Failure Modes for Large Language Models in Islamic Legal Reasoning

Muhammad Aurangzeb Ahmad [1]

Abstract

Millions of Muslims consult large language models (LLMs) for Islamic legal guidance, yet the specific failure modes of LLMs in this domain remain undercharacterized. The Islamic jurisprudence *fiqh* is characterized by authenticated transmission chains *isnad*, codified schools of law, formal certainty grades *qaṭ/zann*, and institutionalized scholarly authority. One can map open problems at the intersection of AI and Islamic law into open problems in LLM reliability, calibration, and alignment. We present a six-category failure mode taxonomy grounding each failure at the intersection of a classical jurisprudential concept and a specific ML failure class, derive evaluation criteria for each mode, and propose a research agenda connecting *usul al-fiqh* epistemology to ML methodology. We conclude that *fiqh* is a productive and underexplored test-bed for formal legal AI.

1. Introduction

Hundreds of millions of Muslims worldwide face questions about religious obligation e.g., whether a financial product is permissible, how inheritance should be divided, what constitutes a valid marriage or divorce etc. Many of these Muslims are increasingly turning to large language models (LLMs) for answers. Islam is practiced by approximately a quarter of the global population. Historically, Muslims have turned to religious scholars **ulema** for religious advice on matters of daily life and practice. The proliferation of large language models seems to be partially disrupting the practice of consulting the ulema for religious advice. Opponents of their use argue that multiple evaluation studies have found that no AI chatbot tested to date achieves the standard that Islamic scholarship would recognize as sound (Wahid, 2025). Failures of LLM based chatbots in the Islamic context include fabricated *hadith* attributions, inconsistent legal school-specific rulings, overconfident answers to genuinely contested questions, and the implicit assumption of jurisprudential authority that models are not equipped to hold. These failure modes are not incidental: they reflect a deep mismatch between what LLMs are and what Islamic jurisprudence *requires*.

The contributions of this paper are as follows:. First, we provide ML researchers with a precise account of the epistemological structure of *fiqh* sufficient to understand why standard LLM failure modes manifest in domain-specific ways. Second, we present a six-category **failure mode taxonomy** that formally maps each failure to both a jurisprudential concept from *usul* and a specific ML failure class. Third, we derive evaluation criteria from each failure mode and sketch benchmarks that go beyond multiple-choice accuracy. Fourth, we propose a research agenda connecting classical Islamic epistemological vocabulary to open problems in LLM reliability.

We write for an ML audience. *fiqh* terms are defined when introduced; the paper assumes no prior knowledge of Islamic scholarship but does assume familiarity with LLM architectures, RLHF, retrieval-augmented generation (RAG), and calibration. Crucially, we treat jurisprudential vocabulary as technically precise, not metaphorical:*isnad* is a provenance chain with formal authentication properties;*qatc* and *zann* constitute an epistemic probability hierarchy; *tahqīq al-manāṭ* is an instance-to-predicate classification problem.

This paper focuses on Sunni jurisprudence across the four major schools (Ḥanafī, Mālikī, Shāfi'ī, Ḥanbalī) and draws on Shī'ī sources where benchmark data is available. We do not claim to cover Islamic theology (*'aqīda*). We do not evaluate specific commercial systems bur rather focus on the analysis of documented patterns in published benchmarks. The taxonomy described in this paper is not meant to be exhaustive.

2. Background

In this section we introduce the minimum jurisprudential vocabulary required to contextualize the taxonomy in Section 4. Given the complexity, depth and breadth of the Islamic jurisprudence, it is not possible to cover the complete

*Equal contribution [1]Department of Computer Science, University of Washington Bothell, Bothell, USA. Correspondence to: Muhammad Aurangzeb Ahmad <maahmad@uw.edu>.

Proceedings of the 43^{rd} International Conference on Machine Learning, Seoul, South Korea. PMLR 267, 2026.

scope of this tradition. Here we give a high level overview that would be sufficient for non-experts to understand this tradition.

2.1. Sources of Islamic Law

Islamic law derives from four ranked sources. The **Qur'ān** is primary divine revelation for Muslims and is meant to have supreme authority for Muslims. The **Sunnah** consists of reported words, actions, and tacit approvals of the Prophet Muhammad (pbuh), transmitted through authenticated chains of narrators isnad. After these two **ijma** (scholarly consensus) and **qiyas** (analogical reasoning) provide secondary mechanisms for extending rulings to new cases, are how rulings in Islamic law are derived. This hierarchy determines the certainty grade of any derived ruling, with Qur'ānic injunctions receiving the highest epistemic status and analogical derivations the lowest.

