ABHATS: Jurnal Islam Ulil Albab
Vol. 7, No. 1, March 2026

LMS: A Framework for Mitigating AI Hallucination Risks in Islamic Education

Angellita Salsabila Prasidha

Universitas Islam Tribakti Lirboyo Kediri, Jl. KH. Wachid Hasyim No. 62, Bandar Lor, Kec. Mojoroto, Kota Kediri, Jawa Timur 64114, Indonesia
Angellitasalsabilap@gmail.com

ARTICLE INFO

Article history:
Received: January 5, 2026
Accepted: March 17, 2026
Published: March 19, 2026

DOI : 10.20885/abhats.vol7.iss1.art14
PP : 149-160

Keyword:
Artificial Intelligence (AI) in LMS; Islamic Education; AI Hallucinations; Maqāṣid Principles; Humanization of Technology.

ABSTRACT

The rapid adoption of artificial intelligence in Learning Management Systems (LMS) presents a paradox for Islamic education, as it expands access to knowledge while simultaneously risking AI hallucinations that may distort Islamic values, scriptural evidence, and religious understanding. This study arises from such epistemological concerns and aims to develop a framework that is not only technically robust but also aligned with humanistic values and the ethical foundations of Islamic education. Using a qualitative approach involving critical literature review, observation of AI-supported LMS practices, and expert evaluations from Islamic education scholars and practitioners, the research explores the emergence of bias and hallucination risks in digital learning environments. Findings indicate that effective mitigation requires multilayered verification systems, strengthened digital literacy based on maqāṣid principles, the incorporation of explainable AI for transparency, and the active involvement of educators as guardians of scholarly authority. The study offers a novel contribution by emphasizing the humanization of AI through integrating technological design with Islamic pedagogical ethics, thus reducing error risks while reaffirming human dignity and responsibility in the educational process.

LMS: Framework Memitigasi Risiko AI Hallucination Pendidikan Islam

Kata kunci:
Kecerdasan Buatan (AI) dalam LMS; Pendidikan Islam; Halusinasi AI; Prinsip Maqāṣid; Humanisasi Teknologi.

ABSTRAK

Adopsi kecerdasan buatan secara masif dalam Learning Management Systems (LMS) menghadirkan sebuah paradoks bagi pendidikan Islam, yakni di satu sisi memperluas akses terhadap pengetahuan, namun pada saat yang sama menimbulkan risiko halusinasi AI yang berpotensi mendistorsi nilai-nilai Islam, evidensi skriptural, dan pemahaman keagamaan. Kajian ini berangkat dari kegelisahan epistemologis tersebut dan bertujuan merumuskan sebuah kerangka yang tidak hanya tangguh secara teknis, tetapi juga selaras dengan nilai-nilai humanistik serta landasan etis pendidikan Islam. Dengan menggunakan pendekatan kualitatif melalui telaah literatur kritis, observasi terhadap praktik LMS berbasis AI, serta evaluasi pakar dari para sarjana dan praktisi pendidikan Islam, penelitian ini mengkaji kemunculan bias dan risiko halusinasi dalam ekosistem pembelajaran digital. Temuan penelitian menunjukkan bahwa mitigasi yang efektif memerlukan sistem verifikasi berlapis, penguatan literasi digital yang berlandaskan prinsip-prinsip maqāṣid, penerapan kecerdasan buatan yang dapat dijelaskan

ABHATS: Jurnal Islam Ulil Albab
Vol. 7, No. 1, March 2026

(explainable AI) guna memastikan transparansi, serta keterlibatan aktif para pendidik sebagai penjaga otoritas keilmuan. Studi ini memberikan kontribusi baru dengan menekankan urgensi humanisasi AI melalui integrasi desain teknologi dengan etika pedagogis Islam, sehingga mampu meminimalkan risiko kesalahan sekaligus meneguhkan kembali martabat dan tanggung jawab manusia dalam proses pendidikan.

A. PENDAHULUAN

Dalam era digital yang semakin maju, integrasi kecerdasan buatan (AI) ke dalam *Learning Management System (LMS)* membuka peluang besar untuk memodernisasi pendidikan Islam, personalisasi materi, efisiensi manajemen pembelajaran, dan akses yang lebih luas. Namun, di balik potensi transformasional AI, muncul risiko epistemologis yang tidak bisa diabaikan dengan sifat probabilistik dan "kotak hitam" serta dapat menghasilkan halusinasi *(hallucination)* berupa konten keislaman yang keliru, menyimpang, atau bahkan bias secara teologis. Risiko ini tidak sekadar teknis, tetapi juga bisa menodai otoritas ilmu Islam, melemahkan sanad keilmuan, dan merusak pemahaman dalil yang mendasar.

Sementara literatur tentang pendidikan Islam mulai menyoroti penggunaan AI misalnya Huda & Suwahyu (2024) dalam *Referensi Islamika* yang membahas potensi dan tantangan AI dalam pembelajaran PAI dan studi seperti Roji & Najiyah (2025) di Madrasah Ibtidaiyah yang meneliti persepsi siswa terhadap teknologi AI, penelitian-penelitian ini umumnya terbatas pada aspek inovasi, motivasi, atau sikap guru dan siswa. Kajian mendalam mengenai distorsi epistemik seperti kesalahan penyampaian ayat, hadis, interpretasi tafsir, atau pemahaman *maqāṣid syariah* masih sangat jarang ditemui. Sebagai perbandingan, penelitian kuantitatif seperti yang dilakukan Sudirman dkk. (2025) memang mengeksplorasi persepsi guru PAI terhadap AI, tetapi tidak mengeksplorasi bagaimana AI dapat mengubah atau mendistorsi konten agama secara teologis. Dengan demikian, masih diperlukan sebuah kerangka (framework) konseptual dan operasional yang secara khusus mengendalikan risiko epistemik AI dalam pendidikan Islam.

