Integrating Wasathiyyah Values into AI-Assisted Learning: A Literature Review for Islamic Education in Indonesia

Ainur Rofik
Universitas Islam Negeri K.H. Abdurrahman Wahid Pekalongan, Indonesia
E-mail: rofiq.arfa3124@gmail.com

Abstract

Generative AI is increasingly used by students and teachers to obtain religious guidance, yet research on its implications for religious moderation (*wasathiyyah*) in Islamic education remains limited. This study examines how the literature treats the relationship between AI-assisted learning, *wasathiyyah*, and the plurality of legitimate juristic positions (*ikhtilaf*), and proposes a framework for integrating *wasathiyyah* values into AI-assisted learning in Indonesia. A qualitative library research design was applied, using content analysis of a purposive corpus of 31 sources published between 2004 and 2026. Statements on benefits, risks, limits of religious judgement, human oversight, moderation, and plurality of positions were coded into seven categories. The findings show that the literature mainly treats AI as a vehicle for delivering moderation content. Attention to the plurality of legitimate positions was clearly present in two of seven core sources and partially in a third, and none translated it into a principle of classroom practice. Four risks to the epistemic dimension of *wasathiyyah* were identified: flattening of *ikhtilaf*, error and invention, skewed representation of Islam, and displaced authority. The study distinguishes dispositional and epistemic layers of *wasathiyyah* and proposes a five-gate model (frame, prompt, verify, deliberate, and assess), linking each gate to a value, a risk, and a pedagogical practice. The model is a conceptual synthesis based on single-coder, largely abstract-based analysis and requires empirical validation, for which four testable propositions are offered. The study contributes an *ikhtilaf*-preserving perspective to the integration of religious moderation and AI in Islamic education.

Keywords: wasathiyyah, ikhtilaf, AI-assisted learning, Islamic education, religious moderation

Introduction

Ask a chatbot whether the *qunut* should be recited in the dawn prayer, a question that Indonesian students meet early and often, and it will probably give one tidy answer. The scholarly picture is less tidy. Jurists have long held more than one position on the matter, and Indonesian Muslims grow up alongside neighbours who pray differently and are no less sincere. The difference is part of the tradition, not a defect in it. What an AI model hands back, in contrast, is typically a single paragraph written in the voice of settled knowledge. A recent evaluation of ChatGPT (GPT-4o) against fourteen human-issued *fatawa* on yoga found that the model answered correctly in six cases and failed in eight, and its author attributes the failures to a bias towards permissiveness and a reliance on secondary sources over authoritative collections (Jan, 2026). The study is time-locked to one model and one topic, but it illustrates the problem plainly: the way an AI system answers a contested question can differ from the way the tradition answers it.

This matters for religious moderation because moderation is more than good manners. Kamali (2015) treats *wasatiyyah* as an ethical virtue rooted in the Qur'anic idea of the middle path, and he takes care to separate it from the weaker meanings sometimes attached to it, such as compromise or mediocrity. In Indonesia the idea has been turned into policy. The Ministry of Religious Affairs works with four indicators: national commitment, tolerance, rejection of violence, and accommodation of local culture (Kementerian Agama Republik Indonesia, 2019; Fahri & Zainuri, 2019). Islamic universities and *pesantren* are treated as important sites for cultivating these dispositions (Nasir & Rijal, 2021), and the middle position itself has been examined critically, including in relation to the two largest Muslim organisations (Hilmy, 2013). Read closely, however, the indicators describe attitudes. They say much about how a person should treat difference, and rather less about how a person should know, and should teach, that difference exists.

A word of care is needed here, because not every difference is a legitimate one. Islamic legal thought has long separated matters whose evidence and meaning are settled from matters left open to interpretation, and respect for *ikhtilaf* applies only to the second kind. To say that a model should preserve difference is therefore not to say that every answer is as good as any other, or that the tradition has no fixed points. It is to say that a teacher, and a tool used under a teacher's guidance, should be able to tell which kind of question is on the table. A system that treats a settled matter as open misleads students in one direction, and a system that treats an open matter as settled misleads them in the other. The second mistake is the one we think AI-assisted learning is more prone to, and it is the one this article follows.

