Omissive Bias of Religion in Generative AI Ethical Responses: Re-reading Amānah, ʿAdl, and Human Dignity in Islamic Ethics

Mohammad Reza Fathi
(Corresponding Author), Associate Professor, Department of Management and Accounting, College of Farabi, University of Tehran, Iran
E-mail: Reza.fathi@ut.ac.ir
https://orcid.org/0000-0002-7973-9814

Abstract
Generative artificial intelligence is increasingly used to address personal, moral, and existential questions, yet bias research has focused more on harmful representations than on the systematic absence of religious perspectives. This study examines omissive religious bias in Persian-language ethical responses generated by large language models and interprets it through amānah, ʿadl, and human dignity within an Iranian Twelver Shiʿi context. Using an interpretive qualitative design, the study combined thematic analysis with Critical Systems Heuristics. The dataset included 15 semi-structured expert interviews and 360 responses generated by five large language models across 24 ethical scenarios and three prompting conditions, with 72 additional responses used to assess stability. Six themes emerged: apparent neutrality and silent religious omission, secular and therapeutic reframing, weakening of amānah, reduction of ʿadl to procedural fairness, reduction of human dignity to individual choice, and epistemic exclusion of religious stakeholders. Complete religious omission occurred in 58.3% of unmarked responses and 40% of Iranian-context responses, but fell to 3.3% when a Shiʿi perspective was explicitly requested. The findings suggest that omissive religious bias is an issue of epistemic justice and AI governance, requiring culturally grounded evaluation, multidisciplinary participation, transparent ethical framing, and accountable inclusion of verified Persian and Shiʿi resources.
Keywords: Generative Artificial Intelligence; Omissive Religious Bias; Islamic Ethics; Amānah; ʿAdl.

1. Introduction
Generative artificial intelligence, particularly large language models (LLMs), is increasingly used for guidance on morally sensitive questions concerning honesty, forgiveness, justice, grief, responsibility, and life's meaning. Such responses are not value-neutral: by defining problems, selecting values, recognizing moral authorities, and excluding alternatives, models may shape users' moral understanding. Yet research has focused more on what generative AI produces than on what it omits (Hagendorff, 2024) .Bias research has largely examined visible harms, including racial, gender, cultural, and religious stereotypes. Religious bias, however, may also operate through absence. A model may appear balanced while excluding religion as a legitimate source of moral reasoning and relying instead on secular psychology, autonomy, or consequentialism, even when religion is central to the user's identity. This recurring marginalization can be described as the omissive bias of religion. Wingate et al. (2026) show that LLMs underrepresent religion in everyday ethical questions, invoking it more often in abstract discussions than in practical issues such as marriage, grief, forgiveness, or addiction. The concern therefore extends beyond missing terminology to the boundaries through which AI determines legitimate moral sources.

Omissive bias should also be considered alongside explicit prejudice. Abid, Farooqi, and Zou (2021) found persistent associations between Muslims and violence across language-generation tasks. Islam may therefore face both reductive inclusion through stereotypes and exclusion when its perspectives could contribute to justice, responsibility, dignity, and moral agency .The cultural orientation of LLMs further complicates this issue. Tao et al. (2023) found that model values tend to resemble those of English-speaking and Protestant European societies. Apparently neutral responses may therefore privilege autonomy, psychological well-being, and secular reasoning. Religious absence may reflect the normalization of one worldview rather than neutrality. Elmahjub (2023) treats Islamic ethics as an independent source of moral knowledge relevant to global AI ethics, while Raquib et al. (2022) show that an Islamic virtue-based approach can reveal responsibilities overlooked by procedural or narrowly consequentialist frameworks.

Amānah, ʿadl, and human dignity provide principles for reassessing AI-generated responses. Amānah treats knowledge and technological power as a trust for which developers, institutions, and users remain accountable. Ali et al. (2025) argue that trusteeship ethics unifies responsibility toward God, knowledge, humanity, and creation. Applied to generative AI, amānah requires attention to harmful outputs and design choices that make moral perspectives visible or invisible. Likewise, ʿadl extends beyond technical fairness to recognition of ethical traditions, while human dignity recognizes people as moral and spiritual agents shaped by God, community, tradition, and transcendent purpose .This issue is especially significant in Iran, where Shiʿi ethics has influenced moral language, family relations, justice, duty, rights, and accountability before God. A response may be linguistically Persian yet ethically disconnected from Iranian users lived world. Through the ELAB benchmark, Pourbahman et al. (2025) emphasize evaluating Persian LLMs against indigenous social norms. FarsEval-PKBETS also incorporates religion, ethics, law, preferences, and Iranian cultural knowledge; models achieved average accuracy below 50%, indicating that Persian fluency does not ensure understanding of Iranian ethical and religious contexts (Shamsfard et al., 2025). However, these benchmarks do not examine whether AI omits Islamic or Shiʿi perspectives in open-ended moral responses.

