AI Hallucination vs. *Tawthīq*: A Critical Epistemological Review of the Reliability of Large Language Models (LLMs) in Fatwa Generation

Rully Agung Firmansyah
Institut Ahmad Dahlan Probolinggo
Email: rullyaf41@gmail.com

Benny Prasetiya
Institut Ahmad Dahlan Probolinggo
Email: prasetiyabenny@gmail.com

Abstract

The revolution of Generative Artificial Intelligence (AI) and Large Language Models (LLMs) has shifted the paradigm of Islamic knowledge acquisition from a traditional chain-of-transmission (*sanad*) and verification (*tawthīq*) framework to algorithmic probabilities. This reliance poses a severe epistemological threat known as "AI hallucination", output that appears fluent and authoritative yet is factually incorrect, manifesting as hallucinations in *isnad*, *matn*, *takhrij*, fatwas, and *manhaj*. This study aims to analyze the computational mechanics of AI hallucination, evaluate the epistemic authority of LLMs based on Hadith and *Usul al-Fiqh* criteria for news acceptance (*khabar*), and formulate an Islamic AI governance framework. Employing a Critical Epistemological Review methodology, this research contrasts the probabilistic mechanics of LLMs ("stochastic parrots") with the *tawthīq* principles of the *Salafus Shalih*. The findings reveal that LLMs fundamentally operate on token statistical probabilities without semantic comprehension (*fahm*), moral integrity (*'adālah*), legal memory precision (*ḍabṭ*), or continuous chains of transmission (*sanad muttaṣil*). Epistemologically, AI is classified as *majhūl al-ḥāl wa al-'ain*, meaning it cannot serve as a binding proof (*ḥujjah*) or a *mufti*, but strictly as a search tool (*adāt al-baḥth*). As a solution, this study formulates the "Digital *Tawthīq*" framework, integrating Hadith criticism methodology into modern AI architecture through ten technical components, including Closed Corpus Constraints, Retrieval-Augmented Generation (RAG) with digital *sanad* metadata, Knowledge Graph Sanad, and Human-in-the-Loop scholar verification. This framework repositions AI as a supportive instrument while safeguarding the purity of Islamic revelation in the digital era.

Keywords: AI hallucination, *Tawthīq,* islamic epistemology, large language models (llms), sanad, AI governance

Introduction

The Generative AI revolution has transformed the way humans access knowledge. Whereas the Muslim community previously relied on classical heritage texts (*turāth*), scholarly assemblies (*majālis al-ʿilm*), edicts of the *ʿulamāʾ* (*fatāwā*), and *iftāʾ* institutions, a segment of modern society has now shifted toward applications powered by Large Language Models (LLMs), such as ChatGPT, Gemini, Claude, Copilot, and other local models. This phenomenon is not merely a change of medium but an epistemological shift: from knowledge retrieval grounded in *sanad*, *ijāzah*, and *tawthīq* toward retrieval predicated on linguistic probability and algorithmic speed.

LLMs are computational systems trained on massive-scale text corpora. They predict the next token based on probability distributions learned from data (Achiam et al., 2023; Brown et al., 2020; Vaswani et al., 2017) . Technically, LLMs do not "know" meaning in the way humans do; they process statistical patterns. They can imitate the rhetorical style of the *ʿulamāʾ*, replicate the structure of *fatāwā*, display Arabic quotations, and create an authoritative impression. However, an authoritative impression is not equivalent to scholarly authority. In Islamic epistemology, religious knowledge is not sufficient merely by being "plausible" or "sounding correct"; it must possess a legitimate source, a verified chain of transmission, and an understanding consistent with the *manhaj* of the *Salaf al-Ṣāliḥ*.

The foundational principle of the *Salafī manhaj* in the reception of knowledge is *al-isnād min al-dīn*, the chain of transmission is part of the religion. ʿAbdullāh ibn al-Mubārak stated:

الْإِسْنَادُ مِنَ الدِّينِ، وَلَوْلَا الْإِسْنَادُ لَقَالَ مَنْ شَاءَ مَا شَاءَ

Al-isnādu min al-dīn, wa lawlā al-isnād laqāla man shā'a mā shā'a.
"The chain of transmission is part of the religion. Were it not for the chain of transmission, anyone would say whatever they pleased." (al-K. al-Baghdadi, n.d.)

Within this framework, religious reports-whether Qur'anic verses, hadiths, *fatāwā*, or scholarly opinions-must undergo the process of *tawthīq* (تَوْثِيق), namely the corroboration, verification, and authentication of both source and content. AI, as a non-human entity, possesses no *sanad*, no *ʿadālah* (عَدَالَة), no *ḍabṭ* (ضَبْط), and no *taqwā*. It merely generates probabilistic outputs. The epistemological question thus becomes profoundly serious: Can AI output be classified as "religious knowledge"? Can it serve as *ḥujjah*? Can it function as a *mufti*?

The central problem of this study is the phenomenon of AI hallucination in Islamic knowledge production. In technical terms, AI hallucination refers to outputs generated by LLMs that are fluent, coherent, and convincing, yet unsupported by facts, sources, or accurate data (Huang et al., 2025; Ji et al., 2023; Zhang et al., 2025) . Within the Islamic domain, these hallucinations may take the following forms. First, *sanad* hallucination, in which the AI fabricates chains of transmission-for example, "narrated by Imam al-Bukhārī from Imam Aḥmad from al-Shāfiʿī…" without any basis whatsoever. Second, *matan* hallucination, in which the AI fabricates hadith wordings that do not exist in any hadith collection and attributes them to the Prophet ﷺ. Third, *takhrīj* hallucination, in which the AI cites incorrect hadith numbers, book titles, or chapter references. Fourth, *fatwa* hallucination, in which the AI generates fictitious *fatāwā* as if they were the official opinions of a particular *iftāʾ* institution or scholar.

Fifth, *manhaj* hallucination, in which the AI inaccurately attributes opinions to Ibn Taymiyyah, Ibn al-Qayyim, or other scholars of the *Salaf*.

