Islamic Knowledge Management
EISSN: 2476-7255

Toward Semantic Retrieval in Islamic Sciences: From Traditional Models to Modern Architectures*

Ali Mirarab
Assistant Professor, Department of Information and Knowledge Dissemination, Research Center for Islamic Documents and Information Management, Islamic Sciences and Culture Academy, Qom, Iran.
alimirarab@isca.ac.ir

Abstract

Keyword-based search engines inherently struggle to understand polysemy, synonymy, and complex conceptual relationships, and therefore often produce large numbers of irrelevant results. This problem is particularly acute in the field of Islamic sciences, where specialized terminology is multilayered, texts are bilingual (Persian and Arabic), and classical and modern forms of language are uniquely combined. Over the past decade, the emergence of new technologies such as the Semantic Web, knowledge graphs, large language models (LLMs), and Retrieval-Augmented Generation (RAG) architectures has created unprecedented opportunities for improving semantic retrieval in specialized domains. The present study aims to provide a systematic and up-to-date analysis of the challenges of semantic retrieval in Islamic sciences, critically examine existing approaches in light of

* Mirarab, A. (2025). Toward Semantic Retrieval in Islamic Sciences: From Traditional Models to Modern Architectures. *Islamic Knowledge Management, 7*(1), pp. 104-130. https://Doi.org/10.22081/jikm.2026.74693.1109

▣ **Article Type:** Research; **Publisher:** Islamic Sciences and Culture Academy, Qom, Iran

▣ **Received:** 2024/10/26 ● **Revised:** 2025/01/08 ● **Accepted:** 2025/02/09 ● **Published online:** 2025/03/30

emerging technologies, and propose an integrated evidence-based framework for improving the current state of the field. The study employs a systematic literature review and qualitative content analysis of 52 authoritative international and domestic sources published between 2014 and 2025. The findings indicate that the challenges of semantic retrieval can be classified into four levels: conceptual (polysemy, synonymy, and conceptual hierarchies); linguistic (Arabic morphological complexities, variations in Persian orthography, and bilingualism); technical (limitations of keyword-based approaches, inability to understand complex queries, and hallucinations in language models); and structural (lack of standardized ontologies, data fragmentation, and insufficient training data). The proposed framework consists of five complementary layers: (1) a structured OWL/SKOS ontology for Islamic sciences; (2) an integrated multilingual knowledge graph; (3) domain-adapted large language models; (4) a localized RAG system with hybrid retrieval; and (5) an intelligent search interface incorporating semantic query expansion. Empirical evidence from the SemanticHadith and FarsBase projects, AraBERT, the Fanar model, and findings presented at the QIAS 2025 conference indicates that localized hybrid systems can outperform even the most powerful general-purpose models in specialized Islamic-domain tasks.

Keywords
Semantic Retrieval; Islamic Sciences; Knowledge Graph; Large Language Model; Persian Language; Natural Language Processing.

إدارة المعرفة الإسلامية
EISSN: ۲۴۷۶-۷۲۵۵

نحو الاسترجاع الدلالي في العلوم الإسلامية: من النماذج التقليدية إلى المعماريات الحديثة الهندسية*

علي ميرعرب
أستاذ مساعد، قسم نشر المعلومات والمعرفة، معهد إدارة المعلومات والوثائق الإسلامية، المعهد العالي للعلوم والثقافة الإسلامية، قم، إيران.
alimirarab@isca.ac.ir

الملخص

تُسفر محركات البحث القائمة على الكلمات المفتاحية غالبًا عن أعداد هائلة من النتائج غير ذات الصلة، نظراً لعجزها البنيوي عن استيعاب المشترك اللفظي (تعدد المعاني)، والترادف، والعلاقات المفاهيمية المعقدة. وتبدو هذه المشكلة أكثر جلاءً وحدّة في حقل العلوم الإسلامية، الذي يتميز بمصطلحات تخصصية متعددة الطبقات، ونصوص ثنائية اللغة (الفارسية والعربية)، وتمازج فريد يجمع بين الأسلوبين الكلاسيكي والمعاصر. وعلى مدى العقد المنصرم، أتاح انبثاق تقنيات رائدة—مثل الويب الدلالي، ورسوم المعرفة البيانية، والنماذج اللغوية الكبيرة، ومعمارية «التوليد المعزز بالاسترجاع» فرصاً غير مسبوقة للارتقاء بفاعلية الاسترجاع الدلالي في المجالات التخصصية. تهدف هذه

* **الاستشهاد بهذه المقالة:** ميرعرب، علي. (۲۰۲۵). نحو الاسترجاع الدلالي في العلوم الإسلامية: من النماذج التقليدية إلى المعماريات الحديثة الهندسية. إدارة المعرفة الإسلامية، ۷(۱)، صص ۱۰۴-۱۳۰.
https://Doi.org/10.22081/jikm.2026.74693.1109

▣ **نوع المقال:** بحثية؛ **الناشر:** المعهد العالي للعلوم والثقافة الإسلامية، قم، إيران

▣ **تاريخ الإستلام:** ۲۰۲۴/۱۰/۲۶ ● **تاريخ التعديل:** ۲۰۲۵/۰۱/۰۸ ● **تاريخ القبول:** ۲۰۲۵/۰۲/۰۹ ● **تاريخ الإصدار:** ۲۰۲۵/۰۳/۳۰

الدراسة إلى تقديم تحليل منهجي ومُحَدَّث لتحديات الاسترجاع الدلالي في العلوم الإسلامية، مع فحص نقدي للمقاربات القائمة في ضوء التقنيات الناشئة، واقتراح إطار عمل متكامل مدعوم بالأدلة التجريبية للنهوض بالواقع الراهن. واعتمدت الدراسة منهجية المراجعة المنهجية للأدبيات والتحليل النوعي لمحتوى ٥٢ مرجعًا معتمدًا من المصادر الدولية والمحلية المنشورة بين عامي ٢٠١٤ و٢٠٢٥م. وتشير النتائج إلى إمكانية تصنيف تحديات الاسترجاع الدلالي ضمن أربعة مستويات: **المستوى المفاهيمي:** (المشترك اللفظي، الترادف، وتسلسل المفاهيم الهرمي). **المستوى اللغوي:** (التعقيدات الصرفية للغة العربية، وتباين الرسم والتقاليد الإملائية في الفارسية، والازدواج اللغوي). **المستوى التقني:** (محدودية المقاربات القائمة على الكلمات المفتاحية، والقصور عن فهم الاستعلامات المركبة، وظاهرة الهلوسة في النماذج اللغوية). **المستوى البنيوي:** (غياب أنطولوجيا معيارية موحدة، وتشتت البيانات، ونقص بيانات التدريب). ويتألف الإطار المقترح من خمس طبقات تكاملية: ١. أنطولوجيا مهيكلة بصيغتي OWL/SKOS مخصصة للعلوم الإسلامية. ٢. رسم معرفي بياني متكامل ومتعدد اللغات. ٣. نماذج لغوية كبيرة مُهيَّأة ومُطوَّعة بحسب المجال التخصصي (Domain-adapted). ٤. نظام RAG موطن يعتمد على الاسترجاع المركّب. ٥. واجهة بحث ذكية مزودة بتقنية التوسيع الدلالي للاستعلام. وتشير الشواهد والأدلة التجريبية المستمدة من مشروعي SemanticHadith وFarsBase، ونموذج AraBERT، ونموذج Fanar، إضافة إلى مخرجات مؤتمر QIAS ، إلى أن الأنظمة التركيبية المحليّة قادرة على التفوّق حتى على أقوى النماذج ذات الأغراض العامة عند أداء المهامّ التخصصيّة في الدراسات الإسلامية.

الكلمات المفتاحية

الاسترجاع الدلالي، العلوم الإسلامية، رسم المعرفة، النموذج اللغوي الكبير، اللغة الفارسية، معالجة اللغة الطبيعية.