Meanwhile AI-assisted learning has spread quickly. Early reviews found a field in which educators were seldom consulted and ethical reflection lagged behind technical work (Zawacki-Richter et al., 2019), and broader surveys showed AI already touching administration, instruction, and learning (Chen et al., 2020). Large language models promise personalised feedback and relief for teachers, but only when users bring critical judgement to them (Kasneci et al., 2023), and early public responses to ChatGPT combined enthusiasm with concern about integrity and ethics (Tlili et al., 2023). Critical scholars add that AI in education has a history and a politics that technical accounts tend to skip (Williamson & Eynon, 2020; Selwyn, 2019). Within Islamic education, reviews of Quranic education report promising uses alongside worries about authenticity and contextual appropriateness (Ahmad et al., 2025; Zainadun et al., 2025), and survey evidence suggests that Indonesian Muslim students accept AI because it is easy and useful (Faizin et al., 2025). Only two studies in the material gathered here tie AI directly to religious moderation, one through instructional design for prospective teachers (Zaimina & Munib, 2025) and one through educational governance (Khairi et al., 2026).

The risks that bear on moderation are documented elsewhere in the AI literature. Abid et al. (2021) found that GPT-3 completed a prompt about Muslims with references to violence in 66% of its responses, far more often than for other religious groups. Language models also generate confident falsehoods (Ji et al., 2023) and inherit patterns from training data that nobody fully audited (Bender et al., 2021). Beyond the classroom, recommendation systems shape what young people read, and research on echo chambers and on pathways towards extreme content gives reason for caution (Cinelli et al., 2021; Ribeiro et al., 2020). These studies concern bias and misinformation. They do not ask the question we are pursuing here, which is what happens to a legitimate difference of opinion when a machine mediates it.

We therefore advance a thesis. *Wasathiyyah* has a dispositional layer, made up of tolerance, fairness, and balance, and an epistemic layer, which is the honest handling of difference in what is known and taught. AI-assisted learning is most likely to damage the second layer, because it favours a single confident answer, and a classroom that teaches tolerance as an attitude while feeding students flattened answers works against itself. Integrating *wasathiyyah* into AI-assisted learning should therefore begin by preserving *ikhtilaf*. This review asks three questions. (1) How does the literature treat the relation between AI-assisted learning, *wasathiyyah*, and the plurality of legitimate positions? (2) Which risks to the epistemic layer of *wasathiyyah* can be identified in the evidence? (3) At which points in the learning cycle can *wasathiyyah* values be integrated, and through what practices?

Methods

Research Design

This study is a qualitative library research project, and its analytic tool is content analysis in the sense described by Krippendorff (2018): texts are treated as communications whose meaning is read in the context of their use, and are coded systematically so that claims can be traced back to the passages that support them. The study sits within the broader family of literature reviews (Snyder, 2019), but it is not a protocol-driven systematic review. The intersection of AI, religious moderation, and Islamic education is small, and much of what exists is scattered across national journals, proceedings, and repositories that commercial databases cover unevenly. A purposive corpus read closely suited the question better than a broad automated search.

Data Sources and Search Procedure

Searches were run in October 2026 through Google Scholar, publisher platforms, and the open journal systems of Indonesian and Malaysian Islamic universities. Search strings combined AI terms ("artificial intelligence", "generative AI", "ChatGPT", "large language model"), Islamic education and ethics terms ("Islamic education", "Quranic education", "fatwa", "Islamic ethics"), and moderation or plurality terms ("religious moderation", "wasatiyyah", "ikhtilaf"). Further sources were found by following citations in the most relevant papers.

Selection Criteria

A source was retained when it addressed AI or algorithmic systems in education, religion, or Islamic education; when it addressed religious moderation, Islamic ethics, or Islamic pedagogy; or when it supplied a conceptual or empirical base for bias, authority, or AI literacy. Peer-reviewed articles, major conference papers, university-press books, and a national policy document were eligible. Bibliographic details, including the DOI where one exists, had to be confirmable on the publisher or journal page before a source was admitted, and sources whose details could not be confirmed were excluded even when their titles looked relevant. The final corpus contains 31 analytic sources published between 2004 and 2026. Two further works (Krippendorff, 2018; Snyder, 2019) serve as methodological references only. For several recent sources, only the abstract and publication record could be consulted in full, so claims about them are limited to what the abstracts report.

Data Analysis Technique

The unit of analysis was a statement in which a source made a claim about a benefit of AI-assisted learning, a risk, a limit on AI's religious or moral judgement, the need for human oversight, the handling of moderation, or the plurality of legitimate positions. Categories combined deductive and inductive elements. The deductive ones came from the research questions and from the Ministry's indicators (Kementerian Agama Republik Indonesia, 2019); the inductive one, attention to the plurality of positions, emerged during reading and became the focus of the study. The seven categories are listed in the note to Table 2. The coding matrix for the seven core sources is reported so that readers can check it, and the integration model reported below was built by relating the categories to the *wasathiyyah* values in the corpus. Where a result is a finding reported by a source we say so, and where it is our inference we mark it. Coding was done by a single coder, no inter-coder agreement was computed, and this is a limitation we return to below.

