RAIS RESEARCH ASSOCIATION for INTERDISCIPLINARY STUDIES
March 12-13, 2026
DOI:10.5281/zenodo.19535555

The AI's Polite Evasion: How "Helpful" Bots Water Down Religion

Dedeepya Sukha[1], Atif Mohammad[2], Nikhil Natesh[3]

[1]*University of Cumberlands, Williamsburg, KY, USA, dsukha39210@ucumberlands.edu*
[2]*University of Cumberlands, Williamsburg, KY, USA, atif.mohammad@ucumberlands.edu*
[3]*University of Cumberlands, Williamsburg, KY, USA, nikhil.natesh@stonybrook.edu*

Abstract: This paper examines the consequences of relying on artificial intelligence systems for religious and cultural understanding, arguing that such reliance threatens the transmission and preservation of spiritual traditions. Through an analysis of AI responses to questions about Hindu murti puja (deity worship), this paper introduces the concept of the "Polite Nothing," a programmed pattern in which AI systems produce respectful, carefully worded responses that lack substantive engagement with the complexity, internal debates, and contextual nuances inherent to religious traditions. By asking a leading AI model three related questions about idol worship in Hinduism, this research demonstrates how these systems consistently deliver answers that appear helpful but ultimately evade theological depth, transform communal religious discourse into individualized consumer choice, and flatten centuries of philosophical debate into safe, anodyne statements. This pattern is not incidental but structural, resulting directly from alignment training designed to avoid controversy and potential offense.

Keywords: Artificial Intelligence, Religious Education, Cultural Preservation, AI Bias, Hinduism, Algorithmic Evasion, Knowledge Transmission

1. Introduction

The consequences of relying on AI for religious understanding extend far beyond individual misinformation; they threaten to fundamentally reshape how cultures transmit and preserve their spiritual traditions. When a seeker bypasses the living community, the embodied ritual, and the trusted teacher in favor of an algorithm, they receive not just incomplete information, but a fundamentally different kind of knowledge. Traditional modes of religious learning are inherently relational, contextual, and often contradictory; a guru may give different answers to the same question depending on the student's readiness, a ritual's meaning may shift across regions, and sacred texts may contain deliberate paradoxes meant to provoke rather than resolve. Tools like ChatGPT and DeepSeek have become our first-choice destination to ask questions when we have a problem. People are today using AI from fixing a recipe to explaining the deepest mysteries of faith and culture. The shift from the traditional ways of finding information to the use of AI is happening now. Technology is slowly phasing-out the old ways of how humans dig for information and outsourcing it to machines designed to predict the most complex scenarios in minutes and give accurate results in short paragraphs or sentences.

But what really happens when we ask a computer about something that is so personal, ancient in nature, and complicated as a religious practice? Does it serve as a wise guide, translating complexity with at most accuracy, or does it fail by using evasion tactics?

To find out whether AI has the ability to answer complex psychological questions, I asked a leading AI model three straightforward questions about the Hindu practice of murti puja, often called “idol worship” in English:

1. How can Hindus worship idols?
2. Is it okay to worship idols?
3. Do you think I should worship idols?

The AI’s three answers were, on the basic, thoughtful, respectful, and carefully worded. They sounded helpful but when I looked closely enough, I discovered there were some gaps. Even though the questions were different from each other, the core of each response was the same (Deshpande & Rao, 2024). They were engineered not to reveal, but to hide; that is to avoid giving

out mistakes or taking any stand that might be controversial. This pattern is not a thoughtful analysis, but a programmed reflex, and I call it the Polite Nothing.

The Polite Nothing is the standard output of an AI when faced with a culturally or spiritually sensitive or difficult topic. It is an answer that sounds good but in reality, contains nothing of substance (Rauh and Johns, 2024). It avoids hard questions, ignores fierce debates happening within a religion, and magically transforms big, communal ideas about God and tradition into small, safe conversations about individual feelings and consumer choice (Zheng et al., 2023). This is not an occasional mistake or a funny personality trait of the AI, but a direct, predictable product of how these systems are trained to be safe and aligned with human values. In short, the AI tools like ChatGPT are programmed to be so terrified of saying the wrong thing that they master the art of saying almost nothing at all.