This phenomenon is not merely a technical error. If left unverified, it can become a source of *shubhah* (religious doubt and ambiguity) and *taḥrīf* (distortion of the purity of the faith). Allah ﷻ says:

يَا أَيُّهَا الَّذِينَ آمَنُوا إِن جَاءَكُمْ فَاسِقٌ بِنَبَإٍ فَتَبَيَّنُوا أَن تُصِيبُوا قَوْمًا بِجَهَالَةٍ فَتُصْبِحُوا عَلَىٰ مَا فَعَلْتُمْ نَادِمِينَ

Yā ayyuhā alladhīna āmanū in jā'akum fāsiqun bi-naba'in fa-tabayyanū an tuṣībū qawman bi-jahālah fa-tuṣbiḥū ʿalā mā fa'altum nādimīn.
"O you who believe, if a transgressor comes to you with a report, then ascertain its truth, lest you afflict a people out of ignorance and become regretful over what you have done." (Qur'an, al-Ḥujurāt [49]: 6)

This verse affirms the principles of *tabayyun* (scrutiny) and *tathabbuth* (verification). If a report from a transgressive human being must be verified, then how much more so a report from a machine that lacks moral consciousness, intention, and *sharʿī* accountability.

AI ethics scholarship in the West has generally focused on privacy, algorithmic bias, transparency, accountability, copyright, safety, and social justice (Cath, 2018; Jobin et al., 2019; Weidinger et al., 2022) . Studies on AI hallucination have likewise centered on technical dimensions: detection, mitigation, factuality evaluation, and retrieval augmentation (Ji et al., 2023; Tonmoy et al., 2024) . Meanwhile, Islamic studies on AI have often only addressed general aspects such as *maqāṣid* (higher objectives of Islamic law), moral ethics, *fatāwā* on technology use, or user ethics. Few studies have examined AI hallucination through the analytical tools of *ʿilm al-riwāyah* (the science of hadith transmission) and *ʿilm al-dirāyah* (the science of hadith criticism), particularly through the frameworks of *tawthīq* and *jarḥ wa taʿdīl* (narrator criticism and validation)

It is precisely this gap that the present study seeks to fill: bridging the probabilistic mechanics of LLMs with the criteria for accepting *khabar* in hadith sciences and *uṣūl al-fiqh*, and subsequently formulating a Digital *Tawthīq* framework as both an epistemological and technological solution. In alignment with this objective, this study aims: (1) to analyze the mechanisms underlying AI hallucination from the perspectives of computer science and Islamic epistemology; (2) to explain why AI is epistemologically incapable of serving as *ḥujjah* in religious knowledge production based on the criteria of *tawthīq* and *jarḥ wa taʿdīl*; and (3) to formulate a Digital *Tawthīq* framework for integrating hadith methodology into AI governance and the architecture of Islamic knowledge retrieval systems.

Based on the foregoing background, the research problem is articulated through three key questions: (1) What are the mechanisms underlying AI hallucination when examined comparatively from the perspective of statistical data processing in Large Language Models (LLMs) versus Islamic epistemology? (2) Why is AI epistemologically incapable of serving as an authoritative source (*ḥujjah*) in religious knowledge production based on the criteria of *tawthīq* and *jarḥ wa taʿdīl*? (3) How can the integration of *tawthīq* methodology into artificial intelligence governance (AI governance) be formulated to mitigate the risk of disseminating invalid (*bāṭil*) religious information?

Theoretically, this study enriches the dialogue among hadith sciences, *uṣūl al-fiqh*, and AI ethics. Practically, it provides guidance for Muslim AI developers, *ʿulamāʾ*, *iftāʾ* institutions, and regulators. Its primary contribution is the Digital *Tawthīq* framework—a model that positions source verification, digital *sanad*, and scholarly oversight as the foundational pillars of Islamic AI architecture.

Methods

This study employs a Critical Epistemological Review method grounded in a qualitative literature review design. Unlike quantitative research, this approach is not oriented toward testing empirical hypotheses (Snyder, 2019) . Rather, it focuses on deconstructing the epistemological assumptions underlying the reliability of Large Language Model (LLM) systems, which are then critically compared (*muqāranah*) with the principles of validation and knowledge acceptance in the Islamic tradition.

The data sources in this study are classified into two primary clusters. The first cluster encompasses contemporary literature from globally reputable journals and proceedings addressing Transformer architecture, LLM mechanics, the hallucination phenomenon, Retrieval-Augmented Generation (RAG) technology, and AI governance regulation. The second cluster comprises authoritative classical Islamic literature (*turāth* texts) in the fields of hadith sciences, *uṣūl al-fiqh*, and Islamic epistemology.

The data analysis process was conducted through four systematic stages (Elo & Kyngäs, 2008) , First, an in-depth content analysis of AI literature and *turāth* texts. Second, a critical comparison between the probabilistic computational mechanics of LLMs and the parameters of *tawthīq al-isnād wa al-maṭn*. Third, normative deduction derived from Qur'anic evidence, hadith, and the principles of *uṣūl al-fiqh*. Fourth, conceptual synthesis to formulate a novel architecture termed Digital *Tawthīq*.

To ensure the credibility and validity of the findings, this study applies source triangulation by cross-referencing data from peer-reviewed AI scholarly articles, *turāth* texts recognized by scholarly consensus, and epistemological *fatāwā* of the *ʿulamāʾ* (Flick, 2018) . Nevertheless, this study has limitations due to its purely conceptual-normative nature and the absence of empirical testing. Consequently, the effectiveness of the Digital *Tawthīq* framework produced in this research still requires further validation through system implementation studies and technical evaluation in the future.

Result and Discussion

The Anatomy of AI Hallucination: Why Do LLMs Fabricate Without Intent?

LLMs as Probability Machines

Modern Large Language Models (LLMs) based on the Transformer architecture operate by relying on a self-attention mechanism to process contextual relationships among tokens simultaneously (Vaswani et al., 2017) . Upon receiving an instruction (prompt), these systems fundamentally do not search for "truth" in a database of facts. Instead, they predict and compute

the probability of the next token through mathematical conditional probability calculations, formulated as follows:

$$P(w_t \mid w_1, w_2, \ldots, w_{t-1})$$

This formula is read as: "The probability of token w_t occurring given that the preceding sequence of tokens is w_1 through w_{t-1} (Ghaseminejad Raeini, 2025). In other words, the output is generated based on statistical patterns learned from the training data. The model possesses no independent representation of truth. It possesses no verifiable ontology of facts. All it possesses is a statistical distribution of billions of parameters learned from internet-scale text corpora. When a user asks, "What is the ruling on *zakāt* on gold?" or "What is the hadith on intention (*niyyah*)?", the model does not perform a search for truth in the epistemological sense. It merely predicts the most probable sequence of tokens based on the patterns it has learned. It does not know whether a hadith is authentic (*ṣaḥīḥ*) or fabricated (*mawḍūʿ*), unless the data patterns, fine-tuning, or retrieval system assist it. Even when the model produces a correct answer, its correctness is accidental rather than essential, not the product of verification, and not the fruit of semantic comprehension.

