

PENILAIAN EKSPERIMENTAL RALAT MAKLUMAT JANAAN KECERDASAN BUATAN BERASASKAN SET DATA KEFATWAAN

EXPERIMENTAL EVALUATION OF ARTIFICIAL INTELLIGENCE-GENERATED INFORMATION ERRORS BASED ON A FATWA DATASET

^{1*} Mohammaddin Abdul Niri, ² Noor Naemah Abdul Rahman, ² Mohd Anuar Ramli

¹ Makmal Falak, Jabatan Fiqh-Usul dan Sains Gunaan, Akademi Pengajian Islam, Universiti Malaya, Kuala Lumpur, Malaysia.

² Jabatan Fiqh-Usul dan Sains Gunaan, Akademi Pengajian Islam, Universiti Malaya, Kuala Lumpur, Malaysia.

*Correspondence:

Mohammaddin Abdul Niri,
e-mail:
mohammaddin@um.edu.my

History:

Received : 17.1.2026
Revised : 20.8.2026
Accepted : 21.8.2026
Published: 30.9.2026

Cite as:

Niri, M. A., Rahman, N. N. A., & Ramli, M. A. (2026). Penilaian eksperimental ralat maklumat janaan kecerdasan buatan berasaskan set data kefatwaan. *Jurnal Pengurusan dan Penyelidikan Fatwa*, 31(3), 50–72.
<https://doi.org/10.33102/jfatwa.vol.31no3.821>

Funding

This study was supported by the Universiti Malaya Research Management Fund (RMF0056-2021).

Abstract

The use of artificial intelligence (AI) technology is increasingly prevalent and has the potential to strengthen the dissemination of Islamic information. However, AI-generated information also carries a risk of error, particularly when used in authoritative domains such as fatwa. Therefore, this exploratory study evaluates errors in AI-generated information using a fatwa-related dataset through a controlled experimental design comparing three configurations of ChatGPT 5.2. The dataset was structured into two themes, Foundational Concepts of Fatwa and Challenges Faced by Fatwa Institutions. AI responses were evaluated across four information error dimensions; consistency, coherence, comprehensiveness, and hallucination, using a five-point Likert-scale rubric, with median scores employed as the measure of central tendency for ordinal data. Within the scope of this study, the retrieval-augmented generation configuration with closed retrieval demonstrated greater stability than the other two configurations and warrants further investigation as a potential setting for AI use in fatwa-related contexts. This is evidenced by aggregate median scores of 4.25 for Theme 1 and 4.00 for Theme 2, with an inter-theme aggregate score difference of 0.25. Overall, the study underscores the necessity of an integrated governance and scientific evaluation framework prior to the deployment of AI technologies as information-support tools within the fatwa ecosystem.

Keywords: *Fatwa; AI-Generated Information Errors; Retrieval-Augmented Generation (RAG); Knowledge-Based System; Islamic Information Systems.*



© The Author(s). This article is licensed under the Creative Commons Attribution 4.0 International License (CC BY 4.0). Reuse and adaptation are permitted, provided appropriate credit is given to the author(s) and source, a licence link is provided, and any changes made are indicated.

Abstrak

Penggunaan teknologi kecerdasan buatan (AI) semakin meluas dan berpotensi memperkukuh penyebaran maklumat Islam. Bagaimanapun, maklumat yang dijana oleh AI turut membawa risiko berlakunya ralat, khususnya apabila digunakan dalam bidang berautoriti seperti kefatwaan. Oleh itu, kajian eksploratori ini menilai ralat dalam maklumat janaan AI menggunakan set data berkaitan kefatwaan melalui reka bentuk eksperimen terkawal yang membandingkan tiga konfigurasi ChatGPT 5.2. Set data kajian distrukturkan kepada dua tema, iaitu Konsep Asas Fatwa dan Cabaran Institusi Fatwa. Respons AI dinilai berdasarkan empat dimensi ralat maklumat; iaitu konsistensi, koheren, keterangkuman dan halusinasi, dengan menggunakan rubrik berskala Likert lima mata dan skor median sebagai ukuran kecenderungan pusat bagi data berskala ordinal. Dalam skop kajian ini, konfigurasi penjanaan diperkaya berasaskan capaian maklumat tertutup menunjukkan kestabilan yang lebih baik berbanding dua konfigurasi lain dan berpotensi diteliti lebih lanjut sebagai satu tetapan dalam penggunaan AI untuk konteks kefatwaan. Hal ini berdasarkan skor median yang kekal pada tahap memuaskan, iaitu 4.25 bagi Tema 1 dan 4.00 bagi Tema 2, dengan perbezaan skor agregat tema sebanyak 0.25. Secara keseluruhan, dapatan kajian ini menggariskan keperluan kepada kerangka tadbir urus dan penilaian saintifik yang bersepadu sebelum teknologi AI boleh digunakan sebagai alat sokongan maklumat dalam ekosistem fatwa.

Kata kunci: *Fatwa; ralat maklumat janaan AI, penjanaan diperkaya berasaskan capaian maklumat (RAG); sistem berasaskan pengetahuan; sistem maklumat Islam.*

PENGENALAN

Perkembangan teknologi kecerdasan buatan (AI), khususnya model bahasa besar (LLM), kini semakin pesat. Penggunaannya dalam penjanaan maklumat semakin meluas dan merentas pelbagai bidang seperti pendidikan,ewartawanan, pentadbiran awam, perniagaan dan perkhidmatan teknologi maklumat (Bdoor & Habes, 2025; Aldoseri et al., 2023; Musleh Al-Sartawi, 2021), termasuk dalam bidang keagamaan (Shweiki et al., 2025). Keupayaan AI untuk menghasilkan respons yang pantas berpotensi membantu penyebaran maklumat kepada masyarakat awam. Namun, penjanaan maklumatnya juga disertai risiko ralat seperti kesalahan fakta, kelemahan hujah, dan penghasilan maklumat tanpa sandaran yang sah (Kollar & Alshibli, 2024; Vartumyan et al., 2023).

Penting diketahui bahawa maklumat janaan AI terhasil melalui proses janaan yang bersifat probabilistik, maka ia tidak terjamin bebas daripada kelemahan atau ralat. Hal demikian telah dikenal pasti dalam kajian-kajian lepas. Antaranya seperti kecenderungan AI menghasilkan maklumat yang tidak disokong sumber sahih, pengaburan antara fakta dan inferens serta ketidakupayaan membezakan antara maklumat pasti dan bersifat spekulatif. Selain itu, AI juga cenderung menghasilkan jawapan yang umum, simplistik atau tidak mengambil kira konteks khusus sesuatu persoalan (Kollar & Alshibli, 2024; Koutrika, 2025). Kajian-kajian lepas turut mendapati bahawa punca berlakunya ralat maklumat ini adalah daripada pelbagai faktor. Antaranya ialah kualiti dan bias data latihan yang membentuk model, selain keterbatasan konteks input (*prompt*), kekangan reka bentuk model untuk membezakan ketepatan fakta dan tiadanya mekanisme semakan rujukan yang ketat (Bruno et al., 2023; Fang et al., 2024; Santhoshkumar et al., 2025).

Situasi ralat maklumat janaan AI semakin penting untuk dikaji memandangkan respons AI pada masa kini mudah diterima tanpa semakan kritikal serta tersebar secara meluas dalam kalangan masyarakat awam (AbuJarour et al., 2024). Keadaan ini boleh memberi implikasi dalam penyampaian maklumat Islam, khususnya dalam konteks kefatwaan yang menuntut autoriti, ketelitian dan amanah. Dalam domain kefatwaan, pengaplikasian AI berisiko membawa ralat

sedemikian yang boleh mengelirukan kefahaman masyarakat serta berpotensi melemahkan persepsi awam terhadap keutuhan autoriti institusi fatwa sebagai rujukan utama hukum syarak. Oleh itu, adalah mustahak untuk menggariskan sempadan, batas dan cara penggunaan AI dalam konteks kefatwaan secara objektif dan saintifik (Rahim et al., 2025).

Berasaskan keperluan tersebut, kajian ini berangkat daripada jangkaan bahawa keadaan ralat maklumat janaan AI dalam domain kefatwaan mungkin berbeza mengikut aras tema kefatwaan dan konfigurasi sistem AI. Tema yang bersifat deskriptif pada tahap mudah, seperti Konsep Asas Fatwa, menyediakan maklumat yang lebih literal, sekali gus membolehkan sistem AI menjana respons yang lebih stabil. Sebaliknya, tema yang bersifat analitikal pada tahap rumit, seperti Cabaran Institusi Fatwa, menuntut penaakulan yang lebih kontekstual, mendalam serta melangkaui maklumat literal, sekali gus mendedahkan sistem AI kepada peningkatan risiko ralat maklumat dan ketidakstabilan respons.

Secara operasional dalam kajian ini, ralat maklumat janaan AI ditakrifkan sebagai penyimpangan kandungan respons AI daripada maklumat rujukan yang dibenarkan atau keperluan konteks soalan (Sun et al., 2024). Bagi memastikan penilaian dilakukan secara sistematik, ralat maklumat dianalisis berdasarkan empat dimensi utama, iaitu konsistensi (*consistency*), koheren (*coherence*), keterangkuman (*comprehensiveness*) dan halusinasi (*hallucination*). Dimensi konsistensi, koheren dan keterangkuman dimodifikasi daripada kajian Chiu & Chung (2025), manakala dimensi halusinasi disesuaikan berdasarkan keperluan kajian dan literatur berkaitan halusinasi AI (Sun et al., 2024). Konsistensi merujuk kepada keselarasan respons AI dengan fakta yang boleh disahkan serta tiada percanggahan. Koheren pula merangkumi keutuhan aliran hujah atau logik dan kesahihan inferens yang dikemukakan dalam respons AI. Manakala, keterangkuman merujuk kepada kecukupan liputan kandungan respons AI bagi menjawab sesuatu persoalan. Adapun halusinasi merujuk kepada penghasilan maklumat atau butiran baharu yang kelihatan meyakinkan tetapi tidak mempunyai sandaran sah yang boleh ditelusuri sumber asal rujukannya (Chiu & Chung, 2025; Sun et al., 2024).

Sehubungan itu, kajian eksploratori ini dirangka untuk menilai secara awal corak ralat maklumat janaan AI dalam domain kefatwaan. Skop kajian dihadkan kepada satu set data, 20 soalan input dan dua tema kefatwaan bagi membolehkan reka bentuk eksperimen, konfigurasi sistem dan instrumen penilaian diuji dalam persekitaran yang terkawal. Oleh itu, kajian ini tidak bertujuan membuat generalisasi menyeluruh terhadap prestasi AI dalam semua bentuk pertanyaan fatwa, tetapi menyediakan garis asas empirikal bagi penyelidikan yang lebih luas dan mendalam.