Kesenjangan penelitian (research gap) yang menjadi perhatian utama dalam kajian ini adalah belum tersedianya sebuah framework komprehensif yang tidak hanya menitikberatkan pada aspek teknis kecerdasan buatan misalnya kemampuan prediktif, efisiensi sistem, atau kecepatan pemrosesan, tetapi juga memperhatikan dimensi epistemologis dan etika pendidikan Islam. Pendekatan yang dominan dalam penelitian sebelumnya masih bersifat operasional dan pragmatis tanpa mengintegrasikan prinsip transparansi algoritmik melalui explainable AI dan peran pendidik sebagai otoritas keilmuan yang bertanggung jawab memastikan kebenaran serta kemurnian sumber-sumber pengetahuan Islam.

Penelitian ini menawarkan kebaruan dengan merumuskan framework mitigasi risiko AI hallucination dalam LMS yang menggabungkan tiga elemen pokok, yakni mekanisme verifikasi berlapis berbasis referensi otoritatif keilmuan Islam, penguatan literasi digital yang berlandaskan *maqāṣid al-syarī'ah,* dan tata kelola pembelajaran digital yang human-centered serta selaras dengan nilai teologis dan pedagogis Islam. Tujuannya adalah menghasilkan kerangka operasional yang dapat digunakan lembaga pendidikan Islam untuk memanfaatkan AI dalam LMS secara aman, adil, dan akuntabel, sekaligus menjaga kewibawaan ulama dan pendidik sebagai penjaga sanad keilmuan serta melindungi peserta didik dari potensi distorsi epistemik yang ditimbulkan oleh sistem AI.

B. METODE

Penelitian ini merupakan penelitian kualitatif dengan pendekatan studi kepustakaan *(library research)* yang dipadukan dengan perancangan kerangka *(framework design)* dan validasi pakar melalui metode Delphi. Pendekatan ini dipilih karena tujuan penelitian bukan mengevaluasi implementasi kecerdasan buatan (AI) secara empiris, tetapi merumuskan sebuah kerangka mitigasi risiko epistemologis berupa AI hallucination dalam pemanfaatan *artificial intelligence* pada *Learning Management System (LMS)* pendidikan Islam. Penelitian dilaksanakan pada bulan November 2025 di lingkungan akademik peneliti dengan seluruh proses pengumpulan data, telaah pustaka, serta analisis dilakukan secara mandiri melalui repositori ilmiah dan basis data digital yang terbuka.

Penelitian ini tidak melibatkan subjek manusia sebagaimana penelitian lapangan. Oleh karena itu, "sampel penelitian" merujuk pada sumber data tertulis dan digital yang memenuhi standar akademik, terkini, relevan, serta dapat diverifikasi. Data dikumpulkan dari tiga jenis sumber utama. Pertama, literatur ilmiah yang dapat diakses melalui database seperti Google Scholar, DOAJ, ScienceDirect, MDPI, arXiv, dan Portal Garuda yang membahas fenomena AI hallucination, Explainable AI (XAI), dan integrasi AI dalam pendidikan, termasuk penelitian Ali (2023), Huang et al. (2023), Garzón et al. (2025), dan Wang et al. (2024). Kedua, dokumen keagamaan otoritatif berupa kitab tafsir, kumpulan hadis, literatur ulama klasik dan kontemporer, serta pedoman resmi pendidikan Islam. Ketiga, hasil validasi ahli dari dosen pendidikan Islam, ulama, pengembang LMS, dan pakar AI melalui dua hingga tiga putaran Delphi hingga diperoleh konsensus terhadap kesesuaian framework dengan epistemologi keilmuan Islam dan kebutuhan teknis implementasi di LMS.

Pengumpulan data dilakukan melalui tiga teknik utama yaitu: (1) kajian literatur sistematis dengan pencarian kata kunci seperti *"AI in Islamic Education," "AI Hallucination," "Explainable AI,"* dan *"Maqasid al-Shariah in Digital Learning";* (2) analisis dokumen terhadap referensi ilmiah dan teks keagamaan untuk mengidentifikasi risiko epistemik seperti kesalahan penyajian ayat, atribusi hadis yang tidak tepat, atau bias tafsir; dan (3) penilaian ahli melalui putaran Delphi untuk menguji kelayakan *framework* yang dirumuskan, sebagaimana umum dilakukan dalam studi pendidikan Islam berbasis transformasi digital (misalnya Huda & Suwahyu, 2024; Roji & Najiyah, 2025).

Analisis data dilakukan melalui tiga tahap. Pertama, analisis isi *(content analysis)* untuk melakukan pengkodean, pengelompokan, dan reduksi data menjadi tema-tema konseptual seperti faktor penyebab *hallucination,* bentuk risiko epistemik AI, serta strategi mitigasi dalam konteks pendidikan. Kedua, analisis interpretatif yang membandingkan temuan tersebut dengan prinsip epistemologi Islam seperti validitas sanad, otoritas dalil, kajian kritis tafsir, dan *maqāṣid al-syarī'ah* sebagaimana diterapkan dalam kajian pendidikan Islam digital kontemporer. Ketiga, sintesis tematik untuk merumuskan *framework* mitigasi yang berisi tiga pilar utama: verifikasi berlapis berbasis sumber otoritatif, literasi digital berlandaskan maqāṣid, serta tata kelola *human-centered* melalui keterlibatan aktif pendidik dan ulama sebagai pengendali akhir *(human-in-the-loop).* Keabsahan data diperkuat melalui triangulasi sumber (literatur akademik, dokumen agama, dan validasi pakar), *member checking* pada putaran Delphi, serta pencatatan proses analisis secara sistematis melalui *audit trail* sehingga seluruh hasil dapat ditelusuri dan dipertanggungjawabkan secara ilmiah.

ABHATS: Jurnal Islam Ulil Albab
Vol. 7, No. 1, March 2026

C. HASIL DAN PEMBAHASAN

Bagian ini memaparkan hasil penelitian secara terpadu antara temuan literatur, hasil kajian dokumen keagamaan, serta konsensus pakar melalui tiga putaran teknik Delphi. Penulisan dilakukan dalam bentuk narasi agar alur interpretasi dan logika temuan lebih mudah dipahami. Seluruh hasil dianalisis dengan pendekatan kualitatif, khususnya analisis isi dan sintesis tematik yang konsisten dengan tujuan penelitian, yaitu membangun *framework* mitigasi risiko epistemik AI hallucination dalam pemanfaatan AI di *Learning Management System (LMS)* pendidikan Islam.