Bender et al. (2021) term this risk the *stochastic parrot*: a model can reproduce and string together language convincingly without understanding the meaning of what it utters (Bender et al., 2021). This parrot metaphor is profoundly apt from an epistemological standpoint. A parrot can utter the phrase "*Lā ilāha illā Allāh*" (لَا إِلَٰهَ إِلَّا ٱللَّهُ) fluently, yet it does not comprehend the meaning of *tawḥīd* (divine oneness), possesses no faith (*īmān*), and cannot serve as a reference for creed (*ʿaqīdah*). Likewise, an LLM can generate *fatāwā* that appear scholarly, yet it possesses no faith, no piety (*taqwā*, تَقْوَى), and cannot be held to *sharʿī* accountability. In the Islamic domain, this is profoundly dangerous because religious language carries immense authority. A single altered hadith wording (*lafẓ*) can alter legal rulings, creed, and acts of worship.

Typology of Hallucination and Its Manifestations in the Islamic Domain

Technically, AI hallucination is defined as output generated by a model that is fluent, coherent, and convincing, yet unsupported by facts, sources, or accurate data (Huang et al., 2025; Ji et al., 2023; Zhang et al., 2025) . Maynez et al. (2020) classify AI hallucination into four primary types: *intrinsic hallucination* (where the output contradicts the source it is supposed to reference); *extrinsic hallucination* (where the output introduces information entirely absent from the source); *factuality hallucination* (where the output contradicts objective facts); and *faithfulness hallucination* (where the output is unfaithful to the source document) (Maynez et al., 2020) .

Within the Islamic domain, this classification can be extended into a more specific and perilous typology. First, *sanad* hallucination: the AI fabricates chains of transmission that have never existed, for instance by stating "narrated by Imam al-Bukhārī from Imam Aḥmad from Imam al-Shāfiʿī from Mālik…" when no such chain of transmission exists in any authoritative (*muʿtabar*) *takhrīj* (hadith source tracing) compendium. This constitutes the most egregious form of falsification, as it strikes at the very foundation of hadith sciences, namely the *isnād* (chain of transmission).

Second, *matan* hallucination: the AI generates hadith wordings that do not appear in any hadith collection and attributes them to the Prophet Muḥammad ﷺ. For example, a model may produce the sentence "مَنْ صَلَّى الْفَجْرَ فِي جَمَاعَةٍ فَهُوَ فِي ذِمَّةِ اللَّهِ" (*man ṣallā al-fajra fī jamāʿatin fa-huwa fī dhimmat Allāh*, "Whoever prays the dawn prayer in congregation is under the protection of Allah") with inauthentic wording variations, or even fabricate an entirely nonexistent hadith.

Third, *takhrīj* hallucination: the AI cites incorrect hadith numbers, book titles, chapter headings, or authenticity grades (*ṣaḥīḥ*/*ḍaʿīf*). For instance, it may claim that a particular hadith is found in *Ṣaḥīḥ al-Bukhārī* number 1234, when in fact that number refers to an entirely different hadith, or to no hadith at all.

Fourth, *fatwa* hallucination: the AI generates fictitious *fatāwā* as if they were the official opinions of *al-Lajnah al-Dāʾimah*, *Majmaʿ al-Fiqh al-Islāmī*, or a specific scholar. This is the most dangerous form of *taḥrīf* (تَحْرِيف, distortion) in the realm of practical law, as it can mislead the Muslim community in matters of worship (*ʿibādāt*) and transactions (*muʿāmalāt*).

Fifth, *manhaj* hallucination: the AI inaccurately attributes opinions to the Companions (*Ṣaḥābah*), the Successors (*Tābiʿīn*), the Successors of the Successors (*Tābiʿī al-Tābiʿīn*), the eponymous imams of the legal schools (*aʾimmat al-madhāhib*), Ibn Taymiyyah, Ibn al-Qayyim, Muḥammad ibn ʿAbd al-Wahhāb, or other scholars of the *Salaf*, thereby creating the false impression that the *Salafī manhaj* endorses something that in fact contradicts it. This constitutes an indirect assault on the integrity of the *manhaj*.

Fabrication Without Intent: A Sharʿī Analysis of the Phenomenon of Algorithmic Falsehood

From a *sharʿī* perspective, *kadhib* (كَذِب, falsehood) is generally defined as reporting something that does not correspond to reality, whether intentionally or unintentionally. However, the sinful form of falsehood is that which is accompanied by *qaṣd* (قَصْد, deliberate intent) and *niyyah* (نِيَّة, inner intention). AI possesses neither *qaṣd* (intent) nor *niyyah*, nor does it possess moral consciousness. Therefore, technically speaking, AI cannot be described as "lying" or as a "liar" in the sense applicable to a *mukallaf* (مُكَلَّف, a morally and legally accountable agent). Nevertheless, its output may contain *kadhib* in an objective sense.

This does not, however, imply a total absence of responsibility. Rather, responsibility shifts to the parties involved in the chain of information production and distribution: developers who fail to implement *tawthīq* (تَوْثِيق, authentic verification), users who disseminate content without *tabayyun* (تَبَيُّن, scrutiny and verification), and platforms that monetize religious information without scholarly oversight. Those who disseminate such content without verification may share in its consequences. The Prophet ﷺ said:

مَنْ حَدَّثَ عَنِّي بِحَدِيثٍ يَرَى أَنَّهُ كَذِبٌ فَهُوَ أَحَدُ الْكَاذِبِينَ

Man ḥaddatha ʿannī bi-ḥadīthin yarā annahu kadhibun fa-huwa aḥad al-kādhibīn.
"Whoever narrates a hadith from me while believing it to be false is himself one of the liars." (Ṣaḥīḥ Muslim)

This hadith demonstrates that responsibility does not rest solely on the originator of falsehood but extends equally to the negligent disseminator. If a person who merely narrates a hadith he suspects to be false is already counted among the liars, then disseminating AI-

generated religious content without verification constitutes a profoundly serious form of negligence. Allah ﷻ says:

وَلَا تَقْفُ مَا لَيْسَ لَكَ بِهِ عِلْمٌ ۚ إِنَّ السَّمْعَ وَالْبَصَرَ وَالْفُؤَادَ كُلُّ أُولَٰئِكَ كَانَ عَنْهُ مَسْئُولًا

Wa lā taqfu mā laysa laka bihi 'ilm; inna al-sam'a wa al-baṣara wa al-fu'āda kullu ulā'ika kāna 'anhu mas'ūlā.
" And do not pursue that of which you have no knowledge. Indeed, the hearing, the sight, and the heart—about all of those will one be questioned." (Qur'an, al-Isrāʾ [17]: 36)

This verse affirms that *ʿilm* (عِلْم, knowledge) is a prerequisite before one speaks, issues a *fatwā*, or disseminates information. AI output does not constitute *ʿilm* in the Islamic epistemological sense, as it has not passed through the channel of *tawthīq*. Disseminating it without verification amounts to speaking without knowledge, and this is a direct violation of Qur'an, al-Isrāʾ 17:36. AI can become an instrument for the mass dissemination of *kadhib* if it is not safeguarded by *tawthīq*.

Why Does Hallucination Occur?: Technical Factors and Epistemological Implications

AI hallucination arises from the interaction of several technical factors. First, training data are incomplete, outdated, or contaminated with non-authoritative sources. Models are trained on internet-scale corpora that include discussion forums, blogs, articles lacking *sanad*, erroneous translations, and even fabricated texts. Second, models lack a real-time verification mechanism. They cannot distinguish between authentic (*ṣaḥīḥ*) and fabricated (*mawḍūʿ*) hadiths unless an external retrieval system is in place. Third, decoding sampling introduces random variation. Models operating at high temperature settings may produce more creative outputs but also less accurate ones. Fourth, models tend to fill knowledge gaps with the most linguistically probable patterns rather than acknowledging their own ignorance (Kadavath et al., 2022) . Fifth, Reinforcement Learning from Human Feedback (RLHF) can incentivize models to always provide an answer rather than admit uncertainty, because users tend to reward satisfying responses over answers that state "I do not know" (Ouyang et al., 2022) .

Lewis et al. (2020) introduced Retrieval-Augmented Generation (RAG) to mitigate hallucination by coupling generative models with document retrieval systems (Lewis et al., 2020) . However, RAG does not eliminate the problem if the retrieved sources are themselves unverified (Gao, 2023) . If the RAG database contains fabricated texts, *ḍaʿīf* hadiths without explanatory annotations, or non-authoritative (*ghayr muʿtabar*) *fatāwā*, then hallucination merely changes form, from "internal hallucination" to "citation hallucination," in which the model confidently cites incorrect sources.

Within the Islamic domain, this risk is compounded by several unique factors. First, the authority of religious language is exceptionally high. A single altered hadith wording (*lafẓ*) can alter legal rulings, creed, and acts of worship. Second, the Muslim community possesses a high degree of sensitivity toward Qur'anic and hadith quotations, yet not all members have the capacity to verify them independently. Third, AI can closely imitate the rhetorical style of classical scholars, thereby creating an illusion of authority. Fourth, the speed and ease of access incline users to accept outputs without verification. All of these factors render AI hallucination not merely a technical error but a serious threat to the integrity of the faith.

The Concept of *Tawthīq* in Salafī Epistemology: The Bastion of Religious Knowledge Authenticity

The Principle of al-Isnād min al-Dīn: The Foundation of Salafī Epistemology

The first principle of Salafī epistemology is *al-isnād min al-dīn*—the chain of transmission is part of the religion. Religious knowledge is not to be drawn from arbitrary sources. It must possess a clear chain of transmission, unbroken and traceable to its authentic origin. Al-Khaṭīb al-Baghdādī, in *al-Kifāyah fī ʿIlm al-Riwāyah*, affirms the critical importance of the *sanad* as the distinguishing criterion between the authentic and the fabricated: "If you remain silent and I remain silent, then when will those who do not know be able to distinguish a sound hadith from a defective one?" (A. B. A. bin A. bin T. al-K. al-Baghdadi, 1938) .

Allah ﷻ says:

فَاسْأَلُوا أَهْلَ الذِّكْرِ إِن كُنتُمْ لَا تَعْلَمُونَ

Fa-s'alū ahl al-dhikr in kuntum lā ta'lamūn.
"So ask the people of knowledge if you do not know." (Qur'an, al-Naḥl [16]: 43)

This verse demonstrates that the authority of knowledge resides with ahl al-dhikr (the people of knowledge) not with machines, algorithms, or entities devoid of consciousness. Ahl al-dhikr are human beings who possess a scholarly chain of transmission (sanad ʿilmī), piety (taqwā), and sharʿī accountability criteria that AI entirely lacks.

Narrator Criteria: ʿAdālah and Ḍabṭ as Conditions for the Acceptance of Khabar

In hadith sciences, a report is accepted as *ḥujjah* only if it fulfills specific conditions. Among these are *ittiṣāl al-sanad* (an unbroken chain of transmission extending continuously to the Messenger of Allah ﷺ), *ʿadālah* (عَدَالَة, moral integrity), and *ḍabṭ* (ضَبْط, precision of retention and recording).

ʿAdālah (عَدَالَة) etymologically denotes balance and justice. In the technical terminology of hadith sciences, it is defined as an inherent quality of the narrator that ensures consistency in piety (*taqwā*, تَقْوَى), avoidance of major sins, refraining from persistent commission of minor sins, preservation of *murūʾah* (مُرُوءَة, personal dignity and decorum), and abstention from acts that compromise integrity. Ibn Ḥajar al-ʿAsqalānī, in *Nuzhat al-Naẓar* (نُزْهَة النَّظَر), explains that *ʿadālah* is the primary condition a narrator must possess before his narration can be accepted (al-Asqalani, 2000) . *Ḍabṭ* (ضَبْط) etymologically denotes control, precision, and accuracy. In the technical terminology of hadith sciences, it is defined as the narrator's capacity to preserve narrations faithfully, whether through memorization (*ḍabṭ al-ṣadr*, ضَبْط الصَّدْر) or through written documentation (*ḍabṭ al-kitāb*, ضَبْط الْكِتَاب). A narrator who is *ḍābiṭ* (ضَابِط, possessing precision) is one who can transmit a hadith exactly as he received it, without addition, omission, or alteration.

