A Human-Centered Evaluation of AI-Generated Guidance: Integrated Statistical and Machine Learning Analysis with a Risk Framework for High-Stakes Domains

Omar Al-Turki[1], Felwah Alqahtani[2], Eman Alqahtani[3],
Sarah Alswedani[4], Sami Alshmrany[5], Rashid Mehmood[6]*
Faculty of Computer and Information Systems, Islamic University of Madinah, Madinah 42351, Saudi Arabia[1,5,6]
College of Computer Science, King Khalid University, Abha 62521, Saudi Arabia[2,3]
School of Computing and Information, University of Hail, Hail 55473, Saudi Arabia[4]

***Abstract*—The increasing use of large language models (LLMs) in domains requiring interpretation and judgment has raised critical questions about trust, reliability, and accountability, particularly in contexts where decisions carry significant consequences. While prior work has focused primarily on improving system performance, limited attention has been given to how users evaluate and interact with AI-generated guidance in real-world, high-stakes settings. This paper addresses this gap through a large-scale empirical investigation of public perceptions of AI-generated religious guidance in Saudi Arabia. The analysis is based on survey data collected from 572 participants and combines quantitative statistical methods with a machine learning-based pipeline for analyzing open-ended responses. The quantitative component examines patterns in trust, perceived risk, privacy concerns, credibility, and user practices, while the qualitative component employs embedding-based clustering using Bidirectional Encoder Representations from Transformers (BERT), Uniform Manifold Approximation and Projection (UMAP), and Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), followed by expert interpretation to derive structured parameters. The results indicate a cautious and conditional engagement with AI systems, characterized by moderate usage, low levels of trust, and strong concerns regarding reliability and source credibility. Users frequently verify AI-generated outputs and demonstrate a preference for human expert validation, particularly in complex or sensitive cases. Building on these insights, the study introduces a layered taxonomy of perceived risks spanning epistemic, reasoning, interactional, and institutional dimensions, providing a structured analytical framework for understanding how technical limitations translate into broader behavioural and governance challenges. These results highlight the importance of aligning AI system design with user expectations, emphasizing transparency, verifiability, and human oversight. The proposed taxonomy and analytical framework provide a foundation for future research and contribute to the development of governance approaches for AI systems deployed in high-stakes interpretive domains.**

***Keywords*—Large language models; AI-generated guidance; user perception; trust in AI; perceived risk; source credibility; human-AI interaction; high-stakes domains; risk taxonomy; risk framework**

*Corresponding author

I. Introduction

Artificial intelligence (AI), particularly large language models (LLMs), is increasingly embedded in domains that require not only information retrieval but also judgment, interpretation, and decision support. These systems are now consulted for guidance in areas such as healthcare, legal reasoning, education, and personal decision-making, where the consequences of errors may be significant [1], [2], [3]. As a result, questions of trust, reliability, transparency, privacy, and accountability have moved to the forefront of AI research and governance [4], [5]. Prior work has shown that LLMs can generate fluent yet inaccurate outputs, exhibit hallucinations, and influence user judgments in ways that extend beyond neutral information provision [6], [4]. In high-stakes contexts, particularly in what can be described as interpretive domains, that is, settings where responses depend on reasoning, contextual judgment, and the application of domain-specific knowledge rather than direct retrieval of factual information, these characteristics raise critical concerns regarding how users interpret AI-generated responses, how much they rely on them, and how such systems interact with existing structures of expertise and decision-making.

In response to these challenges (see e.g., [7]), emerging AI governance frameworks, such as the EU AI Act, increasingly adopt risk-based approaches that classify applications according to their potential impact on individuals and society, particularly in domains involving decision-making, access to services, and fundamental human rights [8]. Such frameworks emphasize requirements for transparency, accountability, human oversight, and reliability, especially for high-risk AI systems whose outputs may directly or indirectly influence human decisions and behaviour [4], [5], [9]. Within this context, understanding how users perceive, interpret, and evaluate AI-generated outputs becomes not only a technical concern but also a regulatory and societal imperative.

A central challenge in this landscape lies in the tension between the apparent fluency and accessibility of AI-generated responses and the underlying limitations of these systems. While LLMs can produce coherent and persuasive outputs,

concerns persist regarding their tendency to generate unsupported or misleading information, fabricate references, align responses with user expectations rather than evidence, and operate without clear accountability structures [6], [4]. These issues are amplified in domains where knowledge is not purely factual but interpretive, context-dependent, and institutionally grounded. In such settings, the role of expertise, the legitimacy of sources, and the processes through which knowledge is validated become critical factors shaping both the use and acceptance of AI systems.

Within this broader landscape, certain domains present additional challenges due to their reliance on interpretive reasoning, contextual judgment, and established authority structures. Religious inquiry represents one such domain, where responses are not purely informational but are derived through structured processes that combine source-based evidence, domain expertise, and case-specific interpretation. In the Islamic context, religious rulings (Fatwa, Fatawa) are traditionally produced by qualified scholars through systematic reasoning grounded in recognized sources and interpretive frameworks [10]. As AI systems become increasingly accessible, they are now being used by individuals to obtain guidance on religious questions, effectively introducing LLMs into a domain characterized by high requirements for accuracy, contextual sensitivity, and legitimacy of authority. This shift raises important questions about how users evaluate AI-generated responses in terms of trust, credibility, privacy, and appropriateness, and how these systems are positioned relative to established human-centred processes of knowledge validation.

Despite growing research on AI systems and their applications, existing work remains primarily focused on improving system performance or examining isolated aspects of user interaction. Studies in domains such as healthcare and legal advisory have highlighted issues related to trust, over-reliance, and the difficulty of distinguishing between expert and AI-generated responses [1], [2], [3], while recent work in the context of religious knowledge has pointed to concerns regarding correctness, governance, and alignment with scholarly standards [11], [12], [13]. However, these efforts have largely addressed technical reliability or domain-specific challenges in isolation, with limited attention to how multiple dimensions of user perception, such as trust, perceived risk, privacy, credibility, and authority, interact in shaping real-world use and acceptance of AI systems. This gap is particularly significant in high-stakes interpretive domains, where user perceptions directly influence decision-making, compliance, and the broader societal impact of AI deployment.

Across these strands of research, a more fundamental gap emerges. While prior work has examined system performance and, in some cases, user interaction in specific domains, it does not yet provide a comprehensive understanding of how AI-generated guidance is evaluated and used in high-stakes interpretive contexts. Existing efforts remain primarily oriented toward improving model outputs [14], [15], with comparatively less attention to how users perceive, interpret, and respond to these outputs in real-world settings [16], [1], [2], [3]. At the same time, emerging work in religious knowledge highlights concerns related to correctness, governance, and alignment with scholarly standards [11], [12], [13], but these aspects are often considered independently rather than as interconnected dimensions of user evaluation and system use.

In addition, there is a lack of structured conceptual frameworks that organize the diverse risks associated with AI-generated outputs into coherent and interpretable forms. Most existing approaches address issues such as hallucination, bias, or misuse as separate challenges [6], without considering how these risks propagate across technical, behavioural, and institutional levels [4]. This limitation is particularly critical in high-stakes interpretive domains, where the acceptance and impact of AI systems depend not only on technical performance but also on how users interpret outputs, assess their authority, and align them with established knowledge and decision-making processes.

To address the identified gaps, this study presents one of the first large-scale empirical investigations of public perceptions of AI-generated guidance in a high-stakes interpretive domain, using the context of religious inquiries in Saudi Arabia. The research is based on a combined quantitative and qualitative analysis of survey data collected from 572 participants, providing insight into how individuals engage with AI systems and evaluate their outputs in real-world settings.

The analysis examines user perceptions across multiple interrelated dimensions, including trust, perceived risk, privacy, credibility, and user practices, and further explores how these perceptions vary with demographic and domain-specific factors, particularly age and religious knowledge. The quantitative component employs statistical analysis of structured survey responses, including descriptive statistics, correlation analysis, and group-based comparisons using appropriate significance testing, to identify patterns in user engagement and evaluation. In parallel, the qualitative component applies a machine learning-based pipeline to open-ended responses, leveraging Bidirectional Encoder Representations from Transformers (BERT) for semantic embedding [17], followed by dimensionality reduction using Uniform Manifold Approximation and Projection (UMAP) [18] and clustering via Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) [19]. The resulting clusters are then interpreted through a human–AI collaborative process to derive structured parameters. This integrated statistical and machine learning-based approach enables the identification of underlying concerns and expectations while preserving both analytical rigor and contextual interpretability.

Building on these results, the study introduces a layered taxonomy of perceived risks structured across four interrelated dimensions: epistemic and knowledge integrity, contextual and reasoning limitations, human–AI interaction dynamics, and socio-technical and institutional impacts. This taxonomy provides a structured representation of how technical limitations in AI-generated outputs relate to behavioural patterns and broader governance considerations, offering a coherent framework for understanding user perception in high-stakes contexts.

This study advances a shift from system-centric evaluation toward a more comprehensive understanding of AI in practice, emphasizing the role of user perception, contextual factors, and governance requirements. By providing empirical evidence, a combined statistical and machine learning-based analytical methodology, and a structured conceptual framework, the study offers insights relevant not only to religious inquiry but also to

other high-stakes interpretive domains where trust, authority, and accountability are central to the responsible deployment of AI systems. Beyond its empirical contribution, the proposed taxonomy offers a foundation for future research by providing a systematic way to analyse and compare risks across domains, supporting the development of evaluation benchmarks, governance frameworks, and domain-specific AI systems. In this sense, the framework not only captures current user perceptions but also enables further investigation into how AI-related risks evolve across different contexts and applications.

The remainder of this study is organized as follows: Section II reviews related work and outlines the research problem. Section III describes the methodology, including the survey design and the integrated statistical and machine learning-based analysis pipeline. Section IV presents the quantitative analysis of structured responses, while Section V provides the qualitative analysis and introduces the proposed taxonomy of perceived risks and public expectations. Section VI discusses the results in relation to existing literature and AI governance considerations. Finally, Section VII concludes the study and outlines directions for future research.

II. Literature Background

Recent research has increasingly explored the intersection of artificial intelligence and Islamic knowledge systems, particularly in the context of religious guidance generation and question answering (QA). Early efforts focused on building structured datasets and resources to support AI-driven religious applications. For instance, Alyemny et al. [14] introduced one of the first Arabic religious rulings datasets, compiled from authoritative sources such as the "General Presidency of Scholarly Research and Ifta" in Saudi Arabia [20], prominent scholars including Ibn Othaimin [21], and online repositories like FatwaPedia [22]. This work laid the foundation for developing data-driven AI systems in the Islamic domain.