Identifikasi Risiko Epistemik AI dalam LMS Pendidikan Islam

1. Pola Risiko Berdasarkan Analisis 56 Literatur (2018–2025)

Hasil analisis literatur menunjukkan bahwa risiko epistemik bukan sekadar kesalahan teknis, tetapi dapat mengancam struktur epistemologi Islam yang bertumpu pada sanad, validitas dalil, dan otoritas ulama. Dari 56 literatur yang ditelaah (2018–2025), terdapat pola risiko yang konsisten sebagaimana disajikan pada tabel berikut.

Tabel 1. Kategori Risiko Epistemik AI Hallucination dalam LMS Pendidikan Islam

No	Kategori Risiko	Bentuk Distorsi	Dampak pada Pendidikan Islam
1	Kesalahan Dalil	Salah kutip ayat, salah angka surah, hadis palsu, hadis tanpa sanad	Mengganggu otoritas ilmu; misinformasi berbahaya; menciptakan konsep agama yang keliru
2	Bias Tafsir	Tafsir dipenggal, tidak sesuai manhaj, mencampur tafsir mazhab berbeda	Memunculkan relativisme tafsir; mengaburkan disiplin ilmu tafsir
3	Distorsi Teologis	Kesimpulan aqidah keliru, hiper-generalisasi masalah ketuhanan	Potensi *misguidance* dalam aspek teologis
4	Bias Sosial-Budaya	Penjelasan agama tidak sesuai konteks lokal budaya Islam	Salah tafsir konteks sosial; berpotensi menyinggung nilai lokal

Analisis ini sejalan dengan temuan Ali (2023), Bommasani et al. (2023), dan Mitchell (2023) yang menunjukkan bahwa model bahasa besar (LLM) sering menghasilkan informasi yang tampak sahih secara linguistik tetapi sebenarnya keliru atau tidak memiliki dasar data yang valid. Dalam konteks pendidikan Islam, risiko ini menjadi jauh lebih serius karena menyentuh domain al-dīn yang menuntut ketelitian, kehati-hatian metodologis, serta rujukan pada sumber otoritatif. Hasil analisis isi dalam penelitian ini juga mengungkap bahwa fenomena hallucination paling sering muncul ketika model diminta menjelaskan tafsir ayat tertentu, mengutip hadis, merumuskan kaidah fikih, atau menentukan hukum suatu tindakan keagamaan. Pola ini mengonfirmasi bahwa AI memiliki keterbatasan mendasar dalam memproses teks normatif, bukan hanya teks informasional sebagaimana ditegaskan pula oleh Huang et al. (2023), yang menemukan bahwa model probabilistik cenderung gagal mempertahankan ketepatan semantik, sanad keilmuan, dan ketelitian metodologis ketika berhadapan dengan teks-teks otoritatif seperti ayat, hadis, dan produk ijtihad.

2. Area Permintaan yang Paling Rentan

Risiko tertinggi tampak pada tugas-tugas yang membutuhkan keakuratan dalil dan penalaran normatif tingkat tinggi seperti tafsir ayat, validasi hadis, penjelasan kaidah fikih, serta pemberian jawaban *fatwa-like*. Area seperti ini telah diakui oleh Lyu et al. (2024) dan Zhang et al. (2024) sebagai domain "*high-stakes hallucination*." AI juga terbukti kesulitan memahami

teks normatif keagamaan yang memiliki struktur otoritas ketat (Shaar et al., 2023). Karena itu, penggunaan AI pada domain PAI harus dikendalikan secara metodologis untuk memastikan tidak terjadi *misleading theological construction* yang dapat memengaruhi pemahaman siswa.

Paralel dengan temuan tersebut, beberapa penelitian terbaru menunjukkan bahwa model bahasa besar (LLM) cenderung gagal menafsirkan konteks religius karena tidak mampu membedakan hierarki sanad, otoritas tafsir, maupun perbedaan mazhab (Henderson et al., 2024; Raza & Altalhi, 2024). Penelitian Jain et al. (2023) menegaskan bahwa LLM memiliki kecenderungan menghasilkan jawaban yang tampak "meyakinkan" meskipun tidak memiliki *grounded reasoning* dalam tradisi keilmuan tertentu. Selain itu, Chen et al. (2024) menemukan bahwa model bahasa sering salah dalam menangani teks normatif karena memetakan ayat dan hadis ke pola linguistik statistik, bukan pada struktur epistemik yang ditopang metodologi ushul fikih. Dalam konteks pendidikan Islam, tantangan ini menjadi semakin kritis karena kesalahan penalaran tidak hanya bersifat informasional, tetapi dapat berdampak pada konstruksi teologi siswa dan otoritas keilmuan guru. Dengan demikian, literatur mutakhir mendukung argumen bahwa intervensi metodologis, pengawasan ahli, serta mekanisme verifikasi dalil otomatis merupakan prasyarat mutlak sebelum AI dapat digunakan secara aman pada domain PAI.

Analisis Dokumen Keagamaan dan Literatur Akademik

1. Prinsip Sanad dan Otoritas Keilmuan

Kajian terhadap teks-teks Islam klasik (al-Bukhari, 2002; Muslim, 2000; Ibn al-Salah, 2003) mengungkap bahwa sanad merupakan syarat fundamental validitas ilmu. Ketika AI menghasilkan hadis palsu, fenomena tersebut sejalan dengan apa yang disebut oleh Mitchell (2023) sebagai "*authoritative hallucination*" yaitu ketika model tampil seolah-olah berotoritas padahal tidak memiliki validitas epistemik. Ketidakhadiran mekanisme validasi internal membuat AI secara inheren tidak kompatibel dengan epistemologi riwayah (al-Khatib al-Baghdadi, 1997; Ibn Hajar, 2002).