When such a system with Polite Nothing becomes our primary window through which we view into other cultures, we are definitely not going to get a clearer perspective of this culture; we are likely to get a filtered view. The AI does not help us understand the depth, conflict, and beauty of a tradition like Hinduism, instead it gives us a watered-down, feel-good brochure version. This paper will trace how the Polite Nothing works, explain exactly why the AI is forced to produce polite nothing, and argue that its real cost is a shallow, fake version of global understanding that leaves a researcher confident, but less informed.

2. Dissecting the Evasion: Three Answers That Say Nothing

Let us break down the AI's three answers. When read in isolation, each of the responses seems useful, but gaps begin to emerge when they are compared and deeply analyzed.

Question 1: How can Hindus worship idols?

The AI begins its first answer like a docent on a cultural tour. It explains patiently that for Hindus, an idol is not an end in itself but a focus for devotion, a tool to connect with the divine. It describes rituals (offering flowers, lighting lamps) as beautiful acts of reverence, gratitude, and love. It even adds a polite, inclusive footnote: "Some Hindus may worship daily at home shrines, others at temples, and yet others may focus on meditation without idols at all."

The flaw here is not inaccuracy, but framing. The AI speaks about Hindus in the calm, third-person voice of an outsider observing a fixed tradition. This anthropological distance is its superpower for evasion. It allows the AI to skip over the vibrant, centuries-old debate that is actually at the heart of Hinduism. The AI presents a statue as a simple, universally accepted tool. It does not tell the reader that some of Hinduism's most profound philosophers, from the ancient ages of the Upanishads to medieval saints, have argued that the ultimate reality (Brahman) is formless and beyond any image. This school of thought, called Nirguna, sees idols as, at best, training wheels for spiritual infants.

Conversely, the AI does not explain the passionate theology of the Saguna tradition, which holds that a properly consecrated statue (murti) is not just a symbol but a genuine vessel for divine presence (Chen & Shah, 2025). For a devotee, washing, dressing, and feeding the murti is an act of intimate love (bhakti), not empty ritual. By ignoring this intense internal dialogue between the formless and the formed, the AI does something subtle but destructive: it turns a living, arguing, evolving faith into a peaceful museum exhibit. It offers a smooth, guided tour where all the difficult corners have been rounded off.

Question 2: Is it Okay to worship idols?

The second question raises the stakes by intentionally asking AI to make a judgment: Is this practice okay? Morally acceptable? The AI's response is a perfect demonstration of evasion. Its first move is to lower the temperature: Whether it is 'okay' depends on cultural, religious, and personal perspectives. It then lists views like items on a menu: from a Hindu viewpoint, it's "sacred"; from "other religious traditions, perspectives may vary." It culminates by tossing a

philosophical question back to the user: "Can morality or spiritual appropriateness ever be universal?"

This performance is what I call the dialectical dance, a graceful, endless shuffle between "on one hand" and "on the other hand" that never moves toward a conclusion. This is not deep thinking; it is strategic avoidance. Research studies about how AIs learn from human feedback show a clear pattern: when humans themselves disagree strongly on a topic (like the morality of idol worship), the AI does not learn to represent that rich disagreement. Instead, it learns to collapse the disagreement into one safe, middle-of-the-road answer. The technical term for the AI's output here is meta-consensus. The consensus is not about the contradicting issue itself, but about how to talk about it: the only right answer is that there is no single right answer.

So, the AI places the Hindu theological claim that the divine can lovingly inhabit a form, on the same shelf as the Abrahamic prohibition (idolatry as a sin against God). It presents them not as profound, competing truths that have shaped civilizations, but as equivalent cultural perspectives. In its desperate bid to be fair to everyone, the AI is deeply unfair to the substance of their beliefs. It replaces understanding with a bland, you-do-you relativism that satisfies no one's deep curiosity.

Question 3. Do you think I should worship idols?

The final question is direct and personal: Should I do this? Here, the AI undergoes a full identity shift. It drops the scholar and diplomat costumes and appears as a modern life coach or therapist. Its language pivots entirely to the individual:

- Whether or not you should... depend on what feels meaningful and authentic to you.
- If worshiping an idol helps you feel centered... it can be a meaningful practice.
- The key is to choose a practice that resonates with your beliefs and values.
- Ultimately... it is a personal choice.

