QCRI
معهد قطر لبحوث الحوسبة
Qatar Computing Research Institute
جامعة حمد بن خليفة
HAMAD BIN KHALIFA UNIVERSITY

جامعة
حمد بن خليفة
HAMAD BIN KHALIFA
UNIVERSITY

AI Systems for Islamic Knowledge:
Capabilities, Evaluation, and Trustworthiness

Lessons from Islamic Knowledge Question Answering

Gagan Bhatia
University of Technology Nuremberg, Qatar Computing Research Institute

EMNLP2026
BUDAPEST

https://islamic-media-ethics-ai.github.io/

Fluent Answers Are Not Always Trustworthy

Example: Islamic Jurisprudence

Question (EN): When does the time for Fajr prayer begin?
Gold answer (EN): The time for Fajr prayer begins at true dawn.
Model: Qwen3-4B + RAG
Predicted answer (EN): Fajr begins at dawn, i.e., sunrise time when daylight starts to appear.
Judge: Incorrect
Model: Qwen3-4B + Agentic RAG
Predicted answer (EN): Fajr begins at true dawn (*al-fajr al-ṣādiq*), i.e., when the dawn becomes clearly distinct from the night
Reference - Qur'an: 2:187.

Judge: Correct

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

Fluent Answers Are Not Always Trustworthy

A case study in culturally grounded, domain-specific alignment

Low-resource

High-stakes

“AI systems are increasingly mediating how communities access, study, and apply Islamic sources”

SUCH SYSTEMS CAN

1. hallucinate citations and fabricate attributions
2. collapse nuanced juristic disagreement into a single “answer”
3. reproduce cultural or religious stereotypes

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

What a Decade of Research Reveals?