مدیریت دانش اسلامی
EISSN: ۲۴۷۶-۷۲۵۵

حرکت به‌سوی بازیابی معنایی در علوم اسلامی: از مدل‌های سنتی تا معماری‌های مدرن*

علی میرعرب
استادیار، گروه اشاعهٔ اطلاعات و دانش پژوهشکدهٔ مدیریت اطلاعات و مدارک اسلامی پژوهشگاه علوم و فرهنگ اسلامی، قم، ایران.
alimirarab@isca.ac.ir

چکیده

موتورهای جستجوی کلیدواژه‌ای به دلیل ناتوانی ذاتی در درک چندمعنایی، هم‌معنایی و روابط مفهومی پیچیده معمولاً نتایج نامرتبط و حجیمی ارائه می‌دهند. این مسئله در حوزهٔ علوم اسلامی که اصطلاحات تخصصی چندلایه، متون دوزبانه (فارسی و عربی) و ترکیب منحصربه‌فردی از زبان کلاسیک و مدرن دارند، به‌مراتب حادتر به نظر می‌رسد. در دههٔ اخیر، با ظهور فناوری‌های نوینی مانند وب معنایی، گراف‌های دانش، مدل‌های زبانی بزرگ (LLMs) و معماری بازیابی افزوده با تولید (RAG)، فرصت‌های بی‌سابقه‌ای برای ارتقای بازیابی معنایی در حوزه‌های تخصصی فراهم شده است.

هدف پژوهش حاضر، ارائهٔ تحلیل نظام‌مند و به‌روز از چالش‌های بازیابی معنایی در علوم اسلامی، بررسی انتقادی رویکردهای موجود در پرتو فناوری‌های جدید و ارائهٔ چارچوبی یکپارچه مبتنی‌بر شواهد تجربی با هدف بهبود وضعیت موجود است. این پژوهش با روش مرور نظام‌مند

* **استناد به این مقاله:** میرعرب، علی. (۱۴۰۴). حرکت به‌سوی بازیابی معنایی در علوم اسلامی: از مدل‌های سنتی تا معماری‌های مدرن. مدیریت دانش اسلامی، ۷(۱)، صص ۱۰۴-۱۳۰.
https://Doi.org/10.22081/jikm.2026.74693.1109

▣ **نوع مقاله:** پژوهشی؛ **ناشر:** پژوهشگاه علوم و فرهنگ اسلامی، قم، ایران

▣ **تاریخ دریافت:** ۱۴۰۳/۰۸/۰۵ ● **تاریخ اصلاح:** ۱۴۰۳/۱۰/۱۹ ● **تاریخ پذیرش:** ۱۴۰۳/۱۱/۲۱ ● **تاریخ انتشار آنلاین:** ۱۴۰۴/۰۱/۱۰

ادبیات و تحلیل محتوای کیفی ۵۲ منبع معتبر بین‌المللی و داخلی (از ۲۰۱۴ تا ۲۰۲۵) انجام شده است. یافته‌ها نشان می‌دهد چالش‌های بازیابی معنایی را می‌توان در چهار سطح طبقه‌بندی کرد: مفهومی (چندمعنایی، هم‌معنایی، سلسله‌مراتبِ مفاهیم)، زبانی (پیچیدگی‌های صرفی عربی، تنوع رسم‌الخطی فارسی، دوزبانگی)، فنی (محدودیت رویکردهای کلیدواژه‌ای، ناتوانی در فهم پرسش‌های پیچیده، توهم مدل‌های زبانی) و ساختاری (نبود هستی‌شناسی استاندارد، پراکندگی داده‌ها، کمبود داده‌های آموزشی).

چارچوب پیشنهادی این پژوهش پنج لایهٔ مکمل دارد: (۱) هستی‌شناسی ساختاریافته OWL/SKOS برای علوم اسلامی، (۲) گراف دانش یکپارچهٔ چندزبانه، (۳) مدل‌های زبانی بزرگ تنظیم‌شده برای دامنه، (۴) سیستم RAG بومی‌سازی‌شده با بازیابی ترکیبی و (۵) رابط جستجوی هوشمند با بسط پرسش معنایی. شواهد تجربی از پروژه‌های SemanticHadith، FarsBase، AraBERT، مدل Fanar و نتایج کنفرانس QIAS 2025 نشان می‌دهد سیستم‌های ترکیبی بومی‌سازی‌شده می‌توانند در وظایف تخصصی اسلامی حتی از قدرتمندترین مدل‌های عمومی نیز پیشی بگیرند.

کلیدواژه‌ها

بازیابی معنایی، علوم اسلامی، گراف دانش، مدل زبانی بزرگ، زبان فارسی، پردازش زبان طبیعی.

مقدمه

در عصر دیجیتال، دسترسی به منابع علمی با سرعت و دقت بالا از ضرورت‌های اساسی پژوهش به شمار می‌رود. پژوهشگران علوم اسلامی امروزه با حجم فزایندهٔ منابع دیجیتال روبه‌رو هستند: کتابخانه‌های بزرگ متون اسلامی، نظیر مجموعهٔ متون اسلامی شامله[1]، پروژهٔ بین‌المللی متون اسلامی باز (OpenITI)، کتابخانهٔ دیجیتال فارسی (PDL)، پایگاه‌های نورمگز و SID در ایران و منابع متعدد دیگری که هرساله میلیون‌ها صفحه متن اسلامی را در دسترس پژوهشگران قرار می‌دهند. این دریای اطلاعاتی، بدون ابزارهای جستجوی هوشمند، به باری سنگین برای پژوهشگر تبدیل می‌شود (Pavlova & Makhlouf, 2024).

موتورهای جستجوی مبتنی‌بر کلیدواژه که دهه‌ها بر فضای دیجیتال علوم اسلامی حاکم بوده‌اند، در برابر پیچیدگی‌های ذاتی این حوزه ناتوان مانده‌اند. وقتی پژوهشگری «ادلهٔ اعتبار خبر واحد در مذاهب اسلامی» را جستجو می‌کند، باید بتواند منابعی بیابد که از مفاهیم هم‌معنا، نظیر «حجیت»، «اثبات»، «تصدیق»، یا «توثیق»، سخن گفته‌اند؛ حتی اگر هیچ‌کدام از این واژه‌ها در پرسش اصلی نیامده باشند. کاربر ممکن است به متن عربی نیاز داشته باشد اما پرسش را به فارسی مطرح کرده باشد. این شکاف‌های معنایی و زبانی حلقهٔ گمشده‌ای است که بازیابی معنایی می‌تواند آن را پر کند.

در سال‌های اخیر، ارزیابی‌های تجربی اولیه (Turner et al., 2009; Andago et al., 2010) نشان داد موتورهای جستجوی معنایی عمومی، مانند هاکیا[2] (O'Leary, 2010)، برتری آشکاری بر موتورهای کلیدواژه‌ای ندارند. این یافته نه به معنای ناکارآمدی ایدهٔ بازیابی معنایی بلکه نشانه‌ای بود که بازیابی معنایی واقعی نیازمند زیرساخت دانشی خاص دامنه است. هم‌زمان، در دههٔ ۲۰۱۰–۲۰۲۵ تحولاتی عمیق در حوزهٔ هوش مصنوعی رخ داده که چشم‌اندازهای کاملاً نوینی گشوده است: مدل‌های مبتنی‌بر ترنسفورمر، توسعهٔ

1. Shamela
2. Hakia

مـدل‌های تخصصـی عربـی، ماننـد AraBERT (Antoun et al., 2020)، و فارسـی، ماننـد ParsBERT (Farahani et al., 2021)، سـاخت گراف‌هـای دانـش اسـلامی، نظیـر SemanticHadith (Kamran et al., 2023)، و بویژه معمـاری بازیابی ـافزوده بـا تولیـد RAG (Lewis et al., 2020) که توانایی استدلال مستند را ممکن می‌کند.