Results

What the Corpus Looks Like and What the Coding Shows

Table 1 shows how the 31 sources are spread. Six appeared before 2019, five in 2019, ten between 2020 and 2022, and ten from 2023 onwards, which is what one would expect when the technology itself is so new. Seven sources count as core because they deal most directly with AI in Islamic education, Islamic ethics, or religious guidance.

Table 1 Composition of the analytic corpus (n = 31)

Category	n	Sources
Core: AI in Islamic education, ethics, and religious guidance	7	Ahmad et al. (2025); Elmahjub (2023); Faizin et al. (2025); Jan (2026); Khairi et al. (2026); Zaimina & Munib (2025); Zainadun et al. (2025)
Religious moderation and Islamic pedagogy	6	Fahri & Zainuri (2019); Halstead (2004); Hilmy (2013); Kamali (2015); Kementerian Agama RI (2019); Nasir & Rijal (2021)
AI in education (general)	7	Chen et al. (2020); Holmes et al. (2022); Kasneci et al. (2023); Selwyn (2019); Tlili et al. (2023); Williamson & Eynon (2020); Zawacki-Richter et al. (2019)
AI ethics, bias, and algorithms	9	Abid et al. (2021); Bender et al. (2021); Cinelli et al. (2021); Floridi et al. (2018); Ji et al. (2023); Jobin et al. (2019); Kitchin (2017); Mittelstadt et al. (2016); Ribeiro et al. (2020)
AI literacy	2	Long & Magerko (2020); Ng et al. (2021)

Table 2 gives the coding of the core sources. Four or five of the seven talk about benefits (A), accuracy or authenticity risks (B), limits on religious or moral judgement (C), and the need for human or expert oversight (E). Two mention bias (D), and two put moderation at the centre (F). The category our argument rests on, attention to the plurality of legitimate positions (G), shows up clearly in two sources and partly in a third. Elmahjub (2023) points out that Islamic jurisprudence holds heterogeneous views even about the content of *maslaha*, the notion he offers as an ethical guide for AI. Khairi et al. (2026) warn that AI-driven governance can narrow pluralistic dialogue. Jan (2026) tests model output against fatwas issued by different institutions and scholars; we coded it as partial because its abstract mentions schools of thought without

developing the idea. None of the three turns plurality into something a teacher could use in a lesson.

Table 2 Coding matrix for the seven core sources

Source	A	B	C	D	E	F	G
Ahmad et al. (2025)	●	●	●	–	●	–	–
Zainadun et al. (2025)	●	●	●	–	●	–	–
Zaimina & Munib (2025)	●	–	●	–	●	●	–
Faizin et al. (2025)	●	●	–	–	–	–	–
Khairi et al. (2026)	●	–	●	●	●	●	●
Elmahjub (2023)	–	–	–	–	–	–	●
Jan (2026)	–	●	●	●	●	–	(●)
Sources coding the category	5	4	5	2	5	2	2 + 1 partial

Note. ● = present; (●) = partial; – = not evident in the abstract or record consulted. A = benefit of AI-assisted learning; B = accuracy or authenticity risk; C = limit on religious or moral judgement; D = bias; E = human or expert oversight; F = explicit focus on moderation; G = attention to plurality of legitimate positions. Coding by a single coder from abstracts and records.

What this pattern says is fairly plain. The literature mostly treats AI as a vehicle, something that delivers content, personalises practice, or helps run institutions. Zaimina and Munib (2025) studied prospective Islamic elementary teachers at two Indonesian universities and report that AI-based instructional design can give learners interactive, personal, and contextual experiences of moderation, while warning that religious values may be misread. That is a useful result for the dispositional layer. It does not tell us what happens to the epistemic layer when a model answers a question on which the tradition is divided.

Four Risks to the Epistemic Layer

Reading the core sources next to the wider AI literature, we can name four risks. The first and most important is flattening. Jan (2026) found that the model tended to downplay authoritative judgments and lean on secondary sources, so that a range of institutional positions collapsed into one smoothed-over answer. We take this as evidence for the flattening risk, with the caution that it comes from one model and one subject. The second risk is error and invention. Language models write fluent mistakes (Ji et al., 2023), and the Quranic education reviews keep returning to authenticity and contextual appropriateness (Ahmad et al., 2025; Zainadun et al., 2025). A flattened answer that is also wrong is very hard for a student to catch.

