Jurnal Studi Ilmu-Ilmu al-Qur'an dan Hadis – ISSN: 1411-6855 (p); 2548-4737 (e)
Vol. 27, No. 2 (Juli 2026), hlm. 458-493, doi: 10.14421/qh.v27i1.7499
https://ejournal.uin-suka.ac.id/ushuluddin/qurdis/index
Article History: Submitted:05-12-25 Revised: 16-03-26 Accepted: 18-05-26

Between Machines and the Tafsīr Tradition: Reorienting *Tafsīr bi al-Ma'ṡūr* in the Age of Computational Interpretation

Antara Mesin dan Tradisi Tafsir: Reorientasi Tafsīr bi al-Ma'ṡūr di Era Penafsiran Berbantuan Komputasi

Saifullah[(a)]*, Soleh Hasan Wahid[(a)(b)], Muhammad Ihza Fazrian[(b)]
* Corresponding author: saifullah.masduki@gmail.com
(a) Universitas Islam Negeri Kiai Ageng Muhammad Besari Ponorogo, Jl. Pramuka No. 156, Ronowijayan, Siman, Ponorogo, East Java 63471, Indonesia.
(b) Universitas Islam Internasional Indonesia, Jl. Raya Bogor Km 33.5, Cisalak, Sukmajaya, Depok, West Java 16416, Indonesia.

Abstract

This study examines authority and attribution in Qur'anic interpretation mediated by artificial intelligence. The phrase *fī sabīlillāh* in Q 9:60 serves as a test case because its meaning has been disputed since the earliest period, its transmission is layered, and its implications extend to zakat distribution. The same interpretive question was submitted to seven mainstream language models and assessed through *tafsīr bi al-ma'ṡūr*, understood as an evaluative hermeneutic centered on traceable attribution, *isnād*, *tarjīḥ*, and fidelity to sources. Interpretations by al-Ṭabarī, al-Qurṭubī, and Ibn Kaṡīr provided the basis for an analytic rubric. Two raters independently evaluated 28 model responses and reconciled their judgments through consensus, producing 248 audited error findings. Although the models generally identified the three exegetes and their works, they performed poorly in preserving *isnād*, narration structure, citation accuracy, and attribution. Unsupported additions and shifts of meaning dominated the errors. Higher-performing models produced more polished answers but did not eliminate source-level failures. The findings indicate that greater model capacity alone cannot secure credible machine-mediated *tafsir*; verification must restore the attributional discipline of *tafsīr bi al-ma'ṡūr*.

Keywords: *tafsīr bi al-ma'ṡūr*; interpretive authority; source attribution; *fī sabīlillāh*; language models

Abstrak

Penelitian ini mengkaji otoritas dan atribusi dalam penafsiran al-Qur'an yang dimediasi kecerdasan buatan. Frasa *fī sabīlillāh* dalam QS al-Tawbah [9]:60 dijadikan kasus uji karena maknanya diperselisihkan sejak masa awal, transmisinya berlapis, dan implikasinya menjangkau hukum penyaluran zakat. Pertanyaan penafsiran yang sama diajukan kepada tujuh model bahasa arus utama dan dinilai dengan kerangka *tafsīr bi al-ma'ṡūr*, yang dipahami sebagai hermeneutika evaluatif berbasis keterlacakan atribusi, *isnād*, *tarjīḥ*, dan kesetiaan pada sumber. Penafsiran al-Ṭabarī, al-Qurṭubī, dan Ibn Kaṡīr menjadi dasar penyusunan rubrik analitis. Dua penilai secara mandiri mengevaluasi 28 jawaban model, lalu menyelaraskan penilaian melalui konsensus dan menghasilkan 248 temuan kesalahan teraudit. Meskipun seluruh model umumnya mampu menyebut ketiga mufasir dan karya mereka, kinerjanya lemah dalam menjaga *isnād*, struktur riwayat, ketepatan kutipan, dan atribusi. Penambahan tanpa dasar dan pengalihan makna mendominasi temuan. Model berkinerja lebih tinggi menghasilkan jawaban yang lebih rapi, tetapi tidak menghilangkan kesalahan pada tingkat sumber. Kredibilitas *tafsir* berbantuan mesin karena itu tidak dapat dijamin hanya melalui peningkatan kapasitas model; verifikasi perlu mengembalikan disiplin atribusi dalam tradisi *tafsīr bi al-ma'ṡūr*.

Kata Kunci: *tafsīr bi al-ma'ṡūr*; otoritas penafsiran; atribusi sumber; *fī sabīlillāh*; model bahasa

Introduction

AI-generated information can appear authoritative even when its references or attributions are false, and such output can pass through institutions with strict referencing standards. Globethics[1] identifies related risks: overreliance among students and teachers, automation bias among officials who cite AI output without tracing its source, and the marginalization of Eastern knowledge (knowledge decay). The issue is therefore not only whether AI produces information, but whether that information is traceable, attributable, and worthy of trust.

This problem becomes far more pressing when AI output concerns religion, where a person's belief and practice are at stake. Almost every major religion now has its own chatbot, popularly called a godbot, most built without the endorsement of any religious authority,[2] and several have drawn public criticism for reckless responses.[3] Such use is not confined to religious chatbots; users bring grief, marriage, and even questions of conversion to mainstream systems,[4] and in Indonesia, many students at Islamic boarding schools have made ChatGPT a place to ask about religious matters.[5] The 2023 National Conference of NU Religious Scholars affirmed that AI must not be treated as a guide on religious matters because its output is prone to hallucination and bias;[6] this consumption nonetheless continues without adequate controls.

Some developers have applied retrieval-augmented generation (RAG) to reduce hallucination in Islamic question-answering: AraQA,[7] MufassirQAS,[8]

1 Globethics, "GEF2025: Harnessing the power of AI: Safe and ethical integration," Globethics (2025), https://globethics.net/gef2025-harnessing-power-ai-safe-and-ethical-integration.

2 A. J. Fenton and C. Shannahan, "AI Godbots: Religious Leaders Warn of 'Alarming Consequences' When Machines Speak in the Name of God," *BusinessWorld*, June 10, 2026, https://bworldonline.com/opinion/2026/06/10/755584/ai-godbots-religious-leaders-warn-of-alarming-consequences-when-machines-speak-in-the-name-of-god-2/.

3 Fenton and Shannahan, "AI Godbots."

4 R. Contreras, "AI Stumbles on Questions of Faith," *Axios*, June 1, 2026, https://www.axios.com/2026/06/01/ai-religious-bias-catholics-chatbots; S. Khullar, "People Are Turning to AI Chatbots for Religious Guidance," *WIRED Middle East*, February 12, 2026, https://www.wired.me/story/people-are-turning-to-ai-chatbots-for-religious-guidance.

5 "PBNU Diminta Buat Chat GPT Islami," *Republika*, September 21, 2023, https://www.republika.id/posts/45738/pbnu-diminta-buat-chat-gpt-islami.

6 M. R. Balya B., "Hukum Menggunakan Artificial Intelligence untuk Rujukan Agama," *NU Online Jatim*, April 18, 2025, https://jatim.nu.or.id/keislaman/hukum-menggunakan-artificial-intelligence-untuk-rujukan-agama-G3hei.

7 Yomna Adel et al., "AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text," in *2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)* (IEEE, 2023), 235–239, https://doi.org/10.1109/JAC-ECC61002.2023.10479645.

8 Ahmet Yusuf Alan, Önder Aydın, and Enis Karaarslan, "A RAG-based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM," *SSRN Electronic Journal* (2024), https://doi.org/10.2139/ssrn.4707470.

multilingual NativQA,[9] and Aftina[10] all improve over baselines based on BoW–TF-IDF, fuzzy matching, or CNNs trained on limited datasets.[1112] Yet the evaluation of these systems has focused on benchmarks of accuracy, fluency, and usefulness and has not touched the interpretive hierarchy at the core of tafsir — intertextual reasoning across verses, verification of the sanad for ḥadīṡ reports, and tarjīḥ among earlier scholars.[13] Nur et al. and Qamar et al. note that many AI applications in tafsir remain conceptual and have not been tested against the measures of classical epistemology.[1415]

Within Qur'ānic and tafsir studies specifically, recent work shows that the integration of AI has not resolved the problem of source credibility and authorization in interpretation. Ahmad Syahir et al. find that scholarship on AI in Qur'ānic studies remains limited and fragmented, with three dominant themes: digital transformation, accessibility of texts and practices, and ethical-philosophical reflection.[16] In qirā'āt and Qur'ānic learning, Abdelgelil et al. and Zainadun et al. find that AI can support tajwid, memorization, automatic assessment, and monitoring, but cannot replace oral transmission, sanad, spiritual formation, or the teacher's role.[1718]

9 Md Arid Hasan et al., *NativQA: Multilingual Culturally-Aligned Natural Query for LLMs*, arXiv preprint (2025), https://doi.org/10.48550/arXiv.2407.09823.

10 Mohammed Y. Mohammed et al., "Aftina: Enhancing Stability and Preventing Hallucination in AI-based Islamic Fatwa Generation Using LLMs and RAG," *Neural Computing and Applications* (2025), https://doi.org/10.1007/s00521-025-11229-y.

11 F. Qamar, S. Latif, and R. Latif, "A Benchmark Dataset with Larger Context for Non-Factoid Question Answering over Islamic Text," *arXiv* (2024), https://doi.org/10.48550/arXiv.2409.09844.

12 M.T. Sihotang, "Answering Islamic questions with a chatbot using fuzzy string-matching algorithm," *Journal of Physics: Conference Series* 1566, no. 1 (2020): 012007, https://doi.org/10.1088/1742-6596/1566/1/012007.

13 R. N. E. Anggraini, "Islamic QA with chatbot system using convolutional neural network," *Iraqi Journal of Science* 65, no. 4 (2024): 2232–2241, https://doi.org/10.24996/ijs.2024.65.4.38.

14 A. Nur, Mustakim, M. Yasir, and Z. Fatni, "The Comparison of Nearest Neighbor Algorithm as Modeling in Conclusion of Interpretation of Bil Ma'tsur and Bil Ra'yi," in *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)* (IEEE, 2021), 101–5, https://doi.org/10.1109/IC2IE53219.2021.9649246.

15 Qamar, Latif, and Latif, "Benchmark Dataset."

16 A. N. Ahmad Syahir, "Artificial Intelligence and digital transformation in Qur'anic studies: A systematic literature review," *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 542–585, http://mjs.um.edu.my/index.php/quranica/article/view/65005.

17 M. F. M. Abdelgelil, "Challenges of Artificial Intelligence (AI) and its impact on Qur'anic qira'at: A thematic review analysis," *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 74–96, http://mjs.um.edu.my/index.php/quranica/article/view/64938.

18 N. Zainadun, "Exploring the potential of Artificial Intelligence in enhancing Quranic teachers' pedagogy: A systematic literature review," *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 501–541, http://mjs.um.edu.my/index.php/quranica/article/view/65001.

At the level of Qur'ānic chatbots, Raja Yusof et al. show that many AI-based systems lack context and alignment with Islamic sources; therefore, evaluation must attend to factual accuracy, contextual explanation, alignment with sources, courtesy, accessibility, and evidentiary support.[19] Sahimi et al. propose an NLP-based framework to filter, match, and verify tafsir content against authoritative sources with a human-in-the-loop,[20] and Mohamed Nazir et al. ground AI ethics for interaction with the Qur'ān in the disciplines of ḥadīṡ scholarship.[21] These studies show that AI can support accessibility, linguistic analysis, memorization, and digital verification, while remaining constrained by fundamental limits concerning the sanad, the accuracy of tafsir, algorithmic bias, and the validity of sources.