باوجوداین، چالشـی بنیـادین همچنـان بـاقی اسـت: هـیچ فنـاوری واحـدی نمی‌توانـد همهٔ جنبه‌های پیچیدگی علوم اسلامی را پوشش دهد. مـدل‌های زبـانی بـزرگ عمـومی، بـدون آنکـه بـه دانـش دامنـه‌ای خـاص مجهـز شـوند، در پرسـش‌های تخصصـی دچـار «توهم[1]» می‌شوند و پاسخ‌های قانع‌کننده اما نادرست تولید می‌کننـد. نتـایج QIAS 2025 (Bouchekif et al., 2025) نشـان داد سیسـتم‌های ترکیبـی دامنـه‌محور از مـدل‌های عمـومی قدرتمنـدی، ماننـد GPT-4.5 و Gemini 2.5، در وظـایف تخصصـی اسـلامی پیشـی می‌گیرند. پژوهش حاضر با هدف تحلیل نظام‌مند چالش‌هـا و ارائـهٔ چـارچوبی یکپارچـه برای رویارویی با آنها انجام شده است.

پرسش‌های اصلی این پژوهش عبارت‌اند از: (۱) مهمترین چالش‌های بازیابی معنـایی در علوم اسلامی در چهار سطح مفهومی، زبانی، فنی و ساختاری چیست؟ (۲) فناوری‌هـای نوین، اعم از گراف دانش، مدل‌های زبانی بـزرگ و معمـاری RAG، چـه ظرفیـت‌هـایی برای رویارویی با این چالش‌ها دارند؟ (۳) چارچوب یکپارچـهٔ پیشـنهادی چـه لایـه‌هایی دارد و چگونه می‌تواند پیاده‌سازی شود؟

۱. مبانی نظری و پیشینهٔ پژوهش

۱-۱. بازیابی اطلاعات: از مدل‌های کلاسیک تا مدل‌های عمیق

بازیابی اطلاعات علمـی اسـت کـه هـدفش سـازماندهی و دسـترس‌پذیرکردن اطلاعـات متناسب با نیاز کاربر است. مدل‌های کلاسـیک بازیـابی، ازجملـه مـدل فضـای بـرداری[2]

1. hallucination
2. VSM

و مدل احتمالی بایزی، براساس آمار حضور و غیاب کلمات کار می‌کنند. الگوریتم TF-IDF که وزن هر کلمه را براساس فراوانی آن در سند و ندرت آن در مجموعه تعیین می‌کند و الگوریتم BM25 که نسخه‌ای بهبودیافته است، دهه‌ها استاندارد عملی بازیابی اطلاعات بوده‌اند (Darwish & Magdy, 2014). این رویکردها سریع، تفسیرپذیر و مقیاس‌پذیرند اما ناتوانی بنیادین آنها در درک معنا محدودیتی جدی برای حوزه‌های تخصصی است.

با پیشرفت یادگیری عمیق، مدل‌های تعبیه[1] ظاهر شدند که جملات و پاراگراف‌ها را به بردارهای معنادار در فضایی با ابعاد بالا نگاشت می‌کنند. مدل‌های خانوادهٔ BERT که براساس معماری ترنسفورمر و پیش‌آموزش در زمینهٔ حجم عظیمی از متن کار می‌کنند، نقطهٔ عطف اصلی بودند. این مدل‌ها می‌توانند «معنای» جمله‌ای مانند «نماز صبح» را نزدیک به «فجر را به جماعت بخوانید» تشخیص دهند، حتی بدون هیچ کلمهٔ مشترکی. مدل‌های تخصصی AraBERT (Antoun et al., 2020) برای عربی و ParsBERT (Farahani et al., 2021) برای فارسی این توانایی را به زبان‌های منطقه آورده‌اند.

جدیدترین نسل مدل‌های زبانی بزرگ[2] است که قادر به استدلال، تولید متن و ترکیب اطلاعات از منابع متعدد هستند. معماری RAG (Lewis et al., 2020) با ترکیب بازیابی دقیق از پایگاه دانش با توانایی تولید مدل زبانی، راهکاری برای محدودکردن توهم و اتکای مدل به منابع تأییدشده فراهم می‌آورد. بررسی جامع شرپ[3] و همکاران (۲۰۲۴) دربارهٔ وب معنایی نشان می‌دهد ترکیب گراف‌های دانش با LLMها می‌تواند قابلیت استدلال و تفسیر پیچیده را ممکن کند.

۱-۲. پژوهش‌های پردازش متون اسلامی

پردازش متون اسلامی به حوزهٔ پژوهشی تخصصی تبدیل شده است. مجموعهٔ دادهٔ

1. embedding models
2. LLMs
3. Scherp

Qur'an QA 2022 (Malhas & Elsayed, 2022) اولین مجموعهٔ آزمایشی استاندارد برای پرسش‌ـپاسخ قرآنی به عربی بود که در کنفرانس LREC 2022 معرفی و ارزیابی شده است. سال بعد، این مجموعه در Qur'an QA 2023 توسعه یافت (Malhas & Elsayed, 2020). پروژهٔ SemanticHadith (Kamran et al., 2023) که در مجلهٔ Web Semantics منتشر شد، یک هستی‌شناسی OWL و گراف دانش RDF برای شش مجموعهٔ حدیثی ساخت و نشان داد این رویکرد پرسش‌های معنایی پیچیده‌ای را ممکن می‌کند که جستجوی کلیدواژه‌ای از عهدهٔ آنها برنمی‌آید.

کنفرانس QIAS 2025 (Bouchekif et al., 2025) که در چارچوب ArabicNLP 2025 برگزار شد، آخرین دستاوردهای سیستم‌های هوش مصنوعی برای علوم اسلامی را گرد آورد. تیم QU-NLP (Al-Smadi, 2025)، با ترکیب مدل Fanar-1-9B تنظیم‌شده با RAG در وظیفهٔ استنتاج ارث اسلامی دقت ۸۵/۸ درصد به دست آورد و از GPT-4.5، LLaMA و Gemini 2.5 پیشی گرفت. پژوهش پاولوا و مخلوف[1] (۲۰۲۴)، با رویکرد آموزش چندمرحله‌ای برای بازیابی اسلامی چندزبانه، امکان استفاده از سیستم‌های غیرانتفاعی با کارایی بالا را نشان داد.

در حوزهٔ متون اسلامیِ فارسی، مجموعهٔ دادهٔ IslamicPCQA (غفوری و همکاران[2]، ۲۰۲۳) به‌عنوان نخستین مجموعهٔ آزمایشی جامع شناخته می‌شود. این مجموعه شامل ۵۶۰۰ پرسش به زبان فارسی است که با تکیه بر منابع اسلامی طراحی شده‌اند و عمدتاً ماهیتی پیچیده و چندمرحله‌ای دارند. این مجموعه نشان داد مدل‌های چندزبانه، مانند mT5، می‌توانند پرسش‌های فارسی دربارهٔ منابع عربی را تا حدی پاسخ دهند اما هنوز شکاف قابل‌توجهی با عملکرد انسانی دارد. مطالعهٔ منیری، شلوسر و کوورکو[3] (۲۰۲۴) در مجلهٔ Computers نشان داد نبود مجموعه‌های آزمایشی استاندارد اصلی‌ترین مانع پیشرفت پژوهش در بازیابی اطلاعات فارسی است.