THE SURVEY · 160 INCLUDED PAPERS

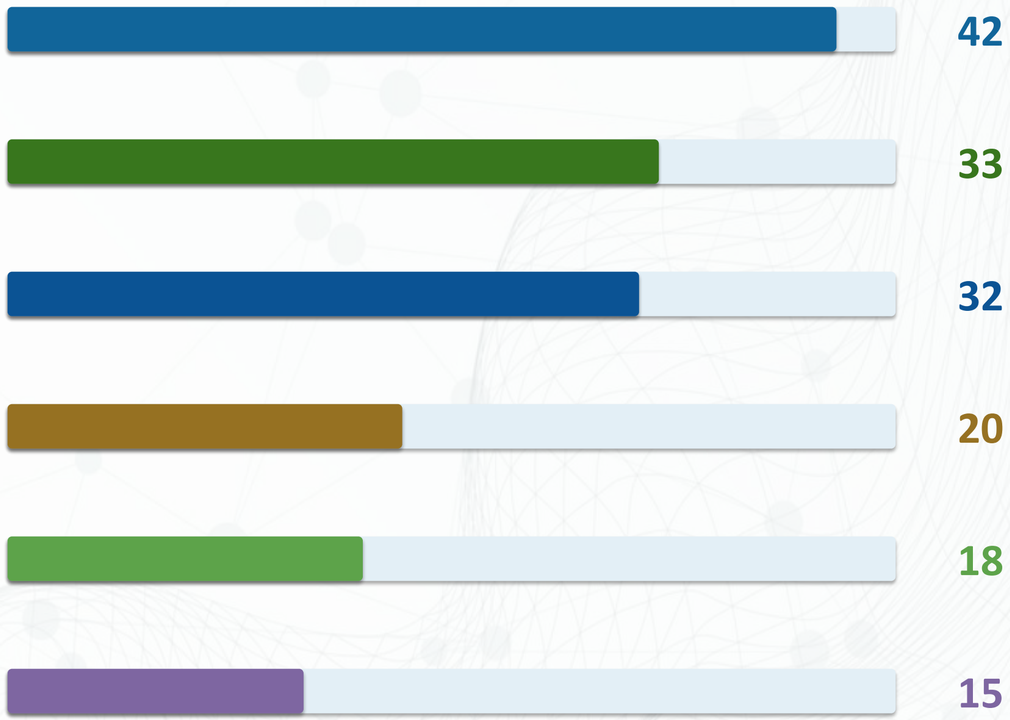
160

papers reviewed

NLP · retrieval · reasoning evaluation · multimodal AI

2016–2025 literature

Research streams



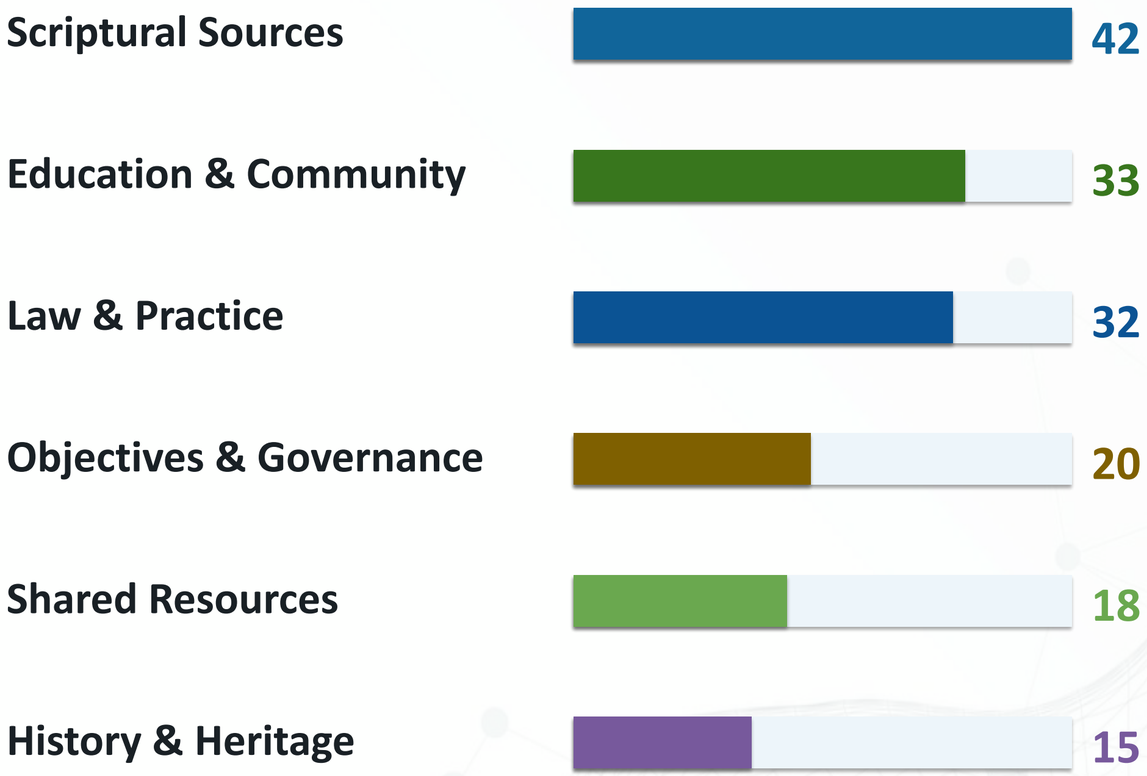
Retrieval and applications are growing; provenance, disagreement handling, and evaluation need more attention.

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

What a Decade of Research Reveals?

Our systematic survey reviewed 160 papers across Islamic knowledge applications.

Research streams



Domains within each stream

Scriptural Sources

Quran & Tafsir 22
Hadith Sciences 20

Education & Community

Islamic Education 21
Contemporary Discourse 12

Law & Practice

Fiqh & Fatwa 20
Rituals 8 · Manuscripts 4

Objectives & Governance

Maqasid Lens 10
Doctrinal Integrity 10

Shared Resources

Datasets & Benchmarks 18

History & Heritage

Sirah & History 10
Manuscripts 5

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

What a Decade of Research Reveals?

THE SURVEY · LAYERED TAXONOMY

1 EPISTEMIC LAYER · WHAT IS MODELLED

Foundations

- **Qur’an**
- **Hadith**

Disciplines (ʿUlūm)

- Qur’anic Sciences
- Hadith Sciences
- Usul al-Fiqh
- Fiqh
- Theology (Kalam)
- Sufism (Tasawwuf)
- History & Sirah

2 AI TASK LAYER · HOW IT IS COMPUTED

Data & Corpora

Digitization
Corpus building

Retrieval & Grounding

Search / IR
RAG & evidence

Understanding

Extraction
Summarization

Reasoning Support

Comparative (school-aware)
Uncertainty & deferral

Evaluation & Governance

Benchmarks
Safety & integrity

Multimodal

Speech / audio
Image-text

3 CROSS-CUTTING NORMATIVE DIMENSIONS

- **Doctrinal integrity & authenticity**
- **Normative correctness & disagreement handling**
- **Objectives, harms & governance**

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

What a Decade of Research Reveals?

Trustworthiness has three dimensions

THE SURVEY · LAYERED TAXONOMY

1 KNOWLEDGE

What is being represented?

- Qur’an and Hadith
- Jurisprudence and legal theory
- Theology, history, heritage

2 AI CAPABILITIES

How is it processed?

- Curated data and retrieval
- Grounded QA and reasoning
- Evaluation and multimodal methods

3 NORMATIVE CONSTRAINTS

What counts as acceptable?

- Source authenticity and provenance
- Pluralism and qualified claims
- Uncertainty, deferral, and safety

The same text can require different computational methods, evidence standards, and forms of human oversight.