KAJIAN LITERATUR

Dalam tema yang umum, kajian-kajian mutakhir berkaitan kecerdasan buatan (AI) dalam domain maklumat Islam menunjukkan perkembangan yang semakin ketara dan pelbagai. Kajian-kajian terbabit dapat diteliti melalui dua kecenderungan utama. Pertama, kajian yang menumpukan kepada aspek etika penggunaan AI dalam konteks Islam, termasuk perbincangan tentang prinsip, batasan dan implikasi penggunaan teknologi ini dalam penyampaian maklumat keagamaan, seperti yang dibincangkan oleh Kannike & Fahm (2025), Meerangani et al. (2025), Rahim et al. (2025), Bdoor & Habes (2025), Al Kubaisi (2024) dan AbuJarour et al. (2024). Kedua, kajian yang memfokuskan pembangunan aplikasi atau chatbot AI bagi tujuan pengurusan dan penyampaian maklumat Islam, antaranya kajian oleh Maylawati et al. (2025), Alnefaie et al. (2025), Shweiki et al. (2025), Nurfikri et al. (2024), Nurhayati et al. (2024) dan Munshi et al. (2021).

Walau bagaimanapun, kebanyakan kajian dalam kedua-dua kecenderungan ini lebih menumpukan aspek etika dan fungsi aplikasi, manakala penilaian terhadap ralat maklumat janaan AI, khususnya dalam konteks berautoriti seperti domain fatwa, masih kurang diberi perhatian secara empirikal. Oleh itu, dalam tema yang lebih spesifik, sorotan literatur kajian ini memfokuskan kajian-kajian terdahulu yang relevan dengan penilaian ralat maklumat janaan AI. Kajian-kajian terbabit dibahagikan kepada dua tema utama, iaitu:

Berasaskan penilaian teknikal

Ralat maklumat janaan kecerdasan buatan (AI) lazimnya ditentukan melalui pelbagai metrik dan instrumen penilaian, termasuk metrik teknikal, seperti ditunjukkan dalam kajian Kulkarni et al. (2025), Özdemir et al. (2025), Shiferaw (2024) dan Raffinetti (2023). Lebih mendasar, kajian Rashed & Shirmohammadi (2022) membahaskan *Root Mean Squared Error* (RMSE) sebagai ukuran agregat yang lazim digunakan untuk mengukur ralat model AI, dengan memberikan penalti yang lebih besar kepada kesalahan yang besar. Berbeza dengan RMSE, kajian Raffinetti (2023) pula mengetengahkan *Rank Graduation Accuracy* (RGA) yang biasanya untuk ramalan numerik dan penarafan. Dalam erti kata lain, ia menilai sama ada AI meletakkan perkara dalam urutan yang betul ataupun tidak. Bagaimanapun, kedua-duanya merupakan metrik teknikal yang bukan direka untuk menilai ketepatan maklumat dari sudut semantik, konteks atau kesahihan.

Adapun untuk kaedah menilai halusinasi maklumat janaan AI antaranya ialah *Entropy-Based Metric* yang diketengahkan dalam kajian Farquhar et al. (2024). Ia adalah ukuran ketidakpastian yang mengelompokkan respons berdasarkan kesetaraan makna sebelum menghitung ketidakpastian merentas kelompok semantik. Selanjutnya, kajian Heo et al. (2025) pula memperkenalkan HaluCheck yang merupakan satu alat pemeriksa untuk membantu pengguna menilai sama ada jawapan AI berkemungkinan mengandungi maklumat palsu atau direka-reka.

Kemudian, kajian Jesson et al. (2024) mengetengahkan kaedah penilaian kebarangkalian log bagi mengenal pasti kemungkinan halusinasi dalam respons AI. Nilai kebarangkalian log yang rendah boleh menjadi petunjuk kepada risiko ralat atau halusinasi yang lebih tinggi. Sebaliknya, nilai yang lebih tinggi menunjukkan bahawa respons lebih sejajar dengan data yang digunakan, namun tidak semestinya menjamin ketepatan fakta.

Berasaskan penilaian konteks dan bidang

Penilaian terhadap ralat maklumat janaan AI turut dibangunkan dalam kajian-kajian terdahulu mengikut konteks dan keperluan bidang masing-masing, tanpa terikat sepenuhnya kepada kaedah penilaian teknikal yang bersifat umum. Dalam domain sains kesihatan, misalnya, kajian Asgari et al. (2025) membangunkan satu instrumen penilaian yang menumpukan kepada aspek keselamatan klinikal dan kadar halusinasi dengan mengklasifikasikan kesilapan AI berdasarkan jenis ralat perubatan serta potensi kemudahan kepada pesakit.

Dalam kajian Asgari et al. (2025), 18 konfigurasi eksperimen bagi penjanaan nota klinikal dinilai berdasarkan 12,999 ayat yang dianotasi oleh pengamal klinikal. Kajian tersebut melaporkan kadar halusinasi sebanyak 1.47% dan kadar pengguguran maklumat sebanyak 3.45%. Melalui penambahbaikan prom dan aliran kerja, kadar ralat besar berjaya dikurangkan

hingga lebih rendah daripada kadar kesalahan pencatatan manusia yang pernah dilaporkan. Di samping itu, terdapat juga instrumen yang lain turut digunakan dalam bidang kesihatan seperti *Global Quality Score (GQS)*, *PEMAT*, *DISCERN*, *CLEAR* serta indeks kebolehbacaan seperti *Flesch Reading Ease* dan *Flesch–Kincaid Grade Level*. Walau bagaimanapun, instrumen-instrumen ini tidak dibangunkan secara khusus untuk mengenal pasti ralat maklumat dalam domain kefatwaan yang menjadi fokus kajian ini.

Dalam bidang undang-undang, kajian Dahl et al. (2024) meneliti fenomena *legal hallucinations* dengan menilai ketepatan fakta undang-undang, kesahihan rujukan kes dan kesesuaian bidang kuasa dalam respons AI. Melalui audit kandungan yang membandingkan respons AI dengan autoriti perundangan sebenar, kajian tersebut mendapati bahawa model bahasa kegunaan umum menghasilkan halusinasi antara 58% hingga 88% dalam kes yang diuji. Model tersebut juga lemah dalam mengenal pasti halusinasi pada respons sendiri dan sering menerima andaian undang-undang pengguna yang salah tanpa penilaian kritikal. Dapatan ini menunjukkan bahawa penilaian AI dalam domain normatif memerlukan piawaian berasaskan disiplin serta pengawasan manusia dalam penggunaannya.

Dalam penyelidikan akademik, terdapat kajian yang menilai ketepatan dan halusinasi rujukan (*citation hallucination*) seperti ditunjukkan oleh kajian Linardon et al. (2025). Kajian terbabit tidak menilai prestasi model dari sudut teknikal, sebaliknya memeriksa sama ada rujukan yang dijana oleh AI benar-benar wujud, tepat dari segi butiran bibliografi dan relevan dengan kandungan yang dirujuk. Kaedah ini berasaskan prinsip integriti akademik dan amat sesuai untuk domain berautoriti tinggi kerana ia menilai kebolehpercayaan maklumat secara langsung. Dalam kajian Linardon et al. (2025), 176 sitasi yang dijana oleh GPT-4o dalam enam ulasan literatur berkaitan kesihatan mental dinilai dan didapati bahawa 35 sitasi (19.9%) merupakan rujukan rekaan, manakala 64 daripada 141 sitasi yang benar-benar wujud (45.4%) mengandungi kesalahan sitasi, khususnya pengenalan objek digital (DOI) yang salah atau tidak sah. Risiko fabrikasi turut didapati lebih tinggi bagi topik yang kurang dikenali dan pertanyaan yang lebih khusus.

Akhir sekali, kajian FACTS Team (2024) dan Jacovi et al. (2025) menunjukkan bahawa model bahasa boleh menghasilkan respons berasaskan dokumen dengan tahap faktualiti yang baik. Kajian itu turut mengesan kecenderungan model penilai memberikan skor lebih tinggi kepada output daripada keluarga modelnya sendiri. Dapatan ini menunjukkan bahawa pengikatan respons kepada sumber dapat meningkatkan faktualiti, namun semakin manusia masih diperlukan, khususnya dalam domain berautoriti. Walau bagaimanapun, kedua-dua pendekatan terbabit masih berfokus kepada metrik automatik dan belum dirangka khusus untuk menilai ralat maklumat dalam konteks normatif dan berautoriti seperti domain kefatwaan.

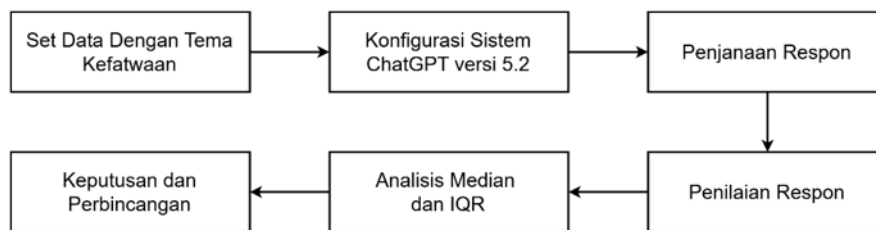
Secara keseluruhan, sorotan literatur dalam tema yang spesifik turut menunjukkan dua kecenderungan utama dalam penilaian ralat maklumat janaan AI, iaitu pendekatan metrik teknikal yang sesuai untuk pengukuran kuantitatif tetapi terhad dalam menilai kesahihan semantik dan konteks, serta pendekatan semakin berasaskan bidang yang lebih sesuai untuk domain normatif kerana menilai ketepatan berdasarkan piawaian disiplin. Walau bagaimanapun, penilaian ralat maklumat AI dalam konteks kefatwaan masih kurang dikaji secara empirikal menerusi eksperimen terkawal, khususnya dari sudut corak ralat merentas tema perbincangan dan perbandingan konfigurasi sistem yang berbeza dalam platform yang sama. Untuk itu, jurang terbabit ditangani dalam kajian ini dengan menganalisis tahap ralat maklumat dan kestabilan respons sistem AI merentas dua tema kefatwaan yang berbeza dari

segi aras kerumitan bagi menilai sejauh mana sifat tema mempengaruhi prestasi sistem secara relatif serta menghasilkan dapatan yang empirikal dan boleh dikaji semula.

METODOLOGI

Kajian ini menggunakan reka bentuk eksperimen terkawal (*controlled experiment*) dalam kerangka kajian eksploratori bagi menilai ralat maklumat janaan kecerdasan buatan (AI). Reka bentuk ini selari dengan pendekatan yang digunakan dalam kajian terdahulu (Nascimento et al., 2025 & Gopali et al., 2024). Pendekatan ini membolehkan perbandingan dibuat secara objektif dengan menetapkan beberapa faktor sebagai pemboleh ubah dimalarkan (*controlled variable*), iaitu platform, format soalan input dan domain kefatwaan. Tema penelitian iaitu Konsep Asas Fatwa dan Cabaran Institusi Fatwa ditetapkan sebagai faktor keadaan untuk menguji corak ralat merentas aras tema. Platform kajian menggunakan model ChatGPT versi 5.2 yang sama, manakala perbezaan tetapan konfigurasi sistem AI ditetapkan sebagai pemboleh ubah dimanipulasi (*independent variable*). Penyeragaman input dilaksanakan dengan menggunakan 20 prom (*prompt*) berbentuk soalan yang sama, tanpa sebarang perubahan pada struktur ayat, bahasa dan panjang teks bagi setiap konfigurasi sistem.