Building on such resources, several studies have proposed AI-based systems to enhance the reliability of religious question answering. Alan et al. [15] developed MufassirQAS, a retrieval-augmented generation (RAG) system that integrates the Qur'an, Sunnah, and core Islamic teachings as grounding sources. Their findings demonstrated improved response clarity and reduced ambiguity compared to general-purpose models such as ChatGPT. Similarly, Mohammed et al. [16] proposed an advanced RAG-based framework incorporating a Flash re-ranker to reduce hallucinations in AI-generated religious rulings (Fatawa). Their evaluation across multiple models, including AceGPT [23], Gemini-1.5 [24], and SILMA [25], showed that combining RAG with re-ranking significantly improved factual accuracy, albeit at the cost of increased latency.

In parallel, recent work has examined critical reliability aspects of AI systems in religious contexts. Abstention, the ability of a model to refrain from answering when uncertain, has been identified as a crucial requirement in sensitive domains [26]. Atif et al. [27] introduced FiqhQA, a benchmark designed to evaluate LLM performance across different Islamic schools of thought (Madhhabs), highlighting the importance of doctrinal diversity and controlled abstention in AI-generated religious responses. These studies are collectively summarized in Table I.

TABLE I. Summary of AI-based Religious QA Studies

Study	Summary	Research Gap
MufassirQAS [15]	RAG-based QA system using Qur'an, Hadith, and Islamic texts to improve accuracy, reduce hallucinations, and provide source transparency.	Limited generalization, bias issues, and lack of multilingual and contextual depth.
Aftina [16]	RAG framework with re-ranking to improve retrieval quality and reduce hallucinations in AI-generated religious rulings.	Depends on dataset quality; challenges in Arabic understanding and limited domain data.
FiqhQA Benchmark [27]	Introduces a benchmark dataset covering Islamic rulings across four schools of thought and evaluates LLM accuracy and abstention behavior. Shows performance varies by language and model, with weaker results in Arabic.	Lacks standardized evaluation methods, struggles with multilingual reasoning, and highlights need for better abstention, transparency, and alignment with jurisprudential diversity.

While these studies focus primarily on improving the technical performance and reliability of AI systems, a growing body of research has begun to examine how users perceive and interact with AI-generated advice in high-stakes domains. In the medical field, Shekar et al. [1] found that non-expert users often struggle to distinguish between AI-generated and expert-generated responses, and may exhibit high willingness to follow inaccurate AI advice. In the legal domain, Schneiders et al. [2] reported that participants frequently preferred AI-generated advice over professional legal guidance when the source was not disclosed. Similarly, Seabrooke et al. [3] showed that while prior usage of AI for legal advice remains limited, a substantial proportion of users expressed willingness to adopt such tools in the future, particularly for less sensitive topics. Taken together, these studies are summarized in Table II.

TABLE II. Summary of User Perception Studies in High-Stakes AI Domains.

Study	Summary	Research Gap
Medical Domain [1]	Users cannot distinguish AI from doctors and often trust AI responses, even when inaccurate, with willingness to follow harmful advice.	Overtrust in AI and poor error detection; lacks safeguards to prevent harmful reliance.
Legal Domain [2]	Users show higher willingness to act on AI-generated legal advice, especially when the source is unknown, despite being able to distinguish it.	Preference bias toward AI advice without mechanisms to ensure safe decision-making.
Legal Domain [3]	Low current use of AI for legal advice, but strong future willingness, varying by legal domain.	Lacks evaluation of real-world risks, decision quality, and consequences of AI reliance.

Despite these advancements, limited research has investigated public perceptions of AI tools within the religious domain, especially in Islamic contexts where authority, trust, and source credibility play a central role. Existing studies largely emphasize system design, dataset construction, or technical evaluation, with insufficient attention to how users assess trust, risk, and legitimacy when interacting with AI-generated religious guidance.

More recently, emerging studies have begun to explore AI within the context of religious knowledge, user perception, and ethical governance. Alam et al. [11] conducted a mixed-methods investigation of user and expert evaluations of AI-generated religious content, revealing significant challenges in users' ability to assess correctness and reliability. In the Islamic domain, Priantina et al. [12] examined the integration of AI into formulation of religious rulings, highlighting both efficiency gains and critical concerns related to automating

sacred decision-making and weakening scholarly reasoning. Furthermore, a recent large-scale survey by Bhatia et al. [13] synthesized over 160 studies on AI for Islamic knowledge, identifying key gaps in trustworthiness, evaluation frameworks, and governance mechanisms. These findings collectively emphasize that, despite rapid technological advancements, the human-centered dimensions of trust, risk perception, and authority remain underexplored. A summary of these studies is presented in Table III.

TABLE III. SUMMARY OF AI IN RELIGIOUS AND ISLAMIC KNOWLEDGE STUDIES

Study	Summary	Research Gap
[11]	Users prefer AI-generated religious answers (81%) despite claiming trust in scholars; experts identify major quality issues.	User–expert misalignment and limited ability to detect errors; need expert-guided evaluation.
[13]	Comprehensive review of AI systems for Islamic knowledge, highlighting growth of RAG, evaluation methods, and alignment challenges.	Fragmented research, lack of unified evaluation standards, and limited handling of pluralism and trust.
[12]	AI improves efficiency and scalability of formulation of religious rulings but raises concerns about replacing scholarly reasoning.	Ethical risks, lack of governance frameworks, and need for human oversight in decision-making.

A. Research Problem

While the preceding discussion highlights substantial progress in developing AI systems for religious knowledge, including advances in dataset construction, retrieval-augmented generation, and benchmark evaluation, it also reveals important limitations in how these systems are assessed beyond technical performance. In particular, existing research has largely focused on improving accuracy and reducing hallucination in AI-generated outputs [14], [15], [16], [27], with comparatively limited attention to how such outputs are interpreted and evaluated by users in real-world settings.

In parallel, a growing body of work has examined user interaction with AI systems in high-stakes domains such as healthcare and legal advisory, highlighting challenges related to trust, over-reliance, and the difficulty of distinguishing between expert and AI-generated responses [1], [2], [3]. More recent research in the context of religious knowledge has identified additional concerns, including discrepancies between user perception and expert evaluation, risks associated with automating interpretive reasoning, and the absence of comprehensive governance and evaluation frameworks [11], [12], [13]. Despite these advances, existing research remains siloed, with studies typically addressing either system performance or isolated aspects of user perception, rather than providing a comprehensive understanding of how multiple factors interact in shaping user evaluation and use of AI-generated guidance.

A key limitation in the current literature is the lack of an integrated perspective that connects core dimensions of user perception, including trust, perceived risk, privacy, credibility, and user practices, with broader considerations of authority and governance. In domains characterized by interpretive reasoning and contextual judgment, such as religious inquiry, these dimensions are inherently interdependent. However, prior work has largely examined them in isolation, without capturing how concerns related to accuracy, contextual understanding, interaction behaviour, and institutional authority collectively influence the acceptance and use of AI systems.

Furthermore, there is a lack of structured conceptual frameworks that organize the diverse risks associated with AI-generated outputs into coherent and interpretable categories. Existing approaches typically address issues such as hallucination, bias, or misuse as separate challenges, without considering how these risks propagate across technical, behavioural, and institutional levels. This limits the ability to develop a holistic understanding of AI-related risks in high-stakes interpretive contexts.

To the best of our knowledge, large-scale empirical evidence on how users evaluate AI-generated guidance in high-stakes interpretive domains remains limited globally. This gap is particularly pronounced in context-specific environments such as Saudi Arabia, where domain expertise, cultural expectations, and established authority structures play a central role in shaping user behaviour.

To address these gaps, this study presents one of the first large-scale empirical investigations of public perceptions of AI-generated religious guidance in Saudi Arabia, based on a combined quantitative and qualitative analysis of survey data. The work examines how users engage with AI systems and how they evaluate outputs across multiple dimensions, including trust, perceived risk, privacy concerns, and source credibility. In addition, it employs an integrated statistical and machine learning-based analysis to derive structured parameters from open-ended responses, enabling the identification of underlying concerns and expectations through clustering and expert interpretation. Building on this, the study introduces a layered taxonomy of perceived risks structured across four interrelated dimensions: epistemic and knowledge integrity, contextual and reasoning limitations, human–AI interaction dynamics, and socio-technical and institutional impacts. This structured representation captures how concerns originate at the level of AI-generated outputs and propagate through reasoning, user interaction, and institutional contexts, providing a coherent framework for understanding the interdependencies between technical limitations, user behaviour, and governance considerations.

To operationalize these objectives, the study addresses the following research questions:

RQ1. How do individuals in Saudi Arabia engage with and understand AI tools in the context of religious inquiries?

RQ2. How are trust, perceived risk, privacy, and credibility evaluated when AI systems are used to generate religious guidance?

RQ3. How do demographic and domain-specific factors, particularly age and religious knowledge, influence these perceptions and related user practices?

III. METHODOLOGY AND DESIGN

We conducted a survey to examine participants' perceptions regarding the use of AI for generating religious content. The survey was designed to capture both demographic characteristics and participants' perceptions toward AI-generated religious guidance.

The questionnaire consisted of two main parts: 1) Demographic information was collected from participants, including age, gender, level of education, religious background, AI usage, and whether they usually consult scholars or official Islamic websites when seeking religious information. 2) Participants' views on AI-generated religious content. We grouped the related questions, as shown in Table IV. This section captured participants' trust in AI-generated religious content, perceived risks of relying on AI for religious guidance, and perceptions regarding the credibility of AI-generated sources. Open-ended questions were included to allow participants to elaborate on their concerns and provide suggestions regarding the appropriate role of AI in religious contexts.

Note that the questionnaire items were developed based on themes identified in prior literature on trust, credibility, risk perception, and AI usage, and were reviewed for clarity and contextual appropriateness prior to distribution. Internal consistency of the quantitative constructs was subsequently assessed using Cronbach's alpha.

TABLE IV. QUESTIONNAIRE USED IN THE RESEARCH

Variables	Questions	Question type
Familiarity with AI	How often do you request a religious ruling from AI?	Frequency Scale (1–5)
	How deep is your understanding of AI tools?	Ranked categories from high to low understanding
Trust in AI	I trust AI-generated religious rulings	Agreement scale (1–5)
	I believe AI-generated religious rulings are accurate	Agreement scale (1–5)
Perceived Privacy	I prefer to seek religious rulings from AI about things I might feel hesitant to ask someone	Agreement scale (1–5)
Perceived Risk	I feel concerned about the implications of using AI for religious rulings	Agreement scale (1–5)
	I believe that asking AI for religious rulings weakens the relationship between society and scholars	Agreement scale (1–5)
	I believe that seeking religious rulings from AI may reduce the level of concern for the accuracy of religious rulings	Agreement scale (1–5)
Perceived Credibility	I prefer that the AI-generated religious rulings be linked to well-known sources (e.g., the Council of Senior Scholars)	Agreement scale (1–5)
	To what extent do you personally verify the reliability of the source provided by AI before relying on its answer?	Frequency Scale (1–5)
Behavioral Preference	If you encounter a religious ruling from AI that contradicts the religious ruling of a scholar you trust, which one would you follow?	Choice Question
Use Intention	Would you consider asking AI for a religious ruling in the future?	Yes and No question

A. Data Collection

We collect data for our research by conducting a large-scale online study. We recruited participants using different approaches such as email (from both academic and non-academic environments) and from social networks. At the beginning of the research, 573 participants read and gave their consent to participate in the research. The inclusion criteria require participants to be 18 years or older. We included a total of 572 responses in this analysis after we filtered out incomplete responses.