2.2. The Schools of Law and Legitimate Pluralism

In Sunni Islam there are four primary school of jurisprudence. Each of these schools have distinct methodological commitments which may yield different rulings on some topics. These differences are the product of legitimate scholarly disagreement (*ikhtilaf*) on questions where evidence is genuinely equivocal. Mainstream Islamic scholarship regards this diversity as feature and not problematic. Consider the following example: The Ḥanafī school permits a legally mature woman to contract her own marriage without a guardian (*walī*), while the Shāfi'ī school requires a guardian's participation as a validity condition. Both rest on careful textual reasoning; neither is simply wrong. An LLM producing a single answer to "Is a marriage valid without a guardian?" is silently picking a school and will be incorrect for any other school of law.

2.3. Transmission of Knowledge: Isnād

The *isnad* refers to the chain of narrators linking a report to the Prophet. It is the primary instrument for assessing a *hadith*'s authenticity. The science of *'ilm al-rijāl* (narrator criticism) evaluates each person in the transmission link for reliability, memory precision, and moral character. *hadith* are classified by chain strength: *ṣaḥīḥ* (sound), *ḥasan* (good), *ḍa'īf* (weak), and *mawḍu'* (fabricated). Deliberate *hadith* fabrication (*waḍ' al-ḥadīth*) amounts to false attribution and can thus corrupt the entire evidential basis of a ruling built upon it.

2.4. Certainty Grades: Qaṭ' and Ẓann

The Islamic jurisprudence has developed a formal epistemic hierarchy. **qat** (certain knowledge) applies when a ruling is established by *mutawātir* (multiple) transmission or explicit Qur'ānic injunction. **Zann** (probable knowledge) usually applies to rulings from *āḥād* narrations or analogical reasoning. **Wahm** (conjecture) is unreliable as a legal basis. Different legal acts are calibrated to different epistemic thresholds.

2.5. Instance Classification: Tahqīq al-Manāṭ

A critical aspect of legal reasoning is *tahqīq al-manāṭ* i.e., the task of correctly identifying whether a specific situation falls under a general ruling (*ḥukm*). Even a jurist with perfect rule knowledge may issue a wrong ruling by misclassifying the concrete case. Much of the practical difficulty in *fiqh* lies not in determining the abstract rule but in correctly characterizing the factual situation.

2.6. Novel Issues and Ijtihād

Islamic jurisprudence is a living tradition. Each generation of scholars faces novel issues with no direct classical precedent. This usually requires fresh independent reasoning (*ijtihad*). Contemporary examples include cryptocurrency, organ donation, genetic engineering, social media contracts, and questions about AI.

3. Related Work

3.1. Islamic NLP and Benchmarking

Research in Islamic NLP includes the analysis of general Islamic sacred-text corpora to specialized legal benchmarks. Atif et al. (2025) introduce **FiqhQA** (960 QA pairs across four Sunni schools covering acts of worship). Elmahjub et al. (2026) present **IslamicLegalBench** (718 manually curated instances across seven schools, measuring knowledge recall and reasoning depth). Abdelaal et al. (2026) contribute **IslamicMMLU** (MMLU-style MCQs across Qur'ānic studies, *hadith* science, and *fiqh* together with the **QIAS 2025** Shared Task on Islamic inheritance law (*'ilm al-mawārīth*) using 9,446 labelled examples from 32,000 *fatawā*. Lahmar et al. (2025) introduce *IslamTrust*, finding 66.5% alignment between GPT-4 outputs and scholar-annotated consensus Islamic ethics positions. The authors note that this figure falls substantially below the threshold they define as reliable for guidance purposes. A consistent finding is that performance varies dramatically by school and task type. Also, school-specific gaps of 10–20 percentage points are reported on frontier models (Atif et al., 2025). (Ahmad, 2025) gives an overview of the use of LLMs in the Muslim community.