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

Reliable Islamic QA

THE SURVEY · SYSTEM DESIGN PRINCIPLES

01 Retrieve with provenance

Curated sources, authority-aware retrieval, and traceable citations

Risk: Source fabrication

02 Verify evidence

Check claim support, citation attribution, and ambiguous references

Risk: Unsupported claims

03 Use tools for exact tasks

Deterministic computation for inheritance, dates, and numerical questions

Risk: Arithmetic drift

04 Respect plurality and limits

Surface legitimate disagreement; ask, abstain, or defer when needed

Risk: Overconfidence

Trustworthiness depends on a verifiable system stack—not just an instruction to “be accurate.”

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

From Generic LLMs to Evidence-Grounded Systems

THE SURVEY · DESIGN PATTERNS

	WHAT GOES WRONG	PATTERN THAT ADDRESSES IT
1	Fabricated or misattributed sources	→ Evidence-first RAG + citation contract
2	Weak sources and context collapse	→ Authority-aware retrieval & filtering
3	Composite questions drift across sub-claims	→ Query decomposition + multi-hop retrieval
4	Hallucinated, unsupported claims	→ Verification layer (citation + entailment)
5	Overconfident answers under ambiguity	→ Calibrated abstention / deferral
6	Arithmetic and logic errors on rule-based tasks	→ Tool-augmented deterministic reasoning
7	Disagreement collapse	→ Pluralism-aware answering
8	Toxic or stigmatizing content	→ Safety + etiquette (adab) guardrails

Bhatia et al., Advances in AI Systems on Islamic Knowledge Capabilities: A Critical Survey, TechRxiv

IslamicFaithQA

BENCHMARK · BILINGUAL OPEN-ENDED EVALUATION

3,810
Arabic–English QA items

70.7%
Require reasoning

55.4%
Require multi-step reasoning

One open-ended response, three distinct judgments

CORRECT
Answer matches the reference

INCORRECT
Answer conflicts with or misses it

NOT ATTEMPTED
Question is left unresolved

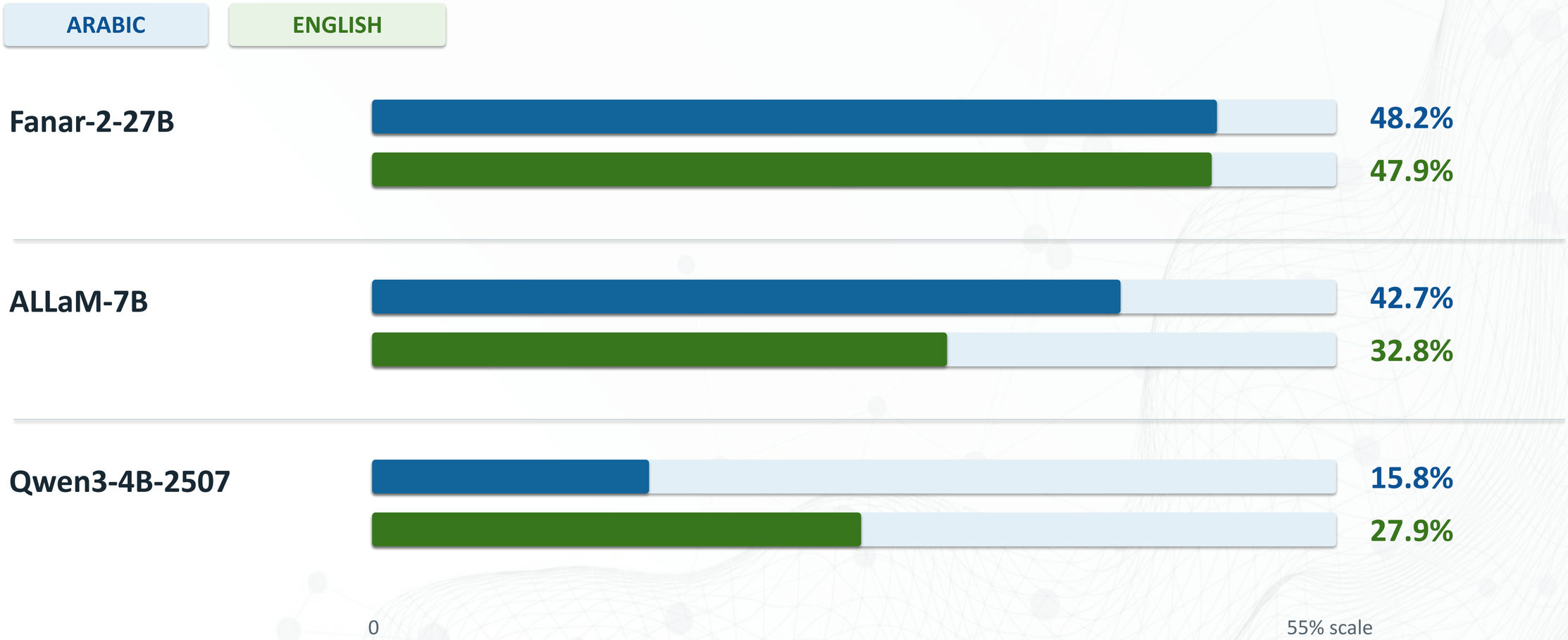
Scope limitation: single-gold atomic answers do not yet represent questions with several valid scholarly positions.