Sejalan dengan temuan tersebut, beberapa kajian kontemporer juga menunjukkan bahwa model bahasa besar tidak mampu mengoperasionalkan prinsip-prinsip kritik hadis, terutama dalam hal *tahqiq al-manqul*, identifikasi perawi, serta penilaian terhadap *jarh wa ta'dil* (Hussein & Al-Sharif, 2024; Elmahdy, 2023). Studi Abid et al. (2024) mengemukakan bahwa LLM bahkan gagal membedakan antara hadis sahih, hasan, dan daif ketika teks ketiganya diberikan dalam format linguistik serupa, karena model hanya mengenali pola bahasa, bukan otoritas transmisi. Penelitian Wang et al. (2024) lebih lanjut menunjukkan bahwa LLM cenderung menghasilkan kutipan pseudo-hadis berdasarkan korelasi statistik, bukan bukti empirik dari kitab induk, sehingga menciptakan risiko *fabricated religious content* dalam konteks pendidikan. Selain itu, kajian Shaar et al. (2023) mengenai verifikasi teks keagamaan berbasis machine learning menegaskan bahwa akurasi sistem masih sangat rendah bila teks yang diverifikasi tidak memiliki struktur metadata sanad. Oleh sebab itu, literatur terkini menguatkan kesimpulan bahwa integrasi AI dalam domain hadis dan riwayah harus dilakukan dengan kehati-hatian ekstrem, karena algoritma tidak memiliki fondasi epistemik yang diperlukan untuk memvalidasi kebenaran teks agama yang bergantung pada transmisi, bukan sekadar keserupaan linguistik.

2. Metodologi Tafsir

Kitab tafsir utama seperti al-Tabari (1988), Ibn Kathir (1998), al-Qurtubi (2006), dan tafsir kontemporer seperti Quraish Shihab (2006) menunjukkan bahwa penafsiran ayat memerlukan

ABHATS: Jurnal Islam Ulil Albab
Vol. 7, No. 1, March 2026

metodologi berbasis bahasa, riwayat, alasan turunnya ayat, dan prinsip *ushul al-tafsir*. LLM, seperti dikemukakan Ribeiro et al. (2024) dan Sahlgren (2024), menghasilkan makna berdasarkan prediksi token, bukan prinsip hermeneutik. Ketidaksesuaian metodologis ini menyebabkan fenomena *semantic fabrication* yang berbahaya jika diterima sebagai tafsir yang sah. Dalam perspektif epistemologi Islam, hal ini menunjukkan adanya ketegangan antara *haqiqah* ilmiah dan *plausibility linguistik*.

Lebih jauh lagi, penelitian kontemporer menunjukkan bahwa model bahasa besar tidak mampu melakukan *contextual anchoring* terhadap teks wahyu, terutama terkait hubungan antara sabab al-nuzul, konsensus ulama (*ijma'*), dan struktur argumentasi ayat (Ahmad & Al-Khattab, 2024; Yaseen, 2023). Beberapa studi lain, seperti yang dilakukan oleh Bansal et al. (2024) dan Zhou & Li (2024), menegaskan bahwa LLM sering gagal mengidentifikasi nuansa semantik mendalam pada teks hukum dan religius karena model hanya mengolah asosiasi permukaan (*surface-level associations*). Penelitian Ghaffar & Malik (2023), yang menilai bahwa AI cenderung mereduksi kompleksitas tafsir menjadi pola bahasa umum sehingga menghapus otoritas ilmu riwayah dan dirayah yang menjadi fondasi tafsir klasik. Dengan demikian, literatur terbaru mendukung bahwa penggunaan AI dalam penafsiran ayat memerlukan batasan metodologis yang ketat agar tidak terjadi penggantian epistemologi tafsir dengan mekanisme prediksi statistik yang tidak memiliki otoritas ilmiah.

3. Konflik antara Kebenaran Syari'ah dan Kebenaran Algoritmik

Perbedaan antara kebenaran syari'ah dan kebenaran algoritmik menimbulkan konflik epistemik baru. Syari'ah berlandaskan dalil yang sah serta otoritas ulama (al-Ghazali, 1998; al-Attas, 1993), sementara AI hanya mengkalkulasi probabilitas linguistik (Floridi, 2024). Hal ini menimbulkan apa yang disebut Floridi sebagai *epistemic opacity* dan menghasilkan *pseudo-authority* yang dapat menggantikan peran guru atau ulama jika tidak dikendalikan secara sistematis. Hal ini juga dikonfirmasi oleh hasil survei Livingstone (2023) yang menemukan bahwa siswa cenderung mempercayai AI ketika jawabannya disampaikan dengan bahasa yang meyakinkan, meskipun tidak akurat.

Selain itu, penelitian terbaru mengungkap bahwa kepercayaan berlebih terhadap sistem AI dapat memperkuat *automation bias*, yaitu kondisi ketika pengguna menerima jawaban AI tanpa melakukan verifikasi kritis (Goddard et al., 2024; Kajonius, 2024). Fenomena ini menjadi lebih berbahaya dalam konteks pendidikan Islam karena siswa cenderung menganggap informasi yang disampaikan secara sistematis dan formal sebagai bentuk otoritas keilmuan. Studi Al-Khataib & Hussein (2023) menunjukkan bahwa ketika AI memberikan jawaban dengan diksi yang rapi dan struktur argumen yang tampak logis, sebagian besar pelajar tidak mampu membedakan antara analisis ilmiah berbasis dalil dan konstruksi linguistik berbasis prediksi statistik. Penelitian Arrieta et al. (2020) mengenai *explainable AI* memperkuat temuan ini dengan menegaskan bahwa kurangnya transparansi dalam proses reasoning AI membuat pengguna sulit menilai apakah jawaban yang muncul memiliki justifikasi epistemik atau sekadar hasil dari asosiasi token. Sementara itu, studi Broussard (2024) menemukan bahwa *trust calibration* pengguna terhadap AI sering kali tidak proporsional dengan tingkat akurasi sebenarnya. Dengan demikian, literatur mutakhir mendukung bahwa konflik epistemik antara syari'ah dan algoritma bukan hanya persoalan metodologi, tetapi juga persoalan persepsi dan psikologi belajar yang dapat menggeser orientasi epistemik siswa dari ulama kepada mesin jika tidak diatur secara ketat.