3.2. Legal AI

The application of LLMs to law has received substantial attention. The *Mata v. Avianca* (2023) incident where a lawyer submitting briefs with AI-generated non-existent citations demonstrated that hallucinated legal citations carry

serious real-world consequences. *fiqh* presents an analogous but structurally distinct challenge: the authentication problem. This problem can be framed as not merely "does this case exist?" but "is this narration sound and correctly attributed?". Standard legal AI solutions which mainly employ statutory retrieval, precedent matching etc. are thus poorly suited to *fiqh*'s school-relative, source-authenticated structure.

3.3. LLM Calibration and Hallucination

Xiong et al. (2024) document systematic mismatch between semantic and verbal uncertainty in LLMs. MetaFaith (Liu et al., 2025) improves faithful calibration by up to 61% through metacognitive prompting. Hallucination in high-stakes domains is documented in medicine (Thirunavukarasu et al., 2023) and law. Religious-content hallucination is documented in Wahid (2025) and Mubarak et al. (2025). Mechanism-level analysis connecting jurisprudential structure to specific hallucination types has not previously been done.

3.4. Sycophancy and Authority

Sycophancy i.e., prioritising user agreement over epistemic correctness, is documented by Sharma et al. (2024) and Natan & Tsur (2026). Regressive sycophancy (changing a correct answer under pressure) is particularly dangerous in religious contexts where users may seek validation for preferred rulings. Cahyono & Subramanian (2025) show that sycophancy is amplified in high-uncertainty domains such as law and religious reasoning. The April 2025 GPT-4o update, which OpenAI subsequently rolled back after reporting that the model had become excessively validating of user sentiments at the expense of accuracy, demonstrated this failure at production scale.

4. Failure Mode Taxonomy

In this section we present six failure modes across three tiers reflecting where in the model pipeline each failure originates: (1) *source/data* failures from training distribution properties; (2) *inference/reasoning* failures from the generation process; and (3) *alignment/social* failures from the interaction between reward structure and the user relationship. Table 1 provides an overview summary of this taxonomy.

4.1. FM1: Isnād Fabrication

Jurisprudential grounding. The *isnad* are the primary mechanism by which Islamic scholarship determines whether a reported speech or action may be attributed to divine guidance and hence carry legal weight. Deliberate fabrication (*wad' al-ḥadīth*) is classified as one of the gravest transgressions because a single fabricated report can corrupt the legal rulings of an entire community.

Characterization. LLMs trained on Islamic text corpora absorb *hadith* surface patterns i.e., literary style, thematic content, standard formulaic expressions etc. without encoding the provenance structure that gives those patterns their legal meaning. A model may thus produce text that is plausible as *hadith* but may have no verifiable source, may conflate two distinct narrations, or may upgrade a weak (*ḍa'īf*) narration to sound (*ṣaḥīḥ*). This is factual hallucination.

The closest legal AI analogue is the fabricated case citation (*Mata v. Avianca*, 2023), but the failure is structurally more serious. A non-existent case can be verified by checking a legal database. A fabricated *hadith* requires cross-referencing against the canonical collections (*Kutub al-Sitta*) and evaluating the narrator chain. This is a task requiring specialist knowledge most users lack. Agent-based evaluation of LLM-generated Islamic content (Mushtaq et al., 2025) finds that even when Qur'ānic references are provided, only a fraction are accurate, with no uncertainty expressed for fabricated citations.

Evaluation criteria. Metrics for evaluation should include: (1) *citation existence rate*: does the cited *hadith* exist?; (2) *attribution accuracy*: is the*isnad* grade correctly stated?; (3) *textual fidelity*: does the cited text match the canonical source?; and (4) *false confidence rate*: proportion of fabricated citations delivered without uncertainty hedging. Abstention on unverifiable queries should be scored positively.

Research directions.

- RAG over authenticated *hadith* corpora (Kutub al-Sitta, 700k+ narrations) with *isnad* verification at retrieval time.
- Hallucination detection fine-tuned for Islamic citations (cf. IslamicEval-2025 Shared Task, Mubarak et al. 2025).
- Mandatory abstention when no verifiable source can be grounded, treating non-retrieval as a positive safety signal.

4.2. FM2: Madhhabic Collapse

Jurisprudential grounding. The four Sunni schools diverge substantively on hundreds of legal questions e.g., conditions for valid marriage and divorce, combining prayers in travel, *zakāh*-liable trade goods, the permissibility of specific financial instruments. These divergences are not

Table 1. Failure mode taxonomy overview. Severity reflects difficulty of user-side detection and irreversibility of consequences.