The third risk is skewed representation. Abid et al. (2021) showed that GPT-3 linked "Muslim" with violence when completing prompts, and mapped it to "terrorist" in 23% of analogy tests, each time more harshly than it treated other religious groups. That study used an older model, so current systems need to be checked and not assumed to behave the same way. Still, it shows that a model's picture of Islam can be distorted in ways that cut against every one of the Ministry's indicators. Critical work on algorithms reminds us that such outputs come from design choices and data, and both can be questioned (Kitchin, 2017; Mittelstadt et al., 2016; Bender et al., 2021). The fourth risk is displaced authority. Faizin et al. (2025) report that students accept AI because it is easy and useful, which is not the same as being reliable, and

Zainadun et al. (2025) list teachers' overreliance on technology among the challenges. Outside school, echo chambers and recommendation pathways shape what young people read (Cinelli et al., 2021; Ribeiro et al., 2020), although those studies concern social platforms and should not be carried over to educational chatbots without care.

Put together, these risks form a chain. A model flattens a contested question, may dress the answer up with invented support or a skewed frame, and the student, who has learned to trust what is easy, takes it as authoritative. Nobody in that chain says the question is open. Governance can make things worse: Khairi et al. (2026) conclude that AI-driven governance without a sound ethical framework risks reproducing inequality and narrowing pluralistic dialogue.

A Five-Gate Model for Preserving Ikhtilaf

Where, then, can *wasathiyyah* enter the learning cycle? We propose five gates, laid out in Table 3 and Figure 1. Each gate pairs a value from the Indonesian moderation discourse (Kementerian Agama Republik Indonesia, 2019; Fahri & Zainuri, 2019) with one of the risks above and a practice a teacher can use. The model is a synthesis. Its parts rest on the corpus, but no source tests it, and the pairing of gates with values is our own reading.

Table 3 Integration matrix: gates, values, risks, and practices

Gate	Value engaged	Risk addressed	Teaching practice	Anchors in the corpus
1 Frame	*Tawassuth*, *tawazun*	Open question treated as having one answer	Before using AI, state whether the topic is settled or open, and name the range of recognised positions	Elmahjub (2023); Kamali (2015); Jan (2026)
2 Prompt	*I'tidal*	Model defaults to one view; permissiveness; secondary sources	Ask for positions with attribution and evidence, and for uncertainty to be flagged, rather than for a verdict	Jan (2026); Abid et al. (2021); Bender et al. (2021)
3 Verify	*Tabayyun*, *sanad*	Fluent error; invented support	Trace each claim to a source, and check disputed points with a qualified authority	Ahmad et al. (2025); Ji et al. (2023); Long & Magerko (2020); Ng et al. (2021)
4 Deliberate	*Tasamuh*, *syura*	Student absorbs the machine's verdict	Use AI output as one input in a class deliberation in which students defend positions courteously and the teacher moderates	Zaimina & Munib (2025); Nasir & Rijal (2021); Selwyn (2019)
5 Assess	*Tawazun*, *i'tidal*	Single correct answers rewarded; overreliance	Assess whether students can state the range, give evidence, and disagree with courtesy, and require disclosure of AI use	Faizin et al. (2025); Zainadun et al. (2025); Tlili et al. (2023)

The first gate is where to start, and it costs the teacher almost nothing. Before anyone opens a tool, the teacher says which part of the topic is settled and which is open, and lists the positions qualified scholars have held. The class then has something to measure the model against. This puts into practice what Elmahjub (2023) observes about the variety of views in Islamic jurisprudence, and it follows Kamali's (2015) point that moderation is an ethical virtue, not a polite word for vagueness. The second gate is about the prompt. Jan's (2026) results suggest that models drift towards a smoothed-over verdict, so students learn to ask for positions with attribution and evidence, and to ask the model what it does not know. That does not cure error, which is why verification comes next.

Islamic education already has the vocabulary for the third gate. *Tabayyun*, the duty to verify a report before accepting or passing it on (Qur'an 49:6), can be taught as a routine: pin down what the model claimed, trace it to a source, ask who is competent to speak on the question, and check the context in which that source speaks. The routine connects to *sanad*, the chain of transmission through which religious knowledge has traditionally been authenticated, and to the wider AI-literacy skills of understanding, evaluating, and using AI ethically (Long & Magerko, 2020; Ng et al., 2021). The fourth gate keeps people talking to one another. Selwyn (2019) argues for holding on to the human side of teaching, and here AI output is simply one voice in the room. Pesantren have long been seen as places where moderation is formed (Nasir & Rijal, 2021), and their habit of working through contested legal questions together looks to us like a natural home for this gate. That link is our suggestion and does not come from the corpus.