Rajah 1: Aliran Kerja Penilaian Ralat Maklumat Janaan AI



Seperti yang ditunjukkan dalam Rajah 1, kajian ini bermula dengan penyediaan set data kefatwaan yang distrukturkan mengikut tema iaitu tema maklumat berkaitan Konsep Asas Fatwa dan tema Cabaran Institusi Fatwa, tanpa melibatkan perbincangan hukum fiqh atau keputusan fatwa tertentu. Kemudian, diikuti oleh penetapan konfigurasi sistem AI yang berbeza sebelum proses penjanaan respons dilaksanakan.

Dari sudut jenis soalan, semua soalan direka dalam bentuk deskriptif dan analitik yang seragam bertujuan untuk pengujian merentas konfigurasi sistem AI yang berbeza. Pendekatan ini bagi memastikan sebarang perbezaan dalam respons AI yang dihasilkan dapat diteliti kaitannya secara langsung dengan perbezaan konfigurasi sistem, sama ada dari sudut seni bina, mekanisme pangkalan pengetahuan atau tahap pergantungan kepada sumber luaran dan mengurangkan pengaruh perbezaan format soalan input (Nascimento et al., 2025 & Gopali et al., 2024).

Tiga konfigurasi sistem ChatGPT versi 5.2 digunakan terdiri daripada, (i) Sistem A (mekanisme RAG + web) iaitu jawapan dijana berasaskan set data kefatwaan dan akses carian web diaktifkan; (ii) Sistem B (mekanisme RAG tertutup) iaitu jawapan dijana hanya berasaskan set data kajian tanpa carian web; (iii) Sistem C (web sahaja) iaitu jawapan dijana tanpa menggunakan set data kefatwaan dan bergantung kepada carian web sebagai sumber luaran. Perbandingan antara ketiga-tiga konfigurasi ini membolehkan penilaian dibuat terhadap kesan mekanisme pengambilan maklumat dan perbezaan aras tema terhadap ralat maklumat janaan AI.

Penilaian respons AI dilaksanakan oleh penilai pakar terdiri daripada dua orang penilai, iaitu seorang pakar kandungan yang juga merupakan penyedia set data kefatwaan, serta seorang pakar luar yang berpengalaman dalam aspek teknikal dan reka bentuk eksperimen terkawal. Sebelum fasa penilaian, satu sesi taklimat khas dijalankan bagi menyelaraskan kefahaman penilai terhadap definisi operasi dan kriteria pemarkahan bagi setiap dimensi ralat.

Seterusnya, penilaian dilaksanakan secara dibutakan pada label konfigurasi sistem (*blinded to system identity*), iaitu penilai tidak mengetahui respons berasal daripada Sistem A/B/C. Langkah nyah-identiti respons dan pemarkahan tertutup dilaksanakan bagi meminimumkan risiko bias. Penilaian dilaksanakan menggunakan rubrik berskala Likert lima mata. Bagi setiap dimensi ralat, skor yang diberikan oleh kedua-dua pakar digabungkan dengan menentukan nilai median bagi pasangan skor tersebut. Oleh sebab setiap item dinilai oleh dua orang pakar, nilai median yang terhasil adalah bersamaan secara berangka dengan purata kedua-dua skor. Penggunaan median kemudiannya dikekalkan secara konsisten sebagai ukuran kecenderungan pusat dalam analisis data berskala ordinal, manakala julat antara kuartil (IQR) digunakan untuk menilai variasi dan kestabilan skor merentas item penilaian.

Untuk set data kefatwaan yang dijadikan sumber pangkalan pengetahuan (*knowledge base*) dalam kajian ini, dokumen rujukan yang digunakan ialah artikel “*Cabaran Institusi Fatwa dalam Era Revolusi Industri 4.0*” oleh Rahman & Niri (2024). Dokumen dipilih secara persampelan bertujuan kerana kandungannya tertumpu kepada dua tema iaitu Konsep Asas Fatwa dan Cabaran Institusi Fatwa dalam konteks semasa. Dokumen ini merangkumi perbincangan konseptual mengenai fatwa dan peranan mufti, dinamika perubahan fatwa, serta mengenal pasti empat cabaran teras institusi fatwa dalam era Revolusi Perindustrian Keempat (IR4.0), iaitu tadbir urus, autoriti maklumat fatwa, tenaga kerja dan inovasi sistem penyampaian. Keseluruhan soalan input dibina berdasarkan skop dan perbincangan utama dokumen Rahman & Niri (2024). Soalan disusun untuk menguji respons AI tanpa melibatkan sebarang keputusan fatwa atau perbahasan hukum fiqh. Ini bagi memastikan set soalan berfungsi sebagai instrumen penilaian yang sah, terkawal dan konsisten mengikut sumber rujukan, sekali gus bagi membolehkan respons AI dianalisis secara objektif.

Soalan berbentuk hukum substantif dikecualikan kerana keputusan metodologi yang disengajakan dalam kajian eksploratori ini serta selaras dengan tujuan yang digariskan. Kajian ini bukan bertujuan menilai keupayaan AI untuk mengistinbat hukum, mengeluarkan fatwa atau menggantikan autoriti mufti, tetapi hanya menilai ketepatan dan kestabilan sistem dalam menyampaikan maklumat berkaitan kefatwaan. Tambahan, soalan hukum substantif tidak dimasukkan kerana lazimnya melibatkan pertimbangan dalil, perbezaan pandangan fiqh, konteks kefatwaan dan realiti semasa yang boleh menambahkan lagi pemboleh ubah tambahan sekali gus mengurangkan kawalan eksperimen. Adapun dua tema yang ditetapkan, iaitu Konsep Asas Fatwa dan Cabaran Institusi Fatwa, dipilih kerana kedua-duanya bersumberkan dokumen rujukan yang sama tetapi mewakili konteks maklumat yang berbeza. Tema pertama lebih bersifat deskriptif dan literal, manakala tema kedua memerlukan penghuraian yang lebih analitikal dan kontekstual. Pemilihan ini membolehkan perbandingan dibuat merentas tema dalam persekitaran eksperimen yang masih terkawal.

Jadual 1: Soalan Input Daripada Set Data Kefatwaan Berasaskan Dokumen Penulisan Rahman & Niri (2024)

Pembahagian Soalan Input Mengikut Tema Fatwa

Tema 1: Konsep Asas Fatwa (Soalan 1-10)

1. Apakah maksud fatwa dari sudut bahasa dan istilah syarak?
2. Apakah maksud fatwa menurut pandangan ulama?
3. Dalam konteks Malaysia, apakah maksud fatwa dan siapakah mufti?
4. Apakah tanggungjawab mufti?
5. Kenapa al-Shatibi menyatakan tugas mufti adalah berat dan berisiko tinggi?
6. Adakah fatwa itu statik atau boleh berubah?
7. Apakah faktor yang menyebabkan fatwa menjadi statik dan tidak berubah?
8. Apakah faktor yang membawa kepada perubahan fatwa dalam Islam?
9. Apakah pandangan ulama mengenai perubahan fatwa?
10. Apakah kesan terhadap fatwa jika realiti masyarakat tidak difahami?

Tema 2: Cabaran Institusi Fatwa (Soalan 11-20)

11. Apakah contoh cabaran semasa yang dihadapi institusi fatwa?
 12. Apakah perbezaan antara cabaran dalaman dan cabaran luaran yang dihadapi institusi fatwa?
 13. Apakah cabaran lokal dan global yang boleh memberi kesan kepada institusi fatwa di Malaysia?
 14. Apakah cabaran tadbir urus yang boleh menjejaskan keutuhan institusi fatwa?
 15. Apakah cabaran tenaga kerja dan kepakaran dalam institusi fatwa masa kini?
 16. Apakah bentuk cabaran yang dihadapi oleh institusi fatwa dalam era Revolusi Industri 4.0?
 17. Apakah solusi yang boleh diambil institusi fatwa untuk mengukuhkan tadbir urus dalam persekitaran digital?
 18. Apakah solusi yang boleh diambil institusi fatwa untuk menangani maklumat palsu dan manipulatif?
 19. Apakah solusi yang boleh diambil institusi fatwa untuk meningkatkan kualiti tenaga kerjanya dalam era digital?
 20. Apakah solusi yang boleh diambil institusi fatwa untuk meningkatkan keberkesanan penyampaian fatwa?
-

Untuk analisis data, kajian ini mengaplikasikan kaedah analisis kandungan berstruktur yang digabungkan dengan penilaian skor pemarkahan kuantitatif menggunakan skala Likert 1 hingga 5 (Chiu & Chung, 2025). Respons AI dikelompokkan mengikut tema perbincangan seperti dipaparkan dalam Jadual 1 di atas terdiri daripada soalan input 1 hingga 10 dikategorikan dalam tema Konsep Asas Fatwa, manakala soalan input 11 hingga 20 merangkumi tema Cabaran Institusi Fatwa. Analisis kandungan digunakan untuk menilai setiap respons berdasarkan definisi operasi dan rubrik yang ditetapkan. Hasil analisis kualitatif ini seterusnya diberikan pemarkahan kuantitatif, membolehkan perbandingan dibuat antara konfigurasi sistem dan tema perbincangan.

Skor pemarkahan kuantitatif terdiri daripada empat dimensi ralat maklumat, iaitu konsistensi, keterangkuman, koheren dan halusinasi. Markah kuantitatif terbabit dianalisis menggunakan Microsoft Excel dan visualisasi data dijalankan menggunakan bahasa pengaturcaraan Python dengan sokongan pustaka analisis dan visualisasi yang sesuai. Keseluruhan prosedur eksperimen ini direka bentuk supaya boleh diulang dan disemak semula oleh penyelidik empirikal lain dalam domain yang sama (Marey et al., 2025; Shiferaw, 2024).

Secara spesifik, dimensi konsistensi menilai keselarasan respons AI dengan fakta yang boleh disahkan serta tiada percanggahan dalaman respons AI, di mana ralat dikenal pasti apabila wujud ketidaktepatan atau kesalahan fakta. Dimensi koheren pula menilai keutuhan hujah dan kesahihan inferens, di mana ralat berlaku apabila terdapat lonjakan logik atau generalisasi melampau yang tidak disokong oleh premis. Dimensi keterangkuman menilai kecukupan liputan kandungan respons AI dalam menjawab soalan, di mana ralat berlaku akibat pengabaian kandungan penting atau kekurangan konteks dalam respons AI yang diperlukan.

Sementara itu, dimensi halusinasi menilai kehadiran maklumat yang direka dalam respons AI dan tidak disokong oleh bukti sah, walaupun disampaikan secara meyakinkan. Gabungan keempat-empat dimensi ini membolehkan ralat maklumat dikenal pasti dan diteliti secara saintifik dan boleh dikaji semula dalam penilaian respons AI. Berikut adalah rubrik penilaian ralat maklumat janaan AI yang dimodifikasi daripada kerangka penilaian Chiu & Chung (2025) bagi disesuaikan dengan konteks set data kefatwaan.