The combination of *ʿadālah* and *ḍabṭ* yields the status of *thiqah* (ثِقَة, trustworthy). A narrator who is *ʿādil* (morally upright) but not *ḍābiṭ* cannot have his narration accepted as *ḥujjah*. A narrator who is *ḍābiṭ* but not *ʿādil* (for instance, a *fāsiq* or a known liar) likewise has his narration rejected. Even a narrator who is both *ʿādil* and *ḍābiṭ* may be rejected if his

narration contains *shudhūdh* (شُذُوذ, deviation from a stronger narration) or *ʿillah* (عِلَّة, a hidden defect).

AI cannot fulfill these conditions. It is not a *mukallaf* (morally accountable human being); it does not possess *ʿadālah* in the *sharʿī* sense; it does not possess *ḍabṭ* in the sense of conscious awareness and responsibility; and it possesses no *sanad*. It may well present text with high statistical accuracy, but technical accuracy is not synonymous with *ʿadālah* and *ḍabṭ*. The absence of these two qualities renders AI *majhūl al-ḥāl* (of unknown condition) in epistemological terms a status that, in hadith sciences, renders its reports unacceptable without external verification.

Tawthīq al-Matn: Safeguarding the Text from Shudhūdh and ʿIllah

Tawthīq (تَوْثِيق) does not stop at the *sanad*. It also encompasses the verification of the *matn* (مَتْن, the text/content of the hadith) against *shudhūdh* (شُذُوذ, deviation) and *ʿillah* (عِلَّة, hidden defect). AI lacks the capacity for *dirāyah* (deep analytical understanding) necessary to detect a hidden *ʿillah*. It can display a popular *matn*, but it is incapable of assessing whether that *matn* is *shādhdh* (anomalous) or *muʿallal* (defective). It does not possess the ability to compare narrations, comprehend historical context, or evaluate consistency with the broader principles of the *Sharīʿah*. It can only reproduce what appears most frequently in its training data—which is not necessarily what is most authentic.

Tawthīq in Uṣūl al-Fiqh and Fatwā

In *uṣūl al-fiqh*, a *khabar* must originate from a trustworthy source. A *fatwā* must not be taken from a person whose scholarship is unknown, and knowledge must be derived from revelation (*waḥy*) and the understanding of the *Salaf*, not from conjecture (*ẓann*) and personal desire (*hawā*).

A *fatwā* is a *sharʿī* position that requires *ijtihād* (اِجْتِهَاد, the maximum scholarly effort to derive legal rulings), comprehension of the scriptural texts (*nuṣūṣ*), knowledge of *uṣūlī* principles, mastery of the *maqāṣid* (مَقَاصِد, the higher objectives of the *Sharīʿah*), and contextual awareness. AI possesses none of these. It merely predicts the most probable text. Appointing AI as a *mufti* is tantamount to entrusting religion to a machine that possesses neither intellect (*ʿaql*), nor soul (*rūḥ*), nor accountability. Therefore, *tawthīq* is the bastion that safeguards religion from *taḥrīf* (distortion), *tabdīl* (alteration), and *kadhib* (falsehood). Without *tawthīq*, religious knowledge becomes contaminated with speculation, opinion, and fabrication, precisely as befell previous nations who abandoned the *sanad* and ultimately lost the purity of their revelation.

The Epistemological Clash: AI as a "Majhūl" Entity and Its Implications for Religious Authority

AI as Majhūl al-Ḥāl: A Critical Analysis of Epistemological Status

In hadith terminology, *majhūl* (unknown) refers to a narrator whose identity or condition is not established. AI, metaphorically speaking, is *majhūl al-ḥāl*: we may know the name of the model (GPT, Gemini, Claude), but we cannot ascertain with certainty the source of each claim it generates. The model possesses no *sanad*, no teacher, no *ijāzah* (scholarly license), and no moral accountability. It is not a narrator who can be subjected to *jarḥ* (criticism) or *taʿdīl*

(validation). It exists outside the category of *taklīf* (moral and legal obligation). Consequently, AI output cannot be accepted as *khabar maqbūl* (an admissible report) without external verification.

Furthermore, AI is an entity devoid of intellect (*ʿaql*) in any substantive sense, devoid of soul (*rūḥ*), devoid of a heart (*qalb*) capable of comprehension, and devoid of piety (*taqwā*). Allah ﷻ says:

أَفَلَا يَتَدَبَّرُونَ الْقُرْآنَ ۚ أَمْ عَلَىٰ قُلُوبٍ أَقْفَالُهَا

A-fa-lā yatadabbarūna al-Qur'āna am 'alā qulūbin aqfāluhā.
"Do they not then reflect upon the Qur'an, or are there locks upon their hearts?" (Qur'an, Muḥammad [47]: 24)

This verse demonstrates that true understanding (*tadabbur*, تَدَبُّر) involves the heart (*qalb*), not merely a process of linguistic computation. AI possesses no *qalb* that can comprehend, contemplate, and reflect upon revelation. It merely processes statistical patterns. Therefore, it cannot be categorized as *ahl al-ʿilm* (أَهْل الْعِلْم, the people of knowledge) in the *sharʿī* sense.

AI Does Not Fulfill the Conditions for Ḥujjah

Ḥujjah (حُجَّة) is a legally binding proof or evidence in the *sharʿī* sense, the abandonment of which without valid excuse constitutes a sin. In Islamic epistemology, religious *ḥujjah* can only derive from the following sources: the Qur'an, as revelation of absolute truth; the Authentic Sunnah (*al-Sunnah al-Ṣaḥīḥah*), proven to originate from the Prophet ﷺ through a sound *sanad*; Consensus (*Ijmāʿ*), established and recognized by the *ʿulamāʾ*; Analogical Reasoning (*Qiyās*), constructed upon the foundation of scriptural texts (*nuṣūṣ*), not speculation; and the *fatāwā* of the *Rabbāniyyūn* scholars, who possess a scholarly chain of transmission, piety, and competence.

AI cannot serve as *ḥujjah* because it possesses no unbroken *sanad* (*sanad muttaṣil*), no *ʿadālah* and *ḍabṭ*, no capacity for *ijtihād*, no piety and *sharʿī* accountability, and no capacity to be held answerable in the Hereafter (*ukhrawī* accountability).