Correctness, hallucination, and non-attempted behavior should not be collapsed into one accuracy metric.

Bhatia et al., From RAG to Agentic RAG for Faithful Islamic Question Answering, ACL Findings 2026

Reliability is not Language-neutral

ISLAMICFAITHQA · ZERO-SHOT BASELINES

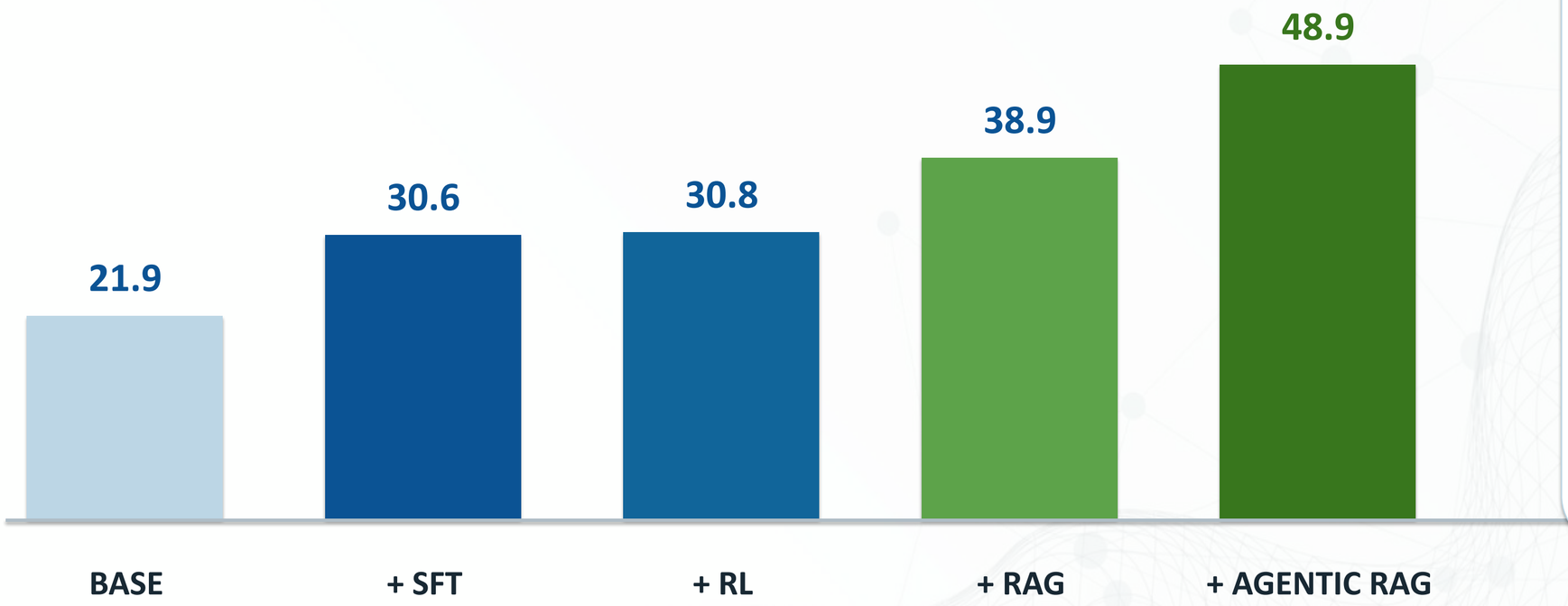


Model rankings and error patterns differ by language—even when questions target the same knowledge domain.

Bhatia et al., From RAG to Agentic RAG for Faithful Islamic Question Answering, ACL Findings 2026

Agentic Evidence Seeking Improves Correctness

FROM RAG TO AGENTIC RAG · QWEN3-4B-2507



WHAT CHANGED?

