., 2024), Gemma-2 and Gemma-3 models (Riviere et al., 2024; Kamath et al., 2025), and Mixtral-8x7B-Instruct (Jiang et al., 2024) were included as strong open-weight baselines spanning a wide range of parameter scales. These models are widely used, well-documented, and provide reference points for general-purpose multilingual performance on Arabic financial and jurisprudential tasks.

Proprietary Models. GPT-5 (OpenAI, 2025), GPT-4o (OpenAI et al., 2024), Claude-4.5 variants (Anthropic, 2025b,c,a), and Gemini-3-Flash (Google, 2024) were evaluated as closed-source upper-bound references. Their inclusion enables contextualization of open and Arabic-focused models against contemporary frontier systems, without implying direct comparability or deployment parity.

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. 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.

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., تمرين ١, تمرين ٢).

Model	Organization	Size	Source / Notes
Arabic-Focused Models			
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)
Strong Multilingual / General Open-Source Models			
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)
Proprietary Models (Upper-Bound References)			
GPT-5	OpenAI	–	(OpenAI, 2025) (API)
GPT-4o	OpenAI	–	(OpenAI et al., 2024) (API)
GPT-4o-mini	OpenAI	–	(OpenAI 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 7: Models evaluated in this study, grouped into Arabic-focused models, strong multilingual open-source baselines, and proprietary frontier models used as upper-bound references.

- 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": {
            :" " "...",
            :" " "...",
            :" " "...",
            :" " "..."
          },
          "correct_answer": "...",
          "is_underlined": true
        }
      ]
    }
  ],
  "page_info": {
    "total_exercises": ...,
    "total_questions": ...,
    "language": "Arabic",
    "subject": "Accounting"
  }
}
```


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"
  }
}
```

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 8), with clear decision rules to ensure consistency across Islamic and conventional financial reporting.

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.

F Arabic Finance Extractive Summarization Annotation Guidelines

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**: الخسائر / الأرباح (profits/losses), الإيرادات (revenues), النسب المئوية (percentages).
- **Performance indicators**: انخفاض / نمو (growth/decline), ارتفاع / هبوط (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.