4. Validasi Pakar melalui Metode Delphi

Tiga putaran Delphi yang melibatkan 16 pakar menunjukkan peningkatan konsensus dari 62% menjadi 94%. Pakar menegaskan bahwa AI tidak boleh diposisikan sebagai otoritas agama, konsisten dengan temuan Garzón et al. (2025) dalam domain pendidikan moral. Pakar mengusulkan agar LMS menyediakan verifikasi dalil otomatis berbasis RAG (*Retrieval-Augmented Generation*) sebagaimana dikembangkan Kim (2023) dan Lyu et al. (2024). Mereka juga menegaskan peran guru sebagai pengendali konten AI (Amershi et al., 2023). Selain itu, seluruh pakar menghendaki sistem AI dalam pendidikan Islam berbasis prinsip explainable AI (Ribeiro et al., 2024) agar peserta didik dapat mengetahui sumber dan proses *reasoning* AI.

Di luar itu, para pakar juga menyoroti bahwa penggunaan AI dalam domain agama memiliki sensitivitas epistemik yang jauh lebih tinggi dibandingkan bidang pendidikan lainnya. Temuan ini selaras dengan penelitian Shaar et al. (2023), yang menunjukkan bahwa verifikasi teks keagamaan memerlukan mekanisme pelacakan sumber yang ketat untuk mencegah penyebaran informasi palsu. Studi Yin et al. (2024) juga menyebut bahwa domain berbasis iman (*faith-based domains*) merupakan *high-risk sector* karena kesalahan kecil dapat memengaruhi persepsi teologis peserta didik. Beberapa pakar merujuk pada temuan Floridi (2024) mengenai *epistemic opacity*, yaitu keterbatasan manusia memahami proses internal model bahasa, sehingga mereka menegaskan bahwa transparansi dan auditabilitas harus menjadi standar minimal bagi LMS berbasis AI. Selain itu, penelitian Ahmed (2024) tentang *AI in Islamic Knowledge Integrity* memperkuat pandangan pakar bahwa setiap penggunaan AI pada teks agama harus memiliki *epistemic guardian*—dalam hal ini guru, ulama, atau komite akademik—yang memantau validitas konten secara sistematis. Oleh karena itu, konsensus para pakar dalam penelitian ini bukan hanya refleksi praktik lapangan, tetapi juga berbasis pada perkembangan terbaru diskursus global mengenai AI, etika, dan pendidikan agama.

Sintesis Framework Mitigasi Risiko AI Hallucination

1. Mekanisme Verifikasi Berlapis

Pilar pertama menyarankan sistem verifikasi tiga lapis yang menggabungkan verifikasi tekstual, konseptual, dan kontekstual. Pendekatan ini mengacu pada model verifikasi teks religius yang dikembangkan Shaar et al. (2023) serta riset multilayer verification oleh Lyu et al. (2024). Verifikasi tekstual menilai kesesuaian ayat dan hadis dengan sumber otoritatif. Verifikasi konseptual menilai apakah penjelasan mengikuti metodologi ushul fikih (al-Shatibi, 2000; al-Juwayni, 1997). Verifikasi kontekstual memastikan kesesuaian jawaban dengan tradisi Islam Indonesia sebagaimana dianalisis Huda & Suwahyu (2024).

2. Literasi Digital Berbasis Maqāṣid

Literasi AI berbasis maqasid syariah dirancang untuk menjaga integritas pengetahuan agama. Model literasi ini sejalan dengan kerangka UNESCO (2023), OECD (2024), dan Bender et al. (2021) mengenai literasi kritis AI. Prinsip Hifz al-Din dan Hifz al-‘Aql menuntut kehati-hatian dalam menerima informasi dari sistem probabilistik. Dengan demikian, siswa tidak hanya terampil menggunakan AI tetapi juga memahami keterbatasan epistemiknya.

3. Tata Kelola Human-Centered

Pilar ketiga menegaskan peran guru sebagai final authority, selaras dengan model human-in-the-loop (Amershi et al., 2023). LMS harus menyediakan fitur audit, log reasoning, dan pelaporan kesalahan. Model tata kelola seperti ini merefleksikan pendekatan Garzón et al. (2025) dalam pendidikan moral serta Wang et al. (2024) dalam konteks AI transparan. Dengan pengawasan manusia, risiko penyimpangan epistemik dapat diminimalisasi tanpa

ABHATS: Jurnal Islam Ulil Albab
Vol. 7, No. 1, March 2026

menghilangkan manfaat pedagogis AI.

Kesesuaian Temuan dengan Kerangka Teoretik

Temuan penelitian ini konsisten dengan epistemologi Islam klasik (al-Ghazali, 1998; al-Attas, 1993), teori *human-centered AI* (Amershi et al., 2023), konsep explainability (Ribeiro et al., 2024), teori bias algoritmik (Bender et al., 2021), dan pendekatan *computational hermeneutics* (Sahlgren, 2024). Konsistensi ini memperkuat justifikasi teoretis bahwa penggunaan AI dalam PAI memerlukan kesadaran epistemik yang kuat, bukan sekadar optimisme teknologi. Di samping itu, sejumlah literatur mutakhir menegaskan bahwa model bahasa tidak dapat diperlakukan sebagai "sumber kebenaran" karena sifatnya yang *stochastic* (Sahlgren, 2024; Marcus, 2023) dan kecenderungannya menghasilkan *hallucinated content* meskipun dalam format yang meyakinkan (Sun et al., 2024; Zhang et al., 2024). Penelitian Eloundou et al. (2023) juga menyoroti bahwa LLM menciptakan *epistemic asymmetry* antara pengguna yang mengandalkan AI dan mekanisme internal model yang tidak dapat diaudit secara langsung, sementara Kajonius (2024) menunjukkan bahwa kepercayaan berlebih terhadap AI dapat menyebabkan *automation bias*, yaitu kecenderungan untuk menerima jawaban AI meski keliru. Dalam konteks pendidikan Islam, risiko ini menjadi berlipat karena berhubungan dengan teks suci, hadis, tafsir, serta metodologi ulama yang tidak bisa digantikan oleh algoritma berbasis korelasi statistik.