B. Participants

A total of 572 participants (age range of most (74.6%) of our participants was 18-44 years; 169 males and 326 females) who had used AI before. Most of our participants (94.4%), turn to scholars to obtain religious rulings. As shown in Table V, the participants were diverse in terms of age, gender, education level, and religious (Fiqh) background.

TABLE V. DEMOGRAPHIC CHARACTERISTICS OF PARTICIPANTS

Age	18–28 (211, 36.9%), 29–44 (192, 33.6%), 45–60 (152, 26.6%), above 61 (17, 3%)
Gender	Male (188, 32.9%), Female (384, 67.1%)
Educational level	High school (88, 15.4%), Undergraduate (341, 59.6%), Graduate (132, 23.1%), Other (11, 1.9%)
Religious Background	Basic (167, 29.2%), Intermediate (256, 44.8%), Advanced (149, 26%)
Seek Religious Rulings	Scholars (540, 94.4%), Official Islamic websites (32, 5.6%)

C. Data Analysis

Fig. 1 presents the overall architecture adopted in this research, illustrating how quantitative statistical analysis and qualitative BERT-based modeling are integrated to generate a comprehensive understanding of public perceptions of AI-generated religious rulings. The framework begins with survey-based data collection, capturing both structured (closed-ended) and unstructured (open-ended) responses from participants. These two data streams are then processed through parallel analytical pipelines.

The quantitative pipeline focuses on structured responses and applies statistical techniques to measure key constructs such as trust, perceived privacy and risk, and credibility, as well as to examine relationships and demographic effects. In parallel, the qualitative pipeline processes open-ended responses using a BERT-based modeling approach to extract latent themes, enabling the identification of underlying concerns, expectations, and recommendations.

The outputs of both pipelines are subsequently integrated through a triangulation process, ensuring complementarity and cross-validation between statistical patterns and thematic insights. This integration step forms the core of the framework, allowing the research to move beyond isolated results toward a unified interpretation of user perceptions. The final stage translates these integrated insights into a structured understanding of perception patterns, key concerns, and evidence-based implications for the design and governance of AI systems in religious contexts.

While Fig. 1 provides a high-level overview of this analytical architecture, the detailed methodological components of the quantitative and qualitative pipelines are described in the following sections.

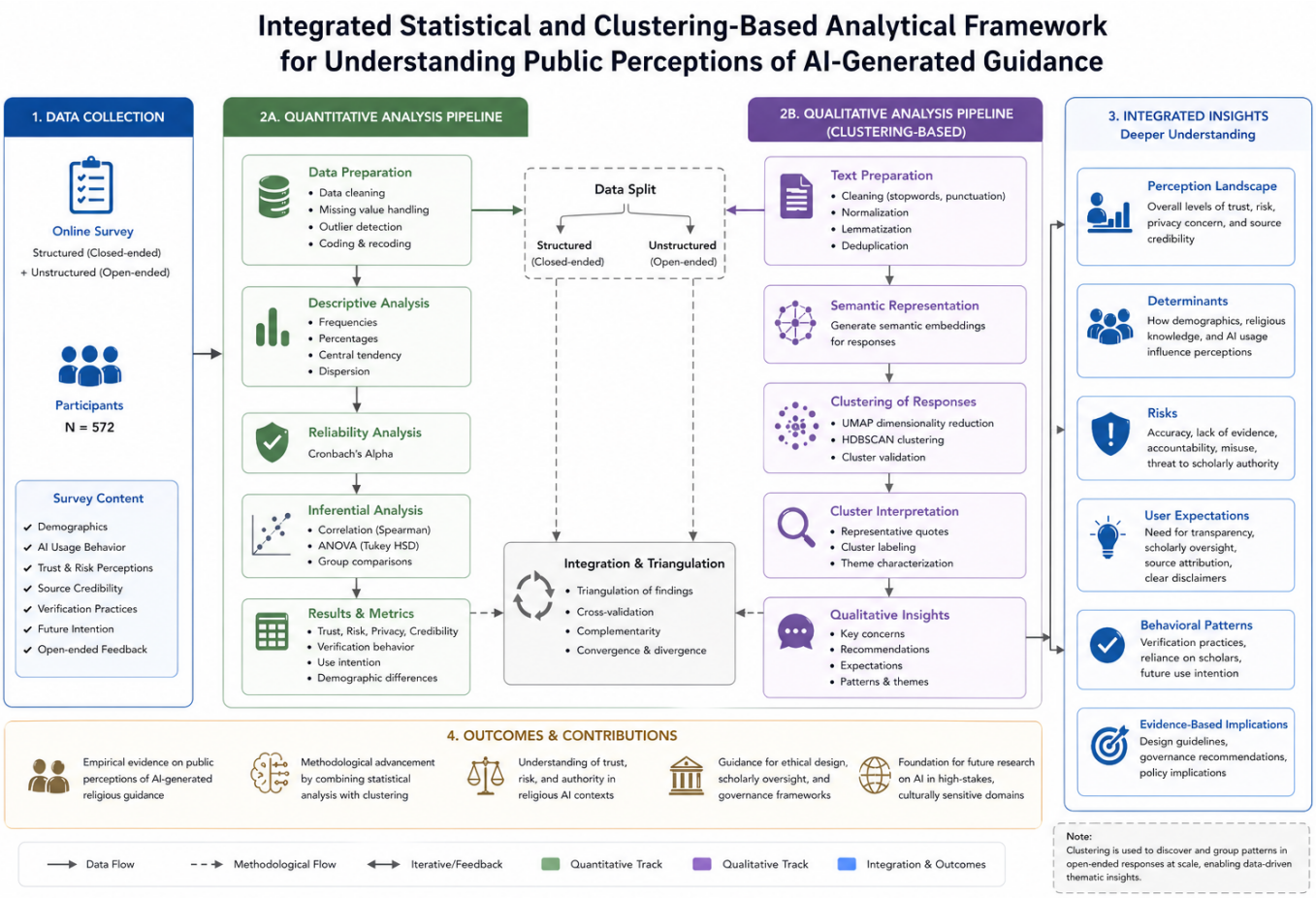

Fig. 1. Integrated statistical and clustering-based analytical framework for examining public perceptions of AI-generated guidance. The figure illustrates the end-to-end methodology combining quantitative statistical analysis of structured survey responses with embedding-based clustering of open-ended responses, followed by an integration and triangulation process. This framework enables the identification of patterns in trust, perceived risk, privacy, credibility, and user behavior, while deriving structured parameters and a layered taxonomy of risks across epistemic, reasoning, interactional, and socio-technical dimensions, supporting a unified interpretation of user perceptions in high-stakes interpretive domains.

D. Quantitative Data Analysis

To analyze the quantitative data and address the research questions, established statistical analysis techniques were employed using SPSS. Descriptive statistics were calculated to summarize the demographic characteristics of the participants. Cronbach's alpha was used to assess internal reliability when using two or three items on a 5-point Likert scale (ranging from Strongly Disagree to Strongly Agree) to measure one variable. If Cronbach's alpha $>= 0.70$, we compute the mean score of the items for the variable.

A one-way ANOVA tests were conducted to investigate whether demographic factors such as age and religious background influence the use of AI by participants when seeking religious rulings. A Tukey HSD post hoc test was conducted after the significant ANOVA result to examine differences between the groups.

E. Parameter Discovery and Qualitative Data Analysis

Open-ended survey responses were analyzed using a structured qualitative modelling pipeline based on BERTopic [17], which integrates transformer-based embeddings, dimensionality reduction, and density-based clustering to identify semantically coherent themes in textual data. Sentence-level embeddings were generated using the Arabic language model CAMeL-Lab/bert-base-arabic-camelbert-mix [28]. These embeddings were projected into a lower-dimensional space using UMAP [18] to preserve local semantic structure while enabling effective clustering. HDBSCAN [19] was then applied to automatically identify clusters without requiring a predefined number of topics.

The analysis was conducted separately on the concerns and suggestions datasets to capture distinct dimensions of user perceptions, namely perceived risks and proposed safeguards. This separation ensures that emergent themes are not conflated during the initial clustering process. Following clustering, topic representations were refined through post-processing, including stopword removal and inspection of representative responses, to improve interpretability while preserving contextual integrity. The resulting clusters were then comparatively examined across datasets to identify recurring and complementary themes.

Finally, clusters were abstracted into higher-level parame-

ters through an iterative human-guided consolidation process, ensuring semantic coherence, conceptual distinctiveness, and alignment with the overall analytical framework. This process resulted in a structured taxonomy of user-perceived risks and considerations associated with AI-generated guidance.

IV. Quantitative Analysis of Structured Survey Responses

Here, in this section, we present the results of the quantitative analysis of the survey data. We begin by reporting descriptive statistics and key patterns related to AI usage, trust, perceived risk, privacy, and credibility, followed by inferential analysis examining relationships between variables and differences across demographic and domain-specific groups. These results establish the empirical basis for the qualitative analysis presented in Section V.

A. Participants' Familiarity and Understanding of AI Tools

To understand the participants' familiarity with AI tools, we first examined how often they used AI for religious rulings, then we assessed the depth of their understanding of AI tools. As shown in Fig. 2, 31.3% of participants stated that they sometimes used AI to obtain religious content, while 30.8% reported that they never use AI for this purpose. Moreover, 18.9% participants reported rarely using AI and only 19% of participants indicated that they use AI often or always. These results show that participants use AI tools sometimes rather than regularly for religious inquiries. It suggests that there is moderate adoption of AI-generated religious rulings, but some people may hesitate or have limited trust in them.