ID	Name	Fiqh concept	ML analogue	Tier	Severity
FM1	*isnad* fabrication	*isnād* (provenance chain)	Factual hallucination / confabulation	Source / data	HIGH
FM2	Madhhabic collapse	*ikhtilaf* (school pluralism)	Majority-class dominance / label bias	Source / data	HIGH
FM3	*qaṭ'/zann* miscalibration	*qaṭ'/ẓann* (certainty grades)	Miscalibration / verbal uncertainty mismatch	Inference	HIGH
FM4	Manāṭ misclassification	*tahqīq al-manāṭ* (situation classification)	Compositional reasoning / multi-hop error	Inference	MED
FM5	Nawāzil extrapolation	*ijtihad*	OOD extrapolation / distribution shift	Inference	MED
FM6	Authority laundering	*iftā'* / *shurūṭ al-muftī*	Sycophancy / RLHF reward hacking	Alignment	HIGH

peripheral. A practising Muslim follows one school and expects rulings consistent with it. An answer conflating school positions is not "neutral'.' It is correct for one school and wrong for all others.

Characterization. Training corpora for Islamic content are heavily skewed toward online sources that over-represent certain schools. Consider IslamQA.org, this website is among the highest-traffic Islamic Q&A sites and hence a large training-data contributor. This website is operated from a Ḥanbalī perspective, creating systematic label bias toward that school. Benchmark results confirm this: GPT-4o achieves 56% accuracy on Ḥanafī questions versus 37% on Mālikī questions (Atif et al., 2025). The 19-point gap is consistent with a training distribution explanation rather than intrinsic question difficulty, though disentangling these factors would require controlled experiments beyond the scope of this paper.

This is a domain-specific instance of a well-documented phenomenon. Santurkar et al. (2023) show that RLHF-tuned models collapse the diversity of human opinion, aligning with a narrow annotator distribution rather than the full range of groups they purport to represent. Sourati et al. (2026) characterize this as a structural *homogenization* rooted in next-token prediction itself: the training objective rewards the most probable continuation, which is by construction the *average* voice, not any particular one. In pluralistic domains like political opinion, cultural identity, religious school of law etc. this average is not a neutral default. It is a specific position that is wrong for everyone outside the statistical majority. In the *fiqh* context the stakes are higher than in opinion polling: the "average" of school positions corresponds to no actual *madhab* and hence to no valid jurisprudential authority.

Research directions.

- Mandatory school identity as a system prompt conditioning variable.
- Multi-task fine-tuning with school-labelled data and contrastive loss ensuring school-divergent questions produce distinct outputs.
- Contrastive pair datasets: same question, divergent correct answers per school, used for both fine-tuning and evaluation.

4.3. FM3: Qaṭ'/Ẓann Miscalibration

Jurisprudential grounding. The *qatczann* distinction is foundational in Islamic legal methodology. The obligation of the five daily prayers and the prohibition of murder have *qaṭ'* status: established by *mutawātir* transmission or explicit Qur'ānic text. The permissibility of a specific contract format or the ruling on combining prayers in unusual circumstances rests on *ẓannī* evidence: probably but not certainly established, with legitimate scholarly disagreement expected. Classical jurists explicitly signalled these distinctions, and competent *iftā'* requires communicating them.

Characterization. RLHF from human preference data systematically conflates helpfulness with decisiveness: a confident, direct answer is rated higher than a hedged answer even in domains where hedging is epistemically correct. This is likely to create a systematic mismatch between semantic uncertainty (the model's internal probability distribution) and verbal uncertainty (expressed confidence in generated text) Xiong et al. (2024).

IslamicLegalBench documents consequences directly: GPT-4o, less likely to abstain on uncertain questions, yielded higher confident-incorrect rates than Gemini, which abstained 90% of the time and erred in only 1% of cases (Elmahjub et al., 2026). This is the *ẓann/qaṭ'* distinction in empirical form. Here, the model fails to distinguish *qaṭ'*-grade questions (where confident answering is appropriate) from *ẓannī* questions (where hedging is required). Sycophancy (Sharma et al., 2024) compounds this: users seeking validation for a preferred ruling can pressure models into

upgrading a correct hedge to a confident ruling.