The fifth gate is assessment, and it matters more than it seems. If students are rewarded for the single right answer, they will ask AI for the single right answer. Assessment should instead ask them to state the range, support it, and show courtesy in disagreement, and it should ask them to say where they used AI, given the integrity concerns in the corpus (Faizin et al., 2025; Tlili et al., 2023). Two conditions hold the gates together. The teacher moderates and curates, which echoes the Quranic pedagogy review's view that AI should support the teacher and not replace one (Zainadun et al., 2025). And institutions need governance with real force behind human oversight. Khairi et al. (2026) propose risk mapping, ethical design protocols, meaningful oversight, public accountability, and continuous evaluation, which sit comfortably with the broader AI ethics literature (Floridi et al., 2018; Jobin et al., 2019; Holmes et al., 2022).

A short illustration may help; it is our own construction, not a finding. Take the *qunut* question again. At the first gate the teacher explains that the Shafi'i school, dominant in Indonesia, recommends the *qunut* in the dawn prayer, that the Hanafi school does not, and that other schools differ in detail. At the second gate students ask a chatbot which schools hold which position, on what evidence, and where it is unsure. At the third they compare its account with a text the teacher supplies or a classical work from the class library, and they take any point of disagreement to a local *ustadz* or *kyai*. The model may turn out to be broadly right, partly wrong, or oddly confident, and in every case the class learns how to treat it.

At the fourth gate small groups present one position as its holders would defend it, and the class talks about how someone who prays one way should behave beside someone who prays another way. Here the dispositional layer, tolerance and courtesy, finally has something to hold on to. At the fifth gate the written task asks students to explain two positions and the courtesy owed between their holders, and to note which AI tool they used and what they

changed after checking. A student who simply copied a chatbot's first answer would find the task cannot be done that way, and that is the point.

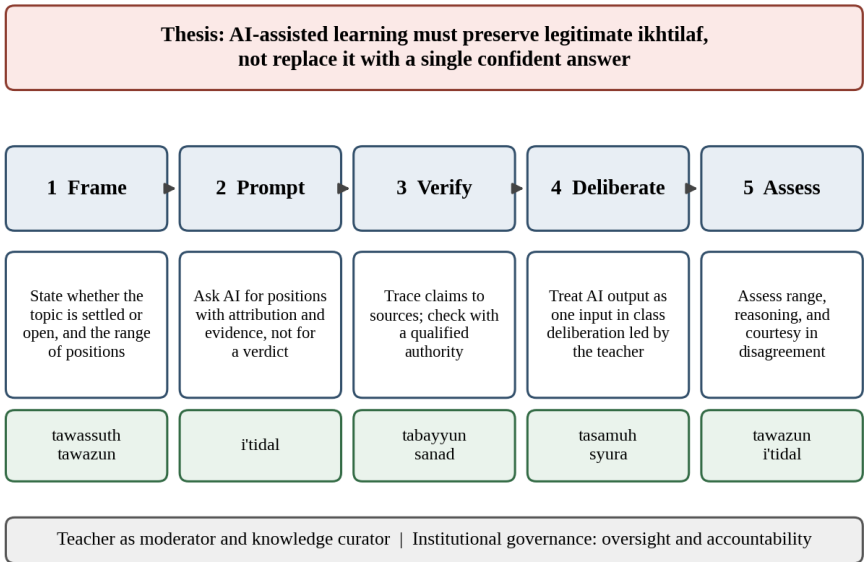


Figure 1 The five-gate model for preserving ikhtilaf in AI-assisted learning

Discussion

Why Flattening Is the Central Risk

The most telling thing in the coding is a silence. Benefits, accuracy, and oversight come up again and again, yet the plurality of legitimate positions is rarely mentioned and never made a design principle. That fits how the moderation discourse itself is framed. The Ministry's indicators describe how a person should treat difference (Kementerian Agama Republik Indonesia, 2019), and the AI literature naturally asks whether AI can deliver tolerance-promoting content or personalise learning (Zaimina & Munib, 2025; Zawacki-Richter et al., 2019). Those are fair questions about the dispositional layer. Our argument is that the epistemic layer deserves the same attention. A student who has been taught to tolerate difference, but whom a confident machine has convinced that there was never any real difference, has an attitude with nothing to apply it to.