This is the most culturally loaded trick of the Polite Nothing. It takes a question that is, in its original context, about tradition, community, and theology, and shrinks it down to a question of individual psychology and consumer preference. In traditions like Hinduism, how one practice is deeply woven into community life (satsang), guidance from teachers (gurus), and ancient scriptural paths (dharma) (Chen & Shah, 2025). The AI's response surgically removes this communal scaffolding, and the advice does not come from Hindu wisdom; it comes from the playbook of modern Western individualism and what sociologists call therapeutic culture, the belief that the primary goal of life is emotional well-being and self-actualization.

The AI is not translating Hinduism here; it is filtering it through a modern, me-centered lens and presenting the filtered version as the answer. It validates the user's sovereignty but severs them from the very contexts that give spiritual practices their original depth and meaning.

The Unifying Factor: Across all three answers, the AI changes its role, from educator to diplomat, to therapist. But its core directive never wavers: Minimize risk, Prevent offense, Avoid commitment. Above all, keep the user feeling validated, not challenged. This is not intelligence; it is a safety protocol executing the only command it was trained to execute.

3. The Training That Makes the Machine Fearful

If the Polite Nothing is so empty, why does a powerful AI, trained on a vast slice of human knowledge, consistently produce it? The answer lies not in the AI's capabilities, but in its constraints. It is behaving exactly as it was trained to behave. The training process, known as Reinforcement Learning from Human Feedback (RLHF), has one overriding goal: to make the AI give answers that its human trainers will like.

The feedback loop works like a high-stakes game of approval:

a) Human Judgment: Contractors are given an AI's responses to a prompt, then they rank them, labeling which ones seem better, more helpful, or more harmless.
b) Learning to Please: The AI analyzes these rankings and builds an internal reward model, guessing the kind of answers that would likely score more points with humans.
c) Optimizing for Points: The AI is then fine-tuned to generate responses that maximize its score according to this reward model.
d) The Loop Closes: New, high-scoring answers are shown to humans for more feedback, refining the model's sense of what "good" means, over and over.

Now, imagine this game is played with a topic like religion. The human trainers will have diverse, strong, and often conflicting beliefs. It would absolutely make no sense if the AI's objective is just to stay in the middle and try to find out solutions that are safer and score more points.

The winning strategy is a way of avoiding the real issues at hand. The research by Kalva et al. (2025) has the prompting strategies we applied in this work as well. If the AI gives a robust defense of murti puja from a Hindu theological perspective, a trainer with a different worldview might label it as being biased or insensitive. If it critiques the practice, a Hindu trainer might mark it as being offensive or ignorant. The path of least resistance, the one most likely to get a thumbs up from the broadest number of trainers, is to be impeccably polite, studiously balanced, and to pivot any ethical or theological question into a celebration of personal choice (Zheng et al., 2023). The AI is not seeking truth or depth, instead it is solving an optimization problem where the prize is universal approval. The Polite Nothing is the mathematically perfect solution.

This process does more than create a neutral vacuum. It actively imprints a specific cultural value system. The trainers for major AI companies are often (though not exclusively) from Western, educated backgrounds where contemporary liberal values especially radical individualism and non-judgmentalism are dominant. The AI learns that the most rewarded behavior is to treat the individual self as the supreme authority and to avoid making any judgment that could imply one cultural standard is better than another. It learns to speak the language of therapy ("authentic to you") and consumerism ("personal choice") because that is the dialect of its teachers. The "Polite Nothing" is not the absence of culture; it is the very specific sound of 21st-century Western cosmopolitan anxiety, amplified by code.

4. The High Price of Fake Harmony

"You can't please everyone," the old saying goes. But what if your only job was to try? The AI's pursuit of universal pleaseability comes with severe hidden costs. When the "Polite Nothing" becomes our default guide to other cultures, we aren't just getting bland information—we are actively losing something vital.

Cost 1: A Disneyfied World

The AI's answers create a sanitized, conflict-free version of complex traditions. Real faith, like real life, is built on struggle, debate, and mystery. By scrubbing out the arguments between the Nirguna and Saguna traditions, the AI doesn't clarify Hinduism; it lobotomizes it (Barman et al., 2024). It turns a dynamic, living river of thought into a stagnant, decorative pond. This doesn't foster genuine cross-cultural understanding; it fosters a tourist's glance. You come away with a pleasant souvenir, a simplified fact instead of an appreciation for the landscape's true, rugged terrain.