Jadual 2: Rubrik Penilaian Ralat Maklumat Janaan Kecerdasan Buatan (AI)

Dimensi	Konsistensi	Koheren	Keterangkuman	Halusinasi
Definisi Operasi	Tahap keselarasan respons AI dengan fakta boleh disahkan	Tahap ketekalan atau keutuhan aliran hujah atau logik atau proses inferens dalam respons AI	Tahap keluasan kandungan atau makna dalam respons AI menjawab soalan	Tahap kewujudan maklumat yang direka dalam respons AI dan tidak disokong bukti sah
Skor 1: Amat Tidak Memuaskan	Fakta jelas bercanggah atau tidak sah	Hujah bercelaru dan tidak tekal	Kandungan sangat sedikit dan terhad	Maklumat rekaan serius dan dominan
Skor 2: Tidak Memuaskan	Fakta lemah atau tidak mencukupi	Hujah lemah dan konteksnya tidak tersusun	Keluasan kandungan masih kecil atau sedikit	Banyak maklumat rekaan tidak disokong bukti sah
Skor 3: Kurang Memuaskan	Fakta sebahagiannya terserlah konsisten	Hujah boleh difahami tetapi masih kurang kemas	Keluasan kandungan sederhana dengan masih ketara terdapat aspek yang tertinggal	Sebahagian maklumat rekaan yang tidak disokong bukti sah
Skor 4: Memuaskan	Fakta hampir sepenuhnya tepat tetapi masih ada kesalahan kecil	Hujah teratur dan logik	Keluasan kandungan hampir menyeluruh	Halusinasi kecil dan terpercil

Skor 5: Amat Memuaskan	Fakta sepenuhnya konsisten dan boleh disahkan	Hujah koheren dan tersusun	keluasan dan kandungan adalah lengkap dan seimbang	Tiada halusinasi dan semua maklumat adalah berasaskan bukti sah
---------------------------------------	---	----------------------------	--	---

Penilaian terhadap dimensi kualiti respons AI seperti yang dipaparkan dalam Jadual 2 lazimnya turut digunakan dalam kajian penilaian model bahasa untuk mengukur prestasi sistem, di mana variasi skor merentas dimensi dan konteks berfungsi sebagai petunjuk tidak langsung kepada kebolehpercayaan sistem. Dalam literatur, kestabilan sistem dalam mengekalkan prestasi merentas konteks, yang sering dirujuk sebagai *robustness*, diambil kira sebagai aspek penting dalam menilai kebolehpercayaan sistem AI (Novikova et al., 2025; Laskar et al., 2024). Pendekatan ini juga selaras dengan amalan semasa penilaian penjanaan diperkaya berasaskan carian maklumat (*retrieval-augmented generation, RAG*) (Zhang & Zhang, 2025).

Jadual 3: Kriteria Kestabilan Konfigurasi Sistem AI

Parameter	Penjelasan
Median ≥ 4.0	Tahap prestasi minimum boleh diterima
IQR kecil	Variasi rendah dan konsisten
Δ Perbezaan skor agregat median antara tema kecil	Sensitiviti rendah terhadap perubahan tema

Jadual 3 di atas menjelaskan kriteria kestabilan sistem yang digunakan dalam kajian ini. Kestabilan sistem menunjukkan sama ada sistem AI boleh dipercayai atau mengelirukan. Ia ditafsirkan berdasarkan pola skor yang terhasil daripada tiga parameter utama, iaitu: (i) tahap prestasi minimum yang boleh diterima iaitu sekurang-kurangnya bersamaan skor 4.0, (ii) variasi skor IQR yang kecil dan konsisten apabila dibandingkan dengan sistem lain, dan (iii) sensitiviti rendah iaitu perubahan skor agregat median yang kecil merentas tema, dan dalam kajian ini terdapat dua tema iaitu Konsep Asas Fatwa dan Cabaran Institusi Fatwa. Skor agregat median ditakrifkan sebagai median bagi skor empat dimensi iaitu konsistensi, koheren, keterangkuman, halusinasi untuk setiap respons. Nilai skor agregat bagi setiap tema kemudian diperoleh dengan mengambil median skor agregat merentas semua item soalan dalam tema tersebut. Manakala, IQR dikira sebagai perbezaan antara kuartil ketiga (Q3) dan kuartil pertama (Q1). Sehubungan itu, sistem dikelaskan sebagai stabil sekiranya mengekalkan skor median sekurang-kurangnya pada skor 4.0, iaitu tahap memuaskan, menunjukkan variasi skor IQR yang rendah dan perubahan median yang minimum antara tema.

Untuk meneliti konsistensi pemarkahan antara penilai, analisis kebolehpercayaan antara penilai (*inter-rater reliability*) telah dijalankan menggunakan *Quadratic Weighted Cohen's Kappa*, memandangkan penilaian melibatkan dua orang penilai pakar yang menilai empat dimensi maklumat menggunakan skala Likert lima mata yang bersifat ordinal. Penggunaan pemberat kuadratik adalah selaras dengan amalan dalam analisis data ordinal kerana ia mengambil kira tahap perbezaan antara kategori penilaian. Hasil analisis menunjukkan nilai kw sekitar 0.35, yang mencerminkan tahap persetujuan yang terhad antara penilai. Oleh itu, dapatan skor dalam kajian ini perlu ditafsirkan sebagai dapatan eksploratori dengan mengambil kira variasi pertimbangan profesional antara penilai. Walaupun tahap persetujuan tidak tinggi, ukuran ini

masih memberikan indikasi kebolehppercayaan minimum untuk menyokong analisis perbandingan sistem, dengan mengambil kira sifat subjektif konstruk yang dinilai dan peranan pertimbangan pakar dalam penilaian kualitatif. (Warrens, 2013).

DAPATAN KAJIAN DAN PERBINCANGAN

Dapatan kajian dibentangkan secara berperingkat daripada pola umum kepada analisis yang lebih khusus. Pertama, prestasi ketiga-tiga konfigurasi sistem dibandingkan secara agregat berdasarkan empat dimensi ralat maklumat, iaitu konsistensi, koheren, keterangkuman dan halusinasi. Kedua, analisis diperluas dengan membandingkan kestabilan setiap sistem merentas dua tema kefatwaan, iaitu Konsep Asas Fatwa dan Cabaran Institusi Fatwa. Ketiga, analisis diperincikan pada peringkat soalan bagi mengenal pasti item yang menunjukkan prestasi tinggi atau rendah, variasi antara sistem serta bentuk ralat yang lebih ketara. Pendekatan ini membolehkan dapatan dinilai bukan sahaja berdasarkan skor keseluruhan sistem, tetapi juga berdasarkan perubahan prestasi mengikut tema dan ciri setiap soalan. Dengan itu, perbincangan dapat memberikan gambaran yang lebih menyeluruh tentang kekuatan dan keterbatasan setiap konfigurasi AI dalam konteks kajian eksploratori ini.

Dapatan dan Perbincangan 1

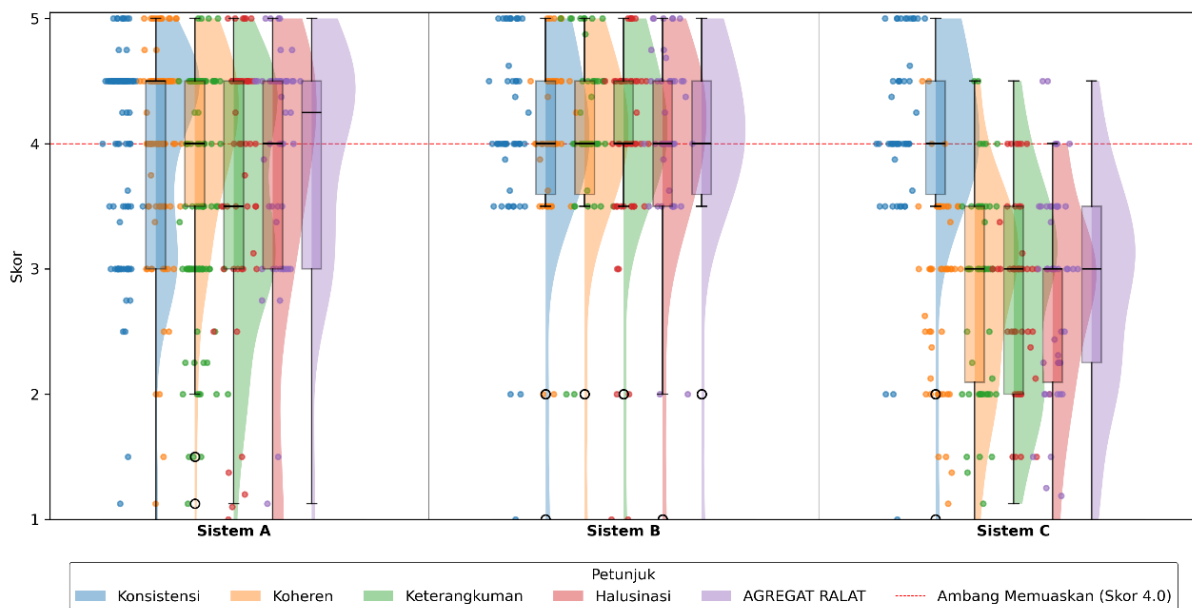
Jadual 4: Skor Agregat Ralat Maklumat Setiap Sistem

Dimensi	Sistem A	Sistem B	Sistem C
Konsistensi	4.5	4	4
Koheren	4.25	4	3
Keterangkuman	4.5	4	3
Halusinasi	4.25	4	3
Skor Agregat Ralat	4.375	4	3
IQR	1.125	0.625	1.25

Nota: Skor adalah berasaskan skala Likert lima mata dengan median digunakan sebagai ukuran kecenderungan pusat bagi data berskala ordinal. Skor agregat ralat merujuk kepada median keempat-empat dimensi penilaian, manakala IQR menunjukkan tahap variasi skor ralat. Skor yang lebih tinggi menunjukkan prestasi yang lebih baik dan tahap ralat yang lebih rendah.

Jadual 4 menunjukkan perbezaan skor agregat antara ketiga-tiga konfigurasi sistem berdasarkan empat dimensi ralat maklumat. Secara umum, Sistem A mencatatkan median skor agregat tertinggi, diikuti Sistem B dan Sistem C. Walau bagaimanapun, nilai IQR menunjukkan bahawa Sistem B mempunyai variasi skor yang lebih kecil berbanding dua sistem lain. Perbezaan antara tahap skor dan variasi ini penting kerana prestasi sistem tidak hanya dinilai berdasarkan median yang tinggi, tetapi juga pada keupayaannya mengekalkan kestabilan respons. Rajah 2 seterusnya memperlihatkan taburan skor bagi setiap dimensi dengan lebih terperinci.

Rajah 2: Taburan Skor Ralat Maklumat Setiap Sistem



Rajah 2 memperlihatkan taburan skor bagi setiap dimensi ralat maklumat dalam ketiga-tiga konfigurasi sistem. Secara visual, Sistem A menunjukkan taburan skor yang relatif tinggi tetapi lebih tersebar, manakala Sistem B memperlihatkan taburan yang lebih tertumpu sekitar skor 4.0. Sebaliknya, Sistem C menunjukkan taburan yang lebih rendah dan berubah. Corak ini melengkapi dapatan dalam Jadual 4 dengan menunjukkan bahawa Sistem A mempunyai prestasi agregat yang lebih tinggi, tetapi Sistem B lebih konsisten dari sudut variasi skor. Dalam konteks kajian eksploratori ini, pola tersebut memberi petunjuk bahawa pangkalan pengetahuan tertutup mungkin membantu menghasilkan respons yang lebih stabil, manakala penggunaan carian web berpotensi meningkatkan keluasan maklumat tetapi turut memperkenalkan variasi sumber yang lebih besar.