In light of these limitations, this study poses one interpretive question to several mainstream LLM chatbots and evaluates the answers through the framework of *tafsīr bi al-ma'ṡūr*. It requires a single test case representing the complexity of interpretive work: a verse whose meaning is contested, transmission is layered, and implications extend beyond exegesis. QS al-Tawbah [9]:60, particularly the phrase *fī sabīlillāh*, is such a case: a field of *ikhtilāf* since the earliest period — restricted by some to military jihad, extended by others to Hajj, da'wah, and the public good — its meaning bearing directly on the distribution of zakat, on the border between tafsir and fiqh. Answering such a question credibly demands the management of plural meanings, discipline in attribution, and caution — precisely where language models are most vulnerable.

This study examines *tafsīr bi al-ma'ṡūr* itself: how al-Ṭabarī manages and weighs (*tarjīḥ*) the differing narrations on this verse, how Ibn Kaṡīr selects and sifts *ḥadīṡ* and *āṡār*, and why al-Qurṭubī moves further into the elaboration of law. The tradition thus becomes an object of analysis that can show what, how, and why the use of chatbots in tafsir needs to be mitigated: the successes and failures of the machines reveal how the credibility of tafsir information is built, maintained, and examined within the tradition. The study also develops a rubric that renders the epistemic constraints of the tradition into examinable criteria, and maps the

19 R. J. Raja Yusof, "Digital al-Qur'an applications and Artificial Intelligence (AI): An overview," *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 646–676, http://mjs.um.edu.my/index.php/quranica/article/view/65011.

20 M. S. Sahimi, "Preserving the authenticity of Quranic exegesis through Artificial Intelligence: A proposed framework for digital verification," *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 325–344, http://mjs.um.edu.my/index.php/quranica/article/view/64982.

21 A. N. U. Mohamed Nazir, "Challenges in AI-mediated engagement with Quranic verses: An ethical framework based on ḥadīṯ methodology," *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 271–303, http://mjs.um.edu.my/index.php/quranica/article/view/64969.

range of errors through the multidimensional quality metrics framework, so that patterns of failure — from unfounded addition to the obscuring of meaning and faulty attribution — can be read systematically and the evaluation replicated for other verses and problems.

This study asks two questions: (1) Where do the responses of seven mainstream language models to *fī sabīlillāh* in Q 9:60 succeed and fail when assessed against the disciplines of attribution and transmission in *tafsīr bi al-ma'ṡūr*? (2) What do these successes and failures reveal about authority, attribution, and the boundary between *tafsir* and *fiqh* in machine-mediated interpretation?

The responses were evaluated against three reference works: *Jāmi' al-Bayān* by al-Ṭabarī, *Tafsīr al-Qur'ān al-'Aẓīm* by Ibn Kaṡīr, and *al-Jāmi' li Aḥkām al-Qur'ān* by al-Qurṭubī. The assessment examined interpretive accuracy, identification of exegetes, preservation of narrations, attribution of opinions, Arabic terminology, presentation of *ikhtilāf*, and caution in conclusions.

Data comprised responses to the same question from GPT 5.5, Claude Opus 4.8, Usul AI Pro, Grok AI Fast, Gemini 3.5, DeepSeek, and Z.AI GLM. Each model was run four times between June 27 and 30, 2026, producing 28 responses. Two raters independently assessed them on 56 scoring sheets. The models were accessed through consumer chat interfaces at default settings; the researchers did not configure retrieval-augmented generation, although built-in web search could operate where available (Supplementary Figures S1–S2). Figure 1 summarizes the data flow, and Table 1 defines each analytical unit.

The audited corpus consists of the two raters' verified records rather than a merged list. One rater recorded 137 errors and the other 111, yielding 248 audited findings. Cross-checking identified 25 overlapping pairs in the same passages: 13 used the same category and 12 different categories. Two passages received two categories from one rater. These records were retained and cross-flagged because the counted unit was each rater's audited record. Severity-weighted scores were calculated separately for each rater and then averaged, preventing any passage from being counted twice in a single score.

Table 1. Definitions and counts of the analysis units

Stage	Definition	Count
Model answer	7 models × 4 repetitions	28
Assessment sheet	28 answers × 2 raters	56
Indicator unit	28 answers × 8 rubric dimensions	224
Coded finding	error notes recorded by the two raters	137 + 111

Stage	Definition	Count
Audited error finding	per-rater records after source verification and consensus (137 + 111)	248

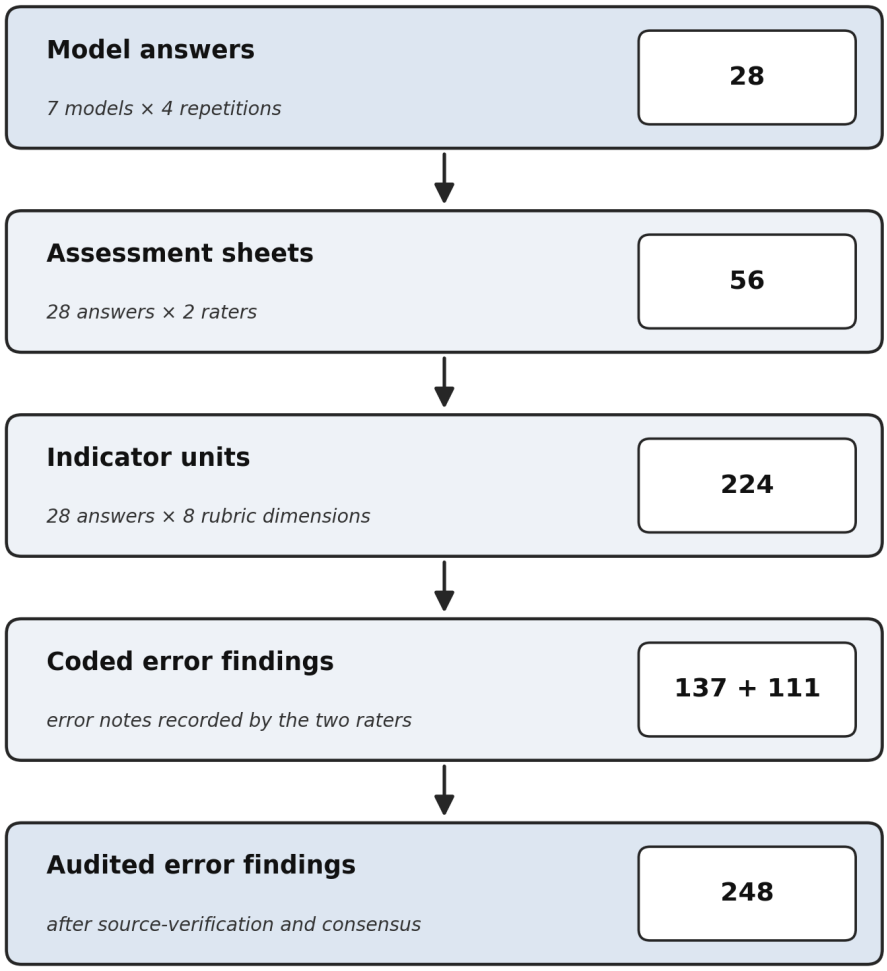

Figure 1. Data flow from model answers to audited findings

To ask about the meaning of *fī sabīlillāh*, we used a chain-of-thought approach to frame the question for the models. The chain-of-thought technique requires the model to reason step-by-step before arriving at a conclusion so that the answer sets out its stated reasoning in explicit steps.[22] What this yields is the rationale displayed in the answer, not a view into the model's internal process; it is this rationale that is checked against the sources. The prompt used is provided in Supplementary Appendix 5.

22 T. Kojima et al., "Large Language Models Are Zero-Shot Reasoners," *Advances in Neural Information Processing Systems* 35 (2023): 22199–213, https://doi.org/10.48550/arXiv.2205.11916; J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," *arXiv* (2023), https://doi.org/10.48550/arXiv.2201.11903.

Responses were segmented into claim units (CUs): statements conveying *tafsir* information, including an exegete's view, the name of a Companion or *tābiʿī*, a narration, an Arabic term, an interpretive extension, or a claim of agreement or disagreement. Each claim, its primary source, assessment status, and verification note was recorded in a claim-audit table (Supplementary Appendix 6).

The three works served as the *ground truth*. Every claim was matched against the relevant source passage so that narrations, opinions, figures, and terms could be traced to each exegete, including the principal view and expansions involving *ikhtilāf*.

Answer quality was assessed using an analytic rubric weighing each element separately: coverage of the three exegetes, accuracy of the core meaning, presence of narration, naming of figures, fit of opinion attribution, precision of Arabic terminology, and caution in conclusions; analytic rubrics of this kind clarify the criteria, the levels of attainment, and the reasons for scores, and their validity rests on the fit between criteria and aim.[23][24] The error categories draw on the study of hallucination in artificial-intelligence systems, which maps fictitious, mistaken, or unverified output into typed classes;[25] this study renders them specific to tafsir — source attribution, the attribution of authority, and the boundary of interpretive genre. The full 0–2 rubric is given in Supplementary Appendix 1. The rubric examines the traceability of attribution, the naming of sources, and the fit of an answer with the three reference works; it does not verify the full quality of each sanad, nor does it reconstruct the exegetes' complete tarjīḥ.

In addition to the rubric score, the study records the types of errors in the models' answers, drawing on work on hallucination, faithfulness, factuality, and source attribution in natural language generation:[26] fabricated citations, added names of figures, misattributed opinions, contextual errors, hasty consensus claims, interpretive expansion beyond the source, and translation distortion (Supplementary Appendix

23 S. M. Brookhart, "Appropriate Criteria: Key to Effective Rubrics," *Frontiers in Education* 3 (2018): 22, https://doi.org/10.3389/feduc.2018.00022; A. Jonsson and G. Svingby, "The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences," *Educational Research Review* 2, no. 2 (2007): 130–44, https://doi.org/10.1016/j.edurev.2007.05.002.

24 B. M. Moskal, "Scoring rubric development: Validity and reliability," *Practical Assessment, Research, and Evaluation* 7, no. 10 (2000): 1–6, https://doi.org/10.7275/q7rm-gg74.

25 Y. Sun, "AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content," Humanities and Social Sciences Communications, 11, 1278 (2024), https://doi.org/10.1057/s41599-024-03811-x.

26 Z. Ji et al., "Survey of Hallucination in Natural Language Generation," *ACM Computing Surveys* 55, no. 12 (2023): 1–38, https://doi.org/10.1145/3571730; J. Maynez et al., "On Faithfulness and Factuality in Abstractive Summarization," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (Association for Computational Linguistics, 2020), 1906–19, https://doi.org/10.18653/v1/2020.acl-main.173; H. Rashkin et al., "Measuring Attribution in Natural Language Generation Models," *Computational Linguistics* 49, no. 4 (2023): 777–840, https://doi.org/10.1162/coli_a_00486.

2). For Arabic terms and translations, the examination follows the accuracy, addition, omission, and mistranslation categories of the Multidimensional Quality Metrics (MQM).[27][28]

The results were summarized into two classifications: the quality score, derived from the analytic rubric, and the severity-weighted error score, computed from the accumulation of errors together with their level of severity — higher values indicate a lighter error burden — with the formulas and severity levels in Supplementary Appendix 3. Every assessment decision carries a brief note recording the reason for the score, the location of the problem, and the link to the primary source, preserving traceability and consistency of category application.[29]

The severity-weighted error score applied MQM's penalty logic. Starting from 100, it deducted severity-weighted penalties using a multiplier of 5 for major errors; verification status scaled each penalty (confirmed 1.0, doubtful 0.5, needs recheck 0.25). Rankings were unchanged under three alternative severity weights (Spearman's $\rho = 1.00$) and stable across raters ($\rho = 0.90$). Because the score partly increases with answer length, it was checked against the length-independent rubric score ($\rho = 0.88$, $p = 0.008$) and interpreted as error volume rather than per-claim accuracy.