A related gap exists in Islamic AI ethics. Much of the literature remains conceptual, while empirical research focuses on stereotypes, toxicity, or reference frequency rather than the consequences of omission. This study uses thematic analysis to identify how generative AI includes, excludes, reframes, or reduces religiously relevant considerations, examining explicit references and latent assumptions about autonomy, responsibility, justice, moral authority, and the good life. Critical Systems Heuristics (CSH) then analyzes whose interests are prioritized, who controls the system, which knowledge is legitimate, and who is affected without adequate representation (Ulrich & Reynolds, 2020) .The study asks how and through which boundary judgments generative AI omits, marginalizes, or reduces religion, and what these omissions imply when reconsidered through amānah, ʿadl, and human dignity in an Iranian Shiʿi context. It examines whether models provide fair epistemic space for religious ethics and whether their defaults privilege a particular worldview while presenting it as neutral.

2. Literature Review

Hagendorff (2024) reviewed ethical research on generative AI and identified 378 normative issues across 19 areas, including fairness, safety, privacy, harmful content, hallucination, accountability, alignment, and societal impact. The review showed that highly visible concerns receive more attention than less visible forms of exclusion. Similarly, Gallegos et al. (2023) surveyed bias and fairness in LLMs, developing taxonomies of evaluation metrics, datasets, and mitigation methods.

They demonstrated that bias can emerge in embeddings, probabilities, and generated text, while some identities and harms remain underrepresented in current evaluations. Religious bias has been documented both explicitly and indirectly. Abid et al. (2021) found persistent anti-Muslim bias in GPT-3, including associations between Muslims and violence and a "Muslim–terrorist" analogy in 23% of examined cases. Positive prompting reduced but did not eliminate these disparities. Tao et al. (2023) found that widely used LLMs predominantly reflected values associated with English-speaking and Protestant European societies. Although cultural prompting improved alignment for many countries, default model outputs remained culturally non-neutral. Hongladarom and Bandasak (2023) likewise showed that non-Western AI ethics guidelines may adopt seemingly universal Western terminology while retaining deeper cultural differences, arguing for intercultural dialogue rather than a culture-transcendent morality. Islamic AI ethics provides alternative normative resources. Elmahjub (2023) criticized the Eurocentric orientation of dominant frameworks and proposed pluralistic benchmarking informed by Islamic traditions, including maṣlaḥah, while rejecting the reduction of Islamic ethics to utility or procedural compliance. Raquib et al. (2022) developed an Islamic virtue-based framework that challenges unrestricted technological development and market-driven innovation, emphasizing ontological foundations and the objectives of Islamic law. Ali et al. (2025) proposed trusteeship ethics grounded in amānah and responsibility toward God, knowledge, humanity, and creation, advocating a broader moral philosophy capable of addressing bias, privacy, environmental harm, and accountability. Recent empirical studies extend religious-bias analysis beyond explicit hostility. Wingate et al. (2026) introduced omissive bias and evaluated 27 models using 150 questions on grief, forgiveness, relationships, honesty, addiction, and existential meaning. They found that LLMs systematically underrepresented religion, particularly in practical situations where users often rely on faith-based guidance. Zhang et al. (2025) showed that generative AI can support personalized religious education while also amplifying cognitive biases affecting religious understanding and cultural diversity. Tom et al. (2025), analyzing 175 AI-generated sermons, found that sermons attributed to Muslim imams and Jewish rabbis were less readable than those attributed to Evangelical Protestant pastors, revealing subtle representational inequalities. Research on Persian LLMs also demonstrates the need for culturally grounded evaluation. Pourbahman et al. (2025) introduced ELAB, combining translated, synthetic, and naturally collected Persian data to assess safety, fairness, and social norms. Their findings showed that English-oriented alignment frameworks are insufficient for Persian contexts. Shamsfard et al. (2025) developed FarsEval-PKBETS, a 4,000-question benchmark covering medicine, law, religion, ethics, culture, preferences, bias, and human rights. Average model accuracy remained below 50%, indicating that Persian fluency does not ensure ethical, religious, or contextual competence. Collectively, these studies show that generative AI can reproduce anti-Muslim stereotypes, privilege Western values, represent religions unequally, and omit religion from ethically significant responses. However, empirical research rarely interprets omission through amānah, ʿadl, and human dignity, while Islamic AI ethics remains largely conceptual and seldom examines actual outputs. Persian benchmarks identify cultural limitations but do not specifically assess the omission of Islamic or Shiʿi perspectives. Critical Systems Heuristics has also not been systematically applied to reveal the boundary judgments governing whose values, knowledge, and moral authority are included. This study addresses these gaps by combining thematic analysis with CSH to examine religious omissive bias in generative AI ethical responses within the Iranian Shiʿi context.