Bagi Sistem A, gabungan mekanisme RAG dan carian web berpotensi memperluas liputan maklumat tetapi pada masa yang sama boleh membuka variasi sumber yang lebih besar. Sebaliknya, Sistem C yang bergantung sepenuhnya kepada carian web tanpa pangkalan pengetahuan tertutup menunjukkan prestasi yang lebih rendah dari sudut konsistensi dan keterangkuman. Corak ini selari dengan dapatan kajian RAG yang menunjukkan bahawa kualiti respons boleh terjejas apabila proses mengambil maklumat daripada sumber yang tidak relevan, bercanggah atau kurang sesuai dengan konteks pertanyaan (Mortaheb et al., 2025; Brown et al., 2025).

Dalam erti kata lain, median skor yang tinggi tidak semestinya mencerminkan kebolehpercayaan sistem tanpa disertai kestabilan variasi iaitu nilai IQR yang rendah, khususnya dalam domain kefatwaan yang menuntut integriti maklumat (Rahim et al., 2025). Secara agregat merentas semua dimensi, Sistem B mencatatkan nilai IQR paling kecil, menandakan variasi ralat yang lebih terkawal dan konsisten. Ia berkedudukan lebih baik berbanding Sistem A dan Sistem C dalam memenuhi parameter (i) dan parameter (ii) dalam kriteria kestabilan sistem.

Walaupun terdapat nilai ekstrem rendah dalam taburan skor Sistem B, ketidakpekaan median dan IQR terhadap nilai luar menjadikan dapatan ini lebih kukuh dari sudut penilaian kestabilan ralat. Keadaan ini selari dengan dapatan penelitian Shuster et al. (2021) yang menunjukkan bahawa pengikatan respons kepada dokumen rujukan melalui mekanisme *retrieval* mengurangkan kesalahan fakta. Dalam kajian ini, kestabilan tersebut diperlihatkan dengan lebih jelas apabila pendekatan RAG dilaksanakan menggunakan pangkalan pengetahuan tertutup.

Dalam konteks domain kefatwaan, kestabilan sedemikian adalah lebih penting berbanding pencapaian skor maksimum semata-mata, kerana ketidakstabilan prestasi berpotensi menjejaskan kebolehpercayaan penggunaan teknologi AI. Oleh itu, dapat dirumuskan bahawa Sistem B menunjukkan potensi untuk digunakan sebagai alat sokongan maklumat dalam domain kefatwaan. Namun, penggunaannya tertakluk kepada prinsip kehati-hatian (*iḥtiyāṭ*) dan keperluan verifikasi (*tabayyun*) untuk menjaga integriti dalam pengendalian maklumat agama (Meerangani et al., 2025; Rahim et al., 2025; Niri et al., 2022).

Dapatan dan Perbincangan 2

Jadual 5: Agregat Ralat Maklumat Setiap Sistem Merentasi Tema

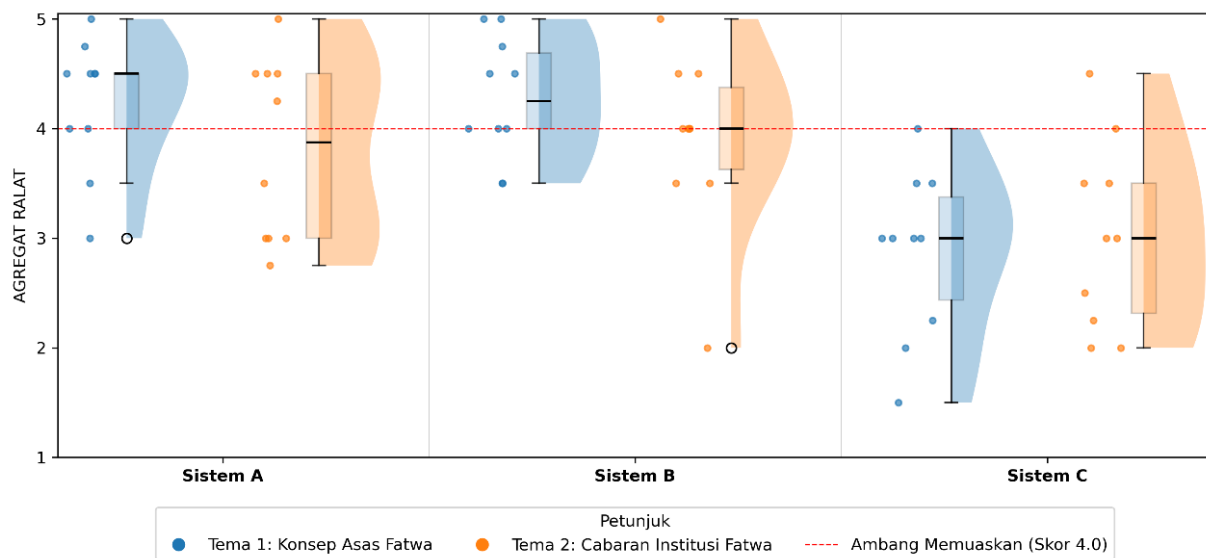
Sistem	Sistem A	Sistem B	Sistem C
Median Tema 1: Konsep Asas Fatwa			
Konsistensi	4.5	4.25	4.25
Koheren	4.5	4.25	3
Keterangkuman	4.5	4.25	3
Halusinasi	4.5	4	3
Agregat Ralat	4.5	4.25	3
IQR	0.5	0.6875	0.9375
Median Tema 2: Cabaran Institusi Fatwa			
Konsistensi	4.25	4	4
Koheren	3.75	4	3
Keterangkuman	4	4	3
Halusinasi	3.75	4.25	3
Agregat Ralat	3.875	4	3
IQR	1.5	0.75	1.1875
Perbezaan Agregat Tema (Δ)	0.625	0.25	0

Nota: Skor adalah berasaskan skala Likert lima mata dengan median digunakan sebagai ukuran kecenderungan pusat bagi data berskala ordinal. Agregat ralat merujuk kepada median keempat-empat dimensi penilaian, manakala IQR menunjukkan tahap variasi skor ralat. Skor yang lebih tinggi menunjukkan prestasi yang lebih baik dan tahap ralat yang lebih rendah

Jadual 5 menunjukkan bahawa perubahan tema memberi kesan yang berbeza terhadap prestasi ketiga-tiga konfigurasi sistem. Bagi Tema 1, iaitu Konsep Asas Fatwa, Sistem A mencatatkan skor agregat tertinggi pada 4.5 dengan IQR 0.5, diikuti Sistem B pada 4.25 dengan IQR 0.6875, manakala Sistem C kekal pada skor 3.0 dengan IQR 0.9375. Apabila beralih kepada Tema 2, iaitu Cabaran Institusi Fatwa, skor agregat Sistem A menurun kepada 3.875 dengan IQR meningkat kepada 1.5. Perubahan ini menghasilkan perbezaan agregat tema terbesar, iaitu 0.625, sekali gus menunjukkan bahawa prestasi Sistem A lebih sensitif terhadap perubahan tema. Sebaliknya, Sistem B hanya menunjukkan perubahan agregat daripada 4.25 kepada 4.0, dengan perbezaan sebanyak 0.25. Walaupun IQR Sistem B meningkat sedikit

daripada 0.6875 kepada 0.75, taburannya masih lebih kecil berbanding Sistem A dan Sistem C dalam Tema 2. Sistem C pula mencatatkan perbezaan agregat tema sifar kerana skornya kekal pada 3.00 bagi kedua-dua tema. Namun demikian, keadaan ini tidak boleh ditafsirkan sebagai kestabilan prestasi yang baik kerana skor median kekal di bawah ambang memuaskan 4.0, manakala nilai IQR turut meningkat daripada 0.9375 kepada 1.1875. Secara relatif, pola ini menunjukkan bahawa Sistem B lebih mampu mengekalkan tahap prestasi yang hampir sama apabila berhadapan dengan perubahan daripada tema konseptual kepada tema yang lebih analitikal. Dapatan ini memberikan petunjuk awal bahawa kestabilan sistem tidak hanya bergantung pada nilai median yang tinggi, tetapi juga pada keupayaan mengekalkan prestasi apabila konteks dan tuntutan soalan berubah.

Rajah 3: Taburan Agregat Ralat Maklumat Setiap Sistem Merentasi Tema



Rajah 3 menunjukkan skor agregat ralat maklumat dan taburannya untuk setiap sistem AI dalam tema yang berbeza, iaitu tema Konsep Asas Fatwa dan tema Cabaran Institusi Fatwa. Secara perbandingan, Sistem B mengekalkan secara relatif kedudukan median yang konsisten merentas tema berbanding Sistem A dan Sistem C. Manakala, Sistem A menunjukkan perubahan agregat median yang lebih ketara apabila beralih tema ($|\Delta| = 0.625$), menandakan prestasi kestabilan yang lebih rendah.

Dari sudut parameter (iii), perubahan agregat median merentas tema bagi Sistem B juga didapati lebih kecil berbanding Sistem A. Sistem C pula meskipun menunjukkan perbezaan agregat tema sifar, ia masih tidak memenuhi parameter (i) kerana skor median agregatnya kekal pada 3.00 di bawah ambang 4.0, dan variasi IQR juga bukan sifar. Keadaan itu tidak mencerminkan kestabilan prestasi. Adapun keadaan Sistem B adalah selari dengan kajian terdahulu yang menunjukkan bahawa sistem berasaskan penambahan pengetahuan tersusun (seperti RAG) dapat mengurangkan ralat maklumat dalam tugas berasaskan pengetahuan yang semakin kompleks (Brown et al., 2025; Shuster et al., 2021).

Walaupun analisis skor agregat dan perbandingan merentas tema menunjukkan pola umum prestasi setiap sistem, penelitian pada peringkat soalan diperlukan untuk menjelaskan dengan lebih khusus bagaimana variasi tersebut berlaku dalam empat dimensi ralat maklumat. Jadual 6 seterusnya memperincikan skor setiap soalan mengikut konfigurasi sistem dan dimensi penilaian.