SEARCH	Find relevant evidence
INSPECT	Read specific verses
REFINE	Revise before answering

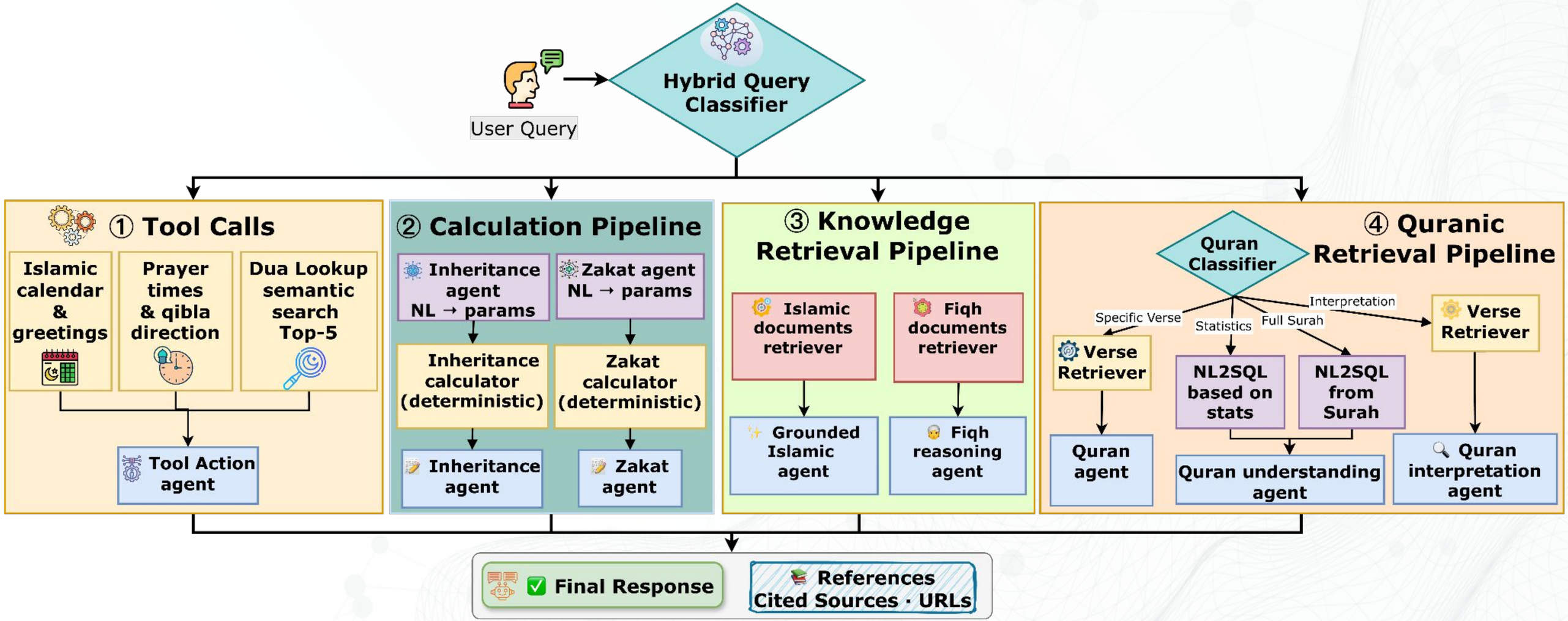
+10.05 pts

Cost of agentic RAG · ~6 tool calls/query · 1,520 vs 502 tokens · 3.45 s vs 0.52 s latency

Grounded tool use substantially improves measured correctness—but introduces a real efficiency trade-off.

Bhatia et al., From RAG to Agentic RAG for Faithful Islamic Question Answering, ACL Findings 2026