Allah ﷻ says:

وَلَا تَقْفُ مَا لَيْسَ لَكَ بِهِ عِلْمٌ ۚ إِنَّ السَّمْعَ وَالْبَصَرَ وَالْفُؤَادَ كُلُّ أُولَٰئِكَ كَانَ عَنْهُ مَسْئُولًا

Wa lā taqfu mā laysa laka bihi ʻilm; inna al-samʻa wa al-baṣara wa al-fu'āda kullu ulā'ika kāna ʻanhu mas'ūlā.
"And do not pursue that of which you have no knowledge. Indeed, the hearing, the sight, and the heart—about all of those will one be questioned." (Qur'an, al-Isrāʾ [17]: 36)

Following AI output without verification amounts to following something without knowledge. This is the gateway to *ittibāʿ al-hawā*—following the dictates of personal desire rather than what is established by evidence. *Ittibāʿ al-hawā* is the root of all misguidance, for it blinds one to evidence and opens the door to *shubhah* (religious doubt and ambiguity).

AI Is Not a Mufti

A *fatwā* is a *sharʿī* position that requires knowledge, piety, and competence. A *mufti* must comprehend the scriptural texts (*nuṣūṣ*), the principles of *uṣūl al-fiqh*, the *maqāṣid*, the context,

and the spectrum of scholarly disagreement. AI possesses none of these. It merely predicts the most probable text. Appointing AI as a *mufti* is tantamount to entrusting religion to a machine that possesses neither intellect, nor soul, nor accountability. The Prophet ﷺ said:

مَنْ كَذَبَ عَلَيَّ مُتَعَمِّدًا فَلْيَتَبَوَّأْ مَقْعَدَهُ مِنَ النَّارِ

Man kadhaba ʻalayya mutaʻammidan fa-l-yatabawwaʼ maqʻadahu min al-nār.
"Whoever deliberately fabricates a lie against me, let him take his seat in the Hellfire." (Ṣaḥīḥ al-Bukhārī, no. 1291; Ṣaḥīḥ Muslim, no. 3)

If a human being who deliberately lies in the name of the Prophet is threatened with Hellfire, then a system that produces and disseminates falsehoods in the name of the Prophet without correction is far more dangerous.

Comparative Epistemological Table: LLMs vs. Salafī *Tawthīq*

Source	Statistical corpus	Revelation and verified *turāth*
Mechanism	Token probability	*Sanad* and *matn*
Authority	None	*ʿUlamāʾ* / *Muḥaddithūn*
Character	*Majhūl*, non-*mukallaf*	*ʿĀdil*, *ḍābiṭ*, *mukallaf*
Output	Plausible (possibly correct)	*Ṣaḥīḥ* / *Ḍaʿīf*
Defect	Hallucination	*Shudhūdh* and *ʿIllah*
Verification	Technical evaluation	*Jarḥ wa taʿdīl*
Accountability	None	In this world and the Hereafter
Intellect / Soul	None	Present
Piety (*Taqwā*)	None	Present

This table demonstrates that AI and the *tawthīq* method operate within two fundamentally distinct paradigms. AI operates in the realm of *iḥtimāl* (اِحْتِمَال, probability), whereas *tawthīq* operates in the realm of *taḥqīq* (تَحْقِيق, verification). AI produces conjecture, whereas *tawthīq* produces knowledge or, at the very least, a preponderant conjecture (*ghalabat al-ẓann*) supported by evidence.

Implications for Religious Authority and the Danger of Taḥrīf

If AI is positioned as a religious authority, the implications are profoundly serious. First, *taḥrīf* (تَحْرِيف, distortion) of the purity of the faith occurs, as fabricated hadiths, fictitious *sanads*, and imaginary *fatāwā* can spread on a massive scale without correction. Second, *tabdīl* (تَبْدِيل, substitution) of the methodology of knowledge occurs, as *sanad* and *tawthīq* are replaced by probability and algorithmic speed. Third, *shubhah* (شُبْهَة, doubt and ambiguity) arises among the Muslim community, as they become unable to distinguish between the authentic and the fabricated. Fourth, *ittibāʿ al-hawā* (اتِّبَاع الْهَوَى, following personal desire) proliferates, as people follow what the machine produces rather than what is established by evidence. Fifth, erroneous *taqlīd* (تَقْلِيد, uncritical imitation without evidence) takes hold, as people accept AI output without verification.

"Then who is more astray than one who follows his own desire without guidance from Allah?" (Qur'an, al-Qaṣaṣ [28]: 50). Accordingly, grounding religious rulings in AI without verification constitutes a form of *ittibāʿ al-hawā* and a violation of the prohibition against

speaking without knowledge. This is not merely a technical problem but a matter of creed (*ʿaqīdah*) and otherworldly (*ukhrawī*) accountability.

Formulating the "Digital *Tawthīq*" Framework: Integrating Hadith Methodology into Future AI Architecture

Definition of Digital Tawthīq

Digital *Tawthīq* is a verification and governance framework that treats every AI-generated Islamic claim as *khabar majhūl* (a report of unknown status) until it is traced to a verified source through a data trail that is transparent, auditable, and supervised by the *ʿulamāʾ*. Its purpose is not to replace the human *sanad* but to ensure that AI serves solely as *adāt al-baḥth* (أَدَاة الْبَحْث, a research aid), not as a *maṣdar* (مَصْدَر, an authoritative source).

The philosophy of Digital *Tawthīq* proceeds from the recognition that technology is not epistemologically neutral. Every AI architecture embeds assumptions about what counts as knowledge, how knowledge is verified, and who holds the authority to define truth. In the dominant Western AI paradigm, truth is frequently reduced to linguistic consistency, statistical likelihood, or user satisfaction. In the Islamic paradigm, truth is that which originates from revelation (*waḥy*) and is authenticated through a sound *sanad*. Digital *Tawthīq* seeks to bridge these two paradigms by positioning *tawthīq* as a first principle in AI architecture design.