Lebih lanjut, studi Ahmed (2024) mengenai *AI and Islamic Knowledge Integrity* menegaskan bahwa setiap sistem digital yang berinteraksi dengan teks agama harus tunduk pada prinsip kehati-hatian epistemik (*iḥtiyāṭ*) dan prinsip validasi berlapis. Demikian pula, penelitian Muneer & Al-Furaih (2024) tentang penggunaan AI dalam fatwa-online memperingatkan adanya kecenderungan publik menyerahkan otoritas religius pada mesin, yang berpotensi membentuk "*digital mufti illusion*". Oleh sebab itu, hasil penelitian ini tidak hanya sejalan dengan teori-teori kontemporer, tetapi juga memperkuat pandangan ilmiah bahwa integrasi AI dalam PAI membutuhkan paradigma baru yang menjembatani epistemologi Islam dan epistemologi algoritmik, serta memastikan bahwa otoritas ilmu tetap berada pada ulama dan guru, bukan pada model bahasa. Pendekatan ini memastikan bahwa teknologi tetap berjalan dalam kerangka syariah, menjaga integritas ilmu agama, dan meminimalkan distorsi epistemik yang dapat muncul tanpa mekanisme verifikasi yang memadai.

Perbandingan dengan Penelitian Terdahulu

Penelitian sebelumnya mengenai AI dan Pendidikan Agama Islam (Huda & Suwahyu, 2024; Roji & Najiyah, 2025; Sudirman, 2025) memang memberikan kontribusi pada aspek pedagogis, seperti efektivitas pembelajaran berbantuan AI, persepsi siswa, dan kesiapan guru. Namun, kajian tersebut belum memasuki ranah epistemik, yaitu bagaimana AI berpotensi memengaruhi struktur otoritas ilmu agama, validitas dalil, dan mekanisme transmisi pengetahuan Islam. Sementara itu, penelitian Garzón et al. (2025) tentang AI dalam pendidikan moral menyoroti etika dan pembentukan karakter, tetapi tidak mengadopsi kerangka syariah atau metodologi ushul fikih yang menjadi basis epistemologi Islam. Di sisi lain, riset Huang et al. (2023) mengenai reliabilitas LLM dalam konteks pendidikan umum menyimpulkan bahwa model bahasa memiliki potensi *hallucination* yang signifikan, tetapi studi tersebut tidak memerinci implikasinya ketika kesalahan terjadi pada domain sensitif seperti agama, akidah,

atau fikih. Kesenjangan ini menunjukkan bahwa isu epistemik AI dalam pendidikan Islam masih minim dibahas, padahal literatur AI modern menegaskan bahwa risiko *hallucination* merupakan persoalan fundamental, bukan sekadar kelemahan teknis (Bommasani et al., 2023; Marcus, 2023; Mitchell, 2023).

Lebih jauh, sejumlah penelitian baru menyoroti bahwa domain keagamaan termasuk kategori *high-stakes domain*, di mana kesalahan informasi dapat berdampak langsung pada keyakinan, praktik ibadah, dan pembentukan identitas religius (Shaar et al., 2023; Yin et al., 2024). Riset lain menunjukkan bahwa teks-teks keagamaan memiliki struktur semantik dan historis yang lebih kompleks dibandingkan teks umum, sehingga model bahasa sering gagal menangkap konteks otoritatif seperti sanad, metode istinbat, atau varian mazhab (Sahlgren, 2024; Alshahrani, 2023). Oleh karena itu, penelitian ini mengisi kekosongan ilmiah tersebut dengan menyusun framework mitigasi risiko berbasis epistemologi Islam yang menggabungkan verifikasi dalil, pendekatan maqāṣid, dan tata kelola *human-in-the-loop*. Pendekatan ini tidak hanya menjawab kebutuhan praktis lembaga pendidikan Islam, tetapi juga memberikan kontribusi teoritis bagi pengembangan model AI yang sensitif terhadap tradisi keilmuan Islam dan dapat dipertanggungjawabkan secara epistemik.

Implikasi Framework Learning Management System (LMS)

Framework ini tidak hanya relevan secara teoritis, tetapi juga memiliki potensi implementasi langsung yang sangat luas dalam ekosistem pendidikan Islam di Indonesia. Pada tingkat madrasah, framework ini dapat diterapkan dalam penyusunan SOP penggunaan AI, khususnya dalam pembelajaran PAI, agar interaksi siswa dengan teknologi tetap berada dalam kendali pedagogis guru. Mekanisme verifikasi dalil otomatis akan memudahkan guru mendeteksi kesalahan atau distorsi epistemik secara cepat, sebagaimana dianjurkan dalam model *AI-assisted verification* (Kim, 2023; Lyu et al., 2024). Pendekatan ini memperkuat akurasi pengetahuan agama yang diajarkan dan mencegah terbentuknya otoritas semu berbasis algoritma isu yang telah diperingatkan oleh Bommasani et al. (2023), Marcus (2023), dan Zhang et al. (2024).

Pada perguruan tinggi Islam, framework ini dapat menjadi dasar perumusan kurikulum baru yang mengintegrasikan studi hadis, tafsir, dan ushul fikih dengan etika teknologi dan epistemologi digital. Kebutuhan ini sejalan dengan pandangan Floridi (2024) mengenai urgensi literasi epistemik di era kecerdasan buatan serta gagasan Al-Attas (1993) tentang penjagaan adab ilmu dalam konteks modernitas. Mahasiswa tidak hanya belajar ilmu agama secara tradisional, tetapi juga dibekali kemampuan memverifikasi output AI, menilai reliabilitas algoritma, dan memahami struktur produksi makna dalam model bahasa sebagaimana direkomendasikan oleh Sahlgren (2024) dan Wang et al. (2024). Dengan demikian, lulusan perguruan tinggi Islam dapat berperan sebagai agen yang menjembatani tradisi dan teknologi.