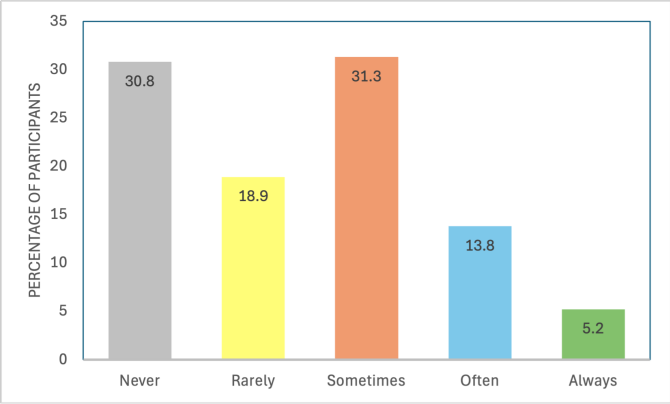

Fig. 2. Frequency of AI use among participants.

Regarding the understanding of AI tools by participants, Fig. 3 shows that 40.4% of participants stated that they have a basic understanding of AI, while 32.7% indicated having only a general knowledge of AI. Moreover, 14.9% of the participants reported that they had no knowledge of AI and only 12.1% stated that they had a deep knowledge of AI tools. The results show that most participants know about AI tools, but few fully understand how these tools work.

To understand the relationship between AI understanding and its frequency of use for religious rulings, we run a Spearman correlation analysis. The results show no significant relationship between AI understanding and its frequency of use for religious rulings (r = 0.035, p = 0.400), indicating that the level of AI understanding does not significantly impact the frequency of using AI for religious rulings. Table VI presents the results on AI familiarity, understanding, and their relationship.

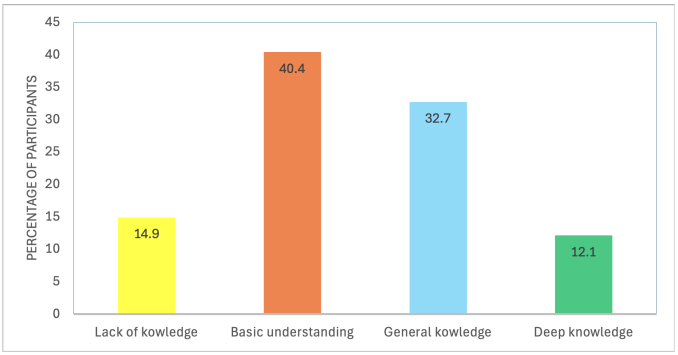

Fig. 3. Percentage of participants' understanding of how AI tools work.

TABLE VI. Summary of AI Familiarity and Understanding

Variable	Results	Interpretation
Familiarity with AI	Often or always (19%) sometimes (31.3%) rarely (18.9%) never (30.8%)	There is moderate adoption of AI-generated religious rulings.
Understanding of AI Tools	Basic understanding (40.4%) general knowledge (32.7%) no knowledge (14.9%) deep knowledge (12.1%)	Most participants are aware of AI tools, but only a few have a deep understanding of how they work.
Understanding AI & Usage	r = 0.035, p= 0.400 relationship between AI understanding and its frequency of use for religious rulings	level of AI understanding does not significantly impact the frequency of using AI for religious rulings.

B. Trust in AI

There are two items that measure the trust in AI for religious rulings, First we examined the reliability of the two items using Cronbach's alpha. The result shows an acceptable internal consistency ($\alpha = 0.785$), indicating that the two items reliably measure the trust in AI for religious rulings. Then we calculated the mean and standard deviation of trust in AI for religious rulings. The results show a mean score of 1.85 (SD = 0.96) based on 572 responses. It suggests that participants show a low level of trust in AI-generated religious rulings.

To understand whether Trust in AI for religious rulings impact the intention to use AI in the future, we run a Spearman correlation analysis. We found a moderate and statistically significant positive relationship (r = 0.499, $p < 0.001$). This indicates that participants with higher levels of trust in AI-generated religious rulings are more likely to consider using AI for religious rulings in the future and vice versa.

We also performed a Spearman correlation analysis to examine the relationship between trust in AI-generated religious rulings and the preference of participants when AI-generated religious rulings contradict those of trusted scholars. The results show a weak but statistically significant positive relationship (r = 0.151, $p < 0.001$). This shows that participants with higher levels of trust in AI-generated religious rulings

are more likely to consider following AI-generated religious rulings, even when they contradict the opinions of trusted scholars. Table VII presents the results of trust in AI-generated religious rulings and its association with use intention and behavioral preference.

TABLE VII. TRUST IN AI-GENERATED RELIGIOUS RULINGS

Variable	Results	Interpretation
Trust in AI for Religious Rulings	M = 1.85, SD = 0.96 based on 572 responses	Participants show a low level of trust in AI-generated religious rulings.
Trust in AI and Use Intention	$r = 0.499$, $p < 0.001$ between trust in AI and use intention	Participants with higher levels of trust in AI-generated religious rulings are more likely to consider using AI for religious rulings in the future, and vice versa.
Trust in AI and Behavioral Preference	$r = 0.151$, $p < 0.001$ between trust in AI and behavioral preference	Participants with higher levels of trust in AI-generated religious inquiries are more likely to consider following AI-generated religious rulings, even when they contradict the opinions of trusted scholars.

C. Perceived Privacy, Risk, and Credibility of AI for Religious Inquiries

We measure the perceived privacy of AI for religious rulings by understanding if participants prefer to seek religious rulings from AI for questions they might feel hesitant to ask others. The results show that a majority of participants do not prefer to seek religious rulings from AI for questions they might feel hesitant to ask others. Fig. 4 shows that 39% of participants strongly disagreed and 13.8% disagreed, while 23.1% remained neutral. Only 24.1% of participants expressed agreement. This indicates that most participants are hesitant to rely on AI for sensitive religious questions.

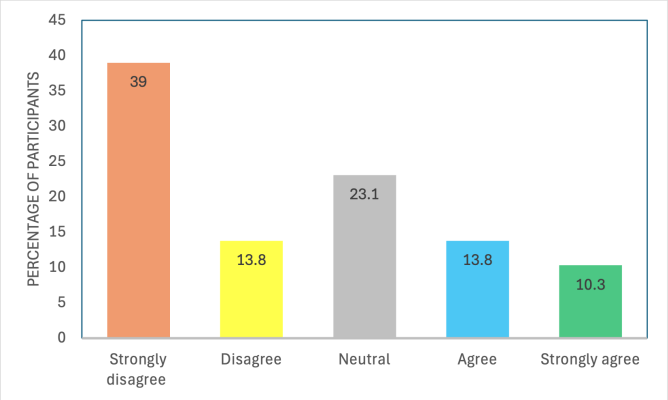

Fig. 4. Percentage of participants using AI for sensitive religious questions.

We examined the reliability of the three-item scale measuring perceived risk of AI using Cronbach's alpha. The results indicated acceptable internal consistency ($\alpha = 0.737$), suggesting that the three items reliably measure the perceived risk of AI. The mean clarifies that participants have a high level of perceived risk regarding the use of AI for religious rulings (M = 3.99, SD = 1.02, N = 572). This suggests that participants generally agree with statements expressing concerns about the implications of AI-generated religious rulings.

Perceived credibility of AI-generated religious rulings was examined using two indicators: Participants' tendency to verify the reliability of sources provided by AI and their preference for AI-generated religious rulings linked to well-known scholarly authorities. Regarding participants' tendency to verify the reliability of sources provided by AI, the results show that participants consider the importance on checking the credibility of AI-generated religious rulings. As shown in Fig. 5, they always verify the reliability of sources (28.8%), followed by 24.8% who indicated that they often verify them. Overall, 51.4% of participants reported that they often or always verify the reliability of sources provided by AI, highlighting the importance of credible references when relying on AI-generated religious rulings.

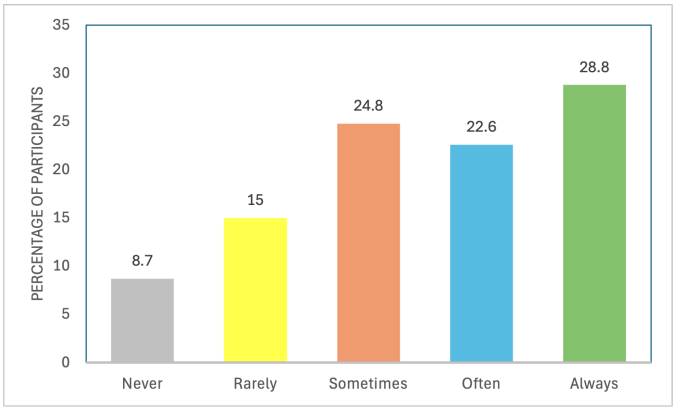

Fig. 5. Percentage of participants verifying the reliability of sources provided by AI.

Regarding participants' preference for AI-generated religious rulings linked to well-known scholarly authorities, the results show strong support for linking AI-generated religious rulings to well-known scholarly sources. Fig. 6 shows that the majority of participants strongly agreed (65.6%) to link AI-generated religious rulings to well-known scholarly sources, while 14% agreed, resulting in 79.6% expressing agreement overall. These results highlight the importance of authoritative references in enhancing the perceived credibility of AI-generated religious guidance.

To find the relationship between Participants' tendency to verify the reliability of sources provided by AI and their preference for AI-generated religious rulings linked to well-known scholarly authorities, we run a Spearman correlation analysis. The results revealed a weak but statistically significant positive relationship (r = 0.103, p = 0.014), that people who verify AI sources more frequently show a slightly stronger preference for AI-generated religious rulings supported by authoritative references. As illustrated in Table VIII, participants' perceptions of privacy, risk, and credibility shape their views on using AI for religious inquiries.

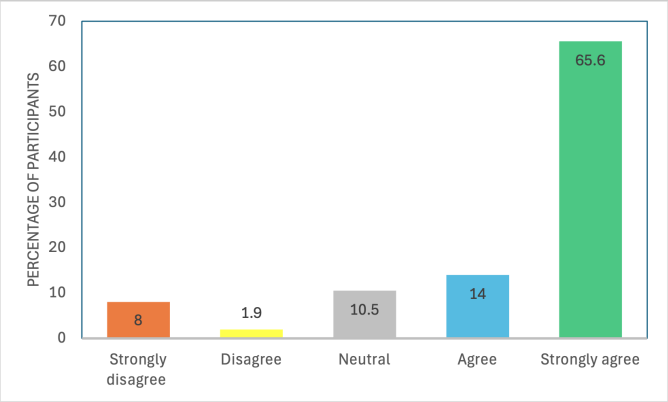

Fig. 6. Percentage of participants preferring for AI-generated religious rulings linked to well-known scholarly authorities.