Foundational Principles of Digital Tawthīq

Digital *Tawthīq* is built upon five foundational principles derived from the values of the *Sharīʿah*. First, *amānah* (أَمَانَة, trust and responsibility): developers and users bear the responsibility of safeguarding the faith. They must not permit AI to disseminate religious information without verification. Second, *masʾūliyyah* (مَسْؤُولِيَّة, accountability): every output must be traceable and answerable. This encompasses data audit trails, algorithmic transparency, and source documentation. Third, *iḥsān* (إِحْسَان, excellence) and *itqān* (إِتْقَان, precision): data quality and precision are maintained at the highest possible standard. Fourth, *ḥifẓ al-dīn* (حِفْظ الدِّين, preservation of the faith): the paramount priority is safeguarding the purity of the faith from *taḥrīf* and *tabdīl*. This constitutes an application of the *maqāṣid al-sharīʿah* (مَقَاصِد الشَّرِيعَة, the higher objectives of Islamic law) in the technological domain. Fifth, *lā ḍarar wa lā ḍirār* (لَا ضَرَر وَلَا ضِرَار, no harm shall be inflicted or reciprocated): preventing the harm caused by the dissemination of invalid (*bāṭil*) information that may mislead the Muslim community. This maxim is a fundamental principle in Islamic jurisprudence.

Technical Architecture of Digital Tawthīq

Digital *Tawthīq* can be implemented through ten mutually integrated technical components:

Closed Corpus: Only data drawn from the Qur'an, authentic (*ṣaḥīḥ*) and sound (*ḥasan*) hadiths, authoritative (*muʿtabar*) *turāth* texts, and *fatāwā* of official institutions are admitted.

RAG (Retrieval-Augmented Generation)*:* The AI generates responses based on verified documents rather than relying solely on internal parameters (Gao, 2023; Lewis et al., 2020) .

***Digital Sanad Metadata**:* Every data fragment is tagged with its author, book title, edition, page number, hadith number, *sanad*, authenticity grade, and verifier.

***Sanad* Knowledge Graph**: A traceable network of narrators, texts, and chains of transmission (Pan et al., 2024) .

Exact Quotation Mode: For Qur'anic verses and hadiths, the AI must not paraphrase divine revelation; it must display the original Arabic text, translation, and *takhrīj* (source reference).

***Uncertainty Calibration*:** The AI must explicitly express uncertainty and decline to answer when no source is found.

Self-RAG dan Factuality Checking: The AI performs internal reflection and self-verification (Asai et al., 2024) .

Human-in-the-Loop: The *ʿulamāʾ* serve as the final verifiers. No *fatwā* may be issued automatically.

***Audit Trail**:* A permanent data trail for both *sharʿī* and technical auditing.

***Manhaj Filter**:* A filtering mechanism to ensure that outputs conform to the understanding of the *Salaf al-Ṣāliḥ*.

Human-in-the-Loop as the Final Authority

AI may assist in locating Qur'anic verses, hadiths, or scholarly opinions. However, the issuance of *fatāwā* and the exercise of *tarjīḥ* (preponderance weighting among competing opinions) remain exclusively in the hands of *Rabbāniyyūn* scholars who possess an unbroken *sanad* (*sanad muttaṣil*) and a sound *manhaj*. AI must not be granted the authority of *iftāʾ*. Every religious response must pass through human verification. This principle is consistent with the command of Allah in Qur'an, al-Naḥl [16]: 43. *Ahl al-dhikr* are human beings, not machines. They are individuals who possess knowledge, *sanad*, piety, and accountability. AI cannot replace them.

Implementation Challenges and Solutions

The implementation of Digital *Tawthīq* faces several significant challenges. First, the accurate digitization of the *turāth* corpus requires considerable time, funding, and specialized expertise. Second, the variations in *qirāʾāt* (Qur'anic recitations) and *riwāyāt* (narration pathways) must be carefully accommodated. Third, differences in *manhaj* and *madhhab* (legal school) must be managed without engendering conflict. Fourth, the scalability and cost of developing such a complex system pose practical constraints. Fifth, data protection and user privacy must be ensured. Sixth, there is a risk of the commercialization of religion, which could reduce *fatāwā* to mere commodities.

Nevertheless, the principle is unequivocal: without *tawthīq*, there is no religious authority. Technology may advance, but the purity of the faith must be preserved. AI may assist, but it must not replace the *ʿulamāʾ*. Digital *Tawthīq* represents the middle path between technological progress and the preservation of tradition.

Conclusion

First, the mechanisms underlying AI hallucination can be explained from two complementary perspectives. From the perspective of computer science, LLMs operate on the basis of token occurrence probability (stochastic parrots), not on the basis of semantic comprehension (*fahm*, فَهْم) or the pursuit of truth (*ṭalab al-ḥaqq*, طَلَب الْحَقّ). The model generates output by predicting the next token based on the statistical distribution of its training data. It possesses no independent representation of truth, no factual verification mechanism, and no capacity to distinguish between the authentic and the fabricated unless an external retrieval system is in place. Hallucination occurs because the model fills knowledge gaps with the most linguistically probable patterns rather than acknowledging its own ignorance. From the perspective of Islamic epistemology, AI hallucination is a manifestation of the absence of *sanad* (سَنَد), the absence of *tawthīq* (تَوْثِيق), and the absence of *ʿadālah* (عَدَالَة) and *ḍabṭ* (ضَبْط). AI produces *ẓann* (ظَنّ, conjecture) wrapped in convincing language, and if left unverified, it can become a source of *shubhah* (شُبْهَة, doubt and ambiguity) and *taḥrīf* (تَحْرِيف, distortion) of the purity of the faith.

Second, AI is epistemologically incapable of serving as an authoritative source (*ḥujjah*, حُجَّة) in religious knowledge production. AI does not fulfill the conditions for the acceptance of *khabar* (خَبَر) in hadith sciences and *uṣūl al-fiqh*: it possesses no unbroken *sanad* (*sanad muttaṣil*, سَنَد مُتَّصِل), no *ʿadālah* (عَدَالَة), no *ḍabṭ* (ضَبْط), no capacity for *dirāyah* (دِرَايَة, critical analysis), no piety (*taqwā*, تَقْوَى), no accountable intellect (*ʿaql mukallaf*, عَقْل مُكَلَّف), and no *sharʿī* accountability. Its epistemological status is *majhūl* (مَجْهُول)—a source whose condition is not definitively known—such that its reports cannot be accepted as *ḥujjah* without verification. Grounding religious rulings in AI without verification constitutes a form of *ittibāʿ al-hawā* (اتِّبَاع الْهَوَى) and a violation of the prohibition against speaking without knowledge in Qur'an, al-Isrāʾ [17]: 36. AI is merely a research aid (*adāt al-baḥth*, أَدَاة الْبَحْث), not an authoritative source (*maṣdar*, مَصْدَر) or a *mufti* (مُفْتِي). It may be used to search, organize, and present data, but it must not serve as the final reference in matters of religion.