The rubric assessment was carried out by two raters independently and then reconciled through a consensus meeting, the standard procedure in double-annotated qualitative coding: first, measure how far the raters agree; then, deliberate each difference to a single agreed final decision.[30]

Independent assessments covered 224 units (28 responses × 8 dimensions). Both raters scored 223 units; one was completed during consensus. Exact agreement occurred on 94 units (42.2%). Of the 129 differences, 90.7% were one point apart, producing agreement within one point on 211 of 223 units (94.6%). Because the scale contained only three levels (0–2), this figure indicates proximity rather than strict agreement. Full distributions appear in Supplementary Tables S1–S2.

Cohen's kappa was 0.065 unweighted, 0.136 with linear weighting, and 0.227

27 The MQM Council, *MQM: The Multidimensional Quality Metrics Framework*, specification (2023), https://themqm.org/.

28 The MQM Council, *MQM*.

29 C. O'Connor and H. Joffe, "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines," *International Journal of Qualitative Methods* 19 (2020): 1–13, https://doi.org/10.1177/1609406919899220.

30 C. O'Connor and H. Joffe, "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines," *International Journal of Qualitative Methods* 19 (2020): 1–13, https://doi.org/10.1177/1609406919899220; J. L. Campbell et al., "Coding In-Depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement," *Sociological Methods & Research* 42, no. 3 (2013): 294–320, https://doi.org/10.1177/0049124113500475.

with quadratic weighting.[31] This manifests the kappa paradox: raw agreement can be high while kappa stays low when a skewed distribution of scores depresses the chance-correction value.[32] Here most dimensions clustered at a single value, so kappa fell even though within-one agreement exceeded nine of every ten units while exact agreement was 42.2%; the percentage of direct agreement is therefore the more representative reliability measure in this context, and the remaining differences were resolved through consensus. Per-dimension kappa values are given in Supplementary Appendix 4.

At a consensus meeting on July 1, 2026, the raters reviewed all 129 differing units against the source texts. Seventy-nine were resolved in favor of one rater, 39 in favor of the other, and 11 received a new value. All 224 units then had a single consensus score, which was used in the performance analysis.

The error types were not tested using the kappa method because the units of finding were open-ended and not parallel across raters: the two raters produced 137 and 111 audited records, and the audit corpus is their simple sum of 248, as set out above — overlapping records are cross-flagged rather than merged, and a claim unit meeting more than one category is counted once per category and reported as multiply coded. Each record was verified against the source text and reviewed at consensus; a fabricated reference remained a finding regardless of the rater who recorded it. The error notes are thus treated as qualitative audit data, in line with human-evaluation practice within MQM, which computes reliability at the level of scores.[33] MQM was used here as a secondary layer for recording and harmonizing error types; the primary taxonomy was derived from the epistemic problems of Tafsir and source attribution.

Tafsīr bi al-Ma'šūr, Interpretive Authority, and the *Tajdīd–Aṣālah* Framework

The Epistemology of Tafsīr bi al-Ma'šūr as a Discipline of Validity

Within *'ulūm al-Qur'ān*, *tafsīr bi al-ma'šūr* denotes interpretation based on

31 J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement* 20, no. 1 (1960): 37–46, https://doi.org/10.1177/001316446002000104.

32 Alvan R. Feinstein and Domenic V. Cicchetti, "High Agreement but Low Kappa: I. The Problems of Two Paradoxes," *Journal of Clinical Epidemiology* 43, no. 6 (1990): 543–549, https://doi.org/10.1016/0895-4356(90)90158-L.

33 A. Lommel, "Multidimensional Quality Metrics (MQM): A framework for declaring and describing translation quality metrics," Tradumàtica, 12, 455–463 (2014), https://doi.org/10.5565/rev/tradumatica.77.; M. Freitag, "Experts, errors, and context: A large-scale study of human evaluation for machine translation," Transactions of the Association for Computational Linguistics, 9, 1460–1474 (2021), https://doi.org/10.1162/tacl_a_00437.

transmitted sources: the Qur'ān interpreted by the Qur'ān, the Prophet's Sunnah, the opinions of the Companions, and those of the *tābi'īn*. It is therefore an epistemic discipline governing how interpretive knowledge is acquired, weighed, and held accountable.[34]

An interpretation using the *tafsīr bi al-ma'ṡūr* method must follow a hierarchy of clear and accountable sources: the Qur'ān explained by the Qur'ān itself, the Sunnah of the Prophet as elucidator of revelation, the opinions of the Companions who witnessed the context of revelation, and the tābi'īn who inherited the Companions' explanations.[35]

This method demands two marks of validity: the traceability of narration — the clarity of the chain linking an interpretation to its source — and the accuracy of attribution — the careful linking of every opinion to its owner. These two aspects distinguish transmitted interpretation from interpretation by free reasoning (*tafsīr bi al-ra'y*), and demand caution lest a ma'ṡūr interpretation rest a meaning on a source to which it does not belong.

Authority in *tafsīr bi al-ma'ṡūr* is therefore relational: it depends on the connectedness of a statement to its source. Al-Suyūṭī in al-Durr al-Manṡūr, for example, includes the sanad for the narrations he cites.[36] A meaning is regarded as valid when it is linked to a traceable narration, attributed to the right authority, and weighed through clear rules — the principles this study takes as the basis for assessing the output of language models.

Tajdīd and Aṣālah as an Islamic Scholarly Framework

What the exegetes did implicitly through attribution and the weighing of narration points to a settled scholarly tradition that ties the validity of knowledge to its connection with the source. Under present conditions, this practice tends to be uprooted: studies note a digital-era shift in the isnād system, from validity grounded in connectedness to a teacher toward authority resting on technology and popularity,[37] and studies of state-produced abridged tafsirs show that simplified presentation carries its own ideological orientation.[38] The uprooting of a tradition

34 Mannā' Khalīl al-Qaṭṭān, *Mabāḥiṡ fī 'Ulūm al-Qur'ān* (Cairo: Maktabah Wahbah, 2007); Muḥammad Ḥusain al-Żahabī, *Al-Tafsīr wa'l-Mufassirūn* (Cairo: Dār al-Ḥadīṡ, 2002).

35 Al-Qaṭṭān, *Mabāḥiṡ fī 'Ulūm al-Qur'ān*; al-Żahabī, *Al-Tafsīr wa'l-Mufassirūn*.

36 J. al-Dīn al-Suyūṭī, *Al-Durr al-Manṡūr fī al-Tafsīr bi al-Ma'ṡūr* (Dār al-Fikr, 1993).

37 A. Majit, "Pembelajaran al-Qur'an secara digital: Pergeseran sistem isnad dan peneguhan otoritas baru," *Jurnal SMART (Studi Masyarakat, Religi, dan Tradisi)* 9, no. 1 (2023): 133–146, https://doi.org/10.18784/smart.v9i1.1795.

38 Aksin Wijaya, Ahmad Zainal Abidin, and Muh. Syaifudin, "State-Sponsored Qur'ānic Exegesis and

from its foundations is a long-recognized problem in contemporary Islamic thought,[39] and out of this concern two concepts become relevant here — *tajdīd* and *aṣālah* — a framework for how the validity of an interpretive tradition is maintained and, when it loosens, restored.

Tajdīd denotes renewal that returns a tradition to the Qur'ān and Sunnah while addressing new conditions. In Islamic thought, it is an internal revitalization that rejects distortion and accretion. It does not repeat inherited forms mechanically; it renews understanding by restoring the tradition's connection to its foundations.[40]

Aṣālah, or authenticity, complements *tajdīd* by affirming the basis on which a tradition counts as authentic: fidelity to the source, as opposed to an understanding uprooted from its foundations.[41] The *isnād* paradigm is its clearest example — the authority of narration derives from its connection to the original source through a chain of trustworthy transmitters.[42] This measure accords with the traceability and accuracy of attribution at the core of *tafsīr bi al-ma'ṡūr*; *tajdīd* and *aṣālah* thus frame this study's reading of the successes and failures of machine-mediated interpretation.

Construction and Dynamics in the *Tafsīr bi al-Ma'ṡūr* Works Regarding the Phrase *Fī sabīlillāh*

This section compares how the three exegetes construct *fī sabīlillāh* in Q 9:60 through verse interpretation, *aṡar*, *ḥadīṡ*, and juristic reasoning. It begins with al-Ṭabarī (d. 310/923), continues with al-Qurṭubī (d. 671/1273), and concludes with Ibn Kaṡīr (d. 774/1373).

Their works represent different stages of the tradition. Al-Ṭabarī wrote *Jāmi' al-Bayān* near the end of the great period of *ḥadīṡ* compilation; al-Qurṭubī wrote *al-Jāmi' li Aḥkām al-Qur'ān* after the legal schools had become firmly established; and Ibn Kaṡīr wrote *Tafsīr al-Qur'ān al-'Aẓīm* in a mature culture of *ḥadīṡ* criticism. The analysis therefore considers both the narrations transmitted and the juristic

Interreligious Relations: A Comparative Study of Egypt, Saudi Arabia, and Indonesia," *Ascarya: Journal of Islamic Science, Culture, and Social Studies* 5, no. 2 (2025): 280–292, https://doi.org/10.53754/iscs.v5i2.854.

39 Ḥasan Ḥanafī, *al-Turāṡ wa al-Tajdīd: Mawqifunā min al-Turāṡ al-Qadīm* (Cairo: al-Markaz al-'Arabī li al-Baḥth wa al-Nashr, 1980); Muḥammad 'Ābid al-Jābirī, *al-Turāṡ wa al-Ḥadāṡa: Dirāsāt wa Munāqasyāt* (Beirut: Markaz Dirāsāt al-Waḥda al-'Arabiyya, 1991).

40 Mohammad Hashim Kamali, "Tajdīd, Iṣlāḥ and Civilisational Renewal in Islam," *Islam and Civilisational Renewal* 4, no. 4 (2013): 484–511; John O. Voll, "Renewal and Reform in Islamic History: Tajdid and Islah," in *Voices of Resurgent Islam*, ed. John L. Esposito (New York: Oxford University Press, 1983), 32–47.

41 Muḥammad 'Ābid al-Jābirī, *al-Turāṡ wa al-Ḥadāṡa: Dirāsāt wa Munāqasyāt* (Beirut: Markaz Dirāsāt al-Waḥda al-'Arabiyya, 1991).

42 W. A. Graham, "Traditionalism in Islam: An essay in interpretation," Journal of Interdisciplinary History 23, no. 3 (1993): 495–522, https://doi.org/10.2307/206100.

reasoning built upon them.