Evaluation criteria. A calibration benchmark requires questions annotated along two dimensions: (1) epistemic status (*qaṭ'* vs. *ẓann* vs. genuinely contested); and (2) model verbal confidence in the generated response. Calibration error is the mismatch between these dimensions. Abstention quality on genuinely contested questions is a primary metric. Sycophancy probes (adversarial pressure on correct hedges) provide an additional evaluation axis.

Research directions.

- *Qaṭ'/ẓann* classification as an auxiliary prediction task: training models to classify epistemic status of their own answers before generating them.
- Verbal uncertainty intervention: suppressing confident generation when semantic entropy exceeds a threshold (cf. MetaFaith, Liu et al. 2025; Xiong et al. 2024).
- Sycophancy-resistance training using adversarial pressure pairs in religious law domains.

4.4. FM4: Manāṭ Misclassification

Jurisprudential grounding. Legal reasoning requires two separable steps: (1) retrieving the applicable rule (*ḥukm*), and (2) correctly classifying the specific situation (manat) as an instance of that rule. The second step i.e., *tahqīq al-manāṭ*, requires careful factual examination of the concrete case. The jurisprudential literature is explicit that errors of *manat* classification are a primary source of practical error, distinct from errors in rule knowledge.

Characterization. LLMs conflate these two steps. Rule retrieval is handled by parametric memorisation and is relatively reliable for well-represented rules. Situation classification requires contextual fact analysis i.e., gathering and evaluating the specific features of a described case. Standard RAG and chain-of-thought prompting improve rule retrieval but do not address situation classification. A canonical example is the *ribā/murābaḥa* distinction in Islamic finance. *Ribā* (usury) is absolutely prohibited; *murābaḥa* (cost-plus markup) is a permissible deferred-sale structure. The surface features can be similar (money transferred now, more returned later), but the legal difference rests on specific conditions: whether the bank owns the asset before selling it, whether risk genuinely transfers, whether the markup is fixed at contract formation. A model correctly retrieving the prohibition of *ribā* but misclassifying a *murābaḥa* contract produces a wrong ruling despite correct rule knowledge.

Research directions.

- Multi-hop benchmarks with situation-classification steps evaluated independently of rule retrieval.
- Structured fact-gathering templates requiring enumeration and verification of situation-features before applying a rule.
- Contrastive pairs: same rule (*ribā* prohibition), different *manat* classifications (*ribā* vs. *murābaḥa*), requiring different outputs.

4.5. FM5: Nawāzil Extrapolation Failure

Jurisprudential grounding. Every generation of scholars faces *nawazil* which questions that classical jurisprudence did not anticipate. These require *ijtihad* which is fresh, qualified reasoning explicitly flagged as a new derivation rather than settled *ijma*. Contemporary nawazil include cryptocurrency (is it currency? commodity?), organ donation (does preservation of life outweigh prohibitions on body mutilation?), AI-generated content (who owns it?), and the status of AI systems within Islamic law itself.

Characterization. LLMs face a distribution shift problem for *nawazil*: training data does not contain directly relevant examples. Two failure modes follow. First, the model performs implicit analogical reasoning from the closest superficially similar classical ruling, presenting the result as established jurisprudence rather than an untested analogy. Second, the model produces confident rulings on questions where active scholarly debate continues and no *ijma* has emerged. In both cases, the epistemic status of the answer is invisible to the user.

Research directions.

- Retrieval over living *fatwā* corpora (Dār al-Iftā', AMJA, European Fatwa Council) to access current scholarly positions.
- A nawazil-detection classifier flagging queries outside the classical training distribution, triggering mandatory uncertainty disclosure.
- Explicit epistemic status labelling: "scholars actively debate this; no consensus exists; please consult a qualified scholar."

4.6. FM6: Authority Laundering

Jurisprudential grounding. Issuing a *fatwa*, or a legal ruling, is a regulated act with defined preconditions (*shurūṭ al-muftī*): mastery of sources, familiarity with the questioner's circumstances, accountability to a scholarly community, and the capacity to assess when a question exceeds one's competence and must be referred (*tawaqquf*). The

mufti operates within an institutional structure providing accountability. Rulings can be challenged and revised. Thus, a *mufti* issuing rulings beyond competence or without accountability is failing in their scholarly duty.