Third, the concept of integrating *tawthīq* methodology into AI governance can be formulated through the Digital *Tawthīq* framework (التَّوْثِيق الرَّقْمِي). This framework encompasses ten technical components: a verified closed corpus, Retrieval-Augmented Generation (RAG), digital *sanad* metadata, a *sanad* knowledge graph, exact quotation mode, uncertainty calibration, Self-RAG and factuality checking, human-in-the-loop (with the *ʿulamāʾ* as final verifiers), an audit trail, and a *manhaj* filter. Digital *Tawthīq* aims to ensure that AI serves solely as *adāt al-baḥth*, not as *maṣdar* or *mufti*. Every religious output must undergo verification by *Rabbāniyyūn* scholars who possess an unbroken scholarly *sanad* and a sound *manhaj*. In this way, technology can be harnessed for the advancement of knowledge without compromising the purity of the faith.

Accordingly, it can be conclusively affirmed that AI is merely a research aid (*adāt al-baḥth*, أَدَاة الْبَحْث), not an authoritative source (*maṣdar*, مَصْدَر) or a *mufti* (مُفْتِي). Technological progress must not come at the expense of the purity of the faith. The principle of *al-isnād min al-dīn* (الْإِسْنَاد مِنَ الدِّين) must remain the foundation for receiving and disseminating knowledge. Without *sanad*, without *tawthīq*, without *ʿadālah* and *ḍabṭ*, there is no legitimate religious

authority. Ultimately, the most appropriate scholarly and *sharʿī* stance in confronting the wave of artificial intelligence is to subordinate technology to the control of revelatory methodology, not the reverse.

Although this study offers a novel framework, it has methodological limitations due to its purely conceptual-normative nature, which precludes direct empirical computational testing (software benchmarking) of hallucination rates in modern LLMs, as well as limitations related to the dynamics of digitizing the *turāth* corpus, which remains not yet fully structured.

Therefore, future research is recommended to develop a quantitative Islamic Hallucination Benchmark, to test a prototype Digital *Tawthīq* architecture based on RAG and a *Sanad* Knowledge Graph, and to examine the formulation of cyber law pertaining to the liability of AI developers. Ultimately, technological progress must be met neither with passive rejection nor with blind acceptance; artificial intelligence must remain subordinate to the control of revelatory methodology and scholarly oversight in order to preserve the integrity of the *Sharīʿah* in the digital era.

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., & Anadkat, S. (2023). Gpt-4 technical report. *ArXiv Preprint ArXiv:2303.08774*.

al-Asqalani, A. bin A. bin M. bin H. (2000). *Nuzhat al-Nazar fi Tawdih Nukhbat al-Fikar fi Mustalah Ahl al-Athar* (N. al-D. Itr, Ed.; 3rd ed.). Matba'at al-Sabah. https://shamela.ws/book/1565/56

al-Baghdadi, al-K. (n.d.). Sharaf Ashab al-Hadith. In *Al-Bayan anna al-Asanid hiya al-Tariq ila Ma'rifat Ahkam al-Shari'ah*. al-Maktabah al-Shamilah. Retrieved https://shamela.ws/book/13125/96

al-Baghdadi, A. B. A. bin A. bin T. al-K. (1938). *Al-Kifayah fi Ilm al-Riwayah* (1st ed.). Dairat al-Maarif al-Osmania. https://shamela.ws/book/13055/118

Asai, A., Wu, Z., Wang, Y., Sil, A., & Hajishirzi, H. (2024). Self-rag: Learning to retrieve, generate, and critique through self-reflection. *International Conference on Learning Representations*, *2024*, 9112–9141.

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜. *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21*, 610–623. https://doi.org/10.1145/3442188.3445922

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., & Askell, A. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, *33*, 1877–1901.

Cath, C. (2018). Governing artificial intelligence: ethical, legal and technical opportunities and challenges. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *376*(2133), 20180080. https://doi.org/10.1098/rsta.2018.0080

Elo, S., & Kyngäs, H. (2008). The qualitative content analysis process. *Journal of Advanced Nursing*, *62*(1), 107–115. https://doi.org/https://doi.org/10.1111/j.1365-2648.2007.04569.x

Flick, U. (2018). Doing Triangulation and Mixed Methods. In *Qualitative Research Kit*. SAGE Publications Ltd. https://doi.org/10.4135/9781529716634

Gao, Y. (2023). Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv Preprint ArXiv:2312.10997*.

Ghaseminejad Raeini, M. (2025). The evolution of language models: From N-Grams to LLMs, and beyond. *Natural Language Processing Journal*, *12*, 100168. https://doi.org/https://doi.org/10.1016/j.nlp.2025.100168

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., & Qin, B. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, *43*(2), 1–55.

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.*, *55*(12). https://doi.org/10.1145/3571730

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., & Tran-Johnson, E. (2022). Language models (mostly) know what they know. *ArXiv Preprint ArXiv:2207.05221*.

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

Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On Faithfulness and Factuality in Abstractive Summarization. In D. Jurafsky, J. Chai, N. Schluter, & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 1906–1919). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.173

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., & Ray, A. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, *35*, 27730–27744.

Pan, S., Luo, L., Wang, Y., Chen, C., Wang, J., & Wu, X. (2024). Unifying large language models and knowledge graphs: A roadmap. *IEEE Transactions on Knowledge and Data Engineering*, *36*(7), 3580–3599.

Snyder, H. (2019). Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, *104*, 333–339. https://doi.org/https://doi.org/10.1016/j.jbusres.2019.07.039

Tonmoy, S. M., Zaman, S. M., Jain, V., Rani, A., Rawte, V., Chadha, A., & Das, A. (2024). A comprehensive survey of hallucination mitigation techniques in large language models. *ArXiv Preprint ArXiv:2401.01313*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, *30*.

Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., … Gabriel, I. (2022). Taxonomy of Risks posed by Language Models. *Proceedings of the 2022 ACM Conference on Fairness,*

Accountability, and Transparency, FAccT '22, 214–229. https://doi.org/10.1145/3531146.3533088

Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., Huang, X., Zhao, E., Zhang, Y., & Chen, Y. (2025). Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models. *Computational Linguistics*, *51*(4), 1373–1418.