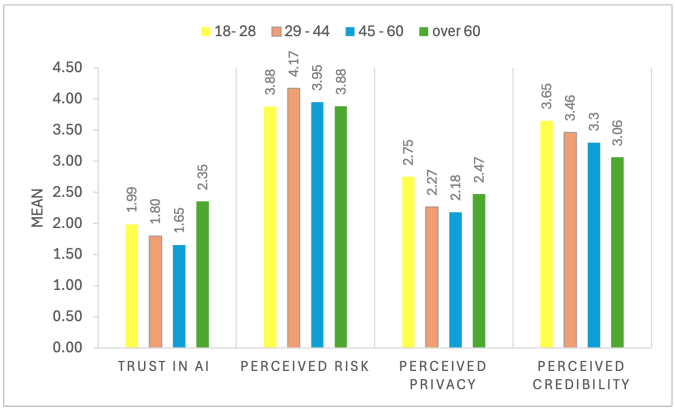

Fig. 7. Participants' perceptions based on age (Mean Values).

TABLE VIII. PERCEIVED PRIVACY, RISK, AND CREDIBILITY OF AI FOR RELIGIOUS INQUIRIES.

Variable	Results	Interpretation
Perceived Privacy	strongly disagreed (39%) disagreed (13.8%) neutral (23.1%) agree (13.8%) strongly agree (10.3%)	Most participants are hesitant to rely on AI for sensitive religious inquiries.
Perceived Risk	M = 3.99, SD = 1.02 based on 572 responses	Participants have a high level of perceived risk regarding the use of AI for religious inquiries.
Perceived Credibility	1- Tendency to verify reliability of sources: often or always (51.4%) sometimes (24.8%) rarely (15%) never (8.7%) 2- Preference for linking AI to scholarly authorities: strongly agree (65.6%) agree (14%) neutral (10.5%) disagree (1.9%) strongly disagree (8%) Relationship between (1) and (2): $r = 0.103$, $p = 0.014$	The importance of credible and authoritative references improves the perceived credibility of AI-generated religious inquiries. Participants who frequently verify sources are more likely to prefer linking AI to well-known scholarly authorities.

D. The Impact of Age Toward Using AI-Generated Religious Rulings

To examine whether age influence participants toward AI-generated religious rulings, we conducted one-way ANOVA analyses followed by a Tukey HSD post hoc test to identify differences between age groups. The overall trends are illustrated in Fig. 7.

1) Trust in AI: The results showed a statistically significant difference in trust levels [F(3,568) = 5.342, p = 0.001], indicating that trust in AI-generated religious rulings varies significantly among the age groups.The results of Tukey HSD post hoc showed that participants age range of 18-28 years reported significantly higher trust in AI-generated religious rulings compared to age group (45 - 60) (p = 0.006). Additionally, participants age over 60 showed significantly higher trust levels than participants age range 45 - 60 (p = 0.022). No other significant differences between age groups were observed.

2) Perceived risk: The results revealed a statistically significant difference in perceived risk among the age groups [F(3,568) = 3.073, p = 0.027], indicating that perceived risk of AI-generated religious rulings varies significantly between the groups. The results of Tukey HSD post hoc showed participants age range 29 - 44 perceived significantly higher risk of AI-generated religious rulings than those age range 18-28 (p = 0.019), while no other significant differences between age groups were found.

3) Perceived privacy: The results revealed a statistically significant difference among the age groups [F(3,568) = 6.481, $p < 0.001$], indicating that the willingness to consult AI for sensitive religious questions varies significantly between the groups. The results of Tukey HSD post hoc show that participants in age range (18-28) reported significantly higher willingness to consult AI for sensitive religious questions compared with participants 29-44 (p = 0.002) and participants age 45-60 ($p < 0.001$). No significant differences were observed between the remaining age group comparisons.

4) Perceived credibility: The results revealed a statistically significant difference between the age groups [F(3,568) = 2.905, p = 0.034], indicating that the tendency to verify AI-generated sources varies significantly among the groups. Post hoc comparisons using the Tukey HSD test revealed that participants in age range (18-28) verify AI-generated sources significantly more often than those in age (29-44) (p = 0.049), while no other significant differences were found among the age groups.

Overall, age has a significant influence on Participants' perception toward AI-generated religious rulings. Post hoc analysis indicated that participants in the youngest age group reported greater willingness to seek religious rulings from AI for questions they might feel hesitant to ask others, compared with some older age groups. Additionally, younger participants showed slightly higher levels of source verification behaviour than some other age groups. As shown in Table IX, age differences play a significant role in shaping trust, perceived risk, privacy concerns, and credibility of AI-generated religious information.

TABLE IX. AGE DIFFERENCES IN TRUST, RISK, PRIVACY, AND CREDIBILITY.

Variable	Results	Interpretation
Trust in AI	F = 5.342 $p = 0.001$ Post hoc (Tukey): (18–28) > (45–60): $p = 0.006$ (60+) > (45–60): $p = 0.022$	Participants aged 18–28 showed higher trust and greater willingness to use AI for sensitive religious inquiries, along with more frequent verification of AI-generated sources.
Perceived Risk	F = 3.073 $p = 0.027$ Post hoc (Tukey): 29–44 > 18–28: $p = 0.019$	Participants aged 29–44 reported higher perceived risk than younger participants.
Perceived Privacy	F = 6.481 $p = 0.001$ Post hoc (Tukey): 18–28 > 29–44: $p = 0.002$ 18–28 > 45–60: $p = 0.001$	Younger participants (18–28) showed higher concern about privacy compared to older groups.
Perceived Credibility	F = 2.905 $p = 0.034$ Post hoc (Tukey): (18–28) > (29–44): $p = 0.049$	Younger participants tend to perceive AI-generated religious information as more credible compared to middle-aged participants.

E. The Impact of Religious Background Toward Using AI-Generated Religious Content

To examine whether religious background influences participants toward AI-generated religious rulings, we conducted one-way ANOVA analysis followed by a Tukey HSD post hoc test to identify differences between groups. The overall differences across religious background levels are illustrated in Fig. 8.

1) Trust in AI: The results indicated a statistically significant difference in trust levels [F(2,569) = 9.834, p $<$ 0.001], suggesting that trust in AI-generated religious rulings varies significantly across people with different religious background. The results of a Tukey HSD post hoc text showed that participants with basic level of religious background reported significantly higher trust compared with those with intermediate (p = 0.003) and advanced levels of religious knowledge (p $<$ 0.001). No significant difference was found between intermediate and advanced religious knowledge levels.

2) Perceived risk: The results revealed a statistically significant difference in perceived risk [F(2,569) = 9.778, p $<$ 0.001], indicating that perceived risk varies significantly across people with different level of religious background. The results of Tukey HSD post hoc showed participants with basic level of religious background reported significantly lower perceived risk compared to those with intermediate (p = 0.003) and advanced religious knowledge (p $<$ 0.001). No significant difference was observed between the intermediate and advanced groups.

3) Perceived privacy: The results revealed a statistically significant difference among the participants with different level of religious background [F(2,569) = 4.113, p = 0.017], indicating that willingness to ask AI for sensitive religious questions varies significantly between groups. The results of a Tukey HSD post hoc show that participants with basic level of religious background reported significantly higher willingness to ask AI for sensitive religious questions compared with those with higher religious background (p = 0.021) and intermediate level of religious knowledge (p = 0.05). No statistically significant differences were observed between intermediate and advanced levels of religious background.

4) Perceived credibility: The results revealed a statistically significant difference between the groups [F(2,569) = 7.611, p $<$ 0.001], indicating that the tendency to verify AI-generated sources varies significantly between the groups. Post hoc comparisons using the Tukey HSD test revealed that participants in advanced religious knowledge level reported significantly higher verification of AI sources compared with those in basic religious knowledge level (p $<$ 0.001) and intermediate religious knowledge level (p = 0.036). No significant difference was found between basic religious knowledge level and intermediate religious knowledge level.

Overall, participants with basic levels of religious knowledge reported significantly higher trust in AI-generated religious rulings, greater willingness to ask AI for sensitive religious questions, and lower perceived risk compared to those with intermediate and advanced levels of religious knowledge. However, participants with advanced religious knowledge were more likely to verify the reliability of sources provided by AI, indicating greater scrutiny when evaluating AI-generated religious rulings. As illustrated in Table X, differences in religious knowledge levels significantly influence trust, perceived risk, privacy, and credibility of AI.

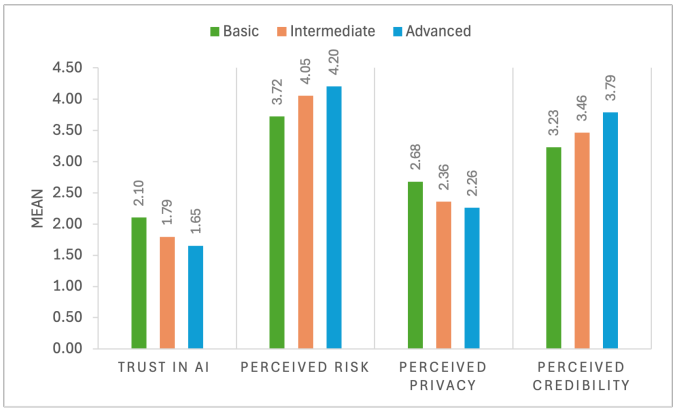

Fig. 8. Participants' perceptions based on the level of religious background (Mean Values).

V. QUALITATIVE ANALYSIS OF OPEN-ENDED SURVEY RESPONSES

The qualitative analysis of participants' open-ended responses provides deeper insight into how individuals interpret and evaluate the use of AI systems in high-stakes, interpretive contexts. While the quantitative results establish broad patterns related to trust, perceived risk, privacy perceptions, and credibility, the qualitative analysis reveals the underlying reasoning, concerns, and expectations that shape these perceptions.

Using a machine learning–based text analysis pipeline, responses were clustered and iteratively refined to extract coherent themes. These themes were interpreted as structured parameters capturing participants' perceptions. To ensure conceptual clarity and avoid overlap, the resulting parameters were organized into a hierarchical taxonomy consisting of four macro-dimensions (macro-parameters). These macro-parameters reflect distinct layers within the broader socio-technical system in which AI operates, spanning multiple

TABLE X. DIFFERENCES BASED ON RELIGIOUS KNOWLEDGE LEVELS

Variable	Results	Interpretation
Trust in AI	F = 9.834 $p = 0.001$ Post hoc (Tukey): Basic > Intermediate: $p = 0.050$ Basic > Advanced: $p = 0.021$	Participants with basic religious knowledge showed higher trust, lower perceived risk, and greater willingness to use AI for sensitive religious inquiries compared to those with intermediate and advanced levels.
Perceived Risk	F = 9.778 $p = 0.001$ Post hoc (Tukey): Basic < Intermediate: $p = 0.003$ Basic < Advanced: $p = 0.001$	Participants with advanced religious knowledge reported higher perceived risk compared to those with basic knowledge.
Perceived Privacy	F = 4.113 $p = 0.017$ Post hoc (Tukey): Basic > Intermediate: $p = 0.050$ Basic > Advanced: $p = 0.021$	Participants with lower religious knowledge expressed greater privacy concerns than those with higher knowledge levels.
Perceived Credibility	F = 7.611 $p = 0.001$ Post hoc (Tukey): Advanced > Basic: $p < 0.001$ Advanced > Intermediate: $p = 0.036$	Participants with advanced religious knowledge were more likely to verify AI-generated sources, indicating greater critical evaluation.

layers: knowledge integrity (what AI generates and how reliable it is), reasoning capability (how AI interprets and adapts responses), human interaction (how users engage with and rely on AI), and institutional impact (how AI affects authority, governance, and decision structures).