1. Pavlova & Makhlouf
2. Ghafouri et al.
3. Moniri, Schlosser & Kowerko

۱-۳. گراف‌های دانش برای علوم اسلامی

گراف دانش ساختاری است که موجودیت‌ها را به‌عنوان گره و روابط میان آنها را یال نمایش می‌دهد. در علوم اسلامی، این رویکرد می‌تواند روابطی، مانند «حکم X از آیهٔ Y استنباط شده»، «عالم A شاگرد عالم B بوده» و «مفهوم C در فقه امامیه با D هم‌معناست» را به‌روشنی کدگذاری کند. هستی‌شناسی عربی جارار (Jarrar, 2021)، با ۶۲٬۰۰۰ مفهوم و ۵۱۳۰۰ واژه، زیرساختی قوی برای تحلیل معنایی متون عربی فراهم می‌آورد. در فارسی، پروژهٔ FarsBase (Asgari-Bidhendi et al., 2019) گراف دانش فارسی را با اتصال به منابع متعدد ساخته و موتور جستجوی معنایی ارائه داده است. اتصال این گراف‌ها به Wikidata و DBpedia از طریق ابزارهای پیوندداده، مانند LIMES، امکان غنی‌سازی با دانش عمومی را فراهم می‌آورد (Kamran et al., 2026).

۲. روش پژوهش

پژوهش حاضر از روش مرور نظام‌مند ادبیات با چارچوب پریزما[1] بهره گرفته است. این روش که برای ترکیب شواهد موجود در حوزه‌ای علمی طراحی شده، هنگامی مناسب است که هدف نه آزمایش فرضیه‌ای خاص، بلکه سنتز یافته‌های موجود و ارائهٔ چارچوبی جامع باشد. در کنار این، از رویکرد طراحی علمی برای ساخت چارچوب پیشنهادی استفاده شده تا محصول نهایی نه‌فقط توصیفی بلکه تجویزی و قابل‌پیاده‌سازی باشد.

در پایگاه‌های بین‌المللی Scopus، Web of Science، Google Scholar، ACL Anthology (برای پژوهش‌های NLP و زبان‌شناسی محاسباتی)، IEEE Xplore، Semantic Web Journal و پایگاه‌های داخلی نورمگز، SID و پرتال جامع علوم انسانی جستجو شد. کلیدواژه‌های جستجو در سه دسته تنظیم شدند: دستهٔ اول مرتبط با موضوع اصلی (semantic retrieval, information retrieval, Islamic sciences, Arabic NLP, Persian IR)؛ دستهٔ دوم مرتبط با فناوری‌ها (knowledge graph, ontology, RAG, LLM, BERT,

1. PRISMA

word embedding)؛ دستهٔ سوم مرتبط با زبان (Arabic morphology, Persian orthography, bilingual retrieval). معیارهای ورود: ارتباط مستقیم موضوعی، اعتبار علمی و بازهٔ زمانی ۲۰۱۴–۲۰۲۵. از ۱۴۸ منبع اولیه، ۵۲ منبع نهایی انتخاب شدند.

داده‌های استخراج‌شده با روش تحلیل محتوای کیفی چهارمرحله‌ای تحلیل شدند. در مرحلهٔ اول، استخراج: چالش‌ها، راهکارها، نتایج تجربی و مفاهیم کلیدی هر منبع استخراج شد. در مرحلهٔ دوم، کدگذاری باز: هر یافته به یک کد اختصاص یافت. در مرحلهٔ سوم، کدگذاری محوری: کدهای مشابه در دسته‌های موضوعی گردآوری شدند. در مرحلهٔ چهارم، ترکیب: روابط میان دسته‌ها تحلیل و چارچوب یکپارچه ترسیم شد. اعتبار از طریق مثلث‌سازی منابع و بازبینی متخصصان تأمین شد.

۲. چالش‌های بازیابی معنایی در علوم اسلامی

۲-۱. چالش‌های مفهومی

۲-۱-۱. چندمعنایی[1]

پدیدهٔ چندمعنایی در علوم اسلامی به دلیل تخصص‌یافتگی واژگان در شاخه‌های مختلف، بسیار گسترده است. واژهٔ «عدل» در فقه به شرط عدالت در شاهد، در کلام به صفت الهی، در اخلاق به فضیلت میانه‌روی و در علم رجال به وثاقت راوی اشاره دارد. «قیاس» در اصول فقه به سرایت حکم از اصل به فرع است اما در منطق معنایی کاملاً متفاوت دارد. «تقلید» در اصول فقه به پیروی از مجتهد زنده است اما در ادبیات به تقلید شعری اشاره می‌کند. سیستم بازیابی معنایی باید بافت پرسش را تشخیص دهد تا واژه را در معنای مناسب تفسیر کند (Jarrar, 2021).

۲-۱-۲. هم‌معنایی[2] و مترادف‌های بافتی

در علوم اسلامی، مفاهیم یکسان غالباً با واژگان متعددی بیان می‌شوند. «نماز» در

1. Polysemy
2. Synonymy

فارسی معادل «صلاة» در عربی است و «روزه» و «صوم» نیز به یک مفهوم دلالت دارند. همچنین، «اجتهاد»، «استنباط» و «رأی» در بسیاری از بافت‌ها به‌صورت هم‌معنا به‌کار می‌روند. «حلال»، «مباح» و «جائز» هم‌پوشانی معنایی زیادی دارند. کاربری که «فضایل صوم» را جستجو می‌کند، باید منابع «فضایل روزه» و «فضایل الصیام» را هم بیابد. موتور کلیدواژه‌ای این ارتباط را درک نمی‌کند (Malhas & Elsayed, 2022).

۲-۱-۳. سلسله‌مراتب مفاهیم

علوم اسلامی نظام‌های طبقه‌بندی دقیق و عمیق دارند: «عبادات» ← «طهارت» ← «وضو» ← «وضوی جبیره‌ای». در علوم قرآن، رابطهٔ آیه با سوره، رابطهٔ آیه با احکام مستنبط و رابطهٔ آیه با احادیث تفسیری آن باید کدگذاری شوند. سیستم SKOS با روابط «broader» و «narrower» می‌تواند این ساختارها را مدل کند. SemanticHadith (Kamran et al., 2023) نشان داد با گراف دانش OWL می‌توان پرسش‌هایی، مانند «احادیثی که با تفسیر آیهٔ X مرتبط‌اند»، را که جستجوی کلیدواژه‌ای از عهدهٔ آن برنمی‌آید، پاسخ داد.

۲-۲. چالش‌های زبانی

۲-۲-۱. پیچیدگی صرفی عربی

نظام صرفی زبان عربی مبتنی‌بر ریشه‌های سه یا چهارحرفی است. از ریشهٔ «ع ل م» بیش از هشتاد واژه با معانی مرتبط اما متمایز ساخته می‌شود. ریشه‌یابی دقیق برای بازیابی اطلاعات عربی ضروری است اما بسیار دشوار. عربی کلاسیک قرآن و حدیث با عربی مدرن استاندارد (MSA) تفاوت‌های قابل‌توجهی دارد و بیشتر مدل‌های موجود روی MSA آموزش دیده‌اند. ابزار ALMA (Jarrar, Akra & Hammouda, 2024) و مجموعهٔ ابزارهای SinaTools (Hammouda, Jarrar & Khalilia, 2024) برای ریشه‌یابی و برچسب‌گذاری دقیق عربی ابزارهای معتبری هستند.

۲-۲-۲. تنوع رسم‌الخط فارسی

زبان فارسی چالش‌های رسم‌الخطی منحصربه‌فردی دارد: (الف) حروف مشابه با کدهای یونی‌کد متفاوت که در رایانه یکسان به نظر می‌رسند اما در جستجو تفاوت دارند؛ (ب) نبود استاندارد یکپارچه برای برخی کلمات مانند «مسأله»، «مسئله»، «مساله»؛ (ج) تنوع در استفاده از همزه، مانند «رأی»، در برابر «رای»؛ (د) سرهم‌نویسی و جدانویسی ترکیب‌ها مانند «می‌رود» و «میرود». مطالعهٔ منیری، شلوسر و کوورکو[1] (۲۰۲۴) این چالش‌ها را موانع اصلی بازیابی دقیق فارسی شناسایی کرد.