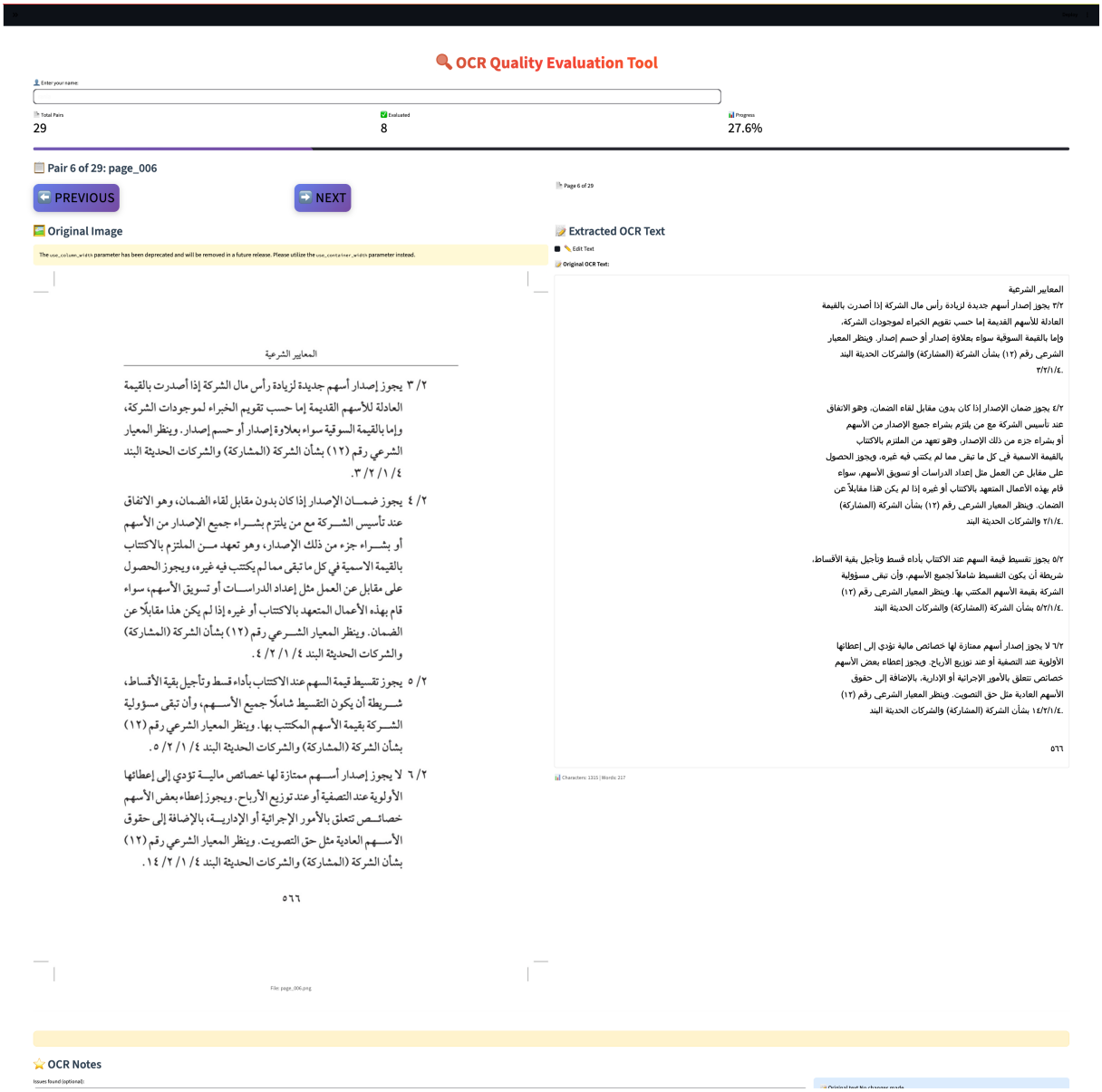

Figure 10: OCR quality evaluation interface for the Shari'ah Standards QA dataset. The tool displays each scanned page from the AAOIFI Shari'ah Standards (left) alongside the OCR-extracted Arabic text (right) to support manual quality verification. Annotators compare the original page with the extracted 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.

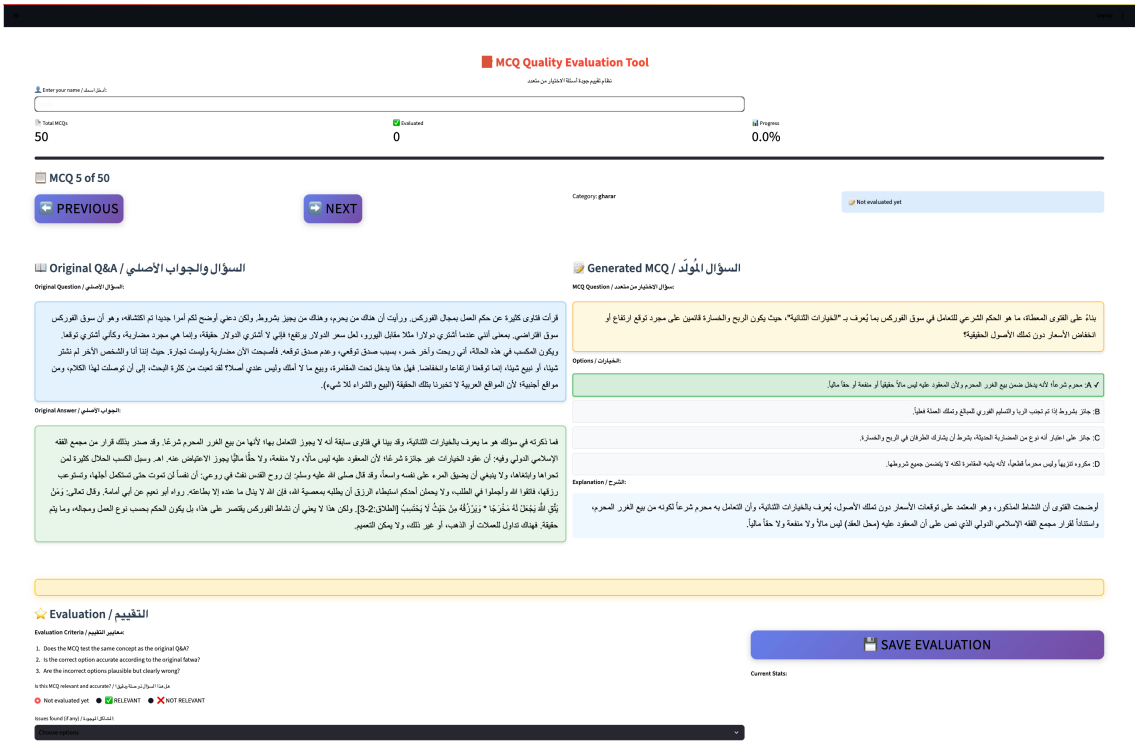

Figure 11: Custom annotation interface used to validate automatically generated multiple-choice questions (MCQs) for the Islamic Finance Fatwa Q&A dataset. The interface displays each original question–answer pair on the left and the corresponding AI-generated MCQ on the right, including the question, 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.

G Event–Cause Reasoning Annotation Guidelines

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 κ, 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.

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).