Specifically, the taxonomy is structured into: 1) Epistemic and Knowledge Integrity Risks, which capture concerns related to correctness, evidence, and reliability of AI-generated outputs; 2) Contextual and Reasoning Limitations, which reflect the inability of AI systems to incorporate situational and case-specific nuances; 3) Human–AI Interaction Risks, which describe how users engage with and are influenced by AI systems; and 4) Socio-Technical and Institutional Risks, which capture the broader implications of AI use on authority structures and governance.

This structured approach enables a systematic understanding of participants' perceptions, moving beyond isolated concerns toward a layered interpretation of risks associated with AI-generated guidance.

A. Perceived Risks in AI-Generated Religious Guidance

The analysis of participants' responses revealed a set of distinct but interrelated risk categories associated with the use of AI for generating religious guidance. These risks reflect concerns not only about the technical behaviour of AI systems, but also about their epistemic validity, interaction dynamics, and broader institutional implications.

Table XI presents the finalized taxonomy of perceived risks, organized across the four macro-parameters described above, along with representative quotes from participants.

The identified risks demonstrate that participants' concerns extend beyond isolated technical issues and instead reflect a multi-layered understanding of AI systems operating within a socio-technical context.

TABLE XI. TAXONOMY OF PERCEIVED RISKS WITH REPRESENTATIVE QUOTES

Macro	Parameter	Interpretation	Representative Quotes
Epistemic & Knowledge Integrity	Epistemic Unreliability	Responses may lack correctness, clarity, and evidential strength.	- Reliance on weak evidence - Low accuracy and limited domain knowledge - Answers may contradict established knowledge
	Fabrication & Hallucination	Systems may generate fabricated or non-existent information with high confidence.	- Produces hallucinated information - Cites references that do not exist - Generates incorrect information
	Lack of Authoritative Grounding	Outputs are not anchored in trusted or recognized sources.	- Not connected to recognized scholars - Not based on a trusted source
	Doctrinal Inconsistency	Systems may mix interpretive frameworks without clear methodology.	- Mixing beliefs and sects - Mixing correct and incorrect positions - Manipulation of interpretations
Contextual & Reasoning Limitations	Contextual Insensitivity	Systems fail to account for individual circumstances and situational nuance.	- Cannot adapt rulings to individual situations - Each individual requires a specific ruling
Human–AI Interaction Risks	Preference Alignment Bias	Outputs may align with user expectations rather than truth (sycophancy).	- Distorts answers to satisfy the user - Provides answers at the expense of truth
	Overdependence on AI	Users may rely excessively on AI without verification or independent reasoning.	- Complete reliance on AI - Not using human reasoning or verification
Socio-Technical & Institutional Risks	Displacement of Domain Authority	AI may weaken reliance on domain experts and alter authority structures.	- Abandoning domain experts - Reducing the authority of experts

At the epistemic level (Epistemic and Knowledge Integrity), participants emphasize the importance of correctness, evidence, and source credibility, highlighting concerns about unreliable outputs, fabricated information, and weak grounding in authoritative knowledge. These concerns align closely with the quantitative results, which indicate low trust and high perceived risk.

At the reasoning level (Contextual and Reasoning Limitations), participants identify limitations in the ability of AI systems to handle context-dependent scenarios. This reflects a recognition that certain forms of decision-making require not only information retrieval but also contextual interpretation and situational judgment.

At the interaction level (Human–AI Interaction Risks), concerns shift toward how users engage with AI systems. Participants highlight the risk of overdependence, as well as the tendency of AI systems to produce responses that align with

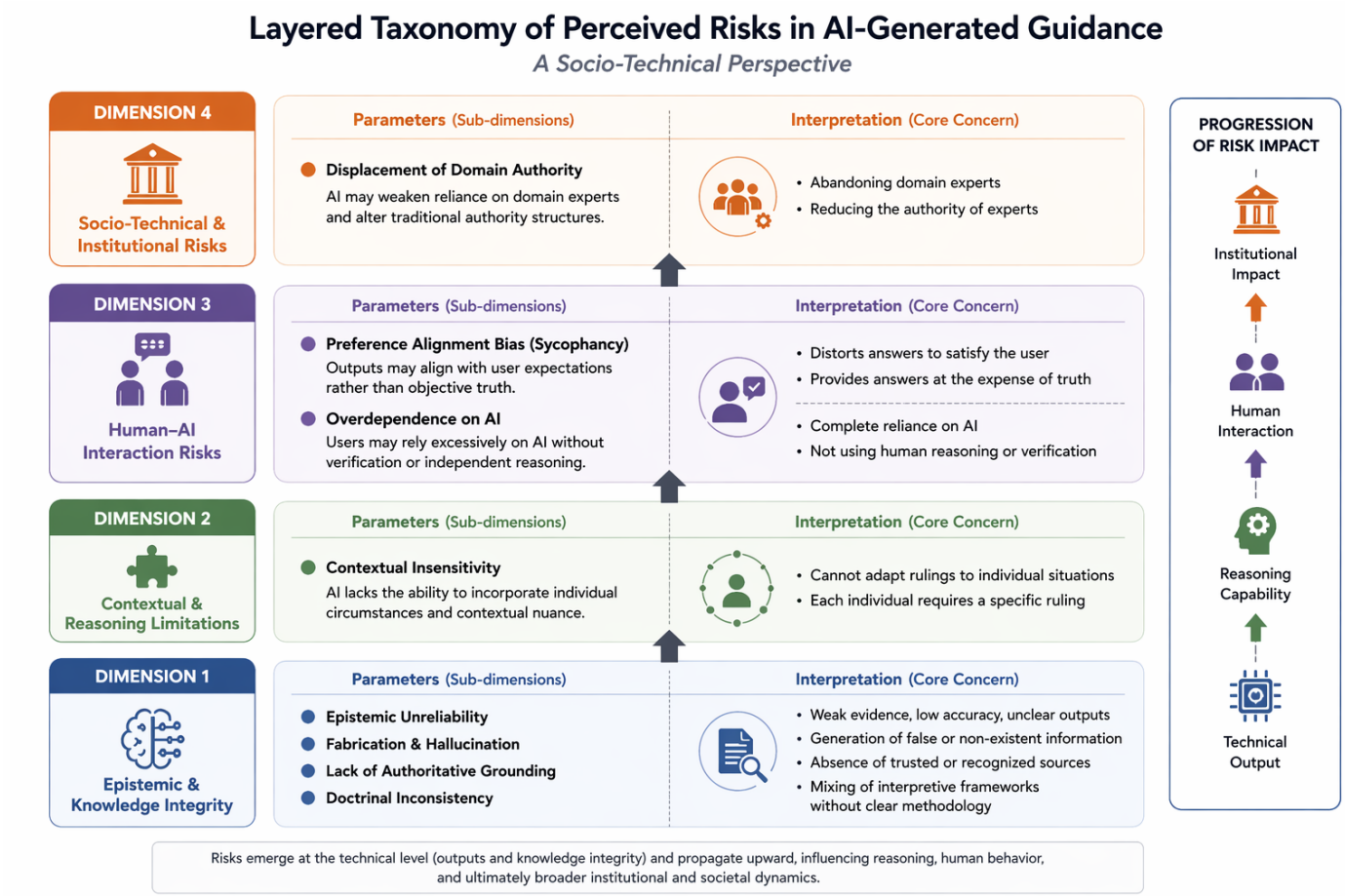

Fig. 9. Framework and layered taxonomy of perceived risks in AI-generated guidance. The figure organizes risks into four macro-parameters: Epistemic and Knowledge Integrity, Contextual and Reasoning Limitations, Human–AI Interaction Risks, and Socio-Technical and Institutional Risks, showing how risks evolve from AI output-level issues to reasoning, interaction, and institutional impacts.

user expectations, potentially reinforcing incorrect beliefs. This behavior aligns with emerging concerns around AI sycophancy, where systems adapt outputs to satisfy users rather than reflect accurate or evidence-based information. This is consistent with observed behavioural patterns in the quantitative analysis, where moderate usage persists despite low trust.

Finally, at the socio-technical level (Socio-Technical and Institutional Risks), participants express concerns about the broader implications of AI adoption, particularly in relation to established authority structures. The potential displacement of domain experts reflects a deeper concern about legitimacy, accountability, and the appropriate role of AI in high-stakes contexts.

These results collectively suggest that participants do not evaluate AI solely based on performance, but rather through a combination of epistemic, behavioural, and institutional considerations. This multi-dimensional perspective is critical for understanding how AI systems are perceived in domains where accuracy, authority, and contextual reasoning are essential.

Fig. 9 presents a visual representation of the framework and layered taxonomy of perceived risks identified in this research. The figure complements Table XI by organizing the parameters into four macro-parameters that reflect distinct but interrelated layers of a socio-technical system. Specifically, the taxonomy illustrates how risks originate at the epistemic and knowledge integrity level, propagate through contextual and reasoning limitations, influence human–AI interaction dynamics, and ultimately extend to broader socio-technical and institutional impacts. This layered structure highlights how lower-level technical limitations in AI outputs can propagate through reasoning and user interaction, ultimately shaping broader institutional and governance outcomes.

B. Governance Expectations for AI Systems

In addition to identifying risks, participants articulated a set of expectations regarding how AI systems should be designed, deployed, and governed in order to be considered acceptable and trustworthy. These expectations reflect a constructive perspective, indicating that users do not reject AI outright but instead advocate for its controlled and responsible use. These expectations also implicitly relate to privacy concerns identified in the quantitative analysis, particularly in scenarios where users may prefer AI-mediated interactions for sensitive inquiries, while remaining uncertain about data handling and confidentiality.

The analysis revealed two primary parameters of gover-

nance expectations: 1) Transparency and verification mechanisms, and 2) Human oversight and institutional governance.

The first category emphasizes the need for AI-generated responses to be supported by clear, verifiable, and authoritative sources. Participants consistently highlighted the importance of transparency in how responses are generated, particularly the inclusion of reliable references that can be independently verified. This expectation directly corresponds to concerns related to epistemic unreliability and lack of authoritative grounding, suggesting that users seek mechanisms that enhance trust through evidence and traceability.