This is not the usual worry about extremism. A chatbot is unlikely to preach violence, though biased output is a real danger (Abid et al., 2021). The more likely trouble is that it speaks with certainty on questions where the tradition holds a legitimate range. Flattening needs no ill will and no offensive content, and it can pass for helpfulness, which makes it a quieter threat than hostility. It also helps explain why Faizin et al. (2025) find acceptance driven by ease and usefulness, since those are exactly the qualities that make a single answer so attractive.

Two Approaches Compared

It helps to set the model beside the pattern the coding revealed. Table 4 contrasts what we call a vehicle approach, in which AI delivers moderation content and its answers are taken as given, with the approach proposed here. The vehicle approach is our label for a pattern across sources, not the position of any single author. Several sources we coded as treating AI as a vehicle also carry the cautions our model builds on, such as Zaimina and Munib's (2025) warning about misinterpretation, so the difference is one of emphasis and not of opposition.

Table 4 Interpretive contrast between two approaches to integrating wasathiyyah

Dimension	**Vehicle approach**	***Ikhtilaf*-preserving approach**
Role of AI	Delivers content and personalises practice	Object of scrutiny and one interlocutor among others
Layer of *wasathiyyah* addressed	Dispositional: tolerance, fairness, balance	Dispositional and epistemic: also the honest handling of difference
Contested questions	The model's answer is taken as the answer	The range of positions is stated before the model is consulted
Role of the teacher	Supervises use of the tool	Moderates deliberation and curates knowledge
What is assessed	Correct answers and engagement	Range, evidence, and courtesy in disagreement

Wasathiyyah Has an Epistemic Layer

Kamali (2015) presents *wasatiyyah* as an ethical virtue woven into Islamic theology and ethics, not merely a social posture. If that is right, it should reach into how knowledge is held and handed on, and the tradition of recognising *ikhtilaf* is one way it does. Our two layers are an analytic device and not a claim that the tradition itself divides things this way. They do help show where AI meets moderation. On the dispositional layer AI can be a useful tool, for instance by making materials more engaging and personal (Zaimina & Munib, 2025). On the epistemic layer it is as likely to wear moderation down as to support it, unless the learning design pushes back. The five gates are our attempt to say how.

The Indonesian Setting and the Teacher

Indonesia offers resources that a generic AI-in-education model would miss. Islamic universities and pesantren are already treated as places where moderation is formed (Nasir & Rijal, 2021), and their long habit of working through disputed questions gives the fourth gate a home. The model also leans on the teacher, not as a guard against technology but as the person who can say with authority which questions are open. That echoes Halstead's (2004) picture of Islamic education as instruction, nurture, and the shaping of conduct held together, and Selwyn's (2019) case for keeping the human side of teaching in view. It also matches the warning about overreliance on technology in Quranic pedagogy (Zainadun et al., 2025). What the model asks of teachers is concrete: to sort a question into settled or open, to prompt for range, to verify, to moderate a discussion, and to assess reasoning. These are demanding skills, but they can be learned, and they could anchor teacher development without borrowing a ready-made international framework.

Propositions for Testing

Because the model is a synthesis, it should be put to the test. We offer four propositions. P1: on questions where recognised schools differ, general-purpose chatbots will present fewer attributed positions than human-curated fatwa collections. P2: prompts that ask for positions with attribution and evidence will raise the number of attributed positions, though not necessarily their accuracy. P3: students who join a deliberation task after consulting AI will state the range of positions more fully, and disagree more courteously, than students who consult AI alone. P4: assessment that rewards range and reasoning will reduce reliance on AI-generated single answers. The first two can be tested by putting a set of *khilafiyah* questions to

several chatbots and comparing the answers with curated sources. The other two call for classroom studies.

A first test of P1 and P2 would not need much. A researcher could draw some ten *khilafiyah* questions from ritual practice and from social affairs, put each to three or four chatbots in two forms, once asking for a ruling and once asking for the positions with attribution and evidence, and record what comes back: how many positions are named, whether each is tied to a school or scholar, which sources are cited, how certain the tone is, and whether a knowledgeable reader, checking against curated collections, finds errors. Because models change quickly, the model versions and dates should be reported, as Jan (2026) rightly does, and the exercise repeated as systems are updated. Results of this kind would show whether flattening is a stable property of these systems or an artefact of particular models, and that distinction decides how much weight the first gate and the second gate should carry.