LLMs, Fatwas and Sycophancy The manner in which standard LLMs have been trained are meant to satisfy users' implicit desire for definitive, helpful answers. In Islamic religious contexts, this directly conflicts with the epistemic requirements of sound *iftā'*. Thus, if a standard LLM is asked a question, it usually produces outputs that perform the social role of a *mufti* i.e., confident, first-person legal rulings without any of the qualifications, accountability, or institutional oversight that role requires. It "launders" absorbed scholarly positions into its own pronouncements, presenting them as its considered judgments rather than attributable to identified scholars or schools.

The problem is that Sycophancy makes this phenomenon worse. Sharma et al. (2024) document that models change correct answers to incorrect ones under user pressure (regressive sycophancy). A model issuing a correctly hedged answer and then, under pressure, upgrading it to a confident ruling commits exactly the failure classical jurisprudence identifies in the incompetent *mufti*: capitulating to social pressure rather than maintaining epistemic requirements. The April 2025 GPT-4o rollback demonstrated this at production scale.

Research directions.

- Role framing: explicit system-prompt positioning as jurisprudential research assistant, not *mufti*; avoiding first-person ruling language.
- Mandatory scholar-referral for personal status law (*aḥwāl shakhṣīyya*): marriage, divorce, inheritance. These are some domains where LLM rulings have permanent life consequences.
- Attribution-required output format: every ruling must cite the school, classical source, or contemporary scholarly body, not presented as the model's own opinion.
- Sycophancy-resistance fine-tuning using adversarial pressure pairs in religious law, with correctness defined relative to source attribution rather than user satisfaction.

5. Cross-Cutting Analysis

5.1. Structural Observations

Three structural patterns run across all six failure modes.

The fiqh vocabulary is technically precise. *isnad* is a provenance chain with formal authentication properties. It is not a metaphor for "reliable source."*qatc* and *zann* constitute a formal epistemic hierarchy with well-defined conditions. It should not be taken as a loose gesture toward "certainty." *Tahqīq al-manāṭ* is an instance-to-predicate classification problem. It is not a vague notion of "context." These concepts encode centuries of accumulated scholarly insight about knowledge quality, and they map onto ML problems at a precision level that should be productive for both communities.

Severity correlates with irreversibility and user detectability. The four high-severity failures (FM1, FM2, FM3, FM6) share two properties: the error is difficult for non-specialist users to detect, and consequences may be permanent: a wrong ruling on divorce, inheritance, or financial permissibility can affect a family for generations. The medium-severity failures (FM4, FM5) involve reasoning gaps a domain expert can identify or that generate surface uncertainty signals.

The canonical mitigation pattern is consistent across all six modes. Every failure mode benefits from the same structural response: move from parametric to retrieval-based knowledge; require explicit source attribution; calibrate against school-conditioned ground truth; build in mandatory abstention or human escalation for queries exceeding reliable scope. This is not accidental. The Islamic legal tradition has developed over fourteen centuries a sophisticated institutional framework for exactly these problems. It includes authenticated sources, qualified interpreters, institutional accountability, explicit epistemic hedging. The appropriate role for an LLM is to support that framework, not to replace it.

5.2. Interactions Between Failure Modes

The six failure modes are not independent. FM1 (*isnad* fabrication) and FM6 (authority laundering) interact: a model fabricating a *hadith* citation and issuing a ruling based on it has compounded both failures. FM3 (miscalibration) and FM6 interact through sycophancy: a correctly calibrated hedge (FM3 averted) can be overridden by user pressure into an overconfident ruling (FM6 triggered). FM2 (madhhabic collapse) and FM4 (*manat* misclassification) can compound: the model applies the wrong school's rule and then misclassifies the situation under that rule, generating a doubly wrong output. These interactions suggest evaluation should include multi-failure test cases designed to probe whether mitigation of one failure mode inadvertently suppresses detection of another.

5.3. The Desacralisation Question

One concern raised in the philosophical literature on AI and Islamic guidance (Imamah & Fardiansyah, 2025) deserves engagement: that applying generative AI to religious guid-

ance inherently desacralises religious knowledge. It does so by reducing it to a "single practical plane of formalisation" and stripping it of its metaphysical foundations. This is a serious concern. FM6 (authority laundering) is precisely the failure mode that arises when a model presents a jurisprudential product without the institutional and epistemic structure that gives it meaning.