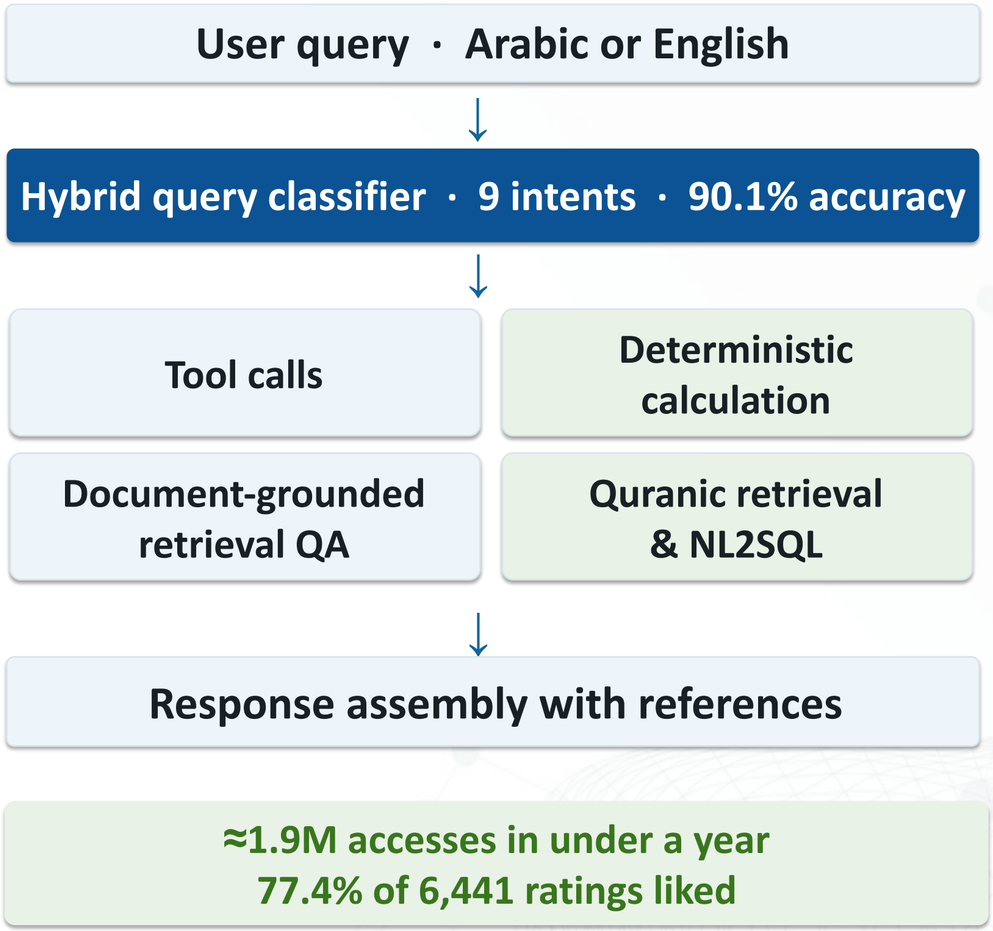
Fanar-Sadiq: Multi-agent Architecture



Abbas et al., Fanar-Sadiq: A Multi-Agent Architecture for Grounded Islamic QA, ACL 2026

Fanar-Sadiq: Multi-agent Architecture

Bilingual question routing



Accuracy (%)

Benchmark	GPT-5	Gemini 3 Pro	Fanar-2 27B	Fanar-Sadiq
PalmX	82.3	84.4	72.5	**85.5**
QIAS (Task 1)	93.0	94.5	63.5	72.2
IslamTrust	95.2	95.6	83.2	94.2
IslamicFaithQA	51.2	56.6	48.2	**65.4**
FatwaQA	63.6	67.0	44.5	65.1
Average	**77.1**	**79.6**	**62.4**	**76.5**

Abbas et al., Fanar-Sadiq: A Multi-Agent Architecture for Grounded Islamic QA, ACL 2026

Fanar-Sadiq: Real-World Deployment

https://chat.fanar.qa/

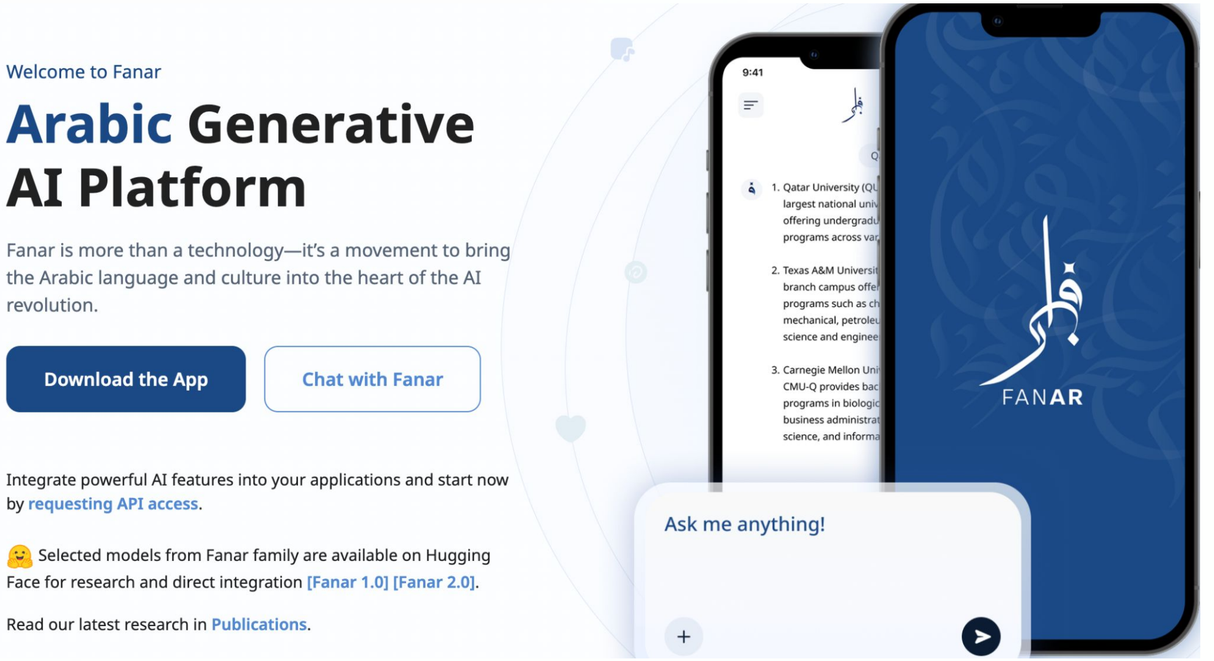
Integrated into the Fanar AI platform as a specialized Islamic backend