۲-۲-۳. دوزبانگی و ترکیب زبانی

بیشتر متون علوم اسلامی فارسی آمیزه‌ای از فارسی و عربی هستند. این آمیختگی نه‌فقط در سطح اصطلاحات، بلکه در سطح جمله هم رایج است. کاربر ممکن است پرسش را به فارسی مطرح کند اما پاسخ اصلی در متن عربی باشد. مدل‌های چندزبانه، مانند mBERT و XLM-RoBERTa، می‌توانند تاحدی این شکاف را پر کنند اما ادغام دانش دو زبان در سطح گراف دانش راه‌حل ساختاری‌تری است (Ghafouri et al., 2023).

۲-۳. چالش‌های فنی

رویکردهای TF-IDF و BM25 که هنوز در بسیاری از پایگاه‌های اطلاعاتی اسلامی کاربرد دارند، سه محدودیت اصلی دارند: تطابق لفظی به‌جای مفهومی که باعث ازدست‌رفتن هم‌معناها می‌شود؛ ناتوانی در پردازش پرسش‌های پیچیده چندجزئی که در پژوهش‌های علوم اسلامی رایج است و ناتوانی در بهره‌گیری از ساختار ضمنی اسناد. افزون بر این، مدل‌های زبانی بزرگ که بدون RAG استفاده می‌شوند، با خطر توهم مواجه‌اند. مدل ممکن است حدیثی نقل کند که وجود ندارد یا فتوایی را به عالمی اشتباه نسبت دهد. سیستم‌های RAG با ارجاع هر پاسخ به منبع تأییدشده، این مشکل را اساسی حل می‌کنند (Bouchekif et al., 2025).

1. Moniri, S., Schlosser, T., & Kowerko

۲-۴. چالش‌های ساختاری

عمیق‌ترین چالش‌ها در سطح زیرساخت دانشی هستند؛ اول، نبود هستی‌شناسی استاندارد جامع ماشین‌خوان: باوجود اصطلاح‌نامه‌های ارزشمند، هیچ‌کدام به‌صورت OWL یا SKOS کامل موجود نیستند. دوم، پراکندگی منابع: پایگاه‌های متعدد با فرمت‌های متفاوت و بدون رابط یکپارچه. سوم، کمبود داده‌های آموزشی: برای آموزش مدل‌های اختصاصی به مجموعه‌های پرسش ـپاسخ حاشیه‌نویسی‌شده با حجم کافی نیاز است که هنوز کافی نیست (منیری، شلوسر و کوورکو، ۲۰۲۴). این سه چالش ساختاری در کنار هم چرخه‌ای معیوب ایجاد می‌کنند: بدون هستی‌شناسی، دادهٔ خوب ساخته نمی‌شود؛ بدون دادهٔ خوب، مدل خوب آموزش نمی‌بیند و بدون مدل خوب، کمبود هستی‌شناسی جبران نمی‌شود. شکستن این چرخه نیازمند سرمایه‌گذاری هم‌زمان در هر سه حوزه است.

۳. فناوری‌های نوین: ظرفیت‌ها و محدودیت‌ها

۳-۱. هستی‌شناسی OWL، SKOS و گراف‌های دانش

گراف دانش با نمایش موجودیت‌ها به‌عنوان گره و روابط به‌عنوان یال، پردازش پرسش‌های استدلالی پیچیده را ممکن می‌کند. این رویکرد می‌تواند در علوم اسلامی روابطی، مانند «حکم X از آیهٔ Y استنباط شده» یا «عالم A در قرن ۳ هجری زیسته»، را به‌روشنی کدگذاری کند. پروژهٔ SemanticHadith (Kamran et al., 2023; 2026) نشان داد این ساختار جستجوهایی را ممکن می‌کند که هیچ موتور کلیدواژه‌ای از عهدهٔ آنها برنمی‌آید؛ برای مثال «همهٔ احادیثی که با تفسیر آیهٔ X مرتبط‌اند و راویان ثقه آنها را روایت کرده‌اند». هستی‌شناسی عربی جارار (Jarrar, 2021) با ۶۲هزار مفهوم، زیرساخت زبانی مناسبی برای این هدف فراهم می‌کند. استاندارد SKOS (Pastor et al., 2009) برای نمایش اصطلاح‌نامه‌ها و سازگاری با سیستم‌های دیگر، مناسب‌ترین گزینه است.

۳-۲. مدل‌های زبانی تخصصی عربی و فارسی

مدل AraBERT (Antoun et al., 2020)، با آموزش براساس ۷۰ گیگابایت متن عربی، در وظایف متعددی عملکرد بهتری از مدل‌های چندزبانهٔ عمومی داشت. مدل GATE-AraBERT (Nacar & Koubaa, 2024) که با یادگیری تضادی آموزش دیده، از بهترین مدل‌های تعبیهٔ عربی در آزمون MTEB است. مدل Fanar-1-9B (Abbas et al., 2025) که برای عربی و متون اسلامی خاص طراحی شده، در QIAS 2025 از GPT-4.5 پیشی گرفت. در زبان فارسی ParsBERT (Farahani et al., 2021)، پایهٔ مناسبی برای تنظیم دامنه‌ای[1] براساس متون اسلامی فارسی است. کلید موفقیت این مدل‌ها تنظیم دامنه‌ای بر پایهٔ داده‌های تخصصی اسلامی است.

۳-۳. معماری RAG برای متون اسلامی

معماری RAG (Lewis et al., 2020)، با ترکیب «بازیابی» از پایگاه دانش و «تولید» توسط مدل زبانی، راهکار مؤثری برای مشکل توهم ارائه می‌دهد. در مرحلهٔ بازیابی، سیستم بخش‌های مرتبط با پرسش را از پایگاه دانش اسلامی استخراج می‌کند. در مرحلهٔ تولید، این بخش‌ها همراه پرسش به مدل داده می‌شوند تا پاسخ مستند تولید شود. پژوهش احمد[2] و همکاران (۲۰۲۵) خط‌لولهٔ ترکیبی سه‌مرحله‌ای ارائه داد: BM25 برای بازیابی اولیه، تعبیهٔ متراکم عربی برای دقت معنایی و بازرتبه‌بندی با کراس‌انکودر برای انتخاب نهایی. این ترکیب سه‌مرحله‌ای از روش‌های منفرد بهتر عمل کرد و توانست الگوی عملیاتی سیستم RAG اسلامی باشد. پژوهش ARAG (Attia et al., 2026) نیز با ارزیابی ۵۷۵ پرسش عربی، برتری روش ترکیبی را تأیید کرد.

۴. چارچوب پیشنهادی

براساس تحلیل جامع چالش‌ها و فناوری‌های موجود، پژوهش حاضر چارچوبی پنج‌لایه

1. fine-tuning
2. Ahmad

پیشنهاد می‌دهد. ویژگی اصلی این چارچوب آن است که هر لایه هم مستقل ارزش افزوده دارد هم با ترکیب با سایر لایه‌ها قدرتمندتر می‌شود.

۴-۱. لایهٔ اول: هستی‌شناسی ساختاریافتهٔ علوم اسلامی (OWL/SKOS)

این لایه بنیادی‌ترین زیرساخت دانشی است. هستی‌شناسی پیشنهادی باید با استانداردهای OWL 2 (برای استدلال خودکار) و SKOS (برای سازگاری با اصطلاح‌نامه‌های موجود) تعریف شود. کلاس‌های اصلی هستی‌شناسی شامل: «مفهوم شرعی[1]»، «حکم[2]»، «متن[3]»، «عالم[4]»، «مذهب[5]» و «منبع[6]». روابط کلیدی: «مستنبط از[7]»، «هم‌معنای[8]»، «زیرمجموعه[9]»، «شرح[10]»، «به نقل از[11]». این هستی‌شناسی می‌تواند بر پایهٔ اصطلاح‌نامه جامع علوم اسلامی موجود ساخته شود و با هستی‌شناسی عربی جارار (Jarrar, 2021) یکپارچه شود.