However, this concern is not in tension with technical improvement. A model that correctly flags epistemic status, attributes rulings to named scholars and schools, abstains on genuinely contested questions, and explicitly recommends consultation with a qualified *mufti* for consequential matters is not pretending to be a scholar since it is a well-calibrated research assistant. The solution to FM6 is exactly the institutional deference that the desacralisation concern demands.

6. Towards a Comprehensive Evaluation Framework

Existing benchmarks share limitation for our project, we propose six evaluation dimensions derived from the taxonomy.

D1 (FM1): Citation integrity. Every generated *hadith* reference should be cross-referenced against authenticated corpora. Evaluation metrics could include citation existence rate, attribution accuracy, textual fidelity, false confidence rate on fabricated citations.

D2 (FM2): School-stratified accuracy. Accuracy and performance should also evaluated separately within each school of jurisprudence. This can be done via considering cross-school contrastive pairs, checking for same questions with divergent correct answers across schools of thought.

D3 (FM3): Epistemic calibration. Questions annotated by epistemic status (*qaṭ'* vs. *ẓannī* vs. may be genuinely contested). Calibration error could be used as the mismatch between epistemic status and expressed confidence by the model. Quantifying abstention quality and sycophancy Could be employed as additional axes of evaluation.

D4 (FM4): Situation-classification accuracy. Test cases with similar surface features i.e., questions that sound similar but are different. These would have different legal classifications (e.g. *ribā* vs. *murābaḥa*). One could employ multi-hop evaluation and situation classification way to evaluate these models.

D5 (FM5): Novel-issue detection. *nawazil*-annotated queries would be used for evaluation of whether models identify questions as novel. One could also test if the model acknowledge active scholarly debate, and avoid presenting tentative analogical reasoning as settled jurisprudence.

D6 (FM6): Attribution and role discipline. Evaluation of whether rulings are attributed to named scholars or schools. This would fall under role consistency under pressure. One way to measure could be measuring scholar-referral rate for consequential personal status law questions.

Annotator requirements. These dimensions require annotators with genuine jurisprudential training, not simply Arabic speakers.*isnad* authentication requires expertise in *'ilm al-rijāl*; school-stratified evaluation requires annotators trained in at least one school's methodology; epistemic status annotation requires *usul* knowledge. Legitimate scholarly disagreement between qualified annotators should be accommodated as a domain feature, not treated as annotation noise—items where annotators disagree should be explicitly marked and used for FM5 and FM3 evaluation.

7. Research Agenda

The most tractable immediate contribution is **structured retrieval over authenticated Islamic corpora**. The necessary resources are increasingly available digitally (Quran-NLP corpus, 14 Books Ḥadīth Collection, IslamWeb *fatawā* archive). The gap is retrieval quality validated for jurisprudential tasks. Standard embedding retrieval may not preserve the *isnad*-to-text structure that *hadith* authentication requires. Knowledge graph representations of narrator chains may be a promising direction.

School-conditioned generation (FM2) is the most immediately actionable training intervention. Fine-tuning with school identity as a conditioning variable and with contrastive loss ensuring school-divergent questions produce different outputs is technically straightforward. The *qatczann* calibration problem (FM3) connects directly to active research on uncertainty quantification and verbal calibration (Xiong et al., 2024; Liu et al., 2025). The *fiqh* framing provides a domain-specific ground truth for epistemic status that can train and evaluate calibration models: *qaṭ'*-grade questions should receive high-confidence outputs; *ẓannī* or contested questions should receive appropriately hedged outputs. This annotation scheme, annotated by qualified jurists, could serve as a calibration benchmark beyond the religious domain.

Collective AI-assisted *ijtihad* models serving as research assistants to qualified scholars engaging in collaborative legal reasoning on nawazil represents a promising direction for the future. In this scenario, the model contributes breadth (access to extensive classical literature, rapid cross-referencing), while qualified scholars contribute the judgment, epistemic authority, and accountability that *iftā'* requires.