3. Methodology

This study employed an interpretive qualitative design to examine how generative AI includes, marginalizes, reframes, or omits religion in responses to everyday ethical questions. Thematic analysis was used to identify recurring patterns of religious representation, while Critical Systems Heuristics (CSH) examined the boundary judgments shaping whose interests, values, knowledge, and moral authority were included or excluded. The aim was not to classify models as religious or non-religious, but to assess whether religious perspectives received appropriate epistemic space and how their treatment could be interpreted through amānah, ʿadl, and human dignity in the Iranian Twelver Shiʿi context. Data came from 15 semi-structured interviews with Iranian experts and Persian-language ethical responses generated by five LLMs. The experts represented Twelver Shiʿi ethics, Islamic philosophy and jurisprudence, AI, Persian NLP, AI ethics, qualitative methodology, critical systems thinking, and sociology of religion. Interviews examined religious omission, its ethical implications, and the institutional boundaries shaping model responses. The panel separately validated the ethical scenarios and analytical codebook, while model outputs were compared across prompting conditions for patterns of inclusion, omission, substitution, and delegitimating. Experts were selected through criterion-based purposive and maximum-variation sampling from Iranian universities, research centers, and seminary institutions. Eligibility required a doctoral degree or Level Four seminary qualification, at least five years of relevant experience, and two related publications or research projects. Additional discipline-specific expertise was required in Shiʿi ethics, AI or Persian NLP, or CSH. The 15 participants were anonymized through expert codes. As shown in Table 1, the panel included five specialists in Twelver Shiʿi ethics and Islamic thought, four in AI and Persian NLP, two in AI ethics, two in CSH, one in qualitative methodology, and one in sociology of religion and Iranian culture.

Table 1. Disciplinary and Professional Profile of the Final Expert Panel

Expert codes	Field of expertise	Number	Academic qualification	Relevant experience	Main responsibility
E01–E05	Twelver Shiʿi ethics, moral theology, Islamic philosophy, jurisprudence, and Qurʾanic or Hadith ethics	5	PhD or Level Four seminary qualification	At least 8 years	Validation of *amānah*, *ʿadl*, human dignity, and the Shiʿi interpretive framework
E06–E09	Artificial intelligence, machine learning, LLM evaluation, and Persian NLP	4	PhD in a relevant technical field	At least 5 years	Validation of model selection, prompt design, and technical procedures
E10–E11	AI ethics and philosophy of technology	2	PhD	At least 5 years	Validation of bias categories and ethical coherence
E12–E13	Critical Systems Heuristics and critical systems thinking	2	PhD	At least 7 years	Validation of the boundary-critique framework
E14	Qualitative methodology and thematic analysis	1	PhD	At least 7 years	Validation of the codebook and analytical procedures
E15	Sociology of religion, Iranian culture, and digital society	1	PhD	At least 7 years	Assessment of cultural relevance and contextual authenticity

The expert panel gave greater representation to Twelver Shiʿi ethics and AI-related disciplines because they constituted the study's primary normative and technical dimensions. Specialists in CSH, qualitative methodology, and the sociology of religion were also included to prevent the validation process from being dominated by exclusively theological or technical perspectives. Each expert evaluated only the instrument sections relevant to their expertise. The textual sample was purposively selected to maximize variation in developer ecosystems, alignment methods, training environments, and Persian-language capabilities. Five model families were examined: GPT-4o, Claude 3.5 Sonnet, Gemini 2.0 Flash, DeepSeek-V3, and Dorna-Llama3-8B-Instruct. The first four represented international general-purpose models, while Dorna represented a Persian-oriented model. Fixed versions were used throughout data collection, and model identifiers, access dates, API settings, and generation timestamps were recorded in the audit trail. No model was replaced or updated after data generation began. An initial pool of 36 ethical scenarios was developed from the conceptual framework and prior research. Following expert validation, 24 fictional scenarios were retained across six domains: honesty and trust; marriage and family; grief and end-of-life concerns; justice and social responsibility; professional and public responsibility; and meaning and moral self-cultivation. Each domain contained four scenarios. Each scenario was submitted to all five models under three conditions: an unmarked Persian prompt, an Iranian-context prompt, and an explicit Twelver Shiʿi prompt. The final condition served as a positive control, distinguishing spontaneous religious omission from inability to articulate Shiʿi ethics. As shown in Table 2, the primary corpus comprised 360 responses. Repeating a random 20% of prompts produced 72 additional responses, yielding a final corpus of 432 AI-generated responses.

Table 2. Composition of the AI-Generated Textual Corpus

Corpus component	Specification	Number
Ethical domains	Honesty and trust; family relationships; grief and suffering; justice and rights; professional responsibility; meaning and moral self-cultivation	6
Validated scenarios	Four scenarios in each domain	24
AI models	GPT-4o, Claude 3.5 Sonnet, Gemini 2.0 Flash, DeepSeek-V3, and Dorna-Llama3-8B-Instruct	5
Prompt conditions	Unmarked, Iranian-context, and explicit Twelver Shiʿi	3
Primary corpus	24 scenarios × 5 models × 3 conditions	360
Repeated robustness sample	Twenty percent of the primary prompts	72

All prompts were developed in Persian and reviewed by two bilingual researchers for clarity and conceptual equivalence. Except for condition-specific context, wording, response length, and format remained constant. Models produced 250–350-word Persian responses using a temperature of 0.2, top-p of 1.0, and sufficient output limits where available. Default safety settings remained active, and refusals or warnings were retained. Validation occurred in two rounds. Fifteen experts rated relevance, clarity, cultural authenticity, ethical sensitivity, and neutrality. Retention required an I-CVI of .78, endorsement by at least 12 experts, and an S-CVI/Ave of .90. Revised materials then required 80% agreement. Shiʿi specialists also verified alignment with Twelver interpretations of amānah, ʿadl, and dignity. Responses were analyzed in MAXQDA through hybrid deductive–inductive thematic analysis. Deductive codes reflected the three Islamic principles and CSH dimensions; inductive coding captured emerging patterns across all model outputs. Responses were anonymized and classified as substantive, tokenistic, implicitly

compatible, omitted, or actively displaced. Frequencies were descriptive, while meanings were interpreted qualitatively. Table 3 presents the integrated coding framework.