Scoring. We compute accuracy as an exact match between the normalized prediction $\hat{y}$ and the gold label y. For example, the output "2 الإجابة هي بسبب صياغة الحكم الخيار هو الصحيح (ج)" is normalized to B, while "الخيار (ج)" 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 8 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 exact Arabic prompt shown to the model (including any report/excerpt and question), (ii) the gold reference answer, and (iii) the model's candidate answer. The judge is blind to model identity and always observes the inputs in fixed, explicitly labeled fields (prompt, ground_truth, candidate_answer) to avoid ordering or positional ambiguity. The judge returns a structured JSON object containing: (a) rubric sub-scores whose sum defines an overall score in $[0, 10]$, (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 explanatory note. We enforce a strict JSON schema during parsing. If a response is invalid JSON or violates the schema, we retry once with the same inputs and an explicit *JSON-only* instruction; persistent failures are marked invalid and excluded from aggregate scores (we report the invalid-rate). We run the judge deterministically (temperature $= 0.0$, greedy decoding, max output tokens $= 4096$), and therefore do not perform repeated judging or score averaging. Full judge prompts, rubrics, and task-specific schemas are provided in the following subsections.

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}>"
}
```

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

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

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}>"
}
```

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}>"
}
```

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). We compare the judge's scores (from Gemini-2.5-Flash) to the mean of the two human scores, obtaining an MSE of 0.41 and a Pearson correlation of $r = 0.92$. Inter-annotator agreement is high ($\kappa = 0.84$ computed on discretized integer scores). These results indicate that the LLM-as-a-judge scores closely track expert human judgments under our rubric.

Dataset	Original Arabic Prompt	English Translated Prompt
Islamic Sharia Standards QA	بناءً على معايير وأحكام التمويل الإسلامي والمعاملات المالية الشرعية، أجب على السؤال التالي بدقة. **Text:** السؤال: {Question}. {Output} **الإجابة وفقاً للضوابط الشرعية:**	Based on Islamic finance standards and Shari'ah-compliant rulings, answer the following question accurately. **Text:** Question: {Question}. **Answer (Shari'ah-compliant):** {Output}
Islamic Fatwa QA	بناءً على أحكام الشريعة الإسلامية والفقه الإسلامي، أجب على السؤال التالي بطريقة مفصلة ومدعمة بالأدلة عند الإمكان. **Text:** {Question}. السؤال: {Output} **Answer:**	Based on Islamic jurisprudence (fiqh) and Shari'ah rulings, answer the following question in a detailed manner, supported by evidence when possible. **Text:** Question: {Question}. **Answer:** {Output}
Islamic Financial Fatwa MCQ	اقرأ السؤال التالي بعناية واختر الإجابة الصحيحة وفقاً لأحكام الشريعة. **Text:** {Question}. السؤال: {Choices}. الخيارات: **Answer:** أخرج حرف الخيار الصحيح فقط.	Read the following question carefully and choose the correct answer according to Shari'ah rulings. **Text:** Question: {Question}. Choices: {Choices}. **Answer:** Output only the correct option letter.
Accounting Exams MCQ	**Text:** اقرأ السؤال التالي بعناية واختر الإجابة الصحيحة. {Question}. السؤال: {Choices}. الخيارات: **Answer:** أخرج حرف الخيار الصحيح فقط.	Read the following question carefully and choose the correct answer. **Text:** Question: {Question}. Choices: {Choices}. **Answer:** Output only the correct option letter.
Business Exams MCQ	**Text:** اقرأ السؤال التالي بعناية واختر الإجابة الصحيحة. {Question}. السؤال: {Choices}. الخيارات: **Answer:** أخرج حرف الخيار الصحيح فقط.	Read the following business/management question carefully and choose the correct answer. **Text:** Question: {Question}. Choices: {Choices}. **Answer:** Output only the correct option letter.
Financial Report Sentiment Analysis MCQ	اقرأ بعناية التقرير المالي التالي واختر التصنيف الصحيح من منظور المستثمر. **Text:** {Input}. التقرير: **Answer:** (إيجابي / سلبي / محايد).	Read the following financial report carefully and choose the correct label from an investor's perspective. **Text:** Report: {Input}. **Answer:** (Positive / Negative / Neutral).
Report Extractive Summarization	قم بتلخيص التقرير المالي التالي باستخدام التلخيص الاستخراجي (Extractive Summarization). اختر الجمل الأكثر أهمية مباشرة من النص الأصلي دون تعديل أو إعادة صياغة، ورتّبها بنفس تسلسلها. اجعل الملخص حوالي 30–40% من حجم النص، وركّز على الأرقام والقرارات والنتائج والتواريخ. **Text:** {Input}. التقرير: **Answer:** أخرج الملخص فقط دون أي 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). **Text:** Report: {Input}. **Answer:** 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. **Text:** {Input}. التقرير المالي: {Question}. السؤال: {Output} **Answer:**	Based on the following financial report, answer the analytical question in a detailed and accurate way, grounded only in the provided text. **Text:** Financial report: {Input}. Question: {Question}. **Answer:** {Output}

Table 8: Instruction templates used for SAHM tasks (Arabic prompts are used in evaluation; English translations document task intent).