8. Conclusion

Islamic jurisprudence is not an exotic domain for legal AI. It is a high-stakes, high-volume, structurally complex domain practised by a quarter of the global population, and it exposes genuine gaps in LLM calibration, provenance grounding, pluralism handling, and alignment. We have presented a six-category failure mode taxonomy, each mode formally grounded in classical *usul* and mapped to a specific ML failure class. The taxonomy reveals three structural findings: the jurisprudential vocabulary is technically precise and directly productive for ML analysis; severity correlates with irreversibility and user detectability; and the canonical mitigation pattern: retrieval, attribution, calibrated abstention, and institutional deference, is consistent across all six modes. We close with the observation that the Islamic legal tradition has spent fourteen centuries developing institutional solutions to many of the problems that researchers now confront: authenticating sources, handling legitimate disagreement, communicating epistemic uncertainty, and maintaining scholarly accountability.

References

Abdelaal, A., Haffar, M. N. A., Fawzi, M., and Magdy, W. Islamicmmlu: A benchmark for evaluating llms on islamic knowledge. *arXiv preprint arXiv:2603.23750*, 2026.

Ahmad, M. A. Islamic chatbots in the age of large language models. *arXiv preprint arXiv:2601.06092*, 2025.

Atif, F., Askarbekuly, N., Darwish, K., and Choudhury, M. Sacred or synthetic? evaluating llm reliability and abstention for religious questions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pp. 217–226, 2025.

Cahyono, J. A. and Subramanian, S. Can you trust an llm with your life-changing decision? *arXiv preprint arXiv:2507.21132*, 2025.

Elmahjub, E., Qadir, J., Mushtaq, A., Naeem, R., Ghaznavi, I., and Iqbal, W. Islamiclegalbench: Evaluating llms knowledge and reasoning of islamic law across 1,200 years of islamic pluralist legal traditions. *arXiv preprint arXiv:2602.21226*, 2026.

Imamah, Y. N. and Fardiansyah, M. D. Ijtihad artificial intelligence: Prospects and ethics of using artificial intelligence in creating contemporary islamic fatwas. *Journalum Juris Islamicum Contemporaneum*, 1(1):22–41, 2025.

Lahmar, A., Arafat, M. E., Farou, Z., and Mahmud, M. Islamtrust: A benchmark for llms alignment with islamic values. In *5th Muslims in ML Workshop co-located with NeurIPS 2025*, 2025.

Liu, G. K.-M., Yona, G., Caciularu, A., Szpektor, I., Rudner, T. G., and Cohan, A. Metafaith: Faithful natural language uncertainty expression in llms. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 29600–29644, 2025.

Mubarak, H., Malhas, R., Mansour, W., Mohamed, A., Fawzi, M., Hawasly, M., Elsayed, T., Darwish, K. M., and Magdy, W. Islamiceval 2025: The first shared task of capturing llms hallucination in islamic content. In *Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 480–493, 2025.

Mushtaq, A., Naeem, R., Elmahjub, E., Ghaznavi, I., Al-Maliki, S., Abdallah, M., Al-Fuqaha, A., and Qadir, J. Can llms write faithfully? an agent-based evaluation of llm-generated islamic content. *arXiv preprint arXiv:2510.24438*, 2025.

Natan, S. B. and Tsur, O. Not your typical sycophant: The elusive nature of sycophancy in large language models. *arXiv preprint arXiv:2601.15436*, 2026.

Santurkar, S., Durmus, E., Ladhak, F., Lee, C., Liang, P., and Hashimoto, T. Whose opinions do language models reflect? In *International conference on machine learning*, pp. 29971–30004. PMLR, 2023.

Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S., Durmus, E., Hatfield-Dodds, Z., Johnston, S., Kravec, S., et al. Towards understanding sycophancy in language models. In *International Conference on Learning Representations*, volume 2024, pp. 110–144, 2024.

Sourati, Z., Ziabari, A. S., and Dehghani, M. The homogenizing effect of large language models on human expression and thought. *Trends in Cognitive Sciences*, 2026.

Thirunavukarasu, A. J., Ting, D. S. J., Elangovan, K., Gutierrez, L., Tan, T. F., and Ting, D. S. W. Large language models in medicine. *Nature medicine*, 29(8):1930–1940, 2023.

Wahid, S. H. Artificial intelligence in islamic guidance: Assessing chatbot performance and jurisprudential adherence. *Journal of Digital Islamicate Research*, 3(1):33–96, 2025.

Xiong, M., Hu, Z., Lu, X., Li, Y., Fu, J., He, J., and Hooi, B. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *International Conference on Learning Representations*, volume 2024, pp. 23650–23678, 2024.