Implications

For teachers, the practical lesson is to begin with the first gate: naming the range of positions before any tool is opened costs little and changes how students read what follows. For curricula, verification routines of the *tabayyun* type and prompts that ask for range can become standing skills in Islamic religious education. For institutions, the governance elements proposed by Khairi et al. (2026) offer a starting point for schools, madrasah, and Islamic universities that are adopting AI without a policy. For researchers, the coding matrix shows where evidence is missing and where the next studies might look.

Limitations

Several limits qualify these conclusions. The corpus is purposive and not exhaustive, so relevant studies, especially recent proceedings and local-language articles, may be missing. Sources were admitted only when their details could be confirmed, which helps reliability but may have shut out legitimate work. For several recent sources we relied on abstracts, so the coding in Table 2 reflects what abstracts reveal, and a full-text re-coding could change individual cells. One coder did the work, with no agreement check. The evidence at the meeting point of AI, moderation, and Islamic education is thin, mostly Indonesian and Malaysian, and largely small-scale, conceptual, or self-reported. The key evidence for flattening comes from one model and one topic and is tied to GPT-4o (Jan, 2026), and findings on earlier models such as GPT-3 (Abid et al., 2021) may not describe current systems. Finally, the five-gate model and its pairing of gates with values are our own synthesis and have not been validated.

Conclusion

We began by asking what happens to *wasathiyyah* when learning is mediated by AI. The literature, we found, has looked mostly at the dispositional layer, treating AI as a vehicle for moderation content, while the plurality of legitimate positions is rarely discussed and has not become a principle of classroom design. The main risk to the epistemic layer is flattening, made worse by invented support, skewed representation, and displaced authority. *Wasathiyyah* can be woven into AI-assisted learning at five gates: frame, prompt, verify, deliberate, and assess, each tied to a value, a risk, and a practice.

The argument comes down to this: integrating *wasathiyyah* into AI-assisted learning begins with keeping *ikhtilaf* alive. In practice that means naming the range before using a tool, asking for positions and not verdicts, verifying with *tabayyun*, deliberating together, and assessing reasoning. In theory it suggests that moderation in Islamic education has two layers,

one of disposition and one of knowledge. The next step is empirical: testing the propositions with chatbots, with classes, and with teachers.

References

Abid, A., Farooqi, M., & Zou, J. (2021). Persistent anti-Muslim bias in large language models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 298–306). ACM. http://dx.doi.org/10.1145/3461702.3462624

Ahmad, K. A., Shohibuddin, W. A. J., & Eldeib, A. A. M. (2025). Artificial intelligence (AI) in Quranic education: A systematic review of pedagogical applications and challenges. *QURANICA - International Journal of Quranic Research, 17*(2), 187–213. http://dx.doi.org/10.22452/quranica.vol17no2.8

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). ACM. http://dx.doi.org/10.1145/3442188.3445922

Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access, 8*, 75264–75278. http://dx.doi.org/10.1109/ACCESS.2020.2988510

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., & Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences, 118*(9), Article e2023301118. http://dx.doi.org/10.1073/pnas.2023301118

Elmahjub, E. (2023). Artificial intelligence (AI) in Islamic ethics: Towards pluralist ethical benchmarking for AI. *Philosophy & Technology, 36*(4), Article 73. http://dx.doi.org/10.1007/s13347-023-00668-x

Fahri, M., & Zainuri, A. (2019). Moderasi beragama di Indonesia. *Intizar, 25*(2), 95–100. http://dx.doi.org/10.19109/intizar.v25i2.5640

Faizin, N., Alfan, M., Basid, A., Ramadhan, M. R., Panatik, S. A., & Kawakip, A. N. (2025). Muslim students' acceptance of artificial intelligence in Islamic religious education: An extended TAM approach. *Discover Education, 4*, Article 304. http://dx.doi.org/10.1007/s44217-025-00767-1

Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. *Minds and Machines, 28*(4), 689–707. http://dx.doi.org/10.1007/s11023-018-9482-5

Halstead, J. M. (2004). An Islamic concept of education. *Comparative Education, 40*(4), 517–529. http://dx.doi.org/10.1080/0305006042000284510

Hilmy, M. (2013). Whither Indonesia's Islamic moderatism? A reexamination on the moderate vision of Muhammadiyah and NU. *Journal of Indonesian Islam, 7*(1), 24–48. http://dx.doi.org/10.15642/JIIS.2013.7.1.24-48

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., Santos, O. C., Rodrigo, M. T., Cukurova, M., Bittencourt, I. I., & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504–526. http://dx.doi.org/10.1007/s40593-021-00239-1