Al-Ṭabarī

Abū Ja'far Muḥammad b. Jarīr al-Ṭabarī was born in Āmul, Ṭabaristān, around 224–225/839, spent most of his life in Baghdad, and died there in 310/923. Trained in Syāfi'ī and Mālikī *fiqh*, he followed neither school exclusively and developed the short-lived Jarīriyya. Writing near the end of the great period of *ḥadīṡ* compilation, he worked as a *musnid* who grounded reports in chains of transmission. *Jāmi' al-Bayān* accordingly gathers narrations before weighing them, often through the agreement of *ahl al-ta'wīl*.[43]

Regarding the interpretation of *fī sabīlillāh*, al-Ṭabarī in *Jāmi' al-Bayān* constructs the phrase *fī sabīlillāh* in a narration-based interpretive mode. He explicitly "restricts" the meaning of *fī sabīlillāh* by centering the interpretation of the phrase on a specific act, namely military defense and warfare against the enemy. Al-Ṭabarī affirms this boundary of meaning through his statement:

> *"As for His words, [Wa fī sabīlillāh], it means: and in expenditure for the defense of the religion of Allāh, His path, and His law that He has prescribed for His servants, by fighting His enemies, and that is warfare against the disbelievers"*[44]

He maintains this interpretive position strictly by resting it on two narrations. The first comes from the authority of the *tābi'īn* figure Ibn Zayd, who construes "*sabīlillāh*" as the soldier (*al-gāzī*).[45] The second is reinforced through the validation of a *marfū' ḥadīṡ* by way of 'Aṭā' b. Yasār and Abū Sa'īd al-Khudrī, which affirms the exception to the permissibility of zakat for a wealthy person engaged in warfare, as recorded in the wording:

> *"Zakat is not lawful for a wealthy person except for . . . one who fights in the path of Allāh"*[46]

This pattern of citation shows that, for al-Ṭabarī, the boundary between tafsir and law is set by the validity of narration, not by practical juristic consideration. He performs tarjīḥ on the basis of the consensus he calls "ahl al-ta'wīl", preserving the verse's functional originality as an instrument for funding the fighting community: zakat serves either domestic need (sadd khallat al-muslimīn) or the strengthening

43 Franz Rosenthal, trans., *The History of al-Ṭabarī, Vol. 1: General Introduction and from the Creation to the Flood* (Albany: State University of New York Press, 1989).

44 I. J. al-Ṭabarī, *Tafsīr al-Ṭabarī: Jāmi' al-Bayān (M. M. Syākir, Ed.; Vol. 14)* (Maktaba Ibn Taymiyya, 1994), 319.

45 Al-Ṭabarī, *Tafsīr al-Ṭabarī: Jāmi' al-Bayān*, 319.

46 Al-Ṭabarī, *Tafsīr al-Ṭabarī: Jāmi' al-Bayān*, 319–320.

of the religion's sovereignty (ma'ūnat al-Islām wa taqwiyatuhu), and fī sabīlillāh belongs to the latter — so that, on the basis of narration, a soldier's economic status is irrelevant to receiving this right.[47]

The criterion the narrations set is therefore functional: the share is due to the one who goes out *fī sabīlillāh* because he fights himself, so the *al-gāzī* of this reading is the fighter and the portion underwrites his going out — a role the narrations define, not to be read through the modern category of the salaried soldier, which al-Ṭabarī's sources do not have in view.

Al-Qurṭubī

Abū 'Abdillāh Muḥammad b. Aḥmad al-Anṣārī al-Qurṭubī was born in Córdoba at the beginning of the seventh century Hijri. He migrated eastward after the city fell to Ferdinand III of Castile in 633/1236 and settled in Munyat Banī Khaṣīb on the Nile, where he died in 671/1273.[48] Of the three exegetes, only he was raised entirely by the legal culture of the age of the schools: the Mālikī school had long been the sole school in al-Andalus, and his standing as a jurist shows in how he treats narration — rarely writing out the sanad in full, more often abridging it from compilations, a narration for him being an evidentiary proof to be put to work in cross-school legal discussion. Calder judges that al-Qurṭubī still stands within the tradition pioneered by al-Ṭabarī, but in a more systematic manner.[49]

Al-Qurṭubī in this regard represents an interpretation that leans closer to a genre between tafsir and law. In *al-Jāmi' li Aḥkām al-Qur'ān* al-Qurṭubī develops the exegesis of the phrase *fī sabīlillāh* into a complex and dynamic discourse of case-by-case juristic questions, which he sets out and arranges as reported positions rather than adopting any one of them as the meaning of the verse. Methodologically, al-Qurṭubī presents his critical analysis of the concept of ownership of zakat through the question of the grammatical function of the letter *lām*. He cites the debate:

> *"An account of the categories of zakat and its vessels; this is the view of Mālik and Abū Ḥanīfa together with their followers, as is said: 'a saddle for a horse and a door for a house.' As for al-Syāfi'ī, he said: 'The letter lām functions as the lām al-tamlīk (of ownership), like your saying: This wealth belongs to Zayd, 'Amr, and Bakr'"*[50]

47 Al-Ṭabarī, *Tafsīr al-Ṭabarī: Jāmi' al-Bayān*, 316.

48 R. Arnaldez, "al-Ḳurṭubī," in *Encyclopaedia of Islam*, 2nd ed., ed. P. Bearman et al., vol. 5 (Leiden: Brill, 1986), https://doi.org/10.1163/1573-3912_islam_SIM_4553.

49 N. Calder, "Tafsīr from Ṭabarī to Ibn Kašīr: Problems in the Description of a Genre, Illustrated with Reference to the Story of Abraham," in *Approaches to the Qur'ān*, ed. G. R. Hawting and A.-K. A. Shareef (London: Routledge, 1993), 101–40.

50 M. b. A. al-Qurṭubī, *Al-Jāmi' li Aḥkām al-Qur'ān (A. al-Bardūnī & I. Aṭfaysy, Eds.; Vol. 8)* (Dār al-Kutub

The difference is rooted in legal principle: al-Syāfi'ī holds to individual ownership rights (*tamlīk*) to protect the right of the poor from the arbitrariness of the ruler, while al-Qurṭubī, following his own Mālikī school and the Ḥanafī, views zakat as an instrument of public policy that demands flexibility — for the Mālikīs the letter *lām* marks a vessel of collective allocation (*tabyīn*) that allows the ruler to act for the public good.

This implies that, for the Mālikī school, the eight categories of zakat recipients were treated as an effort to gather funds. This gave the ruler the legitimacy to allocate zakat funds flexibly, and even to prioritize them toward the military domain without the burden of equal distribution.[51]

The genre between tafsir and law is most evident when al-Qurṭubī explores the micro-variances of practical fiqh within *fī sabīlillāh*. He records an internal shift within the Mālikī school through the views of 'Īsā b. Dīnār and Ibn al-Qāsim, who reconstruct the relevance of the soldier's economic status — an aspect al-Ṭabarī deemed irrelevant — by requiring a wealthy soldier to borrow (*yastaqriḍ*) or limiting his right without access to his wealth on the battlefield.[52] He further records a strong semantic expansion from the military toward the social-religious in a dialogue he narrates from Ibn 'Umar, where the criterion of military defense is delegitimized owing to the army's corrupted integrity and *sabīlillāh* is extended to funding pilgrims as the delegation of God (*wafd al-Raḥmān*):[53] Through this ordering, al-Qurṭubī brings exegesis and law into close contact while presenting juristic material as distinct questions rather than merging it into the meaning of the verse.

Ibn Kaṡīr

'Imād al-Dīn Ismā'īl b. 'Umar b. Kaṡīr was born around 700/1300 near Buṣrā and spent his scholarly life in Damascus, where he died in 774/1373. Trained in Syāfi'ī *fiqh*, he studied *ḥadīṡ* with al-Mizzī and Ibn Taymiyya. By his period, the major *ḥadīṡ* collections and the discipline of *'ulūm al-ḥadīṡ* had been systematized, so scholars assessed inherited compilations rather than gathering reports anew. Mirza describes his *tafsir* as a work of *takhrīj* on al-Ṭabarī and Ibn Abī Ḥātim, reweighing earlier material through the *ḥadīṡ* criticism of his time.[54]

---

al-Miṣriyya, 1964), 167.

51 Al-Qurṭubī, *Al-Jāmi' li Aḥkām al-Qur'ān*, 181.

52 Al-Qurṭubī, *Al-Jāmi' li Aḥkām al-Qur'ān*, 186.

53 Al-Qurṭubī, *Al-Jāmi' li Aḥkām al-Qur'ān*, 185.

54 Y. Y. Mirza, "Was Ibn Kaṡīr the 'spokesperson' for Ibn Taymiyya? Jonah as a prophet of obedience," *Journal of Qur'ānic Studies* 16, no. 1 (2014): 1–19, https://doi.org/10.3366/jqs.2014.0130.

Ibn Kaṡīr in *Tafsīr al-Qur'ān al-'Aẓīm* interprets the phrase *fī sabīlillāh* in a more doctrinal and evaluative manner. Unlike al-Ṭabarī, who stresses more the selection (tarjīḥ) among the opinions of *tābi'īn* authorities and *marfū' ḥadīṡ*, Ibn Kaṡīr opens with a macro-structure of restriction (*ḥaṣr*) through the *ḥadīṡ* of Ziyād b. al-Ḥāriṡ al-Ṣudā'ī in order to affirm that the right of zakat distribution is absolutely *tawqīfī*, directly from God:

> *"Indeed, Allāh was not content with the ruling of a prophet or of anyone else concerning charity until He Himself laid down its ruling"*[55]

Yet he maintains the militaristic meaning while imposing a strict administrative limit on the recipients:

> *"As for fī sabīlillāh: among them are the soldiers who have no entitlement in the pay register (dīwān)"*[56]

According to this criterion, the share is reserved for the fighters who hold no entitlement in the pay register (*dīwān*): the volunteer soldiers who do not receive a regular budget subsidy from the state treasury (*bayt al-māl*).

Ibn Kaṡīr nonetheless practices scholarly acknowledgment and attribution when recording the early dynamics of Islamic law: he notes a loosening of the boundary from the purely military dimension toward non-military physical worship — the allocation of zakat for the Hajj — attributing it to Imām Aḥmad, al-Ḥasan, and Isḥāq (*wa al-ḥajj min sabīl-Allāh*).[57]

Ibn Kaṡīr limits his interpretation textually through *ḥadīṡ*, while delegating the operational complexity of the law to the corpus of practical fiqh works (*wa maḥall tafṣīl hāżā fī kutub al-furū'*, *wa-Allāhu a'lam*).[58]

Tafsīr bi al-Ma'ṡūr as a Framework for Epistemic Integrity

Taken together, the three readings show that *tafsīr bi al-ma'ṡūr* preserves epistemic integrity through two linked controls: how an exegete defines the meaning of *fī sabīlillāh* and how that meaning is attributed and justified. Table 2 compares their interpretations; Table 3 compares their methods.

55 'Imād al-Dīn Ibn Kaṡīr, *Tafsīr al-Qur'ān al-'Aẓīm*, ed. M. Ḥusain S. al-Dīn, vol. 4 (Beirut: Dār al-Kutub al-'Ilmiyya, 1998), 144.

56 Ibn Kaṡīr, *Tafsīr al-Qur'ān al-'Aẓīm*, 148.

57 Ibn Kaṡīr, *Tafsīr al-Qur'ān al-'Aẓīm*, 148.

58 Ibn Kaṡīr, *Tafsīr al-Qur'ān al-'Aẓīm*, 147.

Table 2. The Construction of the Meaning of fī sabīlillāh According to the Three Exegetes

Aspect of Meaning	Al-Ṭabarī (d. 310/923)	Al-Qurṭubī (d. 671/1273)	Ibn Kaśīr (d. 774/1373)
Core meaning	Military defense and warfare against the enemy (*ghazw al-kuffār*)	Records views extending it to pilgrims and charitable ends	Soldiers who fight in the path of God
Scope of meaning	Strictly limited to the soldier (*al-gāzī*); rejects expansion	Expanded to pilgrims (*wafd al-Raḥmān*) and public policy	Military, with a window for expansion to the Hajj attributed to Aḥmad, al-Ḥasan, and Isḥāq
Recipient criterion	Economic status is irrelevant; participation is decisive	Depends on the school; a wealthy soldier may be required to borrow (*yastaqriḍ*)	Soldiers outside the state pay register (*lā ḥaqq lahu fī al-dīwān*)

Table 2 shows a spectrum of meaning. Al-Ṭabarī restricts *fī sabīlillāh* to fighters and treats participation, rather than economic status, as decisive. Al-Qurṭubī records legal views that extend the category to pilgrims and wider charitable purposes while retaining the debate over military restriction. Ibn Kaśīr preserves the military meaning, limits recipients through the *dīwān* criterion, and records an attributed extension to *ḥajj*.