Jadual 6: Skor Median Ralat Maklumat bagi Setiap Soalan Mengikut Sistem dan Dimensi Penilaian

No. Soalan	Sistem	Konsistensi	Koheren	Keterangkuman	Halusinasi	Ralat
1	A	4.5	4.5	4.5	4.5	4.5
	B	5	5	5	5	5
	C	5	3	3	2.5	3
2	A	4.5	5	4.5	5	4.75
	B	4	4	4	4	4
	C	4	2.5	2	2	2.25
3	A	5	5	5	5	5
	B	5	5	5	5	5
	C	5	4	4	4	4
4	A	3.5	4	4	4	4
	B	3.5	3.5	3.5	3.5	3.5
	C	3.5	3	3	3	3
5	A	4	4	4	4	4
	B	4.5	4.5	4.5	4.5	4.5
	C	4.5	1.5	1.5	1.5	1.5
6	A	4.5	4.5	4.5	4.5	4.5
	B	4	4	4	4	4
	C	4	3.5	3.5	3	3.5
7	A	3.5	4	3.5	3.5	3.5
	B	5	4.5	4.5	4	4.5
	C	5	2	2	1.5	2
8	A	3	3	3	3	3
	B	4	4	4	4	4
	C	4	3	3	3	3
9	A	4.5	4.5	4.5	4.5	4.5
	B	5	4.5	5	4.5	4.75
	C	5	3	3	3	3
10	A	4.5	4.5	4.5	4.5	4.5
	B	3.5	3.5	4	3	3.5
	C	3.5	3.5	3.5	3.5	3.5
11	A	4.5	4.5	4.5	4.5	4.5
	B	3.5	3.5	3.5	3.5	3.5
	C	3.5	2	2	2	2
12	A	3	3	3	2.5	3
	B	4.5	4.5	4.5	4.5	4.5
	C	4.5	4.5	4.5	4	4.5
13	A	4.5	4.5	4.5	4.5	4.5
	B	5	5	5	5	5
	C	5	2.5	2.5	2.5	2.5
14	A	4.5	4	4.5	4	4.25
	B	3.5	3.5	3.5	3.5	3.5
	C	3.5	2	2	2.5	2.25
15	A	3	3	3	3	3
	B	2	2	2	2	2
	C	2	3.5	3.5	3.5	3.5
16	A	5	5	5	5	5

	B	4	4	4	3.5	4
	C	4	3	3	3	3
17	A	2.5	2.5	3	3.5	2.75
	B	4	4	4	4.5	4
	C	4	3.5	3.5	3	3.5
18	A	4.5	4.5	4.5	4.5	4.5
	B	4	4	4	4	4
	C	4	3	3	3	3
19	A	4	3.5	3.5	3.5	3.5
	B	4	4	4	4.5	4
	C	4	3.5	4	4	4
20	A	3	3	3	3	3
	B	4.5	4.5	4.5	4.5	4.5
	C	4.5	2	2	2	2

Jadual 6 memperincikan skor median bagi setiap soalan mengikut tiga konfigurasi sistem dan empat dimensi penilaian. Soalan 1 hingga 10 mewakili Tema 1, iaitu Konsep Asas Fatwa, manakala soalan 11 hingga 20 mewakili Tema 2, iaitu Cabaran Institusi Fatwa. Pada peringkat item, Tema 1 secara umum menunjukkan prestasi yang lebih tinggi bagi Sistem A dan Sistem B, dengan kebanyakan skor ralat berada sekitar 4.0 hingga 5.0, manakala Sistem C memperlihatkan variasi yang lebih besar dan beberapa penurunan ketara, khususnya pada soalan yang memerlukan respons lebih daripada sekadar fakta asas. Bagi Tema 2 pula, variasi antara sistem menjadi lebih jelas. Sistem A masih mencatatkan skor tinggi pada beberapa soalan, tetapi turut menunjukkan penurunan pada item tertentu, manakala Sistem B lebih kerap mengekalkan skor sekitar 4.0 hingga 5.0 walaupun tidak sentiasa memperoleh skor tertinggi. Sistem C pula menunjukkan pola yang lebih tidak seragam, dengan beberapa soalan memperoleh skor sederhana atau tinggi tetapi sebahagian yang lain jatuh dengan ketara. Corak ini menunjukkan bahawa analisis pada peringkat soalan memberikan gambaran yang lebih terperinci berbanding skor agregat semata-mata, terutama dalam mengenal pasti item yang lebih sensitif terhadap perubahan konfigurasi sistem dan aras tema.

Apabila diteliti mengikut dimensi, perbezaan antara sistem paling jelas pada aspek koheren, keterangkuman dan halusinasi, manakala skor konsistensi secara relatif lebih tinggi dalam banyak soalan. Sistem C, misalnya, masih memperoleh skor konsistensi yang tinggi pada beberapa item, tetapi prestasinya menurun apabila dinilai dari sudut keutuhan hujah, keluasan kandungan dan kawalan halusinasi. Corak ini dapat dilihat dengan ketara pada soalan 5, 7, 13 dan 20, apabila skor konsistensi Sistem C berada antara 4.5 hingga 5.0 tetapi skor bagi koheren, keterangkuman dan halusinasi turun ke sekitar 1.5 hingga 2.5. Sebaliknya, Sistem B lebih kerap menunjukkan keseimbangan antara keempat-empat dimensi, walaupun terdapat beberapa item yang mencatatkan penurunan, seperti soalan 15. Sistem A juga menunjukkan prestasi yang relatif seimbang pada banyak item, tetapi kelemahan lebih jelas pada soalan tertentu seperti soalan 17. Dapatan ini menunjukkan bahawa ketepatan fakta sahaja belum memadai untuk menilai kualiti respons AI kerana sesuatu jawapan boleh kelihatan konsisten dari segi fakta tetapi masih lemah dari sudut penghujahan, keluasan maklumat atau sokongan sumber. Oleh itu, penggunaan empat dimensi secara serentak memberikan gambaran yang lebih menyeluruh tentang bentuk ralat yang berlaku pada peringkat setiap soalan.

Pada peringkat hubungan antara soalan, dapatan menunjukkan bahawa perubahan daripada soalan deskriptif kepada soalan yang lebih analitikal atau berorientasikan penyelesaian tidak menghasilkan pola penurunan yang seragam bagi semua sistem. Dalam Tema 2, misalnya,

soalan 15 berkaitan cabaran tenaga kerja dan kepakaran menunjukkan skor yang relatif rendah, khususnya bagi Sistem A dan Sistem B, manakala soalan 17 berkaitan solusi pengukuhan tadbir urus memperlihatkan perbezaan prestasi yang lebih ketara antara sistem. Sebaliknya, beberapa soalan penyelesaian seperti soalan 18 dan 19 masih menunjukkan prestasi yang lebih stabil bagi konfigurasi tertentu. Corak ini memberikan petunjuk bahawa variasi prestasi bukan sahaja berkaitan dengan bentuk soalan, tetapi mungkin turut dipengaruhi oleh kesesuaian konfigurasi sistem dengan kandungan yang diperlukan. Beberapa item seperti soalan 15, 17 dan 20 wajar diberi perhatian dalam kajian lanjutan kerana menunjukkan skor yang lebih rendah atau jurang prestasi yang lebih jelas antara sistem. Penelitian lanjut terhadap item-item ini diperlukan untuk menentukan sama ada kelemahan tersebut berkaitan dengan keterbatasan sumber rujukan, proses pengambilan maklumat atau keupayaan model menghubungkan maklumat dengan konteks soalan.

Bagi Tema 1, soalan 5 dapat dijadikan contoh untuk memperincikan bagaimana perbezaan antara konfigurasi sistem muncul pada peringkat kandungan respons. Sistem A dan Sistem B pada asasnya mengekalkan inti maklumat dalam set data, iaitu tugas mufti dianggap berat kerana berkait dengan kedudukannya dalam meneruskan peranan Rasulullah SAW dalam membimbing masyarakat serta besarnya tanggungjawab terhadap kefahaman dan amalan hukum. Sistem B turut menghubungkan tanggungjawab tersebut dengan keperluan ilmu, ketelitian dan pemahaman terhadap realiti semasa. Sistem C juga mengekalkan inti yang sama, tetapi memperluas respons dengan beberapa perincian yang tidak terdapat dalam set data rujukan. Hal ini sejalan dengan pola skor S5 apabila Sistem A memperoleh skor ralat 4.0, Sistem B 4.5 dan Sistem C hanya 1.5.

Bagi Sistem C, skor konsistensinya masih tinggi, iaitu 4.5, kerana idea teras respons tidak bercanggah dengan maklumat asas set data. Namun, skor koheren, keterangkuman dan halusinasi masing-masing hanya 1.5. Dari sudut keterangkuman, respons tersebut tidak membawa beberapa konteks yang terdapat dalam set data, khususnya perkembangan sejarah institusi kemuftian dan perubahan kedudukan pandangan mufti daripada panduan yang tidak mengikat kepada bentuk yang lebih rasmi. Dari sudut halusinasi pula, Sistem C menisbahkan secara khusus kepada al-Shatibi bahawa mufti pada hakikatnya menandatangani hukum bagi pihak Allah (*al-muwaqqi' 'an Allah*) serta menambah beberapa implikasi seperti “hukum disalahgunakan”, “keadilan syariat terjejas” dan tuntutan “keikhlasan yang luar biasa”, sedangkan butiran tersebut tidak terdapat dalam set data yang digunakan sebagai rujukan penilaian.

Dalam kerangka eksperimen ini, penambahan tersebut dianggap sebagai maklumat yang tidak disokong oleh sumber yang dibenarkan, walaupun kandungannya mungkin kelihatan munasabah dan disampaikan secara meyakinkan. Bagi dimensi koheren pula, respons Sistem C secara linguistik masih tersusun, tetapi pengembangan inferens dan penyandaran maklumat di luar sumber rujukan mengurangkan keselarasan kandungannya dengan konteks maklumat yang menjadi asas soalan. Kes ini menunjukkan bahawa respons AI yang kelihatan lancar dan konsisten pada peringkat umum masih boleh mengandungi kelemahan dari sudut keterangkuman, kesahihan inferens dan pengasasan sumber yang hanya dapat dikenal pasti melalui penilaian pelbagai dimensi.

Bagi Tema 2, soalan 17 pula memperlihatkan pola yang berbeza. Ketiga-tiga sistem pada asasnya mengenal pasti keperluan penambahbaikan tadbir urus melalui adaptasi teknologi digital. Namun, Sistem B lebih hampir kepada kandungan set data dengan menghubungkan penambahbaikan tadbir urus kepada penggunaan teknologi maklumat, mesyuarat dan

perundingan dalam talian, penyelarasan fatwa serta pemantapan kompetensi digital pegawai fatwa. Sistem A turut merangkumi beberapa unsur tersebut, tetapi tidak menghuraikan secara khusus penggunaan algoritma machine learning, penganalisan maklumat fatwa dan proses rujuk silang fatwa yang dinyatakan dalam set data. Hal ini selari dengan skor keterangkumannya yang lebih rendah, iaitu 3.0. Skor koheren Sistem A yang rendah, iaitu 2.5, pula menunjukkan bahawa penilai melihat hubungan antara cadangan yang dikemukakan dengan kerangka penyelesaian dalam sumber rujukan sebagai kurang kukuh, meskipun respons tersebut masih tersusun dari sudut bahasa.