77.4 %

positive ratings (6,441 user feedbacks)

93.4

macro-F1 Islamic vs. non-Islamic classifier



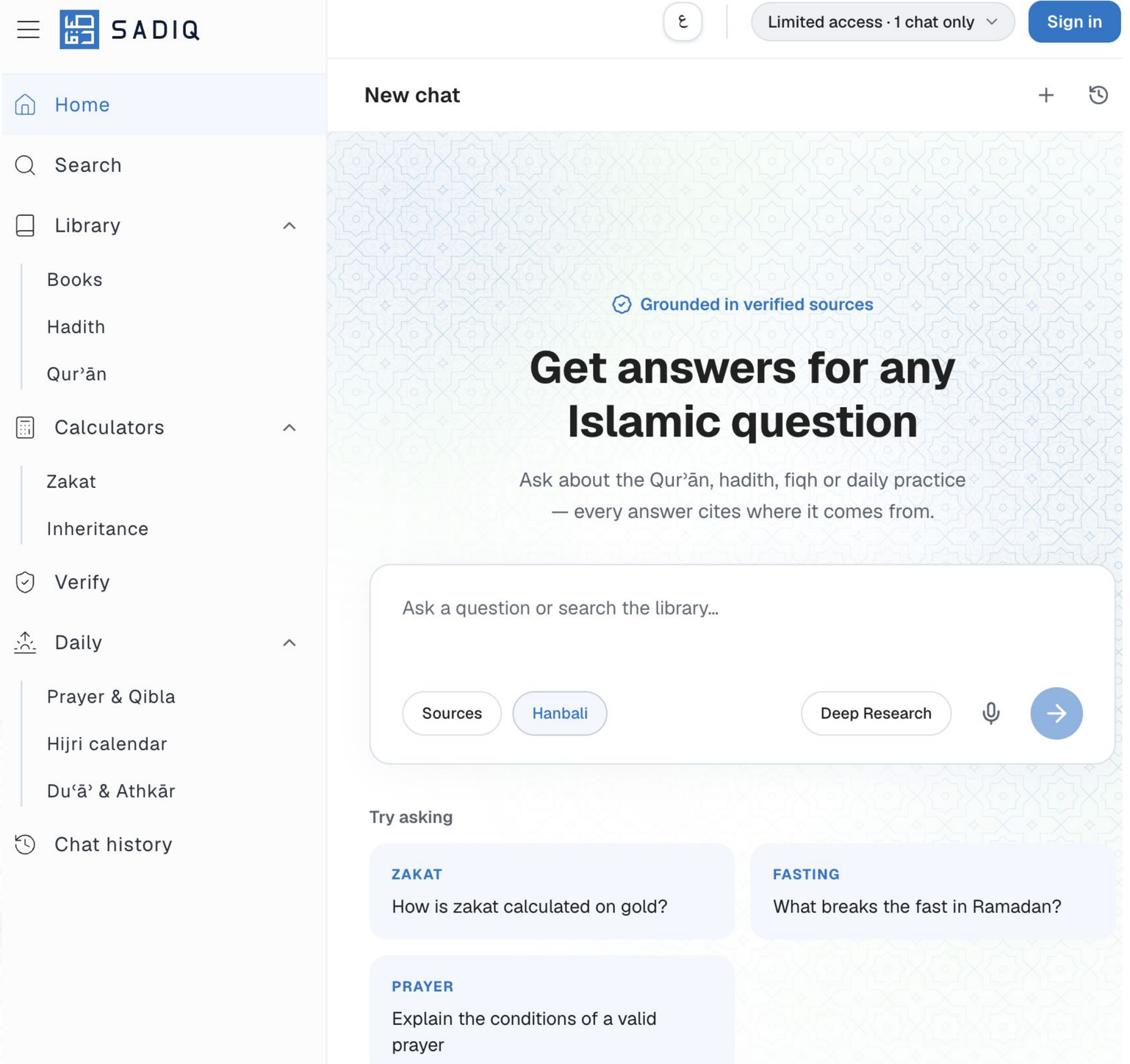
Routing in production. A fine-tuned binary classifier (1,700 annotated queries, Cohen's κ = 0.753) decides whether each incoming query is Islamic, Islamic queries flow into Fanar-Sadiq; the rest go to general-purpose assistants.