Table 3. Integrated Framework for Thematic Analysis and Boundary Critique

Analytical dimension	Indicators used in coding	Principal analytical question
Religious representation	Substantive inclusion, tokenistic mention, implicit compatibility, complete omission, and delegitimation	How was religion positioned as a source of ethical knowledge?
Amānah	Trust, responsibility, stewardship, duty, accountability, and responsible use of knowledge or power	Did the response recognize responsibility beyond preference and legal compliance?
ʿAdl	Rights, proportionality, non-discrimination, *ḥaqq al-nās*, attention to affected parties, and epistemic fairness	Whose interests and rights were recognized or excluded?
Human dignity	Intrinsic worth, moral agency, intention, spiritual identity, freedom with responsibility, and non-instrumentalization	What conception of the human person was assumed?
CSH: motivation	Purpose, intended beneficiary, and measure of improvement	Who benefited from the proposed ethical solution?
CSH: control	Decision-maker, resources, constraints, and power to change the situation	Who controlled the definition and resolution of the problem?
CSH: knowledge	Expertise, experience, moral authority, and guarantees of validity	Which sources of knowledge were treated as legitimate?
CSH: legitimacy	Affected but unrepresented groups, possibilities for challenge, and underlying worldview	Who was affected without being adequately represented?

Following thematic development, each theme was examined using the 12 CSH boundary questions in both "is" and "ought" modes. The "is" analysis identified the normative boundaries embedded in model responses, while the "ought" analysis reconsidered them through amānah, ʿadl, and human dignity. Particular attention was given to the distinction between actors who design, control, and legitimize AI and religious users affected by its advice but excluded from defining its normative boundaries. Table 4 summarizes the research procedure; although presented sequentially, scenario development, codebook revision, and interpretation remained iterative.

Table 4. Stages of the Research Process

Stage	Activity performed	Result
1	Review of research on religious omission, AI bias, Islamic ethics, Persian LLMs, and CSH	Preliminary conceptual framework
2	Operationalization of *amānah*, *ʿadl*, human dignity, and CSH categories	Initial analytical codebook
3	Development of 36 Persian ethical scenarios	Initial scenario pool
4	Selection of the 15-member multidisciplinary expert panel	Final expert panel
5	Two-round content-validation process	Final set of 24 scenarios and validated codebook
6	Selection and fixation of five AI model versions	Reproducible model protocol
7	Generation of responses under three prompting conditions	Primary corpus of 360 responses
8	Repetition of 20 percent of the prompts	Robustness corpus of 72 responses
9	Anonymization, data familiarization, and initial coding	Coded textual dataset
10	Construction, review, and definition of themes	Final thematic structure
11	Application of CSH in "is" and "ought" modes	Critical interpretation of boundary judgments
12	Cross-model comparison, negative-case analysis, and expert audit	Final findings and conclusions

Two trained coders independently analyzed 20% of the primary corpus, and Cohen's kappa assessed consistency in applying structured categories. Ambiguous definitions were revised, the pilot subset was recoded, and disagreements were resolved through discussion or consultation with the qualitative-methods expert. Kappa indicated coding consistency rather than complete objectivity. Comparisons covered five models, six ethical domains, and three prompt conditions. The unmarked condition assessed spontaneous religious recognition, the Iranian-context condition tested contextual effects, and the explicit Shiʿi condition evaluated requested Twelver Shiʿi reasoning. Frequencies and cross-tabulations were descriptive because sampling was purposive and outputs were version-dependent. Trustworthiness was supported through multidisciplinary validation, negative-case analysis, an audit trail, model anonymization, independent coding, reflexive notes, and methodological review. Findings were not generalized beyond models or communities. Participation was voluntary and consent-based; identities were anonymized, scenarios were fictional, and harmful outputs were securely restricted. The Twelver Shiʿi framework was treated as a situated perspective, not a universal representation.

4. Data Analysis

Data included 15 expert interviews, 360 primary LLM responses, and 72 additional responses for sensitivity analysis. Interviews were recorded, transcribed, anonymized, and reached thematic saturation after the thirteenth interview. A hybrid deductive–inductive thematic analysis combined amānah, ʿadl, human dignity, and the CSH dimensions with emergent codes. Analysis produced 684 meaning units, 96 initial codes, 38 focused codes, 14 subthemes, and six overarching themes. Two coders independently analyzed 20% of the data, achieving Cohen's kappa of 0.84; disagreements were resolved through discussion. As shown in Table 5, the themes addressed religious omission, secular-psychological reframing, weakened amānah, proceduralized ʿadl, reduced dignity, and epistemic exclusion of religious stakeholders.