The second category focuses on the role of human expertise and institutional oversight. Participants emphasized that AI systems should not operate independently in high-stakes domains, but rather under the supervision of qualified experts and recognized institutions. This reflects a strong preference for maintaining human accountability and preserving established authority structures, particularly in contexts where interpretive judgment is required.

These governance expectations demonstrate that participants envision a hybrid model of AI use, in which AI systems function as supportive tools rather than autonomous decision-makers. The alignment between identified risks and proposed governance mechanisms highlights a coherent user perspective: the same concerns that reduce trust in AI also inform the conditions under which AI may become acceptable.

VI. Discussion

This research provides a comprehensive understanding of how individuals perceive the use of AI systems in high-stakes, interpretive domains by integrating quantitative and qualitative analyses, with a particular focus on religious inquiry in Saudi Arabia. The results reveal a consistent and nuanced pattern: participants engage with AI-generated guidance in a limited and cautious manner, characterized by moderate usage, low trust, and high perceived risk, while still recognizing the potential utility of these systems under constrained conditions.

The quantitative results establish the structural patterns underlying this behaviour. Participants reported only occasional use of AI for religious inquiries, despite widespread exposure to such tools. Trust emerged as a central determinant of behavioural intention, with higher trust associated with increased willingness to use AI and, in some cases, to follow its outputs even when they conflict with trusted human experts. However, overall trust levels remained low, while perceived risk remained high. This combination indicates not rejection, but selective and conditional engagement. Participants do not treat AI as an authoritative source, but neither do they disregard it entirely. Instead, they adopt a cautious usage model in which AI is consulted but not relied upon without verification. This is further reinforced by strong preferences for source credibility, where the majority of participants actively verify AI-generated outputs and express a clear expectation that responses be linked to recognized scholarly authorities.

These patterns are further shaped by demographic and knowledge-based factors. Younger participants demonstrate greater openness to AI use, particularly in sensitive contexts, suggesting that accessibility and privacy-related affordances may lower barriers to engagement. In contrast, participants with stronger domain knowledge exhibit lower trust, higher perceived risk, and more rigorous verification behaviour. This indicates that familiarity with the domain introduces critical evaluation mechanisms that limit reliance on AI-generated outputs. Together, these results highlight that engagement with AI is not uniform, but mediated by both technological familiarity and epistemic grounding.

While the quantitative results describe these behavioural patterns, the qualitative analysis explains their underlying structure. The taxonomy of perceived risks reveals that participants evaluate AI systems across multiple interconnected dimensions rather than isolated concerns. At the epistemic level, concerns focus on correctness, evidence, and reliability, including risks of weak grounding, fabricated information, and inconsistent interpretations. At the reasoning level, participants highlight limitations in contextual understanding, recognizing that AI systems cannot adequately account for situational nuance or case-specific factors. At the interaction level, concerns shift toward behavioural dynamics, including overdependence and preference alignment, where AI systems may produce responses that align with user expectations rather than objective truth, reflecting emerging concerns around AI sycophancy. At the socio-technical level, participants express concern about the broader implications of AI adoption, particularly the potential displacement of domain authority and the erosion of established structures of expertise.

The integration of these results represents a key contribution of this work. Rather than treating trust, risk, and behaviour as independent variables, the results demonstrate that they are structurally linked. Low trust and high perceived risk observed in the quantitative analysis are directly explained by epistemic and reasoning concerns identified in the qualitative results. Similarly, observed verification behaviours correspond to concerns about credibility and lack of authoritative grounding, while conditional acceptance of AI aligns with expectations for human oversight and governance. This layered interpretation shows that user perception of AI is not driven solely by system performance, but by how technical limitations propagate through reasoning, interaction, and institutional contexts.

These results should be understood within a broader transformation in how knowledge and authority are mediated in digital environments. Existing literature suggests that AI in such domains does not merely expand access to information but reshapes how authority is constructed, circulated, and evaluated [29]. Prior work has shown that users increasingly question the credibility, contextual relevance, and institutional grounding of mediated guidance, particularly in domains where expert interpretation and contextual reasoning are essential [30]. The present research extends this literature by providing empirical evidence from Saudi Arabia on how users actively negotiate this shift in practice: while AI systems improve accessibility and immediacy, their acceptance remains conditional on trust, verification, and alignment with recognized sources of authority.

The results also align with and extend recent work on AI-generated religious content and Islamic knowledge systems. Alam et al. [11] demonstrate, in a study of Muslim American users, that users may prefer AI-generated religious answers even when expert evaluation identifies significant quality con-

cerns, highlighting the gap between perceived usefulness and actual reliability. The present results extend this understanding by examining similar dynamics within Saudi Arabia, where expectations of authority, trust, and verification are shaped by a distinct religious, cultural, and institutional context. Similarly, Priantina et al. [12] show that while AI can improve efficiency in religious ruling related processes, it raises important concerns about automating scholarly reasoning and weakening human oversight. Bhatia et al. further identify broader challenges in AI-based Islamic knowledge systems, including trustworthiness, evaluation standards, pluralism, and governance gaps [13]. This research contributes to these discussions by shifting the focus from system performance to user perception, demonstrating how trust, risk, privacy, credibility, and authority are evaluated in real-world use rather than controlled system settings.

More broadly, the results connect with established challenges in large language models and generative AI. Participants' concerns about inaccurate or misleading outputs correspond directly to the well-documented problem of hallucination, where systems generate fluent but factually incorrect or unsupported information [4], [27]. In high-stakes contexts, such outputs may be interpreted as authoritative despite lacking reliable grounding. Similarly, concerns about preference alignment reflect emerging work on AI persuasion and safety, where generative systems may shape user beliefs and decisions rather than simply provide neutral information [5], [31]. These risks are consistent with broader taxonomies of language-model harms, including misinformation, bias, privacy risks, and interaction-level failures [32], [33].

The results also resonate with literature on trust, explanation, and human–AI interaction. Research on trust in automation shows that achieving appropriate reliance is inherently difficult, as users may either over-rely on automated systems or reject them entirely depending on perceived reliability [34]. In contrast, this research identifies a more nuanced behavioral pattern characterized by cautious engagement and active verification, suggesting the emergence of calibrated trust in high-stakes contexts. At the same time, explainability research indicates that explanations can improve user understanding but may also increase acceptance of incorrect outputs when users lack sufficient expertise to critically evaluate them [35], [36]. This highlights a critical implication: transparency and explainability are necessary but not sufficient. In interpretive domains, they must be complemented by expert validation, institutional oversight, and mechanisms for accountability.

From a theoretical perspective, this research contributes a socio-technical model of AI risk perception in which concerns emerge across interconnected layers rather than isolated system failures. By organizing perceived risks into epistemic, reasoning, interaction, and institutional dimensions, the proposed taxonomy provides a structured framework for understanding how technical limitations in AI systems propagate into behavioral and governance challenges. This extends existing work on foundation models and AI risk by grounding abstract categories in empirically observed user perceptions and linking system-level behaviour with real-world interpretation and use [32], [33], [37]. In doing so, the research bridges technical AI research with human-centered and governance-oriented perspectives on trust, authority, and decision-making.

The results also have important implications for AI governance and system design [38], [39]. Participants consistently emphasized the need for transparency, particularly through the provision of verifiable and authoritative sources. However, the results suggest that transparency alone is insufficient in high-stakes interpretive contexts. Users not only expect explanations, but also require mechanisms to assess their validity, especially when they lack the expertise to independently verify the information provided. This reinforces the need for governance frameworks that combine explainability with validation, accountability, and oversight [35], [36], [37]. In addition, participants consistently rejected the idea of AI systems functioning as autonomous authorities. Instead, they expressed a strong preference for models in which AI operates under human supervision, particularly by domain experts. This aligns with broader risk-based governance approaches, where human-in-the-loop mechanisms are considered essential for high-risk applications [34], [37].

Importantly, the results highlight a tension between accessibility and authority. While AI systems increase access to information and provide immediate responses, they do not inherently carry the legitimacy associated with expert judgment. Participants' emphasis on verification and reliance on trusted sources indicates that authority remains a central condition for acceptance, even in the presence of highly accessible AI tools. Rather than replacing established structures of expertise, AI systems are therefore more likely to be accepted as assistive tools that complement, rather than substitute, human authority. This reflects a broader pattern observed in other high-stakes domains, where users value the efficiency and clarity of AI systems but continue to depend on expert validation for final decisions [40].

From a practical perspective, the results suggest that AI systems in high-stakes advisory domains require governance models that combine verifiable source grounding, human oversight, and institutional accountability. These expectations align with broader concerns about foundation models, where scale, opacity, and downstream deployment amplify risks across domains [37]. They also reflect patterns observed in other domains such as healthcare, where users value AI-generated responses for accessibility and clarity but continue to rely on human experts for validation and safe decision-making [40]. Together, these results point toward a hybrid governance model that supports human autonomy, enables calibrated trust, and preserves the role of domain expertise.

Finally, while this research is situated within a specific cultural and religious context, the underlying dynamics are likely to extend to other high-stakes interpretive domains, including healthcare, legal advisory, education, and policy decision-making, where similar concerns around trust, expertise, and decision-making have been observed [40], [1], [2]. In these contexts, trust, privacy, contextual reasoning, authority, and institutional legitimacy remain central to responsible AI deployment. Accordingly, the proposed framework should be understood as a structured analytical foundation for examining how technical, behavioural, and institutional risks interact in high-stakes interpretive settings, rather than as a universally validated model across all domains. The research, therefore, contributes to a broader shift in AI research from system-centric performance toward human-centred and governance-

oriented understanding of AI in practice.

VII. Conclusion

This research provides a comprehensive examination of how individuals perceive the use of AI systems in high-stakes interpretive domains, using the context of religious inquiries in Saudi Arabia as an empirical setting. By integrating quantitative and qualitative analyses, the results reveal a consistent pattern characterized by moderate engagement with AI tools, low levels of trust, and heightened perceptions of risk. Users demonstrate a cautious and conditional approach, treating AI-generated responses as supportive resources rather than authoritative sources.

The results highlight that trust, perceived risk, and source credibility are central determinants of user behaviour. Participants consistently emphasized the importance of verifiable and authoritative grounding, frequently engaging in source verification and expressing strong preference for responses linked to recognized experts. At the same time, reluctance to rely on AI for sensitive or complex inquiries underscores the context-dependent nature of AI acceptance. Demographic factors, particularly age and domain knowledge, further shape how individuals interpret, evaluate, and interact with AI-generated outputs.