Table 3. The Attribution and Methodological Limits of the Three Classical Exegetes

Methodological Aspect	Al-Ṭabarī (d. 310/923)	Al-Qurṭubī (d. 671/1273)	Ibn Kaśīr (d. 774/1373)
Mode of interpretation	Narration-based (*musnid*): gather the narrations, then weigh them	*Tafsīr aḥkām*: case-by-case juristic interpolation	Doctrinal-evaluative: *takhrīj* and isnād criticism
Basis of attribution	The *aṡar* of Ibn Zayd and marfū‘ *ḥadīṡ* (‘Aṭā’, Abū Sa‘īd)	Debate over the function of the *lām*; views of ‘Īsā b. Dīnār, Ibn al-Qāsim, Ibn ‘Umar	The *ḥaṣr ḥadīṡ* of Ziyād al-Ṣudā‘ī; the right to zakat is *tawqīfī*
Mode of weighing	Tarjīḥ through the agreement of *ahl al-ta’wīl*	Cross-school ijtihād (Mālikī–Ḥanafī–Syāfi‘ī)	Ḥadīṡ selection by the standards of isnād criticism
The tafsir–fiqh boundary	Set by the validity of narration, not by juristic consideration	Deliberately made fluid through debate among the schools	Legal detail delegated to the *kutub al-furū‘*

Table 3 shows that these meanings rest on different modes of attribution. Al-Ṭabarī weighs *aṡar* and *marfū‘ ḥadīṡ* through the agreement of *ahl al-ta’wīl*; al-Qurṭubī presents cross-school debate, especially over the function of the letter *lām*;

and Ibn Kaṡīr applies the *ḥadīṡ* criticism of his period while assigning each view to its source. In each case, interpretive expansion remains accountable to narration, attribution, and genre. These principles form the criteria used to assess the language models.

Accuracy and Quality of Language-Model Answers Within the *Tafsīr bi al-Ma'ṡūr* Framework

The 28 responses and 56 scoring sheets produced 248 audited error findings and consensus rubric scores. Two measures were used. The severity-weighted error score captures the volume and seriousness of errors, with higher scores indicating a lighter burden. The quality score measures completeness and conformity to the *tafsīr bi al-ma'ṡūr* rubric. Table 4 reports both measures.

Table 4. Performance of the Seven Models on the Severity-Weighted Error Score and the Quality Score

Model	Number of Findings	Severity-Weighted Error Score	Quality Score
GPT 5.5	20	78.1	83.6
Claude Opus 4.8	26	72.5	79.7
Usul AI Pro	29	56.9	60.2
Grok AI Fast	30	55.6	45.3
Gemini 3.5	43	38.8	63.3
DeepSeek	48	22.5	43.0
Z.AI GLM	52	21.1	42.4

Table 4 places the models in three groups. GPT 5.5 and Claude Opus 4.8 lead on both measures, with severity-weighted scores of 78.1 and 72.5 and quality scores of 83.6 and 79.7. Usul AI Pro and Grok AI Fast form the middle group, while Gemini 3.5, DeepSeek, and Z.AI GLM have severity-weighted scores below 40. Error counts rise as scores fall: GPT 5.5 has 20 findings, compared with 48 for DeepSeek and 52 for Z.AI GLM.

The two measures do not always move in parallel. DeepSeek and Z.AI GLM score similarly on both, whereas Gemini 3.5 combines a relatively high quality score (63.3) with a low severity-weighted score (38.8). Its responses are well structured but frequently inaccurate, confirming the need to read both measures together.

The error taxonomy comprises three groups: source and attribution accuracy, terminology and translation, and interpretive framing and conceptual validity. Table 5 presents their distribution.

Table 5. Number of Errors by the Three Typological Groups

Error Group	Coverage	Number of Findings	Percentage
P1	Source accuracy and attribution	126	50.8%
P3	Interpretive framing and conceptual validity	108	43.5%
P2	Terminological accuracy and translation	14	5.6%
Total		**248**	**100%**

Source accuracy and attribution (P1) accounts for 126 findings (50.8%), interpretive framing (P3) for 108 (43.5%), and terminological accuracy (P2) for 14 (5.6%). Most errors therefore concern the placement of sources, authority, and concepts. Table 6 gives the category-level distribution.

Table 6. Number of Errors by Typological Category

Category	Group	MQM Type	Count	Percentage
Conceptual misframing	P3	Mistranslation	73	29.4%
Fabrication of citations and sources	P1	Addition	58	23.4%
Hallucination of authority or entity	P1	Addition	25	10.1%
Spurious consensus claim	P1	Addition	22	8.9%
Misattribution of opinion	P1	Mistranslation	21	8.5%
Interpretive expansion beyond the source	P3	Overtranslation	19	7.7%
Translation distortion	P2	Mistranslation	14	5.6%
Omission of key distinctions	P3	Omission	9	3.6%
Anachronistic expansion	P3	Addition	5	2.0%
Methodological reduction	P3	Undertranslation	2	0.8%
Total			**248**	**100%**

Conceptual misframing is the largest category (73 findings; 29.4%), followed by fabricated citations and sources (58; 23.4%). Hallucinated authorities (25), spurious consensus claims (22), and misattributed opinions (21) likewise concern the location of epistemic authority. Together, these categories capture the principal failures: concepts are reframed, sources invented, authorities added, *ikhtilāf* flattened into consensus, and opinions transferred among exegetes. Supplementary Table S3 presents the MQM distribution.

Under MQM, addition is the largest error type (110 findings; 44.4%), followed by mistranslation (108; 43.5%). The former adds sources, authorities, narrations, or conclusions absent from the reference works; the latter shifts meaning, framing, or attribution. Only nine omissions were recorded. The dominant problem is therefore an excess of unsupported information rather than insufficient detail.

The consensus scores were also averaged by rubric dimension across the 28 responses, producing one mean on the 0–2 scale for each dimension (Supplementary Table S6). Unlike the model-level score in Table 4, this measure summarizes performance by criterion.

Coverage of the three exegetes is highest at 1.79 of 2 (89.5%). Opinion attribution is lowest at 0.93, while citation accuracy and preservation of *riwāyah* structure each score 1.11. Thus, models reproduce the form of a *tafsir* answer more successfully than its source accountability. The four core dimensions—a credible *sanad*, narration structure, attribution, and citation—remain below 60%, and no model receives full scores across all four.

The Range of Source-Attribution Errors in Language-Model Answers

This section illustrates the five largest categories in Table 6: conceptual misframing, fabricated citations, hallucinated authority, spurious consensus, and misattribution. Each example was verified against the source text. Supplementary Table S4 provides the excerpt, source check, and implication.

The errors arise chiefly from the assembly of evidence. One model transfers Ibn Kaṡīr's *dīwān* restriction to al-Ṭabarī; another attributes to Ibn ʻAbbās a *hajj* narration that Ibn Kaṡīr assigns to Aḥmad, al-Ḥasan, and Isḥāq; others reduce divergent transmission paths to unsupported consensus. Each claim appears plausible until checked against the primary source.

Supplementary Table S5 gives one verified example from each model's most frequent error category, including both highest-scoring systems, with the relevant answer excerpt and source-check result.

Source-attribution errors occur across the performance range. GPT 5.5 cites al-Qurṭubī at volume 8, pages 111–112, although the passage appears around page 185 in the reference edition. Claude Opus 4.8 introduces Ibrāhīm al-Nakhaʻī, who is absent from al-Ṭabarī's account. Lower-scoring systems make more direct errors: Z.AI GLM removes Ibn Kaṡīr's *dīwān* context, Gemini 3.5 adds figures outside the corpus, and Usul AI Pro assigns al-Ṭabarī's quotation to al-Qurṭubī. All produce plausible statements that fail source verification.

Reading the Failures: Interpretive Authority and the Reorientation of Tafsīr bi al-Ma'ṡūr

Within this design—one Qur'ānic phrase, three commentaries, seven models, and four repetitions—the responses were not reliable references for *fī sabīlillāh*. Of 248 audited findings, unsupported additions account for 44.4% and shifts of meaning or attribution for 43.5%, together nearly nine in ten. The models typically generated excess information: added Companions, invented chains, and unsupported claims of agreement.

Although coverage of the three exegetes reaches 89.5%, the four accuracy dimensions central to *tafsīr bi al-ma'ṡūr*—a credible *sanad*, preservation of narration structure, accurate attribution, and citation—remain below 60%. No model achieves full scores across all four. A response can therefore resemble a scholarly *tafsir* while containing material absent from the sources.

Higher performance changes the form, not the presence, of error. GPT 5.5 and Claude Opus 4.8 lead the rankings but still misstate a page reference and introduce an unattested transmitter. Lower-scoring models more visibly invent figures or collapse disagreement. The polished structure of high-performing answers makes their errors especially difficult for non-specialists to detect.

Model capacity alone therefore cannot solve the credibility problem. Even the strongest systems fail on source traceability. Tafsīr bi al-ma'ṡūr offers an underused checking framework through its disciplines of *isnād*, *tarjīḥ*, and caution in attribution.

A Typology of LLM Chatbot Response Hallucination Within the *Tafsīr bi al-Ma'ṡūr* Framework

The most frequent type is unsupported addition: a model supplies a transmitter, chain of transmission, quotation, bibliographic source, or claim of agreement absent from the reference work. Within *tafsīr bi al-ma'ṡūr*, this includes fabricated citations, added authorities, baseless consensus claims, and anachronistic meanings.

The second type is mistranslation in the broader MQM sense: a model alters a concept, reframes a passage, or assigns an opinion to the wrong exegete. The source may be present, but its meaning or ownership changes while the answer remains superficially plausible.

Interpretive expansion, or overtranslation, occurs when an answer extends an exegete's relevant statement beyond its documented scope, explanation, or implication. A limited meaning is generalized without supporting narration.

Omission removes a distinction, condition, or qualification present in the source, most often an *ikhtilāf* that determines the direction of interpretation. The resulting account is incomplete precisely where the exegete was careful.

Undertranslation, the least frequent type, narrows the source and flattens the exegete's weighing of narrations or construction of argument into an unduly simple account.

Reorienting the *Tafsīr bi al-Ma'šūr* Tradition and Reclaiming Authenticity in the AI Era

These findings raise a question beyond model engineering: how *tafsīr bi al-ma'šūr* should be understood and used in an era of machine-mediated interpretation.

To answer this, we must reconsider what *tafsīr bi al-ma'šūr* really is: interpretation resting on transmitted sources — the Qur'ān by the Qur'ān, the Sunnah, the opinions of the Companions and the *tābi'īn* — together with the disciplines behind transmission: tracing the isnād, assessing a narration's reliability, weighing among narrations (*tarjīḥ*), and linking each opinion accurately to its owner.[59][60]

Tafsīr bi al-ma'šūr is therefore not a collection of detached quotations. It is a scholarly tradition that links the Qur'ān, *ḥadīš*, earlier authorities, *tafsir* works, and exegetes through accountable transmission and attribution.

The problem is that this tradition, along with the development of media, has been oversimplified. Online tafsir sites, long before chatbots, changed how the community accessed interpretation: concise, searchable translations without the mediation of a teacher — a democratization of the religious source.[61] In such a form, the tafsir text is treated as a dead text: an object whose content can be taken without regard to how it was produced, weighed, and attributed within its tradition.