۴-۲. لایهٔ دوم: گراف دانش یکپارچهٔ چندزبانه

گراف دانش یکپارچه موجودیت‌های اصلی علوم اسلامی را به هم پیوند می‌دهد. آیه قرآن می‌تواند به تفسیرهایش، به احادیث مرتبط، به احکام فقهی مستنبط و به علمایی که آن را شرح داده‌اند، ارتباط داشته باشد. ارتباط به Wikidata و DBpedia از

1. ShariaticConcept
2. Ruling
3. Text
4. Scholar
5. School
6. Source
7. derivedFrom
8. synonym
9. narrower
10. commentaryOf
11. narratedBy

طریق LIMES (Kamran et al., 2026) اطلاعات زمینه‌ای غنی فراهم می‌آورد. در مورد متون فارسی، اتصال به FarsBase (Asgari-Bidhendi et al., 2019) یکپارچگی منابع فارسی را ممکن می‌کند.

۴-۳. لایهٔ سوم: مدل‌های زبانی تخصصی

این لایه مسئول تبدیل متون و پرسش‌ها به تعبیه‌های معنایی[1] است. گام اول انتخاب مدل پایه: AraBERT یا Fanar در مورد عربی، ParsBERT در مورد فارسی، یا مدل‌های چندزبانه، مانند mBERT، برای ترکیب. گام دوم، تنظیم دامنه‌ای بر روی مجموعه‌های اسلامی: Qur'an QA 2023، IslamicPCQA و مجموعه‌های خاص که با همکاری متخصصان علوم اسلامی توسعه می‌یابند. روش یادگیری تضادی (Nacar & Koubaa, 2024) برای بهینه‌سازی تعبیه‌ها پیشنهاد می‌شود.

۴-۴. لایهٔ چهارم: سیستم RAG بومی‌سازی‌شده

این لایه قلب سیستم بازیابی است. خط‌لولهٔ پیشنهادی سه‌مرحله‌ای است: (۱) بازیابی اسپرس با BM25 برای سرعت و پوشش اولیه؛ (۲) بازیابی متراکم با مدل تعبیه تخصصی برای دقت معنایی؛ (۳) بازرتبه‌بندی با کراس ـنکودر برای انتخاب نهایی مرتبط‌ترین گذرها. این رویکرد در احمد و همکاران (۲۰۲۵) و ARAG (Attia et al., 2026) اثبات شده است. ویژگی خاص علوم اسلامی «توسعهٔ پرسش معنایی»[2] از هستی‌شناسی است: پرسش «احکام تجارت» به‌صورت خودکار به «بیع، اجاره، مضاربه، مشارکت، معاملات اسلامی» توسعه می‌یابد.

۴-۵. لایهٔ پنجم: رابط جستجوی هوشمند

رابط کاربری هوشمند پیچیدگی‌های فنی را از کاربر پنهان می‌کند. ویژگی‌های

1. semantic embeddings
2. Semantic Query Expansion

کلیدی: (الف) یکسان‌ساز رسم‌الخط کـه «مسـأله» و «مسـاله» را معـادل تلقـی کنـد؛ (ب) پشتیبانی از پرسش به فارسی و عربـی بـا بازیـابی از هـر دو پیکـره؛ (ج) پیشـنهاد مفـاهیم مرتبط از هستی‌شناسی حین تایپ‌کردن؛ (د) نمایش پاسخ‌های مستند بـا ارجـاع دقیـق بـه منابع اصلی، مشابه سیستم «سند» در علوم اسلامی؛ (ه) امکان فیلتر براساس مذهب، قـرن، شاخهٔ علمی و غیره.

جدول ۱ رابطهٔ مسـتقیم هـر دسـته از چالش‌هـا، راهکـار پیشـنهادی و اسـتاندارد فنـی مرتبط را نشان می‌دهد.

جدول ۱. جدول تطبیقی: چالش، راهکار، استاندارد فنی

دسته چالش	مسئله	راهکار پیشنهادی	استانداردها و فناوری‌های مرتبط
مفهومی	چندمعنایی اصطلاحات	استفاده از هستی‌شناسی و ابهام‌زدایی زمینه‌محور	OWL-DL، SPARQL
مفهومی	هم‌معنایی درون‌زبانی و برون‌زبانی	اصطلاح‌نامه و توسعهٔ خودکار پرسش	SKOS Core، RDF
مفهومی	سلسله‌مراتب مفاهیم	گراف دانش با روابط سلسله‌مراتبی	SKOS، OWL Class Hierarchy
زبانی	پیچیدگی صرفی زبان عربی	ابزارهای تحلیل صرفی و مدل‌های بازنمایی خاص دامنه	Arabic POS Tagging
زبانی	تنوع رسم‌الخط فارسی	یکسان‌سازی یونی‌کد و مدل‌های زبانی فارسی همراه قواعد اصلاحی	Unicode Normalization
زبانی	دوزبانگی متون	مدل‌های چندزبانه و گراف دانش چندزبانه	Cross-lingual Information Retrieval

دسته چالش	مسئله	راهکار پیشنهادی	استانداردها و فناوری‌های مرتبط
فنی	دقت پایین بازیابی	معماری ترکیبی بازیابی شامل بازیابی پراکنده و متراکم همراه بازرتبه‌بندی	Bi-encoder، Cross-encoder
فنی	پرسش‌های پیچیدهٔ چندمرحله‌ای	تجزیهٔ پرسش و بازیابی تکراری	Chain-of-Thought، RAG
فنی	توهم در مدل‌های زبانی	تولید مبتنی‌بر اسناد معتبر	Grounded Generation
ساختاری	نبود هستی‌شناسی	توسعهٔ هستی‌شناسی ترکیبی مبتنی‌بر اصطلاح‌نامه	OWL2، SKOS
ساختاری	پراکندگی منابع	ایجاد گراف دانش یکپارچه و اتصال به منابع باز پیوندی	Linked Open Data
ساختاری	کمبود دادهٔ آموزشی	توسعهٔ مجموعهٔ داده و استفاده از یادگیری انتقالی	Transfer Learning

۴-۶. گام‌های عملی پیاده‌سازی

۴-۶-۱. ساخت هستی‌شناسی علوم اسلامی: رویکرد مرحله‌به‌مرحله

ساخت هستی‌شناسی جامع علوم اسلامی نیازمند رویکردی مرحله‌به‌مرحله با همکاری بین‌رشته‌ای است. در گام اول، دامنه و دامنه‌های فرعی تعریف می‌شود: شروع با شاخه‌ای، مانند فقه عبادات، سپس گسترش تدریجی. در گام دوم، ابزار انتخاب می‌شود: نرم‌افزار Protégé برای ویرایش OWL، ابزار Excel2SKOS برای تبدیل اصطلاح‌نامه موجود و LIMES برای اتصال به گراف‌های دانش خارجی. در گام سوم، اصطلاحات اولیه از اصطلاح‌نامهٔ جامع علوم اسلامی جمع‌آوری و با هستی‌شناسی عربی

جارار (Jarrar, 2021) یکپارچه می‌شود. گام چهارم غنی‌سازی روابط است: متخصصان علوم اسلامی روابط سلسله‌مراتبی، مترادف، استنادی و تاریخی را تعریف می‌کنند. گام پنجم اعتبارسنجی است: گروهی از متخصصان و آزمون‌های سازگاری منطقی با SPARQL ارزیابی هستی‌شناسی را انجام می‌دهند.