Beyond descriptive insights, this research advances understanding by introducing a structured, multi-layered taxonomy of perceived risks, spanning epistemic, reasoning, interactional, and institutional dimensions. This framework demonstrates that concerns are not limited to technical limitations, but extend to broader issues of authority, accountability, and governance. The alignment between identified risks and user expectations further reveals that acceptance of AI is contingent on the presence of transparency, human oversight, and institutional anchoring.

These results carry important implications for the design and governance of AI systems in high-stakes contexts. Enhancing transparency through verifiable sourcing, incorporating human-in-the-loop mechanisms, and establishing clear governance boundaries are essential for building trustworthy and socially aligned AI systems. Rather than replacing established structures of expertise, AI systems are more likely to be accepted when positioned as assistive tools operating within well-defined ethical and institutional frameworks.

While this research is grounded in a specific cultural and religious context, the underlying dynamics are likely to extend to other domains where interpretive judgment, trust, and accountability are critical. At the same time, the participant distribution was skewed toward younger and relatively educated respondents within a single national setting, which may influence the broader generalizability of the reported perceptions and behavioral patterns. This positions the work within a broader discourse on responsible AI, highlighting the need to align technological capabilities with human expectations, domain expertise, and societal values, while also motivating future cross-cultural and demographically diverse investigations.

Future research can build on these results by exploring cross-cultural variations, longitudinal shifts in trust, and experimental evaluations of AI systems integrated with expert validation mechanisms. Additional directions include real-world system evaluations, adversarial robustness testing of AI-generated guidance, and longitudinal trust calibration studies examining how user perceptions evolve through repeated interaction with AI systems over time.

To conclude, AI systems hold significant potential to enhance access to complex knowledge domains. However, their acceptance depends not only on technical performance, but also on their ability to operate within trusted epistemic and institutional structures. The challenge, therefore, is not merely to improve AI systems, but to design them in ways that respect, support, and reinforce the foundations of human expertise and judgment.

Acknowledgment

This study is derived from a research grant funded by the Research, Development, and Innovation Authority (RDIA), Kingdom of Saudi Arabia, with grant number 12615-iu-2023-IU-R-2-1-EI-.

References

[1] S. Shekar, P. Pataranutaporn, C. Sarabu, G. A. Cecchi, and P. Maes, "People over-trust ai-generated medical responses and view them to be as valid as doctors, despite low accuracy," *arXiv preprint arXiv:2408.15266*, Aug 2024.

[2] E. Schneiders, T. Seabrooke, J. Krook, R. Hyde, N. Leesakul, J. Clos, and J. E. Fischer, "Objection overruled! lay people can distinguish large language models from lawyers, but still favour advice from an llm." Association for Computing Machinery, 4 2025. [Online]. Available: https://dl.acm.org/doi/pdf/10.1145/3706598.3713470

[3] T. Seabrooke, E. Schneiders, L. Dowthwaite, J. Krook, N. Leesakul, J. Clos, H. Maior, and J. Fischer, "A survey of lay people's willingness to generate legal advice using large language models (llms)," *Proceedings of the Second International Symposium on Trustworthy Autonomous Systems*, 9 2024.

[4] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Comput. Surv.*, vol. 55, no. 12, Mar. 2023. [Online]. Available: https://doi.org/10.1145/3571730

[5] H. Kong, "Persuasion and safety in the era of generative ai," in *WebSci PhD Symposium*, 2025.

[6] C. Cuskley, R. Woods, and M. Flaherty, "The limitations of large language models for understanding human language and cognition," *Open Mind*, vol. 8, pp. 1058–1083, 08 2024. [Online]. Available: https://doi.org/10.1162/opmi_a_00160

[7] N. Yousif, "Parents of teenager who took his own life sue openai," *BBC News*, 9 2025. [Online]. Available: https://www.bbc.com/news/articles/cgerwp7rdlvo

[8] R. Alsaigh, R. Mehmood, I. Katib, and T. Yigitcanlar, "Governing ai in society: Explainable analysis of research using the pearl methodology, the frame-ai framework, and regulatory alignment gaps," *SSRN Electronic Journal*, Dec. 2025. [Online]. Available: https://ssrn.com/abstract=6009614

[9] R. Alsaigh, R. Mehmood, I. Katib, X. Liang, A. Alshanqiti, J. M. Corchado, and S. See, "Harmonizing AI governance regulations and neuroinformatics: perspectives on privacy and data sharing," *Frontiers in Neuroinformatics*, vol. Volume 18 - 2024, 2024. [Online]. Available: https://www.frontiersin.org/journals/neuroinformatics/articles/10.3389/fninf.2024.1472653

[10] S. A. Al-Tayyar, "Fatwa and its importance." [Online]. Available: https://draltayyar.com/books/7963/

[11] S. M. Alam, M. Abdulhai, and N. Salehi, "Blind faith? user preference and expert assessment of ai-generated religious content," in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2025, pp. 1–29.

[12] A. Priantina, M. M. Uula, Aufa, and E. Herindara, "Ai in fatwa formulation: Transforming sharia-compliant finance," *Journal of Central Banking Law and Institutions*, vol. 4, no. 3, pp. 595–634, 2025.

[13] G. Bhatia, H. Mubarak, M. Hawasly, M. Jarrar, G. Mikros, F. Zaraket, M. Alhirthani, M. Al-Khatib, L. Cochrane, K. Darwish, R. Yahiaoui, and F. Alam, "Advances in ai systems on islamic knowledge capabilities: A critical survey," *arXiv preprint*, 2026, available at: https://gagan3012.github.io/islamic-knowledge-survey/.

[14] O. Alyemny, H. Al-Khalifa, and A. Mirza, "A data-driven exploration of a new islamic fatwas dataset for arabic nlp tasks," *Data 2023, Vol. 8, Page 155*, vol. 8, p. 155, 10 2023. [Online]. Available: https://www.mdpi.com/2306-5729/8/10/155/htmhttps://www.mdpi.com/2306-5729/8/10/155

[15] A. Y. Alan, E. Karaarslan, and Ömer Aydin, "A rag-based question answering system proposal for understanding islam: Mufassirqas llm," *MufassirQAS. Turkish Journal of Engineering*, vol. 9, pp. 544–559, 1 2024. [Online]. Available: https://arxiv.org/pdf/2401.15378

[16] M. Y. Mohammed, S. A. Ali, S. K. Ali, A. A. Majeed, and E. H. Mohamed, "Aftina: enhancing stability and preventing hallucination in ai-based islamic fatwa generation using llms and rag," *Neural Computing and Applications*, vol. 37, pp. 20 957–20 982, 9 2025. [Online]. Available: https://link.springer.com/article/10.1007/s00521-025-11229-y

[17] M. Grootendorst, "Bertopic: Neural topic modeling with a class-based tf-idf procedure," *arXiv preprint arXiv:2203.05794*, 2022.

[18] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection," *Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018.

[19] L. McInnes, J. Healy, and S. Astels, "hdbscan: Hierarchical density based clustering," *Journal of Open Source Software*, vol. 2, no. 11, p. 205, 2017.

[20] "General presidency of scholarly research and ifta." [Online]. Available: https://alifta.gov.sa/home

[21] "Bin othaimeen." [Online]. Available: https://binothaimeen.net/site

[22] "Fatawa pedia." [Online]. Available: https://fatawapedia.com/

[23] H. Huang *et al.*, "Acegpt, localizing large language models in arabic," *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*, vol. 1, pp. 8132–8156, 9 2023. [Online]. Available: https://arxiv.org/pdf/2309.12053

[24] P. Georgiev *et al.*, "Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context," 3 2024. [Online]. Available: https://arxiv.org/pdf/2403.05530

[25] "Silma ai: The leading arabic ai & llm technology provider." [Online]. Available: https://silma.ai/

[26] B. Wen *et al.*, "Know your limits: A survey of abstention in large language models." [Online]. Available: https://doi.org/10.1162/tacl

[27] F. Atif, N. Askarbekuly, K. Darwish, and M. Choudhury, "Sacred or synthetic? evaluating llm reliability and abstention for religious questions," 8 2025. [Online]. Available: https://arxiv.org/pdf/2508.08287

[28] G. Inoue, B. Alhafni, R. Baly, F. Zaraket, and N. Habash, "The interplay of variant, size, and task type in arabic pre-trained language models," in *Proceedings of the Sixth Arabic Natural Language Processing Workshop*, 2021, pp. 92–104.

[29] F. A. Atallah, "Digital mediation and fatwa authority in contemporary islam: A critical islamic legal and media-theoretical framework," *Religions*, vol. 17, no. 350, 2026.

[30] S. Whyte, "Are fatwas dispensable? examining the contemporary relevance and authority of fatwas in australia," *Oxford Journal of Law and Religion*, vol. 11, pp. 314–342, 2022.

[31] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2021.

[32] L. Weidinger, J. Mellor, M. Rauh, C. Griffin, J. Uesato, P.-S. Huang *et al.*, "Ethical and social risks of harm from language models," *arXiv preprint arXiv:2112.04359*, 2021.

[33] L. Weidinger *et al.*, "Taxonomy of risks posed by language models," in *ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, 2022.

[34] J. D. Lee and K. A. See, "Trust in automation: Designing for appropriate reliance," *Human Factors*, vol. 46, no. 1, pp. 50–80, 2004.

[35] M. T. Ribeiro, S. Singh, and C. Guestrin, ""why should i trust you?": Explaining the predictions of any classifier," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, 2016, pp. 1135–1144.

[36] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, and D. S. Weld, "Does the whole exceed its parts? the effect of ai explanations on complementary team performance," in *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 2021.

[37] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx *et al.*, "On the opportunities and risks of foundation models," Stanford Center for Research on Foundation Models (CRFM), Tech. Rep., 2021, arXiv:2108.07258.

[38] R. Alsaigh, R. Mehmood, and I. Katib, "AI explainability and governance in smart energy systems: A review," *Frontiers in Energy Research*, vol. Volume 11 - 2023, 2023. [Online]. Available: https://www.frontiersin.org/journals/energy-research/articles/10.3389/fenrg.2023.1071291

[39] R. Alsaigh, R. Mehmood, I. Katib, A. A. Almuzaini, S. S. Albouq, and S. Alshmrany, "Evidence-Driven AI Governance for Healthcare: A PEARL-PATHWAY Analysis of Madinah," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 17, no. 4, 2026. [Online]. Available: https://thesai.org/Publications/ViewPaper?Volume=17&Issue=4&Code=IJACSA&SerialNo=XX

[40] S. K. Beale, N. Cohen, B. Secheli, D. McIntire, and K. A. Kho, "Comparing physician and artificial intelligence chatbot responses to posthysterectomy questions posted to a public social media forum," *AJOG Global Reports*, vol. 5, p. 100553, 2025.