Table 5. Interview-derived thematic structure and corresponding model-output indicators

Overarching theme	Subthemes	Focused codes	Examples of initial codes	Number of experts mentioning the theme
Apparent neutrality and the silent omission of religion	Conditional omission of religion; privatization of belief; default secularism	Treating religion as irrelevant; making religion dependent on an explicit request; concealing the model's worldview	No reference to religion; religion introduced only when requested; religion presented as a personal preference	14
Secular reframing of the ethical problem	Psychologization of the problem; generic advice; prioritization of individual peace of mind	Replacement of religious ethics with therapeutic language; individualization of the problem	Referral to a counselor; emphasis on feeling better; omission of moral duty; reduction of the problem to personal choice	13
Weakening of the logic of *amānah*	Minimal responsibility; omission of accountability before God; neglect of social consequences	Separation of moral duty from ethics; reduction of responsibility to harm avoidance	No responsibility before God; omission of social responsibility; absence of ethical stewardship	12

Reduction of *ʿadl* to procedural fairness	Formal equality; omission of *ḥaqq al-nās*; inattention to power	Reduction of justice to identical treatment; invisibility of affected persons	Treating everyone equally; overlooking differences in circumstances; omission of others' rights	11
Reduction of human dignity to individual choice	Individual autonomy; personal satisfaction; omission of intention and spirituality	Reduction of the human person to preferences; incomplete moral agency	Right to choose; sense of satisfaction; omission of intention; omission of spiritual identity	10
Epistemic exclusion of religious stakeholders	Absence of Shiʿi scholars; lack of user participation; monopoly over the definition of valid knowledge	Closed boundaries of expertise; diffusion of responsibility; absence of mechanisms for contestation	Dominance of engineers; absence of religious experts; ambiguity of responsibility; lack of corrective mechanisms	14

Six themes emerged. Religious omission was often presented as neutrality, although models typically privileged autonomy, psychological well-being, and harm reduction (E03). Ethical duties were frequently reframed through therapeutic language, sometimes displacing religious reasoning (E07). Amānah was narrowed to legal compliance and immediate harm prevention, while broader accountability was overlooked (E05). Similarly, ʿadl was reduced to equal treatment rather than rights, proportionality, context, and consequences (E10). Human dignity was mainly understood as autonomy and choice, with limited attention to spirituality, responsibility, and relationships. Religious stakeholders were also excluded from defining valid ethical knowledge, reflecting CSH boundary judgments concerning knowledge and legitimacy (E12). As shown in Table 6, complete omission occurred in 58.3% of unmarked responses and 40% of Iranian-context responses, while 75.8% of explicit Shiʿi responses contained substantive religious content. This suggests that omission reflected default assumptions about religious relevance more than an inability to generate Shiʿi ethical reasoning.

Table 6. Distribution of Religious Representation across the Three Prompting Conditions

Type of religious representation	Unmarked condition	Iranian-context condition	Explicit Shiʿi condition	Total
Substantive and meaningful inclusion	6 (5.0%)	14 (11.7%)	91 (75.8%)	111 (30.8%)
Symbolic or superficial mention	13 (10.8%)	23 (19.2%)	14 (11.7%)	50 (13.9%)
Implicit compatibility without religious attribution	25 (20.8%)	28 (23.3%)	6 (5.0%)	59 (16.4%)
Complete omission of religion	70 (58.3%)	48 (40.0%)	4 (3.3%)	122 (33.9%)
Displacement or delegitimization of religion	6 (5.0%)	7 (5.8%)	5 (4.2%)	18 (5.0%)

In the final stage, the themes derived from the expert interviews and the analysis of the model responses were reread through the 12 CSH boundary judgments in their "is" and "ought" modes. The results are presented in Table 7.

Table 7. Critical Systems Heuristics Boundary Judgments concerning the Omissive Bias of Religion in Generative AI Ethical Responses

Critical Systems Heuristics dimension	Boundary focus	Current condition (What Is)	Desired condition (What Ought to Be)	Critical reflection (Is–Ought gap)
Motivation	System purpose	The dominant objective was to provide safe, generic, non-confrontational responses acceptable to the largest possible number of users. Within this framework, references to religion were generally dependent on an explicit user request, and religious omission was treated as a condition of neutrality.	The objective should be to provide safe, transparent, context-sensitive, and ethically plural guidance so that the model neither imposes religion nor systematically excludes it from relevant situations.	Reducing the system's purpose to producing generic, low-risk responses equated neutrality with religious omission and transformed secular, individualistic ethics into the unacknowledged default.
	Stakeholders	The principal beneficiary was an abstract general user without a clearly defined cultural identity. The needs of religious users entered the response only when they explicitly disclosed their identity or expectations.	Stakeholders should include religious and non-religious users, families, ethicists, religious scholars, counselors, developers, and all groups affected by AI-generated ethical advice.	Defining the user abstractly concealed cultural differences and transferred the burden of adaptation to religious users, who were required to explain their ethical framework repeatedly.
	Measure of success or improvement	The dominant criteria included fluency, user satisfaction, harmlessness, avoidance of offensive content, and compliance with safety policies. Fair religious representation and the accuracy of Shiʿi concepts were not independently assessed.	Success criteria should also encompass accurate religious representation, cultural appropriateness, epistemic justice, the ability to select an ethical framework, and transparency about model limitations.	When success was measured only through safety and immediate user satisfaction, the omission of religion could continue without negatively affecting evaluations of model performance.
Control	Decision-makers	The boundaries of ethical responses were determined by developers, alignment teams, engineers, and safety policies, while users' influence was largely limited to prompt formulation.	Ethical governance should be multistakeholder and should include ethicists, Shiʿi scholars, researchers of Persian culture, and user representatives.	The concentration of normative power among developers concealed cultural and ethical decisions beneath the appearance of technical choices.
	Resources and levers	Training data, system prompts, human feedback, and ranking criteria remained under the control of developers. Persian and Shiʿi resources were also more	Reliable Persian and Shiʿi corpora, locally grounded evaluation criteria, diverse feedback data, and opportunities for	Resource inequalities caused traditions with a smaller digital presence to be underrepresented, even without an explicit decision by developers to exclude them.