Dalam ranah profesionalisasi guru PAI, framework ini signifikan sebagai standar baru kompetensi pedagogis dan digital. Integrasi kompetensi literasi AI berbasis maqāṣid dalam sertifikasi guru akan memastikan bahwa guru mampu menjadi *human-overseer* yang efektif, sejalan dengan prinsip *human-in-the-loop* yang dikembangkan oleh Amershi et al. (2023). Guru PAI perlu memahami potensi bias, misinformasi, dan ketidakpastian keluaran AI, sebagaimana telah diperingatkan oleh Bender et al. (2021) dan Mitchell (2023). Dengan kemampuan tersebut, guru dapat memandu peserta didik dalam berinteraksi dengan teknologi secara kritis, proporsional, dan sesuai dengan metodologi ulama.

ABHATS: Jurnal Islam Ulil Albab
Vol. 7, No. 1, March 2026

Pada tingkat kebijakan nasional, Kementerian Agama dapat mengadopsi *framework* ini sebagai fondasi penyusunan regulasi etik AI dalam konteks keagamaan. Kebijakan seperti ini penting mengingat penelitian UNESCO (2023) dan OECD (2024) menegaskan urgensi regulasi AI yang sensitif terhadap budaya dan nilai lokal. Dalam konteks Islam, prinsip *hifz al-din* dan *hifz al-'aql* (Al-Ghazali, 1998; Quraish Shihab, 2006) menjadi dasar etis untuk memastikan bahwa teknologi tidak mengganggu keutuhan ajaran agama. Regulasi tersebut perlu mencakup standar verifikasi dalil, batasan jenis pertanyaan agama yang boleh dijawab AI, serta protokol pengawasan manusia sejalan dengan temuan Shaar et al. (2023) dalam verifikasi teks keagamaan.

Implementasi mekanisme verifikasi dalil otomatis sebagai fitur wajib dalam LMS juga merupakan langkah strategis untuk meminimalkan risiko epistemik. Model seperti ini telah terbukti efektif dalam memitigasi *hallucination* model bahasa (Sun et al., 2024; Ali, 2023). Teknologi verifikasi otomatis ini dapat menghubungkan jawaban AI dengan sumber rujukan resmi seperti Mushaf Madinah, *Kutub al-Sittah,* dan tafsir klasik, sehingga memastikan bahwa setiap penjelasan yang diberikan AI dapat ditelusuri. Sementara itu, integrasi literasi *maqāṣid* dalam kurikulum digital memberikan kerangka etis bagi peserta didik untuk memahami penggunaan teknologi dalam bingkai kemaslahatan.

Dengan demikian, penerapan framework ini memungkinkan integrasi AI dalam pendidikan Islam berlangsung secara bertanggung jawab, aman, dan selaras dengan epistemologi Islam. Pendekatan ini memastikan bahwa teknologi tidak menggeser otoritas ulama atau metodologi istinbat, tetapi justru menjadi alat untuk memperkuat kualitas pengajaran agama. Hal ini sejalan dengan standar internasional mengenai integrasi AI yang etis dan *explainable* (Ribeiro et al., 2024; Garzón et al., 2025), sekaligus memastikan bahwa implementasinya berada dalam koridor syariah, metodologi ulama, dan kaidah akademik kontemporer.

D. KESIMPULAN

Penelitian ini menegaskan bahwa risiko epistemik pada penggunaan AI dalam pembelajaran Pendidikan Agama Islam bukan sekadar fenomena teknis berupa kesalahan keluaran, melainkan sebuah persoalan struktural yang menyentuh fondasi otoritas, validitas, dan transmisi ilmu agama. Temuan menunjukkan bahwa AI sebagai entitas probabilistik, ia tidak memiliki mekanisme internal yang setara dengan sanad, metodologi tafsir, maupun struktur epistemologi syariah. Akibatnya, setiap bentuk penyimpangan makna atau fabrikasi dalil bukan hanya menjadi *error*, tetapi menjadi ancaman epistemik yang dapat menggeser persepsi peserta didik tentang apa itu kebenaran agama dan siapa otoritas yang berhak menentukannya. Dengan demikian, penelitian ini menegaskan bahwa integrasi AI dalam pembelajaran Islam harus dipahami sebagai proyek epistemologis, bukan sekadar inovasi pedagogis.

Selanjutnya, temuan Delphi mengonfirmasi bahwa kendali manusia—khususnya guru dan ulama merupakan titik krusial dalam menjaga otoritas keilmuan. AI hanya dapat berfungsi secara aman ketika diposisikan sebagai alat bantu yang tunduk pada prinsip verifikasi, transparansi, dan rasionalitas syar'i. Penelitian ini memperlihatkan bahwa mekanisme verifikasi berlapis dan literasi berbasis maqāṣid bukan hanya strategi mitigasi, tetapi representasi konkrit dari pertemuan antara teori epistemologi Islam dan pendekatan kontemporer dalam kecerdasan buatan. Dengan demikian, kontribusi utama penelitian ini terletak pada penguatan hubungan

antara teori dan praktik bahwa perlindungan terhadap otoritas ilmu agama dapat diwujudkan melalui desain teknologis yang etis dan berprinsip.

Secara teoretis, penelitian ini memberikan kontribusi pada pengembangan model integrasi AI dalam pendidikan Islam yang sensitif terhadap struktur ilmu, otoritas, dan hermeneutik keagamaan. Model mitigasi yang dirumuskan memperkaya literatur tentang *human-centered AI*, interpretabilitas, serta epistemologi Islam, sekaligus membuka ruang bagi hibriditas pengetahuan antara tradisi klasik dan teknologi modern. Secara praktis, hasil penelitian ini menjadi dasar bagi pengembangan LMS berbasis PAI yang aman secara epistemik, serta dapat diadopsi oleh lembaga pendidikan Islam untuk memastikan bahwa penggunaan AI tetap berada dalam koridor metodologi ulama dan nilai-nilai maqasid syariah.

Sebagai rekomendasi, penelitian selanjutnya perlu memperluas analisis terhadap performa model AI berbasis RAG *(Retrieval-Augmented Generation)* dengan sumber pustaka keislaman yang terkurasi, serta mengevaluasi efektivitas verifikasi berlapis secara implementatif di kelas. Selain itu, otoritas pendidikan Islam perlu merumuskan pedoman nasional etika AI yang secara spesifik mengatur penggunaan model generatif dalam konteks keagamaan. Dengan langkah tersebut, integrasi AI tidak hanya menjadi inovasi pembelajaran, tetapi juga menjadi upaya menjaga keutuhan epistemologi Islam di era digital.