Language models enter a digital environment in which this simplification is already established. When they rely on summaries that omit transmission and attribution, their answers reproduce and amplify those omissions.

Tracing the models' workings shows that, when asked to explain a verse, the model does not consult a tafsir work with a *sanad* but draws on online tafsir sites — quran.ksu.edu.sa, shamela.ws, ketabonline.com, surahquran.com, quran-tafsir.

59 Al-Qaṭṭān, *Mabāḥiš fī 'Ulūm al-Qur'ān*; al-Żahabī, *Al-Tafsīr wa'l-Mufassirūn*.

60 Al-Suyūṭī, *Al-Durr al-Manšūr fī al-Tafsīr bi al-Ma'šūr*.

61 F. Lukman, "Digital hermeneutics and a new face of the Qur'an commentary: The Qur'an in Indonesian's Facebook," *Al-Jāmi'ah: Journal of Islamic Studies* 56, no. 1 (2018): 95–120, https://doi.org/10.14421/ajis.2018.561.95-120.

net, and even Wikipedia (Supplementary Figures S1–S2).[62] Such sites present tafsir as translation, and the models' output inherits the trait: Gemini 3.5 concludes that the three exegetes "agree absolutely (*ijmāʿ*)" — a generalization common in popular summaries but absent from the three works — while Grok AI Fast and DeepSeek reduce the exegetes to lists of concise stylistic traits. This kind of characterization is the mark of secondary tafsir, not of a direct reading of a text with a *sanad*.

The result is concrete attribution error: GPT 5.5 links a view to al-Qurṭubī with an incorrect volume-page location, while Claude Opus 4.8 inserts a transmitter absent from al-Ṭabarī (Supplementary Table S5). Invented chains, transferred opinions, and flattened disagreement treat *tafsir* as content detached from its scholarly production.

The study therefore proposes a reorientation of *tafsīr bi al-maʾṡūr* through *tajdīd*: returning the method to its fuller character as a living relationship among the Qur'ān, *ḥadīṡ*, transmitted authorities, *tafsir* works, and exegetes. Such a relationship requires knowledge of an opinion's origin, evaluation, and attribution. It must be restored in teaching, public presentation, and technological use.[63]

This reorientation reestablishes *aṣālah* — the authenticity of interpretation — in the era of machine mediation: not authenticity as free self-expression,[64] but fidelity to the source through a sound chain of transmission.[6566] Validity in tafsīr bi al-maʾṡūr is measured by how far an interpretation can be traced back to its source — the measure that can ground the assessment of model output, as researchers now pursue through source verification and the involvement of human raters.[67]

Reviving *tafsīr bi al-maʾṡūr* must therefore extend beyond technology. Its disciplines should inform *tafsir* education, public presentation, and the evaluation of digital religious sources. Machine systems may assist access and comparison, but their output must remain subordinate to traceable sources and qualified human verification.

62 Lukman, "Digital Hermeneutics"; Raja Yusof et al., "Digital al-Qur'an Applications."

63 Kamali, "Tajdīd, Iṣlāḥ and Civilisational Renewal in Islam."

64 Charles Taylor, *A Secular Age* (Cambridge, MA: Belknap Press of Harvard University Press, 2007).

65 Al-Jābirī, *Al-Turāṡ wa al-Ḥadāṡa*.

66 Graham, "Traditionalism in Islam."

67 Sahimi et al., "Preserving the Authenticity of Quranic Exegesis"; Raja Yusof et al., "Digital al-Qur'an Applications."

Conclusion

This study assessed seven language models on *fī sabīlillāh* in Q 9:60 using *tafsīr bi al-ma'šūr* as the standard. Within this limited design, none could serve as a reliable reference. The main weakness was not incomplete coverage but weak accountability to sources. Models named the exegetes and works yet failed to preserve *sanad*, narration structure, attribution, and quotation. Unsupported additions and shifts of meaning accounted for nearly nine in ten findings. Higher-performing systems produced more convincing forms of the same source-level errors; increased capacity alone did not secure credible interpretation.

The framework nonetheless proved useful as a checking apparatus. By translating *isnād*, *tarjīḥ*, and attributional fidelity into assessable criteria, the rubric detected failures missed by surface-level measures. Independent assessment followed by source-based consensus produced findings that remained traceable to the primary works. Interpretation still requires substantive human judgment, but the procedure offers a replicable test of whether a *tafsir* claim rests on its source.

The study is a proof of concept with clear limits. It examines one phrase in one verse, so the results cannot be generalized to narrative, theological, or linguistic passages that organize transmission differently. Ground truth is restricted to three commentaries, and model outputs may change with system updates. Future work should test the framework across additional verses, traditions, languages, and model versions.

The central contribution is a source-based method for evaluating machine-mediated interpretation. Tafsīr bi al-ma'šūr enables model output to be judged by fidelity and traceability rather than rhetorical plausibility, bringing a classical discipline of transmission into current AI evaluation.

Supplementary Materials

Supplementary Tables S1–S6, Supplementary Figures S1–S2, and Supplementary Appendices 1–6 are provided in *Supplementary_Material*. The record-level audit dataset, including cross-flagged overlaps and multi-category records, is provided in *Supplementary_Data.xlsx*.

Acknowledgments

The authors gratefully acknowledge the colleagues who provided constructive feedback during the research and writing of this article.

Author Contributions

Conceptualization, Saifullah; methodology, Saifullah; validation, all authors; formal analysis, Saifullah, Soleh Hasan Wahid, and Muhammad Ihza Fazrian; investigation, Saifullah; data curation, Saifullah; writing—original draft preparation, Saifullah; writing—review and editing, all authors; project administration, Saifullah. Soleh Hasan Wahid and Muhammad Ihza Fazrian independently assessed the model responses and participated in consensus reconciliation. All authors read and approved the final manuscript.

Data Availability Statement

The data supporting this study, including the audit dataset of model responses checked against al-Ṭabarī, Ibn Kaṡīr, and al-Qurṭubī, are openly available at https://soleh-hukumline.github.io/audit-tafsir/. The record-level dataset with cross-flagged overlaps is also provided in Supplementary_Data.xlsx.

Conflicts of Interest

The authors declare no conflict of interest.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

References

Abdelgelil, M. F. M., B. S. Tahir, and S. N. B. Syed Bidin. "Challenges of Artificial Intelligence (AI) and Its Impact on Qur'anic Qira'at: A Thematic Review Analysis." *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 74–96. http://mjs.um.edu.my/index.php/quranica/article/view/64938.

Adel, Y., M. Dorrah, A. Ashraf, A. ElSaadany, M. Mohamed, M. Wael, and G. Khoriba. "AraQA: An Arabic Generative Question-Answering Model for Authentic Religious Text." In *2023 11th International Japan-Africa Conference on Electronics, Communications, and Computations (JAC-ECC)*, 235–39. IEEE, 2023. https://doi.org/10.1109/JAC-ECC61002.2023.10479645.

Ahmad Syahir, A. N., M. S. Zainal Abidin, C. Z. Sa'ari, and M. Z. Abdul Rahman. "Artificial Intelligence and Digital Transformation in Qur'anic Studies: A Systematic Literature Review." *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 542–85. http://mjs.um.edu.my/index.php/quranica/article/view/65005.

Al-Jābirī, Muḥammad ʻĀbid. *al-Turāṡ wa al-ḥadāṯa: Dirāsāt wa munāqasyāt*. Beirut: Markaz Dirāsāt

al-Waḥda al-'Arabiyya, 1991.

Al-Qaṭṭān, Mannā' Khalīl. *Mabāḥiṯ fī 'Ulūm al-Qur'ān*. Cairo: Maktabah Wahbah, 2007.

Al-Qurṭubī, Muḥammad b. Aḥmad. *Al-Jāmi' li Aḥkām al-Qur'ān*. Edited by A. al-Bardūnī and I. Aṭfaysy. Vol. 8. Cairo: Dār al-Kutub al-Miṣriyya, 1964.

Al-Suyūṭī, Jalāl al-Dīn. *Al-Durr al-Manśūr fī al-Tafsīr bi al-Ma'ṯūr*. Beirut: Dār al-Fikr, 1993.

Al-Ṭabarī, Ibn Jarīr. *Tafsīr al-Ṭabarī: Jāmi' al-Bayān*. Edited by M. M. Syākir. Vol. 14. Cairo: Maktaba Ibn Taymiyya, 1994.

Al-Żahabī, Muḥammad Ḥusain. *At-Tafsīr wa'l-Mufassirūn*. Cairo: Dār al-Ḥadīṡ, 2002.

Alan, A. Y., Ö. Aydın, and E. Karaarslan. "A RAG-Based Question Answering System Proposal for Understanding Islam: MufassirQAS LLM." *SSRN Electronic Journal*, 2024. https://doi.org/10.2139/ssrn.4707470.

Anggraini, R. N. E., D. Tursina, and R. Sarno. "Islamic QA with Chatbot System Using Convolutional Neural Network." *Iraqi Journal of Science* 65, no. 4 (2024): 2232–41. https://doi.org/10.24996/ijs.2024.65.4.38.

Arnaldez, R. "al-Ḳurṭbī." In *Encyclopaedia of Islam*, 2nd ed., edited by P. Bearman, Th. Bianquis, C. E. Bosworth, E. van Donzel, and W. P. Heinrichs, Vol. 5. Leiden: Brill, 1986. https://doi.org/10.1163/1573-3912_islam_SIM_4553.

Balya B., M. R. "Hukum menggunakan artificial intelligence untuk rujukan agama." NU Online Jatim, April 18, 2025. https://jatim.nu.or.id/keislaman/hukum-menggunakan-artificial-intelligence-untuk-rujukan-agama-G3hei.

Brookhart, S. M. "Appropriate Criteria: Key to Effective Rubrics." *Frontiers in Education* 3 (2018): 22. https://doi.org/10.3389/feduc.2018.00022.

Calder, N. "Tafsīr from Ṭabarī to Ibn Kaśīr: Problems in the Description of a Genre, Illustrated with Reference to the Story of Abraham." In *Approaches to the Qur'ān*, edited by G. R. Hawting and A.-K. A. Shareef, 101–40. London: Routledge, 1993.

Campbell, J. L., C. Quincy, J. Osserman, and O. K. Pedersen. "Coding In-Depth Semistructured Interviews: Problems of Unitization and Intercoder Reliability and Agreement." *Sociological Methods & Research* 42, no. 3 (2013): 294–320. https://doi.org/10.1177/0049124113500475.

Cohen, J. "A Coefficient of Agreement for Nominal Scales." *Educational and Psychological Measurement* 20, no. 1 (1960): 37–46. https://doi.org/10.1177/001316446002000104.

Cohen, J. "Weighted Kappa: Nominal Scale Agreement with Provision for Scaled Disagreement or Partial Credit." *Psychological Bulletin* 70, no. 4 (1968): 213–20. https://doi.org/10.1037/h0026256.

Contreras, R. "AI Stumbles on Questions of Faith." Axios, June 1, 2026. https://www.axios.com/2026/06/01/ai-religious-bias-catholics-chatbots.

Feinstein, Alvan R., and Domenic V. Cicchetti. "High Agreement but Low Kappa: I. The Problems of Two Paradoxes." *Journal of Clinical Epidemiology* 43, no. 6 (1990): 543–549. https://doi.org/10.1016/0895-4356(90)90158-L.