Abbas et al., Fanar-Sadiq: A Multi-Agent Architecture for Grounded Islamic QA, ACL 2026

Fanar-Sadiq

Real-World Deployment

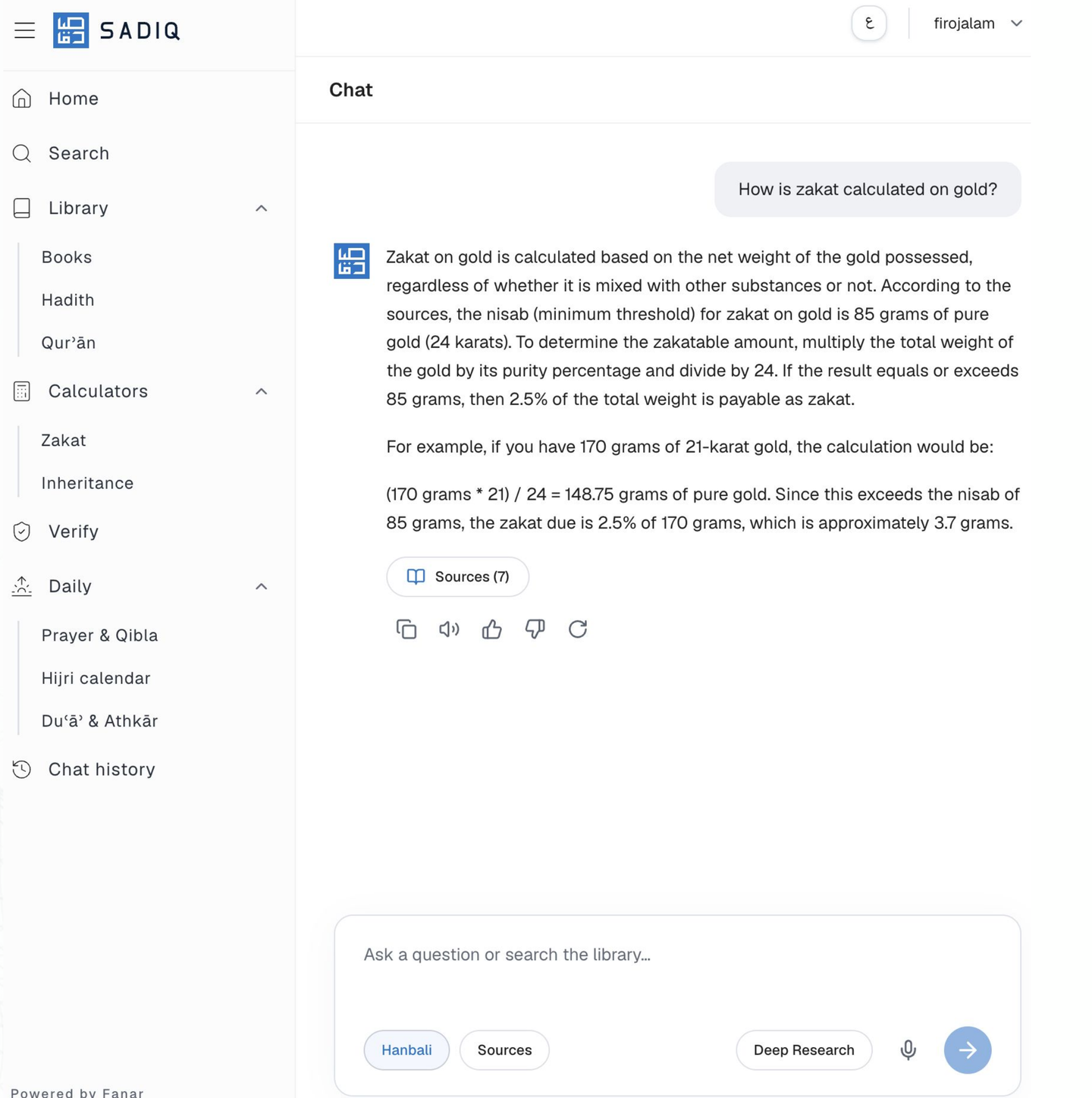




Abbas et al., Fanar-Sadiq: A Multi-Agent Architecture for Grounded Islamic QA, ACL 2026

Fanar-Sadiq

Real-World Deployment



Abbas et al., Fanar-Sadiq: A Multi-Agent Architecture for Grounded Islamic QA, ACL 2026

Takeaways and discussion

SURVEY · KEY TAKEAWAYS

KEY TAKEAWAYS

1. **Reliability over fluency**
2. **Citation is mandatory, not optional**
3. **A shift in evaluation standards**

FOR DISCUSSION

- Who decides which sources count as authoritative?
- How should benchmarks score legitimate disagreement?
- When should a system defer to a human expert?

gagan3012.github.io/islamic-knowledge-survey
huggingface.co/datasets/QCRI/IslamicFaithQA
https://sadiq.chat/