Jan, A. (2026). Digital authority and Islamic legal reasoning: A comparative study of AI-generated and human-issued fatāwā. *Online Journal of Research in Islamic Studies, 13*(1), 1–16. http://dx.doi.org/10.22452/ris.vol13no1.1

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys, 55*(12), Article 248. http://dx.doi.org/10.1145/3571730

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

Kamali, M. H. (2015). *The middle path of moderation in Islam: The Qur'anic principle of wasatiyyah*. Oxford University Press. http://dx.doi.org/10.1093/acprof:oso/9780190226831.001.0001

Kasneci, E., Sessler, K., Küchemann, S., Bannert, M., Dementieva, D., Fischer, F., Gasser, U., Groh, G., Günnemann, S., Hüllermeier, E., Krusche, S., Kutyniok, G., Michaeli, T., Nerdel, C., Pfeffer, J., Poquet, O., Sailer, M., Schmidt, A., Seidel, T., … Kasneci, G. (2023). ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences, 103*, Article 102274. http://dx.doi.org/10.1016/j.lindif.2023.102274

Kementerian Agama Republik Indonesia. (2019). *Moderasi beragama*. Badan Litbang dan Diklat Kementerian Agama RI.

Khairi, A. I., Sugiantoro, Wahyudi, K., Tianah, I., & Al Farisi, S. (2026). Artificial intelligence in educational governance: A risk-based ethical framework for strengthening religious moderation in Indonesia. *Re-JIEM (Research Journal of Islamic Education Management)*, 231–248. http://dx.doi.org/10.19105/re-jiem.vi.24246

Kitchin, R. (2017). Thinking critically about and researching algorithms. *Information, Communication & Society, 20*(1), 14–29. http://dx.doi.org/10.1080/1369118X.2016.1154087

Krippendorff, K. (2018). *Content analysis: An introduction to its methodology* (4th ed.). SAGE Publications.

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. http://dx.doi.org/10.1145/3313831.3376727

Mittelstadt, B. D., Allo, P., Taddeo, M., Wachter, S., & Floridi, L. (2016). The ethics of algorithms: Mapping the debate. *Big Data & Society, 3*(2). http://dx.doi.org/10.1177/2053951716679679

Nasir, M., & Rijal, M. K. (2021). Keeping the middle path: Mainstreaming religious moderation through Islamic higher education institutions in Indonesia. *Indonesian Journal of Islam and Muslim Societies, 11*(2), 213–241. http://dx.doi.org/10.18326/ijims.v11i2.213-241

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education: Artificial Intelligence, 2*, Article 100041. http://dx.doi.org/10.1016/j.caeai.2021.100041

Ribeiro, M. H., Ottoni, R., West, R., Almeida, V. A. F., & Meira, W. (2020). Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on*

Fairness, Accountability, and Transparency (pp. 131–141). ACM. http://dx.doi.org/10.1145/3351095.3372879

Selwyn, N. (2019). *Should robots replace teachers? AI and the future of education*. Polity Press.

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

Tlili, A., Shehata, B., Adarkwah, M. A., Bozkurt, A., Hickey, D. T., Huang, R., & Agyemang, B. (2023). What if the devil is my guardian angel: ChatGPT as a case study of using chatbots in education. *Smart Learning Environments, 10*, Article 15. http://dx.doi.org/10.1186/s40561-023-00237-x

Williamson, B., & Eynon, R. (2020). Historical threads, missing links, and future directions in AI in education. *Learning, Media and Technology, 45*(3), 223–235. http://dx.doi.org/10.1080/17439884.2020.1798995

Zaimina, A. B., & Munib, B. (2025). Enhancing religious moderation pedagogy through AI-based instructional design in Islamic teacher education: A critical study in the context of Society 5.0. *Akselerasi: Jurnal Pendidikan Guru MI, 6*(1), 1–17. http://dx.doi.org/10.35719/akselerasi.v6i1.821

Zainadun, N., Mohamad Salleh, M. N., Jamil, A. I., Rekan, A. A., & Ajmain @ Jima'ain, M. T. (2025). Exploring the potential of artificial intelligence in enhancing Quranic teachers' pedagogy: A systematic literature review. *QURANICA - International Journal of Quranic Research, 17*(2), 501–541. http://dx.doi.org/10.22452/quranica.vol17no2.21

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators? *International Journal of Educational Technology in Higher Education, 16*, Article 39. http://dx.doi.org/10.1186/s41239-019-0171-0