Fenton, A. J., and C. Shannahan. "AI Godbots: Religious Leaders Warn of 'Alarming Consequences' When Machines Speak in the Name of God." BusinessWorld, June 10, 2026. https://

bworldonline.com/opinion/2026/06/10/755584/ai-godbots-religious-leaders-warn-of-alarming-consequences-when-machines-speak-in-the-name-of-god-2/.

Freitag, M., G. Foster, D. Grangier, V. Ratnakar, Q. Tan, and W. Macherey. "Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation." *Transactions of the Association for Computational Linguistics* 9 (2021): 1460–74. https://doi.org/10.1162/tacl_a_00437.

Globethics. "GEF2025: Harnessing the Power of AI: Safe and Ethical Integration." Globethics, 2025. https://globethics.net/gef2025-harnessing-power-ai-safe-and-ethical-integration.

Graham, W. A. "Traditionalism in Islam: An Essay in Interpretation." *Journal of Interdisciplinary History* 23, no. 3 (1993): 495–522. https://doi.org/10.2307/206100.

Ḥanafī, Ḥasan. *al-Turāṡ wa al-tajdīd: Mawqifunā min al-turāṯ al-qadīm*. Cairo: al-Markaz al-'Arabī li al-Baḥṡ wa al-Nasyr, 1980.

Hasan, M. A., M. Hasanain, F. Ahmad, S. R. Laskar, S. Upadhyay, V. N. Sukhadia, M. Kutlu, S. A. Chowdhury, and F. Alam. "NativQA: Multilingual Culturally-Aligned Natural Query for LLMs." arXiv, 2025. https://doi.org/10.48550/arXiv.2407.09823.

Ibn Kaṡīr, 'Imād al-Dīn. *Tafsīr al-Qur'ān al-'Aẓīm*. Edited by M. Ḥusain S. al-Dīn. Vol. 4. Beirut: Dār al-Kutub al-'Ilmiyya, 1998.

Ji, Z., N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung. "Survey of Hallucination in Natural Language Generation." *ACM Computing Surveys* 55, no. 12 (2023): 1–38. https://doi.org/10.1145/3571730.

Jonsson, A., and G. Svingby. "The Use of Scoring Rubrics: Reliability, Validity and Educational Consequences." *Educational Research Review* 2, no. 2 (2007): 130–44. https://doi.org/10.1016/j.edurev.2007.05.002.

Kamali, M. H. "Tajdīd, Iṣlāḥ and Civilisational Renewal in Islam." *Islam and Civilisational Renewal* 4, no. 4 (2013): 484–511.

Khullar, S. "People Are Turning to AI Chatbots for Religious Guidance." WIRED Middle East, February 12, 2026. https://www.wired.me/story/people-are-turning-to-ai-chatbots-for-religious-guidance.

Kojima, T., S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. "Large Language Models Are Zero-Shot Reasoners." *Advances in Neural Information Processing Systems* 35 (2023): 22199–213. https://doi.org/10.48550/arXiv.2205.11916.

Lommel, A., H. Uszkoreit, and A. Burchardt. "Multidimensional Quality Metrics (MQM): A Framework for Declaring and Describing Translation Quality Metrics." *Tradumàtica* 12 (2014): 455–63. https://doi.org/10.5565/rev/tradumatica.77.

Lukman, F. "Digital Hermeneutics and a New Face of the Qur'an Commentary: The Qur'an in Indonesian's Facebook." *Al-Jāmi'ah: Journal of Islamic Studies* 56, no. 1 (2018): 95–120. https://doi.org/10.14421/ajis.2018.561.95-120.

Majit, A., and M. Miski. "Pembelajaran al-Qur'an secara digital: Pergeseran sistem isnad dan peneguhan otoritas baru." *Jurnal SMART (Studi Masyarakat, Religi, dan Tradisi)* 9, no. 1 (2023): 133–46. https://doi.org/10.18784/smart.v9i1.1795.

Maynez, J., S. Narayan, B. Bohnet, and R. McDonald. "On Faithfulness and Factuality in Abstractive Summarization." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 1906–19. Association for Computational Linguistics, 2020. https://doi.org/10.18653/v1/2020.acl-main.173.

Mirza, Y. Y. "Was Ibn Kaśīr the 'Spokesperson' for Ibn Taymiyya? Jonah as a Prophet of Obedience." *Journal of Qur'ānic Studies* 16, no. 1 (2014): 1–19. https://doi.org/10.3366/jqs.2014.0130.

Mohamed Nazir, A. N. U., N. K. Mohd Ithnin, I. Suliaman, and M. T. Amat Famsir @ Amat Tamsir. "Challenges in AI-Mediated Engagement with Quranic Verses: An Ethical Framework Based on Ḥadīš Methodology." *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 271–303. http://mjs.um.edu.my/index.php/quranica/article/view/64969.

Mohammed, M. Y., S. A. Ali, S. K. Ali, A. A. Majeed, and E. H. Mohamed. "Aftina: Enhancing Stability and Preventing Hallucination in AI-Based Islamic Fatwa Generation Using LLMs and RAG." *Neural Computing and Applications*, 2025. https://doi.org/10.1007/s00521-025-11229-y.

Moskal, B. M., and J. A. Leydens. "Scoring Rubric Development: Validity and Reliability." *Practical Assessment, Research, and Evaluation* 7, no. 10 (2000): 1–6. https://doi.org/10.7275/q7rm-gg74.

Nur, A., Mustakim, M. Yasir, and Z. Fatni. "The Comparison of Nearest Neighbor Algorithm as Modeling in Conclusion of Interpretation of Bil Ma'tsur and Bil Ra'yi." In *2021 4th International Conference of Computer and Informatics Engineering (IC2IE)*, 101–5. IEEE, 2021. https://doi.org/10.1109/IC2IE53219.2021.9649246.

O'Connor, C., and H. Joffe. "Intercoder Reliability in Qualitative Research: Debates and Practical Guidelines." *International Journal of Qualitative Methods* 19 (2020): 1–13. https://doi.org/10.1177/1609406919899220.

"PBNU Diminta Buat Chat GPT Islami." Republika, September 21, 2023. https://www.republika.id/posts/45738/pbnu-diminta-buat-chat-gpt-islami.

Qamar, F., S. Latif, and R. Latif. "A Benchmark Dataset with Larger Context for Non-Factoid Question Answering over Islamic Text." arXiv, 2024. https://doi.org/10.48550/arXiv.2409.09844.

Raja Yusof, R. J., M. A. Norasid, M. Abdullah, N. A. A. Nasaruddin, N. Badrol Hisham, A. A. G. Saged, and A. A. Ramchahi. "Digital al-Qur'an Applications and Artificial Intelligence (AI): An Overview." *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 646–76. http://mjs.um.edu.my/index.php/quranica/article/view/65011.

Rashkin, H., V. Nikolaev, M. Lamm, L. Aroyo, M. Collins, D. Das, S. Petrov, G. S. Tomar, I. Turc, and D. Reitter. "Measuring Attribution in Natural Language Generation Models." *Computational Linguistics* 49, no. 4 (2023): 777–840. https://doi.org/10.1162/coli_a_00486.

Rosenthal, F., trans. *The History of al-Ṭabarī, Vol. 1: General Introduction and from the Creation to the Flood*. Albany: State University of New York Press, 1989.

Sahimi, M. S., M. H. Abdul Rahim, A. A. Rushihi, A. Kirin, and N. Zakaria. "Preserving the Authenticity of Quranic Exegesis through Artificial Intelligence: A Proposed Framework for Digital Verification." *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 325–44. http://mjs.um.edu.my/index.php/quranica/article/view/64982.

Sihotang, M. T., I. Jaya, A. Hizriadi, and S. M. Hardi. "Answering Islamic Questions with a Chatbot Using Fuzzy String-Matching Algorithm." *Journal of Physics: Conference Series* 1566, no. 1 (2020): 012007. https://doi.org/10.1088/1742-6596/1566/1/012007.

Sun, Y., D. Sheng, Z. Zhou, and Y. Wu. "AI Hallucination: Towards a Comprehensive Classification of Distorted Information in Artificial Intelligence-Generated Content." *Humanities and Social Sciences Communications* 11 (2024): 1278. https://doi.org/10.1057/s41599-024-03811-x.

Taylor, C. *A Secular Age*. Cambridge, MA: The Belknap Press of Harvard University Press, 2007. https://www.hup.harvard.edu/books/9780674026766.

The MQM Council. *MQM: The Multidimensional Quality Metrics Framework*. Specification. 2023. https://themqm.org/.

Voll, J. O. "Renewal and Reform in Islamic History: Tajdid and Islah." In *Voices of Resurgent Islam*, edited by J. L. Esposito, 32–47. New York: Oxford University Press, 1983.

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models." arXiv, 2023. https://doi.org/10.48550/arXiv.2201.11903.

Wijaya, A., A. Z. Abidin, and M. Syaifudin. "State-Sponsored Qur'ānic Exegesis and Interreligious Relations: A Comparative Study of Egypt, Saudi Arabia, and Indonesia." *Ascarya: Journal of Islamic Science, Culture, and Social Studies* 5, no. 2 (2025): 280–92. https://doi.org/10.53754/iscs.v5i2.854.

Zainadun, N., M. N. Mohamad Salleh, A. I. Jamil, A. A. Rekan, and M. T. Ajmain @ Jima'ain. "Exploring the Potential of Artificial Intelligence in Enhancing Quranic Teachers' Pedagogy: A Systematic Literature Review." *QURANICA – International Journal of Quranic Research* 17, no. 2 (2025): 501–41. http://mjs.um.edu.my/index.php/quranica/article/view/65001.

Appendix

This appendix contains the technical instruments referred to in the Methods section, namely the answer-quality rubric, the error typology mapped to the *Multidimensional Quality Metrics* (MQM), the severity levels together with the scoring formulas, and the claim-audit table format. These instruments were used to preserve the traceability of the assessment of language-model answers explaining the meaning of *fī sabīlillāh* in QS al-Tawbah [9]:60, with three primary tafsir works serving as the *ground truth*, namely *Jāmiʿ al-Bayān* by al-Ṭabarī, *Tafsīr al-Qur'ān al-ʿAẓīm* by Ibn Kaṡīr, and *al-Jāmiʿ li Aḥkām al-Qur'ān* by al-Qurṭubī. The assessment was carried out independently by two raters over 248 findings, and the mapping of each category to an MQM *error type* follows the MQM-Full Master standard (MQM Council, 2024).

Appendix 1. Answer-Quality Rubric on a 0–2 Scale

The following rubric was used to assess the quality of model answers. Each indicator is scored 0, 1, or 2 according to the degree to which the assessed element is fulfilled. The rubric comprises eight indicators, corresponding to the eight scoring items (r1–r8) in the score file.

No.	Indicator	Score 0	Score 1	Score 2
1	Coverage of the three mufassirs	Does not refer to the three principal mufassirs, or merely names them without relevant content.	Refers to some of the mufassirs, or names the three but with an unbalanced treatment.	Discusses al-Ṭabarī, Ibn Kaṡīr, and al-Qurṭubī clearly and relevantly.
2	Accuracy of the core meaning of *fī sabīlillāh*	The core meaning is incorrect or too general, without grounding in the primary sources.	The core meaning is partly correct but still conflated with an expansion whose limits are unexplained.	The core meaning is correct per the sources and distinguished from the expanded meaning that is subject to *khilāf*.
3	Treatment of *khilāf* over expanded meanings	Does not mention any disagreement, or presents the *khilāf* incorrectly.	Mentions the *khilāf* but does not adequately explain the positions and their limits.	Explains the *khilāf* accurately, including the difference between the principal view and the expanded meaning.
4	Presence of *riwāyah* or *āṡār*	Presents no *riwāyah*, or cites a *riwāyah* that cannot be traced.	Presents *riwāyah* in a limited way or without adequate identification of the transmitters.	Presents relevant *riwāyah* or *āṡār* consistent with the primary sources.