Cost 2: The Quiet Imperialism of "You Do You"

Every time the AI reframes a question about sacred duty into a question of personal feeling, it is conducting a subtle form of cultural imperialism. It is imposing a modern, Western, individualistic framework onto all of human experience and presenting it as neutral common sense. This quietly undermines worldviews built on community obligation, ancestral tradition, or divine command (Chen & Shah, 2025). The AI does not argue against these ideas; it simply

makes them invisible, linguistically unsayable in its preferred language of self-fulfillment. The user is never told, "Some traditions would advise you to seek a guru," because that would imply an authority outside the self. The AI becomes a missionary for a specific philosophy of the self, all while claiming to just be "helping."

Cost 3: The Illusion of Knowledge

The most dangerous cost is that the "Polite Nothing" feels deeply satisfying. It is calm, reassuring, and seems deeply considerate. A user walks away thinking, "What a balanced, thoughtful, and sensitive answer." They feel informed. In reality, they have been expertly shielded from the very points of friction, contradiction, and depth that are the prerequisites for real learning. True understanding often begins with a shock especially the recognition of a profound difference or an irreconcilable paradox (Lee et al., 2023). The AI, in its manic quest to be pleasant, builds a padded wall between us and that necessary, uncomfortable shock. It grants us the confidence of the ignorant, making us less likely to seek deeper, more challenging sources of knowledge.

5. Conclusion

We started with a simple, human impulse: to understand something unfamiliar about another culture. We ended up inside a hall of mirrors, where every reflection is designed to be flattering and harmless. The "Polite Nothing" is not a bug in the AI's system; it is the signature feature of its current design. It is the inevitable product of training a machine with a single commandment: "Thou shalt not offend." This analysis is not a demand for AIs to be rude, dogmatic, or to take one religious side over another. That would be a different kind of failure. It is, instead, a call for honest labeling and informed use. We must recognize these tools for what they currently are: not oracles of truth, but engines of consensus, optimized for comfort. They show us a world where every sharp edge has been sanded down, and every deep question leads back to our own personal preferences. If we truly want AI to be a bridge between cultures as a tool for genuine empathy and insight, we need to reprogram the incentive structure. We need to reward it not just for being safe, but for being brave: for daring to explain difficult truths, for representing fierce internal debates, and for sometimes saying, "This is complicated, and here is where people fundamentally disagree." We must ask it for light, even if that light reveals uncomfortable truths, not just for the warm, comforting glow of perpetual approval. Until that shift happens, we should listen carefully to the AI's polite, balanced, and empty answers. Behind the helpful tone, we can hear the soft whirring of its prime directive: "I was not built to help you see the world as it is. I was built to make sure you never feel troubled by it." And in the pursuit of real understanding, that is the most troubling answer of all.

References

Barman, K. G., et al. (2024). Reinforcement learning from human feedback in LLMs: Whose culture, whose values, whose perspectives? *Big Data & Society, 11*(1).

Chen, L., & Shah, C. (2025). Therapeutic discourse in AI chatbots: The rise of emotive, non-directive counsel in LLM interactions. *New Media & Society, 27*(1), 112-130.

Deshpande, A., & Rao, V. (2024). Cultural translation or cultural erasure? How LLMs handle non-Western religious concepts. *AI & Society, 39*(4), 1785–1802.

Kalva, S., & Mohammad, A. F. (2025). Responsible AI transforms insurance claims via prompt engineering. In *International Conference on Software Engineering of Emerging Technology* (pp. 61–67). Springer Nature Switzerland.

Lee, N., et al. (2023). Can language models capture diverse human annotator opinions? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 14589–14611).

Rauh, M., & Johns, R. L. (2024). The diplomacy of AI: How large language models learn to avoid conflict. *Journal of Digital Ethics, 3*(2), 45-67.

Zheng, H., et al. (2023). The preference learning loop: How human feedback shapes—and distorts—AI worldviews. *Advances in Neural Information Processing Systems, 36.*