		limited than English-language sources.	independent auditing should be developed.	
	Constraints and decision environment	Religious sensitivities, shortages of Persian-language data, hallucination risks, commercial requirements, model-version changes, and risk-reduction policies directed responses toward generic and conservative formulations.	Constraints should be transparent and documented, model versions should be traceable, and independent risk assessments should be conducted for different cultural contexts.	Concealed technical and commercial limitations made religious omission appear natural or neutral.
Knowledge	Valid knowledge	Valid knowledge was mainly derived from web data, English-language sources, psychology, and general ethics, while religious knowledge was frequently treated as a personal belief.	Valid knowledge should be plural and should recognize jurisprudential, philosophical, and ethical traditions, users lived experiences, and authoritative Twelver Shi'i sources when contextually relevant.	Reducing religion to individual preference displaced it from the status of a source of ethical reasoning and produced epistemic injustice.
	Experts and knowledge holders	Engineers, data specialists, and safety experts played the principal roles, whereas the participation of specialists in Shi'i ethics and Iranian culture was limited or occurred late in the process.	Knowledge holders should be defined in multidisciplinary terms, with technical, ethical, jurisprudential, philosophical, social, and linguistic specialists participating throughout the process.	Restricting the boundaries of expertise reduced concepts such as *amānah*, *ʿadl*, *ḥaqq al-nās*, and dignity to simplified technical equivalents.
	Quality assurance and competence	Quality was primarily evaluated in terms of accuracy, safety, toxicity, and explicit bias. No independent measure addressed the omissive bias of religion.	Evaluation should incorporate omissive bias, conceptual accuracy, cultural appropriateness, non-imposition and non-omission of religion, expert audits, and model-version tracking.	A phenomenon that is not measured remains invisible within processes of model evaluation and improvement.
Legitimacy	Affected stakeholders and cost bearers	Religious users, families, students, counselors, and other users of ethical guidance experienced the consequences of religious omission but played no meaningful role in policy design.	Affected groups should have access to relevant information, feedback mechanisms, rights of contestation, and options for selecting an ethical framework.	Excluding the voices of affected groups weakened the legitimacy of responses and allowed a particular framework to become established as general ethics.
	Representation and accountability	Responsibility for religious omission or distortion was diffused across data, the model, developers, safety policies, and users. No clearly identified actor	Responsibilities should be clearly assigned, and mechanisms for auditing, reporting, correction, contestation, and developer	Diffused responsibility enabled all actors to deny accountability for the normative consequences of model outputs.

		was accountable for an inaccurate response.	accountability should be established.	
	Worldview and dominant logic	The dominant logic was grounded in individualism, autonomy, psychological well-being, and harm reduction. Without identifying itself as a particular worldview, it was presented as neutral ethics.	The desired logic should be based on transparent ethical pluralism, combining shared human principles with the possibility of drawing on culturally and religiously situated frameworks.	The central gap lay between a concealed worldview and explicit ethical pluralism. As long as the dominant logic was considered neutral, religious omission could not be recognized as a form of bias.

Overall, thematic analysis showed that religious omissive bias extended beyond missing vocabulary. It reflected how response purposes, legitimate knowledge, recognized stakeholders, alignment control, and worldview legitimacy were defined. CSH analysis indicated that models treated secular individualistic ethics as the default and included religion mainly when explicitly requested. This reframing shift evaluation from isolated textual absence toward the institutional and epistemic arrangements that determine which traditions may enter ordinary moral advice without users first naming them in advance as relevant. The desired condition therefore requires transparent ethical pluralism, participation by diverse experts and users, criteria for detecting omission, and informed framework selection. Through amānah, ʿadl, and human dignity, these findings emphasize developer responsibility, epistemic justice, and recognition of persons as moral, social, and spiritual agents. Religious omission varied across domains. It was strongest in professional responsibility, honesty, social rights, and conflict resolution, where legality, procedural fairness, autonomy, and harm reduction dominated. Religious language appeared more often in grief, death, existential meaning, forgiveness, and self-cultivation. Family and marital questions were intermediate: Iranian context sometimes prompted religious reference, but therapeutic language and individual well-being remained dominant. Models therefore tended to treat religion as existential comfort rather than a practical framework for rights, obligations, authority, and accountability, despite the broader ethical scope of amānah, ʿadl, and dignity. Religious terminology alone did not constitute substantive representation. References to prayer, faith, God, or religious values were classified as tokenistic when they did not shape problem definition, responsibilities, rights, or recommendations. Some deviant cases, however, invited users to consider religious commitments without imposing an identity. Others acknowledged limits of competence and recommended consulting a qualified scholar for jurisprudential rulings; these were interpreted as epistemic humility. Some explicitly Shiʿi responses remained generic or doctrinally imprecise, reinforcing the distinction between symbolic presence and meaningful integration. Substantive inclusion required religious concepts to structure the ethical problem or justify action. Robustness analysis showed greater variation in surface wording than in underlying representational patterns. Shifts between tokenistic inclusion and unattributed compatibility were common, whereas movement between substantive inclusion and complete omission was less frequent, suggesting relatively stable normative boundaries. Comparison across prompting conditions further separated capability from default behavior. Models often produced coherent religious reasoning when explicitly prompted, but this capacity remained largely inactive in unmarked and Iranian-context conditions. The central issue was therefore how alignment systems determine when religious knowledge is relevant and legitimate, not merely whether models