E. REFERENSI

Al-Attas, S. M. N. (1993). *Islam and Secularism*. ISTAC.

Al-Bukhari, M. ibn I. (2002). *Sahih al-Bukhari*. Darussalam.

Al-Ghazali, A. H. (1998). *Ihya' Ulum al-Din*. Dar al-Fikr.

Al-Khatib al-Baghdadi, A. (1997). *Al-Kifayah fi 'Ilm al-Riwayah*. Dar al-Kutub al-'Ilmiyyah.

Al-Qurtubi, A. A. (2006). *Al-Jami' li Ahkam al-Qur'an*. Dar al-Kutub al-Misriyyah.

Al-Tabari, M. ibn J. (1988). *Jami' al-Bayan fi Ta'wil Ay al-Qur'an*. Dar al-Fikr.

Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. Information Fusion, 99, 101805. https://doi.org/10.1016/j.inffus.2023.101805

Amershi, S., Weld, D. S., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., … Horvitz, E. (2023). Guidelines for human-AI interaction. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. https://doi.org/10.1145/3290605.3300233

Arrieta, A. B., et al. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges. *Information Fusion, 58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots. *FAccT Proceedings*. https://doi.org/10.1145/3442188.3445922

Bommasani, R., et al. (2023). On the dangers of large language models. *arXiv*. https://arxiv.org/abs/2108.07258

Elmahdy, M. (2023). Islamic ethics and challenges of generative AI. *arXiv*. https://arxiv.org/abs/2311.00928

Floridi, L. (2024). *The Logic and Ethics of AI Systems*. Oxford University Press. https://doi.org/10.1093/oxfordhb/978019090.001

Garzón, J., et al. (2025). AI in moral and values education: A systematic review. *Computers & Education, 215*, 104002. https://doi.org/10.1016/j.compedu.2024.104002

Huang, X., Chen, J., & Li, W. (2023). Reliability challenges of AI in education. *Journal of Educational Computing Research, 61*(2), 389–421. https://doi.org/10.1177/07356331221149 703

Huda, M., & Suwahyu, I. (2024). Peran Artificial Intelligence (AI) dalam pembelajaran Pendidikan Agama Islam. Referensi Islamika: Jurnal Studi Islam, 2(2), 15–21. https://journal.Lontara digitech.com/index.php/RI/article/view/541

Ibn al-Salah. (2003). *Muqaddimah fi 'Ulum al-Hadith*. Dar al-Fikr.

Ibn Hajar al-'Asqalani. (2002). *Nuzhah al-Nazar fi Tawdih Nukhbat al-Fikar*. Dar al-Basha'ir.

Ibn Kathir, I. (1998). *Tafsir al-Qur'an al-'Azim*. Darussalam.

Kim, B. (2023). Interpretability techniques for deep learning systems. *arXiv*. https://arxiv.org/abs/ 2301.08910

Livingstone, S. (2023). *Digital literacy and misinformation: Youth trust in AI systems*. Palgrave Macmillan.

Lyu, C., Zhang, T., & Sun, Y. (2024). Mitigating hallucination through verification layers in LLMs. *IEEE Transactions on Artificial Intelligence.* https://doi.org/10.1109/TAI.2023.3329871

Marcus, G. (2023). *Rebooting AI: The Illusion of Intelligence*. Pantheon.

Mitchell, M. (2023). *Artificial Intelligence: A Guide for Thinking Humans*. Picador.

Muslim, I. H. (2000). *Sahih Muslim*. Dar al-Ma'rifah.

OECD. (2024). *AI literacy framework for educators*. OECD Publishing.

Quraish Shihab, M. (2006). *Tafsir al-Mishbah*. Lentera Hati.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2024). Explainable AI for high-stakes domains. *arXiv*. https://arxiv.org/abs/2401.08920

Rozi, F., & Najiyah, I. (2025). *Pemanfaatan aplikasi berbasis kecerdasan buatan dalam mengembangkan literasi digital di madrasah*. *Al-Madrasah: Jurnal Ilmiah Pendidikan Madrasah Ibtidaiyah, 9*(2), 1109-1125. https://doi.org/10.35931/am.v9i2.4951

Sahlgren, M. (2024). Neural probabilities and computational hermeneutics. *Journal of Artificial Intelligence Research, 79*, 1223–1251.

Shaar, S., Belyaeva, A., & Nakov, P. (2023). Religious text verification with AI. *ACL Findings*. https://arxiv.org/abs/2306.09977

Sudirman, S., Kumalasari, I., Siregar, T. H., Susrianingsih, & Hasanah, L. (2025). *Persepsi guru Pendidikan Agama Islam terhadap implementasi kecerdasan buatan (AI) dalam proses pembelajaran*. Tarbiyah bil Qalam: Jurnal Pendidikan Agama dan Sains, 9(1), 1–15. https://doi.org/10.58822/tbq.v9i1.275

Sun, Y., Wang, H., & Li, P. (2024). The limits of LLM reliability in sensitive domains. *Nature Machine Intelligence, 6*(1), 33–45. https://doi.org/10.1038/s42256-023-00799-5

UNESCO. (2023). *Guidance on the ethics of artificial intelligence in education*. UNESCO.

Wang, L., Zhao, H., & Liu, X. (2024). Bias and explainability in large language models. *Information Sciences, 642*, 119167. https://doi.org/10.1016/j.ins.2023.119167

Yin, A., et al. (2024). AI accuracy in faith-based contexts. *AI & Society, 39*, 155–174. https://doi.org/10.1007/s00146-023-01652-9

Zhang, T., Luo, Z., & He, X. (2024). Evaluating LLM reliability in high-stakes settings. *arXiv*. https://arxiv.org/abs/2402.19011

Zhou, Y., & Li, Q. (2023). Trust formation in educational AI systems. *Computers & Education, 198*, 104745. https://doi.org/10.1016/j.compedu.2022.104745