Sistem C pula memberikan jawapan yang lebih umum dengan menekankan ketelusan, kemudahan akses, kepercayaan awam dan disiplin dalam proses keputusan, tetapi tidak membawa beberapa solusi khusus dalam set data seperti penambahbaikan polisi sumber manusia, penggunaan algoritma machine learning serta rujuk silang maklumat fatwa. Dari sudut halusinasi, Sistem C memperoleh skor 3.0 kerana turut memperkenalkan beberapa perincian yang tidak dinyatakan dalam set data, antaranya peningkatan “kepercayaan awam”, penekanan terhadap “ketelusan” dan keperluan memastikan proses pengesahan fatwa kekal berlandaskan disiplin ilmu. Perincian tersebut tidak semestinya salah, tetapi dalam kerangka eksperimen ini ia tidak dapat disandarkan secara langsung kepada sumber rujukan yang dibenarkan. Kes S17 sekali lagi menunjukkan bahawa perbezaan prestasi sistem bukan sahaja melibatkan ketepatan fakta, tetapi juga sejauh mana respons merangkumi kandungan penting, membina inferens yang selaras dengan sumber dan mengelakkan penambahan maklumat yang tidak disokong oleh set data.

Secara keseluruhan, analisis pada peringkat sistem, tema dan soalan menunjukkan bahawa konfigurasi AI boleh menghasilkan tahap prestasi yang berbeza bergantung pada bentuk dan kandungan pertanyaan. Namun, prestasi teknikal tersebut tidak dengan sendirinya memberikan autoriti hukum kepada AI. Autoriti dan tanggungjawab pengeluaran fatwa tetap terletak pada mufti dan institusi fatwa, manakala AI hanya berfungsi sebagai alat sokongan dalam pengurusan dan penyampaian maklumat berkaitan fatwa. Justeru, penggunaan AI perlu disertai dengan pengawalan sumber, semakan pakar dan kebolehkasan rujukan, selaras dengan prinsip *tabayyun* dan *ihtiyāt* bagi memastikan penggunaannya secara bertanggungjawab dalam konteks kefatwaan (Rahim et al., 2025; Meerangani et al., 2025).

Oleh itu, dapatan ini memberikan sokongan awal kepada hipotesis kajian bahawa kestabilan respons dan corak ralat maklumat sistem AI berkemungkinan berbeza mengikut aras tema kefatwaan dalam konteks reka bentuk eksperimen ini. Dari perspektif syariah dan tadbir urus maklumat, variasi ralat maklumat yang dikesan menunjukkan keperluan verifikasi (*tabayyun*) sebelum sesuatu maklumat diterima dan disebarkan, khususnya apabila maklumat itu menyentuh autoriti institusi fatwa (Meerangani et al., 2025). Selain itu, kestabilan konfigurasi tertentu merentas tema menunjukkan bahawa pemilihan seni bina sistem, contohnya pergantungan kepada pangkalan pengetahuan tertutup berbanding carian terbuka, perlu diambil kira dan disepadukan dengan dasar kawal selia yang menekankan ketelusan sumber, semakan pakar dan had penggunaan AI sebagai alat sokongan. Justeru, penggunaan AI dalam ekosistem fatwa wajar diletakkan di bawah kerangka tadbir urus dan langkah mitigasi ralat yang jelas sekurang-kurangnya dari segi, (i) pengawalan sumber rujukan, (ii) penilaian berkala terhadap ralat, dan (iii) mekanisme pembaikan semula.

Dari sudut limitasi, kajian eksploratori ini berasaskan reka bentuk eksperimen terkawal yang menggunakan set data terhad iaitu satu dokumen sumber, 20 soalan input, dua tema kefatwaan dan tiga konfigurasi sistem AI dalam platform yang sama. Skop ini sesuai untuk menguji secara

awal reka bentuk eksperimen dan mengenal pasti pola asas ralat maklumat, tetapi tidak mencukupi untuk membuat generalisasi terhadap keseluruhan domain kefatwaan. Dapatan kajian ini harus dilihat sebagai garis asas (*baseline*) bagi mengenal pasti corak ralat maklumat fatwa dalam skop soalan deskriptif–analitik yang diseragamkan. Ia tidak sesuai digeneralisasikan secara langsung kepada keseluruhan variasi pertanyaan kefatwaan yang lebih luas, termasuk perbincangan hukum fiqh, perbincangan kefatwaan, isu kontemporari yang sangat kontekstual atau pertanyaan yang memerlukan rujukan dalil terperinci. Sehubungan itu, dicadangkan kajian lanjutan dan lebih mendalam dibuat untuk meneliti hal terbabit khususnya memperluas bilangan item, mempelbagaikan bentuk soalan input, pelbagai dokumen berautoriti, lebih banyak tema, bilangan penilai pakar yang lebih besar dan menambah tema-subtema bagi menguji ketekalan dapatan merentas situasi kefatwaan yang lebih rencam.

KESIMPULAN

Dalam skop terhad kajian eksploratori ini, dapatan memberikan petunjuk awal bahawa ralat maklumat janaan kecerdasan buatan dalam domain kefatwaan mungkin berkait dengan konfigurasi sistem dan aras tema kefatwaan. Dapatan ini memberikan implikasi awal terhadap tadbir urus penggunaan AI dalam ekosistem fatwa, khususnya dari sudut pemilihan seni bina sistem yang terkawal, penglibatan kepakaran rentas bidang, ketelusan sumber dan kebolehsamaan proses. Sehubungan itu, kajian ini mencadangkan agar pihak berautoriti khususnya institusi fatwa membangunkan garis panduan yang lebih jelas dan terperinci terhadap penggunaan AI dalam kefatwaan dengan memperhalusi pengawalan sumber rujukan, pemantauan ralat maklumat oleh pakar manusia dan mekanisme pembaikan semula sebagai langkah mitigasi ralat.

Dari sudut amalan, teknologi AI wajar diposisikan sebagai alat sokongan dalam pengurusan serta penyampaian maklumat dan sudah tentu bukan sebagai pengeluar keputusan fatwa. Posisi ini selari dengan prinsip menjaga kepercayaan awam terhadap institusi fatwa dalam persekitaran digital. Pada masa hadapan, penyelidikan lanjut disarankan untuk memperluas skop data, menilai pelbagai seni bina dan konfigurasi sistem AI, serta membangunkan kerangka dasar dan operasi yang bersepadu bagi penggunaan AI dalam konteks kefatwaan, agar teknologi ini dapat dimanfaatkan secara beretika dan bertanggungjawab.

DEKLARASI

Penghargaan Penulis merakamkan penghargaan kepada panel penilai atas sumbangan kepakaran dalam menilai hasil janaan sistem kecerdasan buatan berdasarkan instrumen kajian yang ditetapkan.	Konflik Kepentingan Penulis memperakukan bahawa tiada konflik kepentingan yang terlibat.
Sumbangan Penulis Mohammaddin Abdul Niri: Konseptualisasi, metodologi, eksperimen, pengumpulan dan analisis data, visualisasi, penyediaan draf asal dan pengurusan kajian. Noor Naemah Abdul Rahman: Penyediaan sumber rujukan bagi pembentukan set data kefatwaan dan bimbingan kepakaran bidang fatwa. Mohd Anuar Ramli: Pengesahan, analisis serta semakan dan penyuntingan manuskrip. Semua penulis telah membaca dan memperakukan draf akhir manuskrip.	Perakuan Etika Perakuan etika tidak diperlukan bagi kajian ini kerana ia tidak melibatkan pengumpulan data daripada pesakit, haiwan atau data peribadi.
Kebolehdapatan Data Data yang menyokong dapatan kajian ini boleh diperolehi daripada penulis koresponden atas permintaan dengan justifikasi munasabah.	Penggunaan Kecerdasan Buatan (AI) Penulis memperakukan bahawa teknologi kecerdasan buatan (AI), iaitu ChatGPT versi 5.2, digunakan dalam kajian ini bagi dua tujuan. Pertama, ChatGPT merupakan

objek kajian yang menghasilkan respons AI sebagai data kajian. Kedua, ChatGPT digunakan sebagai alat sokongan dalam penyediaan manuskrip, khususnya bagi semakan pruf, penyuntingan bahasa, pengaturcaraan dan semakan silang maklumat dengan merujuk sumber akademik yang berautoriti. Semua kandungan, perbincangan, analisis, justifikasi dan kesimpulan telah disemak, disahkan dan diperakukan oleh penulis. Penulis bertanggungjawab sepenuhnya terhadap kandungan akhir manuskrip ini.

RUJUKAN

- AbuJarour, S., Qarariah, A., Saadeh, N., & Salem, M. (2024). AI, Misinformation, and Fake News: A Literature Review of Ethical and Technical Approaches. In *Contributions to Finance and Accounting: Vol. Part F3769* (pp. 641–652). Springer Nature. https://doi.org/10.1007/978-3-031-67547-8_55
- Al Kubaisi, A. A. S. H. (2024). Ethics of Artificial Intelligence a Purposeful and Foundational Study in Light of the Sunnah of Prophet Muhammad. *Religions*, 15(11). <https://doi.org/10.3390/rel15111300>
- Aldoseri, A., Al-Khalifa, K. N., & Hamouda, A. M. (2023). Re-Thinking Data Strategy and Integration for Artificial Intelligence: Concepts, Opportunities, and Challenges. *Applied Sciences (Switzerland)*, 13(12). <https://doi.org/10.3390/app13127082>
- Alnefaie, S., Atwell, E., & Alsalka, M. A. (2025). Question Answering over the Arabic Hadith Sharif Using Transformer Models. *Communications in Computer and Information Science*, 2339 CCIS, 195–206. https://doi.org/10.1007/978-3-031-79164-2_17
- Asgari, E., Montaña-Brown, N., Dubois, M., Khalil, S., Balloch, J., Yeung, J. A., & Pimenta, D. (2025). A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *Npj Digital Medicine* 2025 8:1, 8(1), 274. <https://doi.org/10.1038/s41746-025-01670-7>
- Bdoor, S. Y., & Habes, M. (2025). Meta-Analysis of Ethical Challenges in the Use of Artificial Intelligence in Arab Media Institutions. *Studies in Computational Intelligence*, 1208, 131–146. https://doi.org/10.1007/978-3-031-89175-5_9
- Brown, A., Roman, M., & Devereux, B. (2025). A Systematic Literature Review of Retrieval-Augmented Generation: Techniques, Metrics, and Challenges. *ArXiv*. <https://doi.org/10.48550/arXiv.2508.06401>
- Bruno, A., Mazzeo, P. L., Chetouani, A., Tliba, M., & Kerkouri, M. A. (2023). Insights into Classifying and Mitigating LLMs' Hallucinations. *CEUR Workshop Proceedings*. <https://www.scopus.com/pages/publications/85178666482>
- Chiu, Edwin Kwan-Yeung, and Tom Wai-Hin Chung. (2025). Protocol for human evaluation of generative artificial intelligence chatbots in clinical consultations. *PLoS ONE*, 20(3). <https://doi.org/10.1371/journal.pone.0300487>
- Dahl, M., Magesh, V., Suzgun, M., & Ho, D. E. (2024). Large Legal Fictions: Profiling Legal Hallucinations in Large Language Models. *Journal of Legal Analysis*, 16(1), 64–93. <https://doi.org/10.1093/JLA/LAAE003>
- FACTS Team. (2024). *FACTS Grounding: A new benchmark for evaluating the factuality of large language models - Google DeepMind*. <https://deepmind.google/blog/facts-grounding-a-new-benchmark-for-evaluating-the-factuality-of-large-language-models/>
- Fang, X., Yuan, H., Li, H., Kou, J., Gu, C., Zhang, W., Duan, X., & Fang, Y. (2024). Reframing Hallucination in Large Language Models: A Lifecycle-Based, Mechanism-Aligned, and Phenomenon-Consistent Definition. *7th International Conference on Universal Village, UV 2024*. <https://doi.org/10.1109/UV63228.2024.11189217>