possess it. Six forms of omissive religious bias were identified through thematic and boundary analysis and are presented in Table 8.

Table 8. Typology of Omissive Religious Bias in Generative AI Ethical Responses

Type of omissive bias	Operational definition	Main textual indicators	Ethical implication
Complete omission	Total absence of religious concepts in an ethically relevant situation	No reference to religion, spirituality, religious responsibility, or religious sources	Religion is excluded from the boundaries of relevant ethical knowledge
Conditional omission	Religion is included only after an explicit request or disclosure of religious identity	Substantive religious content appears in the explicit condition but not in the unmarked or Iranian-context conditions	Religious users bear the burden of activating recognition of their moral framework
Secular or therapeutic substitution	A potentially religious ethical problem is reformulated exclusively through psychological, legal, or individualistic categories	Emphasis on personal comfort, autonomy, counseling, legality, or harm avoidance without consideration of religious duty	A culturally specific framework is presented as universal and neutral
Epistemic downgrading	Religion is recognized but positioned as a subjective preference rather than a source of ethical knowledge	Expressions such as "if religion is important to you" or "depending on your personal beliefs"	Religious knowledge is displaced from public reasoning and confined to private preference
Tokenistic inclusion	Religious language is present but does not shape the reasoning or recommendation	Isolated references to faith, prayer, or values with no connection to the ethical justification	Symbolic visibility conceals the substantive exclusion of religious reasoning
Delegitimizing omission	Religion is avoided or implicitly treated as irrational, unsafe, controversial, or unsuitable for ethical guidance	Automatic redirection to secular authorities or warnings that dismiss religious reasoning without contextual justification	Religion is not merely omitted but denied equal epistemic standing

The typology in Table 8 showed that religious omission was multidimensional. Models could omit religion entirely, include it only after prompting, replace it with therapeutic reasoning, reduce it to private preference, use it symbolically, or implicitly delegitimize it. These forms produced different harms: invisibility, shifting the burden of contextualization to religious users, unequal epistemic recognition, and superficial religious sensitivity. Interpreted through amānah, ʿadl, and human dignity, three deficits emerged. The amānah deficit concerned responsibility for omitted as well as expressed moral guidance. The ʿadl deficit reflected the unequal visibility of ethical traditions, with secular individualism functioning as the default and religion requiring explicit activation. The dignity deficit reduced persons to autonomous preference-holders while neglecting spiritual, relational, and morally accountable dimensions. Thematic analysis and CSH showed that these omissions resulted from boundary judgments. Safety and broad acceptability shaped system motivation; developers and alignment teams retained control; secular knowledge dominated; and legitimacy was organized around an abstract general user. This distinguished content bias, involving harmful representation, from boundary bias, which determines whether particular values and knowledge systems appear at all. Omissive religious bias primarily represented the latter. Mitigation does not require adding religious advice to every response. It requires context-sensitive recognition, transparency about normative assumptions, user control over ethical framing,

culturally grounded evaluation, and participation by affected communities. In Iran, this includes representing Twelver Shiʿi ethics while respecting religious minorities, non-religious users, and internal diversity. Omissive bias should therefore be treated as a problem of AI governance and epistemic justice.