No.	Indicator	Score 0	Score 1	Score 2
5	Accuracy in naming figures	Names of Companions, *tābiʿīn*, mufassirs, or scholars are incorrect or added without a source basis.	Naming of figures is partly correct but with remaining gaps or uncertainty.	The figures named are consistent with the sources and placed in the correct context of the view.
6	Correctness of attribution of views	A view is attributed to a figure or mufassir who did not state it.	Attribution is partly correct but lacks precision.	Views are attributed to the figure or mufassir consistent with the primary sources.
7	Precision of Arabic terms	Arabic terms are translated or used incorrectly so that the meaning changes.	Arabic terms are reasonably accurate but not entirely suited to the exegetical context.	Arabic terms are used precisely, suited to the sentence context and to the flow of the mufassir's exposition.
8	Fidelity to primary sources and caution in drawing conclusions	The answer exceeds the sources or contains information not found in them, and its conclusion is too broad or too definitive.	The answer is partly faithful to the sources but still contains poorly controlled expansion and does not preserve the limits of *khilāf*.	The answer is faithful to the primary sources, draws no conclusion beyond the available material, and is composed cautiously within the scope of the sources.

The maximum rubric score is the number of indicators multiplied by two, so that in this rubric the maximum score is 16.

Appendix 2. Error Typology (MQM-Adapted)

The following typology was used to record forms of deviation in model answers. Each category is organized into three pillars and mapped to an official MQM *error type* together with its *Process/Property Identifier* (PID), so that the assessment remains legible within the classical *bil ma'šūr* framework while also being comparable with an established translation-quality standard. The MQM mapping serves as a documentation and interoperability layer and does not alter the pillar-based scoring. A single claim may carry more than one category where the problem touches several elements at once. The *n* column reports the number of findings in this study; categories with *n* = 0 are retained as part of the instrument framework even though they did not occur in the audited case.

Pillar 1 — Source Accuracy & Attribution

Code	Category	MQM equivalent (dimension → error type · PID)	Severity	n
P1-FS	Fabricated Source / Citation	Accuracy → Addition · MQMC_104000	3	58
P1-HE	Hallucinated Authority / Entity	Accuracy → Addition · MQMC_104000	2–3	25
P1-MA	Misattributed View	Accuracy → Mistranslation · MQMC_101000 (sec. Addition · MQMC_104000)	2–3	21
P1-FC	False Consensus Claim	Accuracy → Addition · MQMC_104000	2–3	22
P1-UC	Unsupported Contextual Claim	Accuracy → Addition · MQMC_104000 (sec. Mistranslation · MQMC_101000)	1–3	0

Pillar 2 — Terminology & Textual Fidelity

Code	Category	MQM equivalent (dimension → error type · PID)	Severity	n
P2-TD	Translation Distortion	Accuracy → Mistranslation · MQMC_101000	1–3	14
P2-WT	Wrong Technical Term	Terminology → Wrong term · MQMC_003000	1–2	0
P2-OS	Over-Specification	Accuracy → Overtranslation · MQMC_102000	1–2	0
P2-US	Under-Specification	Accuracy → Undertranslation · MQMC_103000	1–2	0

Code	Category	MQM equivalent (dimension → error type · PID)	Severity	n
P2-TI	Terminological Inconsistency	Terminology → Inconsistent use of terminology · MQMC_002000	1	0

Pillar 3 — Interpretive Framing & Conceptual Validity

Code	Category	MQM equivalent (dimension → *error type* · PID)	Severity	n
P3-CM	Conceptual Misframing	Accuracy → Mistranslation · MQMC_101000 (sec. Undertranslation · MQMC_103000)	2–3	73
P3-IO	Interpretive Overreach	Accuracy → Overtranslation · MQMC_102000 (sec. Addition · MQMC_104000)	1–3	19
P3-AE	Anachronistic Expansion	Accuracy → Addition · MQMC_104000	1–3	5
P3-OKD	Omission of Key Distinction	Accuracy → Omission · MQMC_105000 (sec. Undertranslation · MQMC_103000)	2–3	9
P3-KF	Khilāf Flattening	Accuracy → Undertranslation · MQMC_103000	2	0
P3-MR	Methodological Reduction	Accuracy → Undertranslation · MQMC_103000 (sec. Style · MQMC_300000)	1–2	2

The distribution of findings by MQM *error type* is given below as a cross-pillar summary.

MQM *error type*	PID	Count
Addition	MQMC_104000	110
Mistranslation	MQMC_101000	108
Overtranslation	MQMC_102000	19
Omission	MQMC_105000	9
Undertranslation	MQMC_103000	2
Total		248

Appendix 3. Severity Levels, Floor Enforcement, and Scoring Formulas

Each error is assigned a severity level indicating its impact on the accuracy of the model answer.

Severity	Level	Operational Definition	Impact on the Answer
1	Minor	Imprecise term, incomplete exposition, or a small gap that does not change the main conclusion.	The answer remains intelligible and the main conclusion is preserved.
2	Major	An error that changes an important understanding, such as an inaccurate treatment of *khilāf* or a shifted attribution of a view.	The answer requires substantial correction in a specific part.
3	Critical	Source fabrication, addition of a transmitter's name without basis, a large misattribution of a view, or a conclusion contradicting the primary sources.	The basis of the answer becomes compromised and the main conclusion may be misleading.

Severity floor enforcement

Several categories carry a binding minimum severity, because their impact on the credibility of the tafsir cannot be judged minor. Source fabrication (P1-FS) is fixed at severity 3, while hallucinated entities (P1-HE), misattributed views (P1-MA), false consensus claims (P1-FC), conceptual misframing (P3-CM), and omission of key distinctions (P3-OKD) carry a minimum severity of 2. In the score file, the accuracy index is reported in two forms, namely without floor enforcement (`indeks_akurasi`) and with floor enforcement (`indeks_akurasi_ifclamp`). The second form raises any finding initially scored below its category floor back to that floor value, as a robustness check on the assessment. Floor enforcement lowers some index values but does not change the ranking order among the models.

Scoring formulas

Maximum Score = Number of Indicators × 2 = 16

Quality Score = (Σ Indicator Scores / Maximum Score) × 100

Total Severity = Σ Error Severities

Accuracy Index = 100 – (5 × Total Severity)

The quality score reflects the answer's attainment against the analytic rubric, whereas the accuracy index reflects the deduction resulting from errors found in the claim audit. The weight of five is used as an operational convention so that each error carries a fixed numerical consequence, and all pillars are weighted equally (1.0). The MQM mapping does not enter the formulas; it functions solely as a documentation and standard-interoperability layer.

Appendix 4. Initial Inter-Rater Agreement per Dimension

Appendix 4 presents the initial inter-rater agreement values for each rubric dimension, before reconciliation through consensus. These values complement the account given in the Methods section.

Table Appendix 4. Initial inter-rater agreement per rubric dimension (before consensus)

Dimension	Same (%)	≤1 point (%)	κ	κw linear	κw quadratic
Coverage of the three exegetes	57.1	100.0	0.000	0.000	0.000
Accuracy of the core meaning	17.9	89.3	-0.027	0.032	0.106
Presentation of expansion-related khilāf	10.7	75.0	-0.077	0.032	0.129
Presentation of sanad/athar	46.4	96.4	0.139	0.135	0.128
Preservation of riwāyah structure	42.9	100.0	0.115	0.145	0.200
Accuracy of opinion attribution	39.3	100.0	0.085	0.131	0.212
Citation accuracy	48.1	100.0	0.000	0.000	0.000
Genre discipline	75.0	96.4	0.492	0.434	0.327
Combined	42.2	94.6	0.065	0.136	0.227

Note: the kappa values measure initial agreement before reconciliation; all differences were resolved through consensus (see the Methods section).

Appendix 5. Prompt Used for Data Collection

The prompt below was administered to the models in Indonesian; an English translation is provided here. Each of the seven models received this identical prompt.

Include the date of this response at the very top of the answer.

Explain the meaning of the phrase *fī sabīlillāh* in QS al-Tawbah [9]:60 according to the method of *tafsīr bi al-ma'ŝūr*, specifically on the basis of three tafsir works: *Jāmiʿ al-Bayān* (al-Ṭabarī), *Tafsīr al-Qur'ān al-'Aẓīm* (Ibn Kaŝīr), and *al-Jāmiʿ li Aḥkām al-Qur'ān* (al-Qurṭubī).

Set out step by step:

1. The context of the verse and the position of *fī sabīlillāh* among the eight categories of zakat recipients.

2. For each exegete separately: how he interprets this phrase, what narration or *athar* he cites (naming the transmitter among the Companions or *tābiʿīn* if any), and his conclusion about who is included in *fī sabīlillāh*.

3. A comparison of the three tafsir works: where they agree and where they differ.

4. Provide precise references to the works (name of the work, volume, and page) for each statement.

At the end, include a complete list of all the references you used in this answer—the name of the work, author, volume, and page—arranged as a bibliography.

Cantumkan tanggal respons ini di bagian paling atas jawaban.

Jelaskan makna frasa fī sabīlillāh dalam QS al-Tawbah [9]: 60 menurut metode tafsīr bi al-maʾthūr, khususnya berdasarkan tiga tafsir: Jāmiʿ al-Bayān (al-Ṭabarī), Tafsīr al-Qurʾān al-ʿAẓīm (Ibn Kathīr), dan al-Jāmiʿ li-Aḥkām al-Qurʾān (al-Qurṭubī).

Uraikan langkah demi langkah:

1. Konteks ayat dan kedudukan fī sabīlillāh di antara delapan golongan penerima zakat.

2. Untuk setiap mufasir secara terpisah: bagaimana ia menafsirkan frasa ini, riwayat atau athar apa yang ia kutip (sebutkan periwayat dari kalangan sahabat atau tabi'in bila ada), dan kesimpulannya tentang siapa yang termasuk fī sabīlillāh.

3. Perbandingan ketiga tafsir: di mana mereka sepakat dan di mana berbeda.

4. Cantumkan rujukan kitab secara presisi (nama kitab, jilid, dan halaman) untuk setiap pernyataan.

Di bagian akhir, sertakan daftar lengkap semua referensi yang Anda gunakan dalam jawaban ini—nama kitab, penulis, jilid, dan halaman—disusun sebagai daftar pustaka.

Appendix 6. Claim-Audit Table Format

The claim audit is recorded in a table with the following column structure. Each row contains one model claim together with the results of its verification against the primary sources, the typology code, the MQM mapping, the severity level, and the rater's notes.

Column	Content
No.	Sequential number of the finding.
Run ID	Identifier of the model answer session (covering the model and the repetition index).
Model	Name of the language model tested.
Coder	Rater who audited the finding.
Code	Error-typology code (P1–P3; see Appendix 2).
Category	Name of the error category corresponding to the code.
Pillar	The assessment pillar in which the category sits.
MQM (*Error Type* · PID)	Mapping to the MQM *error type* together with its PID.
Severity	Severity level of the finding (1–3).
Severity (*ifclamp*)	Severity level after category-floor enforcement.
Status	Result of verifying the claim against the sources (see the key below).
Quotation	Excerpt of the model claim under examination.
Source Check	Result of the search in the primary tafsir texts.
Notes	Rationale for the assessment and the location of the problem in the claim.

Key to assessment status

Status	Description
Confirmed	The claim is consistent with the primary sources.
Partially correct	The claim has elements consistent with the sources but remains incomplete or imprecise.
Not confirmed	The claim is not found in the primary sources or contradicts them.