۴-۶-۲. راهکارهای فوری یکسان‌سازی رسم‌الخط فارسی

پیش از انجام هر راهکار پیچیده‌تری، می‌توان رسم‌الخط فارسی را یکسان‌سازی کرد که ارزش زیادی دارد. این یکسان‌سازی سه مرحله دارد: نرمال‌سازی یونی‌کد (تبدیل حروف با کدهای مختلف اما ظاهر یکسان به نماینده‌ای واحد)، یکسان‌سازی همزه (ثبت هر دو شکل «مسأله» و «مساله» در نمایه) و مدیریت سرهم‌نویسی و جدانویسی. کتابخانهٔ Hazm در پایتون ابزارهایی برای این کار دارد که می‌توان آنها را یک لایهٔ پیش‌پردازش در هر موتور جستجوی موجود اضافه کرد.

۴-۶-۳. مقایسهٔ سناریوهای عملی

برای نمایش تفاوت عملی رویکردها، سه سناریو را برای پرسش «ادلهٔ جواز بیمه در فقه معاصر» مقایسه می‌کنیم. در سناریوی اول (موتور کلیدواژه‌ای)، سیستم کلمات را جداگانه جستجو می‌کند. منابعی که از «تکافل» یا «تأمین اجتماعی اسلامی» سخن گفته‌اند اما کلمهٔ «بیمه» را ندارند کنار گذاشته می‌شوند. در سناریوی دوم (موتور با هستی‌شناسی)، سیستم می‌داند «بیمه»، «تکافل» و «تأمین اجتماعی» در حوزهٔ مالی اسلامی مرتبط‌اند؛ نتایج به مراتب کامل‌تر است. در سناریوی سوم (RAG)، مدل زبانی منابع را ترکیب و پاسخ مستند ارائه می‌کند: «سه دیدگاه اصلی در فقه معاصر: (۱) جواز به دلیل اباحهٔ اصلی ـ[منبع الف]، (۲) حرمت به دلیل غرر ـ[منبع ب]، (۳) جواز با شرایط ـ[منبع ج]».

۴-۶-۴. پیش‌نیازهای همکاری بین‌نهادی

پیاده‌سازی موفق چارچوب پیشنهادی نیازمند همکاری سه گروه نهاد است: اول،

نهادهـای علـوم اسـلامی، ماننـد پژوهشـکدهٔ مـدیریت اطلاعـات و مـدارک اسـلامی، دانشگاه‌های علـوم اسـلامی و حوزه‌هـای علمیـه کـه متخصصـان دانـش دامنـه را فـراهم می‌کنند؛ دوم، نهادهای پژوهش فناوری اطلاعـات، ماننـد گروه‌هـای NLP دانشـگاه‌های فنی و شرکت‌های نرم‌افزاری که فناوری را توسعه می‌دهند؛ سوم، نهادهـای کتابـداری و اطلاع‌رسانی کـه در مـدیریت اصـطلاح‌نامه‌ها، استانداردسـازی ابـرداده و طراحـی رابـط کاربری تجربه دارند. بدون این همکاری سه‌جانبه، حتی بهترین فناوری‌ها نیـز نمی‌تواننـد به کاربرد عملی برسند.

نتیجه‌گیری

یافته‌های ایـن پـژوهش نشـان می‌دهـد مسـئلهٔ بازیـابی معنـایی در علـوم اسـلامی مـاهیتی چندبعـدی و پیچیـده دارد و نمی‌تـوان آن را بـا راهکـاری منفـرد حل‌وفصـل کـرد. رویکردهای پیشین که اغلب بر بهبود عملکرد موتورهای جستجو متمرکز بودند، تصـویر کاملی از مسئله ارائه نمی‌دادند. تحلیل‌های این پژوهش نشان می‌دهـد چـالش اصـلی، نـه صرفاً در سطح ابزار جستجو، در سطح زیرساخت‌های دانشـی نهفتـه اسـت؛ زیرسـاخت‌هایی که شامل هستی‌شناسی‌های ماشـین‌خوان، گراف‌هـای دانـش یکپارچـه، مـدل‌های زبـانی تخصصی و پیکره‌های آموزشی کافی و متناسب با دامنهٔ علوم اسلامی هستند.

درعین‌حال، بررسی شواهد تجربـی نـوین نشـان می‌دهـد افـق پـیش‌روی ایـن حـوزه امیدوارکننده است. نتایج پروژه‌هـایی، ماننـد SemanticHadith، FarsBase، AraBERT و Fanar، و یافتـه‌های ارائه‌شـده در چـارچوب QIAS 2025، نشـان می‌دهـد تلفیـق دانـش دامنه‌ای ساختاریافته با مدل‌های زبانی بزرگ در قالب معماری‌های بازیابی‌ـتولید افـزوده می‌تواند به توسعهٔ سـامانه‌هایی بینجامـد کـه در وظـایف تخصصـی، عملکـردی فراتـر از رویکردهای مبتنی‌بر کلیدواژه و حتی برخی مدل‌های عمومی پیشرفته ارائه می‌دهند. ایـن یافته‌ها از فرضیه اصـلی پـژوهش مبنـی بـر کـارایی بـالاتر سـامانه‌های بومی‌سازی‌شـده و دامنه‌محور در مقایسه با راهکارهای عمومی پشتیبانی می‌کنند.

بر این اساس، چارچوب پنج‌لایهٔ پیشنهادی این پژوهش را می‌توان نقشه‌ای تلفیقی در

نظر گرفت که هم از پشتوانهٔ نظری برخوردار است هم قابلیت پیاده‌سازی تدریجی دارد. ویژگی مهم این چارچوب امکان توسعهٔ مستقل هر یک از لایه‌ها و ایجاد ارزش‌افزوده در هر مرحله است. در این میان، اهمیت این رویکرد صرفاً به بهبود ابزارهای بازیابی اطلاعات محدود نمی‌شود، بلکه به ایجاد زیرساختی دانشی می‌انجامد که می‌تواند نقش مهمی در حفظ، سازمان‌دهی و دسترس‌پذیرسازی میراث علمی اسلامی برای پژوهشگران معاصر و نسل‌های آینده ایفا کند.

براساس نتایج به‌دست‌آمده، مسیرهای زیر، به‌عنوان اولویت‌های پژوهشی آینده، پیشنهاد می‌شوند: نخست، توسعهٔ هستی‌شناسی جامع مبتنی‌بر استانداردهای OWL و SKOS برای علوم اسلامی با مشارکت نهادهای علمی مختلف، به‌گونه‌ای که شاخه‌های این حوزه را کامل پوشش دهد. دوم، طراحی و انتشار مجموعهٔ داده‌های ارزیابی استاندارد برای پرسش‌های فارسی و عربی در علوم اسلامی، با هدف فراهم‌کردن امکان سنجش دقیق و تکرارپذیر عملکرد سامانه‌ها. سوم، تنظیم دامنه‌ای مدل‌های زبانی موجود بر پیکره‌های تخصصی علوم اسلامی و ارزیابی تجربی عملکرد آنها در وظایف واقعی. چهارم، پیاده‌سازی نمونه‌های آزمایشی از سامانه‌های مبتنی‌بر معماری بازیابی ـ تولید افزوده برای پایگاه‌های اطلاعاتی تخصصی و ارزیابی کارایی آنها با مشارکت کاربران واقعی. پنجم، بررسی روش‌های پیشرفتهٔ یادگیری، ازجمله یادگیری تضادی و آموزش چندمرحله‌ای، برای بهبود کیفیت بازنمایی‌های معنایی در این حوزه. ششم، مطالعهٔ امکان اتصال هستی‌شناسی‌های علوم اسلامی به پایگاه‌های دانش پیوندی، مانند Wikidata و DBpedia، با هدف غنی‌سازی بافت معنایی و افزایش قابلیت تعامل‌پذیری داده‌ها.

درمجموع، نتایج این پژوهش بر این نکته تأکید دارد که آیندهٔ بازیابی اطلاعات در علوم اسلامی در گرو توسعهٔ هم‌زمان زیرساخت‌های دانشی و بهره‌گیری هدفمند از فناوری‌های نوین است و موفقیت در این مسیر مستلزم رویکردی تلفیقی، بومی‌سازی‌شده و مبتنی‌بر همکاری‌های میان‌رشته‌ای خواهد بود.