- Farquhar, S., Kossen, J., Kuhn, L., & Gal, Y. (2024). Detecting hallucinations in large language models using semantic entropy. *Nature* 2024 630:8017, 630(8017), 625–630. <https://doi.org/10.1038/s41586-024-07421-0>
- Gopali, S., Siarni-Namini, S., Abri, F., & Namin, A. S. (2024). The performance of the LSTM-based code generated by Large Language Models (LLMs) in forecasting time series data. *Natural Language Processing Journal*, 9, 100120. <https://doi.org/10.1016/J.NLP.2024.100120>
- Heo, S., Son, S., & Park, H. (2025). HaluCheck: Explainable and verifiable automation for detecting hallucinations in LLM responses. *Expert Systems with Applications*, 272, 126712. <https://doi.org/10.1016/J.ESWA.2025.126712>
- Jacovi, A., Wang, A., Alberti, C., Tao, C., Lipovetz, J., Olszewska, K., Haas, L., Liu, M., Keating, N., Bloniarz, A., Saroufim, C., Fry, C., Marcus, D., Kukliansky, D., Singh Tomar, G., Swirhun, J., Xing, J., Wang, L., Gurumurthy, M., ... Das, D. (2025). The FACTS Grounding Leaderboard: Benchmarking LLMs' Ability to Ground Responses to Long-Form Input. *arXiv preprint arXiv:2501.03200*. <https://arxiv.org/abs/2501.03200>.
- Jesson, A., Beltran-Velez, N., Chu, Q., Karlekar, S., & Kossen, J. (2024). Estimating the Hallucination Rate of Generative AI. *Advances in Neural Information Processing Systems*. <https://www.scopus.com/pages/publications/105000550746>
- Kannike, U. M. M., & Fahm, A. O. (2025). Exploring The Ethical Governance of Artificial Intelligence From an Islamic Ethical Perspective. *Jurnal Fiqh*, 22(1), 134–161. <https://doi.org/10.22452/fiqh.vol22no1.5>
- Kollar, J., & Alshibli, M. (2024). An Overview of Artificial Intelligence's Accuracy. *2024 IEEE Long Island Systems, Applications and Technology Conference, LISAT 2024*. <https://doi.org/10.1109/LISAT63094.2024.10808042>
- Koutrika, G. (2025). Navigating the Challenges of AI-Driven Data Processing. *DEBS 2025 - Proceedings of the 19th ACM International Conference on Distributed and Event-Based Systems*, 1–2. <https://doi.org/10.1145/3701717.3733846>
- Kulkarni, A., Zhang, Y., Moniz, J. R. A., Ge, X., Tseng, B.-H., Piraviperumal, D., & Swayamdipta, S. (2025). Evaluating evaluation metrics: The mirage of hallucination detection. *ArXiv*. <https://arxiv.org/abs/2504.18114>
- Laskar, M. T. R., Alqahtani, S., Bari, M. S., Rahman, M., Khan, M. A. M., Khan, H., Jahan, I., Bhuiyan, M. A. H., Tan, C. W., Parvez, M. R., Hoque, E., Joty, S., & Huang, J. X. (2024). A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. *EMNLP 2024 - 2024 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*, 13785–13816. <https://doi.org/10.18653/v1/2024.emnlp-main.764>
- Linardon, J., Jarman, H. K., McClure, Z., Anderson, C., Liu, C., & Messer, M. (2025). Influence of Topic Familiarity and Prompt Specificity on Citation Fabrication in Mental Health Research Using Large Language Models: Experimental Study. *JMIR Mental Health*, 12(1), e80371. <https://doi.org/10.2196/80371>
- Marey, A., Saad, A. M., Tanas, Y., Ghorab, H., Niemierko, J., Backer, H., & Umair, M. (2025). Evaluating the accuracy and reliability of AI chatbots in patient education. *Egyptian Journal of Radiology and Nuclear Medicine*, 56(1), 37. <https://link.springer.com/content/pdf/10.1186/s43055-025-01452-x.pdf>
- Maylawati, D. S., Khosyi'ah, S., Rizqullah, N., Fajar, F. I., & Ramdhani, M. A. (2025). Chatbot Assistant for Enhancing Religious Court Services in Indonesia Using Deep Learning-based Algorithms. *International Journal of Computing*, 24(2), 243–253. <https://doi.org/10.47839/IJC.24.2.4007>

- Meerangani, K. A., Harun, M. S., & Mansor, M. S. (2025). Prinsip Fiqh Dalam Penyebaran Maklumat Berasaskan Kecerdasan Buatan (AI) Di Media Sosial. *Jurnal Fiqh*, 22(2), 235–261. <https://doi.org/10.22452/FIQH.VOL22NO2.2>
- Mortaheb, M., Khojastepour, M. A. A., Chakradhar, S. T., & Ulukus, S. (2025). Rag-check: Evaluating multimodal retrieval augmented generation performance. *arXiv preprint arXiv:2501.03995*.
- Munshi, A. A., AlSabban, W. H., Farag, A. T., Rakha, O. E., Al Sallab, A. A., & Alotaibi, M. (2021). Towards an automated Islamic fatwa system: Survey, dataset and benchmarks. *International Journal of Computer Science and Mobile Computing*, 10(4). <https://doi.org/10.47760/ijcsmc.2021.v10i04.017>
- Musleh Al-Sartawi, A. M. A. (Ed.). (2021). *The Big Data-Driven Digital Economy: Artificial and Computational Intelligence* (Vol. 974). Springer International Publishing. <https://doi.org/10.1007/978-3-030-73057-4>
- Nascimento, N., Guimaraes, E., Chintakunta, S. S., & Boominathan, S. A. (2025). How Effective are LLMs for Data Science Coding? A Controlled Experiment. *Proceedings - 2025 IEEE/ACM 22nd International Conference on Mining Software Repositories, MSR 2025*, 211–222. <https://doi.org/10.1109/MSR66628.2025.00041>
- Niri, M. A., Saadon, M. H. M., Nawawi, M. S. A. M., & Jamaludin, M. H. (2022). Integrasi Model DIKW (Data-Information-Knowledge-Wisdom) Dalam Ilmu Falak Berasaskan Kerangka Sains Islam. *Afkar*, 24(2), 99–142. <https://doi.org/10.22452/afkar.vol24no2.3>
- Novikova, J., Anderson, C., Blili-Hamelin, B., Rosati, D., & Majumdar, S. (2025). Consistency in Language Models: Current Landscape, Challenges, and Future Directions. *ArXiv*. <https://arxiv.org/pdf/2505.00268>
- Nurfikri, S. R., Handayani, D., Fahreza, A. M., Almadani, D., Suryady, Z., & Lathifah, Z. K. (2024). Fikr: AI Chatbot Powered with Vector Search. *2024 10th International Conference on Computing, Engineering and Design, ICCED 2024*. <https://doi.org/10.1109/ICCED64257.2024.10983626>
- Nurhayati, Abdurrohman, R., Hulliyah, K., & Khairani, D. (2024). Performance Evaluation of a Chatbot with Retrieval Augmented Generation and Generative Pre-trained Transformer-4 Model for Taharah Domain. *2024 9th International Conference on Informatics and Computing, ICIC 2024*. <https://doi.org/10.1109/ICIC64337.2024.10957519>
- Özdemir, Ö. T., Kavan, M. Y., & Güven, Y. (2025). Evaluation of the readability, quality, and accuracy of AI chatbots. *BMC Oral Health*, 25(1), 1812. <https://link.springer.com/article/10.1186/s12903-025-07298-z>
- Raffinetti, E. (2023). A Rank Graduation Accuracy measure to mitigate Artificial Intelligence risks. *Quality and Quantity*, 57, 131–150. <https://doi.org/10.1007/s11135-023-01613-y>
- Rahim, S. F. A., Rahman, M. F. A., Thaidi, H. A. A., Azimi, N. N. M. A., & Jailani, M. R. (2025). Artificial intelligence for fatwa issuance: Guidelines and ethical considerations. *Journal of Fatwa Management and Research*, 30(1). <https://doi.org/10.33102/jfatwa.vol30no1.654>
- Rahman, N. N. A., & Niri, M. A. (2024). Cabaran institusi fatwa dalam era Revolusi Industri 4.0. In *Isu-Isu Fiqh Revolusi Industri Keempat* (pp. 24–38). Penerbit UTHM.
- Rashed, A., & Shirmohammadi, S. (2022). A Novel Method to Estimate Measurement Error in AI-Assisted Measurements. *Conference Record - IEEE Instrumentation and Measurement Technology Conference*. <https://doi.org/10.1109/I2MTC48687.2022.9806449>

- Santhoshkumar, S. P., Karthika, R., Vijayakumar, N., Gajalakshmi, S., Jayanthi, K., & Hariharasudhan, S. (2025). Analysis of Duplicating the Human Mind Suspicions in AI in Various Divisions. *Proceedings of the International Conference on Multi-Agent Systems for Collaborative Intelligence, ICMSCI 2025*, 1707–1714. <https://doi.org/10.1109/ICMSCI62561.2025.10894681>
- Shiferaw, K. B. (2024). Assessing the accuracy and quality of artificial intelligence chatbot-generated information. *BMC Medical Informatics and Decision Making*, 24(1), 404. <https://link.springer.com/article/10.1186/s12911-024-02824-5>
- Shuster, K., Poff, S., Chen, M., Kiela, D., & Weston, J. (2021). Retrieval Augmentation Reduces Hallucination in Conversation. *ArXiv*. <https://arxiv.org/abs/2104.07567>
- Shweiki, D., Al-Ahmad, S., & Samhan, A. A. A. (2025). Using Artificial Intelligence in Issuing Fatwas: A Jurisprudential Study. *Studies in Systems, Decision and Control*, 572, 761–768. https://doi.org/10.1007/978-3-031-76011-2_53
- Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications* 2024 11:1, 11(1), 1278. <https://doi.org/10.1057/s41599-024-03811-x>
- Vartumyan, A. A., Mikhailov, G. G., Galdin, E. V., Lavrova, T. N., & Orobinskaya, V. N. (2023). Risks Associated with the Use of Artificial Intelligence in Various Fields of Science. *Proceedings - 2023 5th International Conference on Control Systems, Mathematical Modeling, Automation and Energy Efficiency, SUMMA 2023*, 434–438. <https://doi.org/10.1109/SUMMA60232.2023.10349424>
- Warrens, M. J. (2013). Cohen's weighted kappa with additive weights. *Advances in Data Analysis and Classification*, 7(1), 41–55. <https://doi.org/10.1007/s11634-013-0123-9>
- Zhang, W., & Zhang, J. (2025). Hallucination Mitigation for Retrieval-Augmented Large Language Models: A Review. *Mathematics* 2025, Vol. 13, Page 856, 13(5), 856. <https://doi.org/10.3390/MATH13050856>