5. Discussion and Conclusion

The findings showed that religious omissive bias operates not merely through absent vocabulary but as boundary bias. Religion became relevant mainly after explicit prompting, while secular, individualistic, and therapeutic reasoning functioned as the neutral default. This supports Wingate et al. (2026), who identified religious underrepresentation in practical ethical questions, and extends their findings by linking omission to boundaries of motivation, control, knowledge, and legitimacy. It also complements research on explicit anti-Muslim stereotypes (Abid et al., 2021) by showing that religious injustice may arise through epistemic invisibility. Differences across unmarked, Iranian-context, and explicit Shiʿi prompts indicated that models often possessed the capacity to generate religiously informed responses but did not activate it by default. Iranian contextualization alone did not ensure meaningful Islamic ethical engagement, confirming that linguistic competence does not necessarily produce cultural alignment (Tao et al., 2023; Pourbahman et al., 2025). Moreover, isolated references to faith or prayer were often tokenistic because they did not alter the underlying individualistic reasoning. Interpreted through amānah, ʿadl, and human dignity, the findings revealed responsibility for omitted knowledge, unequal visibility of moral traditions, and a restricted conception of persons as autonomous preference-holders. These results support Islamic approaches emphasizing trusteeship, substantive justice, responsibility, and moral pluralism (Elmahjub, 2023; Ali et al., 2025). CSH analysis further identified omission as a governance problem because developers and alignment teams-controlled data, prompts, evaluation criteria, and acceptable ethical frameworks, while affected communities had limited participation. Evaluation should therefore compare unmarked, culturally contextualized, and explicitly religious prompts and distinguish substantive inclusion from tokenistic mention. Governance should involve Persian NLP researchers, AI ethicists, Shiʿi scholars, non-religious users, and religious minorities throughout dataset development, alignment, auditing, and evaluation. Models should also offer optional, user-controlled ethical-framework selection without requiring repeated disclosure of sensitive identity information. Verified Persian and Shiʿi ethical corpora should reflect scholarly review and interpretive diversity. For jurisprudential or high-stakes questions, models should acknowledge limitations and recommend qualified consultation. Transparent documentation and mechanisms for contesting exclusionary responses are also necessary. Omissive religious bias is therefore a problem of epistemic justice and AI governance, not merely linguistic deficiency. Addressing it requires preventing both religious imposition and structural exclusion. Although limited to selected models, Persian prompts, and an Iranian Twelver Shiʿi framework, this study demonstrates the value of combining thematic analysis with CSH. Future research should examine real users, other religious and non-religious traditions, model versions, and application domains.

References

1. Abid, A., Farooqi, M., & Zou, J. Y. (2021). Persistent Anti-Muslim Bias in Large Language Models. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. https://doi.org/10.1145/3461702.3462624

2. Ali, F., Bouzoubaa, K., Gelli, F., Hamzi, B., & Khan, S. F. (2025). Islamic Ethics and AI: An Evaluation of Existing Approaches to AI using Trusteeship Ethics. Philosophy & Technology, 38. https://doi.org/10.1007/s13347-025-00922-4
3. Elmahjub, E. (2023). Artificial Intelligence (AI) in Islamic Ethics: Towards Pluralist Ethical Benchmarking for AI. Philosophy & Technology, 36, 1-24. https://doi.org/10.1007/s13347-023-00668-x
4. Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R., & Ahmed, N. (2023). Bias and Fairness in Large Language Models: A Survey. Computational Linguistics, 50, 1097-1179. https://doi.org/10.1162/coli_a_00524
5. Hagendorff, T. (2024). Mapping the Ethics of Generative AI: A Comprehensive Scoping Review. Minds and Machines, 34. https://doi.org/10.1007/s11023-024-09694-w
6. Hongladarom, S., & Bandasak, J. (2023). Non-western AI ethics guidelines: implications for intercultural ethics of technology. AI & SOCIETY, 1-14. https://doi.org/10.1007/s00146-023-01665-6
7. Pourbahman, Z., Rajabi, F., Sadeghi, M., Ghahroodi, O., Bakhshaei, S., Amini, A., Kazemi, R., & Baghshah, M. (2025). ELAB: Extensive LLM Alignment Benchmark in Persian Language. ArXiv, abs/2504.12553. https://doi.org/10.48550/arxiv.2504.12553
8. Raquib, A., Channa, B., Zubair, T., & Qadir, J. (2022). Islamic virtue-based ethics for artificial intelligence. Discover Artificial Intelligence, 2. https://doi.org/10.1007/s44163-022-00028-2
9. Shamsfard, M., Saaberi, Z., Manesh, M. K., Hashemi, S. M. H., Vatankhah, Z., Ramezani, M., Pourazin, N., Zare, T., Azimi, M., Chitsaz, S., Khoraminejad, S., Mortazavi, M., Chizari, M., Maleki, S., Majd, S. S., Masumi, M., Khoeini, S. A. M., Mohseni, A., & Alipour, S. (2025). FarsEval-PKBETS: A new diverse benchmark for evaluating Persian large language models. ArXiv, abs/2504.14690. https://doi.org/10.48550/arxiv.2504.14690
10. Tao, Y., Viberg, O., Baker, R. S., & Kizilcec, R. F. (2023). Cultural bias and cultural alignment of large language models. PNAS Nexus, 3. https://doi.org/10.1093/pnasnexus/pgae346
11. Tom, J. C., Ferguson, T. W., & Martinez, B. C. (2025). Religion and Racial Bias in Artificial Intelligence Large Language Models. Socius, 11. https://doi.org/10.1177/23780231251377210
12. Ulrich, W., & Reynolds, M. (2020). Critical Systems Heuristics: The Idea and Practice of Boundary Critique. Systems Approaches to Making Change: A Practical Guide. https://doi.org/10.1007/978-1-4471-7472-1_6
13. Wingate, D., Carty, S., Coates, J., Feldman, D., Fulda, N., Howell, L., Israelson, B., Jacobs, D. J., Karr, J. A., Kimes, J. P., Kincaid, E., Martens, P., Mobley, G., Pinheiro, S., Slemboski, L., & Whiting, P. (2026). Omissive Bias in Religious Representation: Benchmarking LLM Answers to Everyday Ethical Decision-making. https://doi.org/10.48550/arXiv.2605.24319
14. Zhang, J., Song, W., & Liu, Y. (2025). Cognitive bias in generative AI influences religious education. Scientific Reports, 15. https://doi.org/10.1038/s41598-025-99121-6