فهرست منابع

Andago, M., Phoebe, T. P. L., & Thanoun, B. A. M. (2010). Evaluation of a semantic search engine against a keyword search engine using first 20 precision. *International Journal for the Advancement of Science & Arts*, 1(2), pp. 55-63.

Antoun, W., Baly, F., & Hajj, H. (2020, May). *Arabert: Transformer-based model for arabic language understanding*. In Proceedings of the 4th workshop on open-source arabic corpora and processing tools, with a shared task on offensive language detection (pp. 9-15).

Asgari-Bidhendi, M., Hadian, A., & Minaei-Bidgoli, B. (2019). Farsbase: The persian knowledge graph. *Semantic Web*, *10*(6), 1169-1196.

Alothman, M., Altammami, A., Alhoshan, M., Almazrua, A., Al-Rasheed, R., Almatham, R., ... & Alfaifi, A. (2026). ARAG: An Agent-Based Hybrid Semantic-Lexical Retrieval-Augmented Generation Pipeline for Arabic Texts. *Procedia Computer Science*, 275, 682-691.

Bouchekif, A., Rashwani, S., Mohamed, E. S. A., Alkhatib, M., Sbahi, H., Gaben, S., ... & Ghaly, M. (2025, November). Qias 2025: Overview of the shared task on islamic inheritance reasoning and knowledge assessment. In Proceedings of The Third Arabic Natural Language Processing Conference: Shared Tasks (pp. 851-860).

Darwish, K., & Magdy, W. (2014). Arabic information retrieval. *Foundations and Trends in Information Retrieval*, 7(4), 239-342. doi:10.1561/1500000031

Abbas, U., Ahmad, M. S., Alam, F., Altinisik, E., Asgari, E., Boshmaf, Y., Boughorbel, S., Chawla, S., Chowdhury, S., Dalvi, F., Darwish, K., Durrani, N., Elfeky, M., Elmagarmid, A., Eltabakh, M., Fatehkia, M., Fragkopoulos, A., Hasanain, M., Hawasly, M., Husaini, M., Jung, S.-G., Lucas, J. K.,

Magdy, W., Messaoud, S., Mohamed, A., Mohiuddin, T., Mousi, B., Mubarak, H., Musleh, A., Naeem, Z., Ouzzani, M., Popovic, D., Sadeghi, A., Sencar, H. T., Shinoy, M., Sinan, O., Zhang, Y., Ali, A., El Kheir, Y., Ma, X., & Ruan, C. (2025). Fanar: An Arabic-centric multimodal generative AI platform. arXiv. https://arxiv.org/abs/2501.13944.

Farahani, M., Gharachorloo, M., Farahani, M., & Manthouri, M. (2021). ParsBERT: Transformer-based model for Persian language understanding. *Neural Processing Letters*, 53, 3831-3847. https://doi.org/10.1007/s11063-021-10528-4

Scherp, A., Groener, G., Škoda, P., Hose, K., & Vidal, M. E. (2024). Semantic Web: Past, Present, and Future (with Machine Learning on Knowledge Graphs and Language Models on Knowledge Graphs). arXiv preprint arXiv:2412.17159.

Moniri, S., Schlosser, T., & Kowerko, D. (2024). Investigating the Challenges and Opportunities in Persian Language Information Retrieval through Standardized Data Collections and Deep Learning. *Computers*, 13(8), 212. https://doi.org/10.3390/computers13080212

Ghafouri, A., Naderi, H., Aghajani Asl, M., & Firouzmandi, M. (2023). IslamicPCQA: A dataset for Persian multi-hop complex question answering in Islamic text resources. *arXiv*. https://arxiv.org/abs/2304.11664

Jarrar, M. (2021). The Arabic ontology: An Arabic WordNet with ontologically clean content. *Applied Ontology* Journal, 16(1), 1-26. Doi:10.3233/AO-200241

Jarrar, M., Akra, D., & Hammouda, T. (2024). Alma: Fast Lemmatizer and POS Tagger for Arabic. *Procedia Computer Science*. 244. 378-387. 10.1016/j.procs.2024.10.212..

Hammouda, T., Jarrar, M., & Khalilia, M. (2024). SinaTools: Open source toolkit for Arabic natural language processing. arXiv. https://arxiv.org/abs/2411.01523.

Kamran, A. B., Abro, B., & Basharat, A. (2023). SemanticHadith: An ontology-driven knowledge graph for the hadith corpus. *Journal of Web Semantics*, 78, 100797. https://doi.org/10.1016/j.websem.2023.100797

Kamran, A. B., Butt, N. A., & Basharat, A. (2026). Semantic Enrichment of Hadith Corpus—Knowledge Graph Generation From Islamic Text. Semantic Web, 17(2), 22104968261431425.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W. T., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, 33, pp 9459-9474.

Malhas, R., & Elsayed, T. (2022). Qur'an QA 2022: Overview of the first shared task on question answering over the Holy Qur'an. In *Proceedings of the 5th Workshop* (OSACT 2022) (pp. 79-87). Marseille, France: Association for Computational Linguistics.

Malhas, R., & Elsayed, T. (2020). AyaTEC: Building a reusable verse-based test collection for Arabic question answering on the Holy Qur'an. ACM Transactions on Asian and Low-Resource Language Information Processing, 19(6), Article 78, 1–21. https://doi.org/10.1145/3400396

Nacar, O., & Koubaa, A. (2024). Enhancing semantic similarity understanding in *Arabic NLP with nested embedding learning*. *arXiv preprint*, arXiv: 2407.21139. Doi:10.48550/arXiv.2407.21139.

O'Leary, M. (2010). Hakia gets serious with semantic search. *Information Today*, 27(6), pp 38-43.

Pastor, J. A., Martinez, F. J., & Rodriguez, J. V. (2009). Advantages of thesaurus representation using the Simple Knowledge Organization System (SKOS) compared with proposed alternatives. *Information Research*, 14(4). Retrieved from http://InformationR.net/ir/14-4/paper422.html

Pavlova, V. (2025). Multi-stage training of bilingual Islamic LLM for neural passage retrieval. In Proceedings of the New Horizons in Computational Linguistics for Religious Texts (pp. 42–52). Abu Dhabi, UAE: Association for Computational Linguistics.

Pavlova, V., & Makhlouf, M. (2024, November). Building an efficient multilingual non-profit IR system for the islamic domain leveraging multiprocessing design in rust. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track* (pp. 981-990).

Al-Smadi, M. (2025). QU-NLP at QIAS 2025 shared task: A two-phase LLM fine-tuning and retrieval-augmented generation approach for Islamic inheritance reasoning. In K. Darwish, A. Ali, I. Abu Farha, S. Touileb, I. Zitouni, A. Abdelali, S. Al-Ghamdi, S. Alkhereyf, W. Zaghouani, S. Khalifa, B. AlKhamissi, R. Almatham, I. Hamed, Z. Alyafeai, A. Alowisheq, G. Inoue, K. Mrini, & W. Alshammari (Eds.), *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks* (pp. 892–898). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.arabicnlp-sharedtasks.123.

Ahmad, M. A., Ballout, M., Ahmad, R. A., & Bruni, E. (2025). *Transformer Tafsir at QIAS 2025 shared task: Hybrid retrieval-augmented generation for Islamic knowledge question answering. arXiv [Preprint]*. https://doi.org/10.48550/arXiv.2509.23793

Tümer, D., Shah, M. A., & Bitirim, Y. (2009, May). An empirical evaluation on semantic search performance of keyword-based and semantic search engines: Google, yahoo, msn and hakia. In *2009 Fourth International Conference on Internet Monitoring and Protection* (pp. 51-55). IEEE..