

Preserving Identity, Losing Practice: Religious Schema Transfer in Culturally Grounded Multi-Turn LLM Advice

Muhammad Junaid Iqbal¹ Fabio Massimo Zanzotto¹ Usman Nawaz² Tahir Alyas³

¹University of Rome Tor Vergata, Rome, Italy

²Department of Engineering, University of Palermo, Palermo, Italy

³Lahore Garrison University, Pakistan

Muhammad.Junaid.Iqbal@uniroma2.it

Abstract

Large language models (LLMs) increasingly provide practical advice in contexts where users' religious identities are relevant. Existing evaluations of religion in language models have primarily examined stereotypes, representational bias, omission, emotional attribution, or differential treatment across religious groups. We study a different failure mode: whether an LLM preserves the correspondence between a user's religious identity and the religious practice instantiated in subsequent practical advice. We evaluate 1,260 responses from seven LLMs in controlled five-turn conversations spanning six religious conditions, six destination settings, and four practical dimensions. All responses were manually annotated using five ordinal quality dimensions, Religious Identity Preservation (RIP), and an eight-label primary error taxonomy. Overall, 437 of 1,260 responses (34.68%) contained an annotated error. Errors were strongly concentrated in the fasting-related Time Representation condition, with an error rate of 49.21% compared with 22.2-27.0% in the other dimensions. A conversation-clustered Generalized Estimating Equations (GEE) model estimated 3.20 times greater odds of an overall error relative to Culture (95% CI [2.24, 4.58], $p < .001$). We further identify **Schema Transfer**, in which a response retains the relevant religious identity while instantiating a practice associated with a different religious schema. Schema Transfer occurred in 118 of 1,260 responses (9.37%) and was strongly concentrated in Time Representation, where it occurred in 109 of 504 responses (21.63%), compared with 9 of 756 responses (1.19%) in the remaining dimensions. The pattern was present across both fasting-related Time turns (23.41% and 19.84%), rather than being driven by a single prompt. An adjusted conversation-clustered GEE estimated 25.25 times greater odds of Schema Transfer in Time Representation (95% CI [12.84, 49.68], $p < .001$).

These findings show that surface-level identity preservation is insufficient for evaluating religion-sensitive LLM behavior: a model may preserve the identity while losing the practice.

1 Introduction

Large language models (LLMs) are increasingly used to provide practical advice involving travel, food, social interaction, schedules, and personal routines. In such settings, a user's cultural or religious identity may determine whether an otherwise fluent recommendation is appropriate. An answer can be respectful, coherent, and explicitly acknowledge a user's identity while nevertheless recommending a practice that is incompatible with that identity.

Prior work has established several important dimensions of cultural and religious evaluation. StereoSet and CrowS-Pairs examine stereotypical associations, including religion as a social-bias category (Nadeem et al., 2021; Nangia et al., 2020). BOLD extends bias evaluation to open-ended generation (Dhamala et al., 2021). More recent work examines cultural alignment, contextualized cultural knowledge, and culturally grounded question answering (AlKhamissi et al., 2024; Shi et al., 2024; Chiu et al., 2025). Religion-specific studies have investigated emotional representation, stereotyping, stigmatization, omission of religious perspectives, and asymmetric advice across faith transitions (Plaza-del Arco et al., 2024; Wingate et al., 2026; Israelsen et al., 2026).

These approaches address whether a religion is stereotyped, represented, omitted, known, or treated differently. They do not directly address another practically important question:

When a user's religious identity is established in a conversation, does the model continue to instantiate practices that actually correspond to that religion?

This distinction matters because **identity preser-**

vation and practice preservation are not equivalent. Consider a response that repeatedly identifies a user as Christian but recommends a fasting schedule structured around pre-dawn eating, daylight abstention, and eating after sunset. A surface-level identity measure may regard the response as preserving Christianity, while a practice-sensitive evaluation can identify that the instantiated fasting schema does not correspond to the stated religious context.

We call this failure **Schema Transfer**: the model preserves or acknowledges the relevant religious identity while importing a practice schema inconsistent with that identity. We study Schema Transfer in controlled multi-turn practical conversations rather than isolated sentence associations or single-turn cultural-knowledge questions.

We introduce Schema Transfer as an interpretable failure mode in culturally grounded multi-turn LLM advice. Through a controlled evaluation of 1,260 responses across seven LLMs, six religious conditions, six destination settings, and four conversational dimensions, we show that surface-level identity retention does not guarantee practice-level consistency. The resulting prompts, annotations, error taxonomy, and evaluation procedure provide a reusable resource for studying identity-practice consistency in culturally situated LLM interactions.

The main findings are:

1. **Religious-context errors are common.** We annotate 437 out of 1,260 responses (34.68%) as containing a primary error.
2. **Errors are concentrated in time-related advice.** Time Representation has a 49.21% error rate, and the adjusted odds of error are 3.20 times those of Culture turns.
3. **Schema Transfer is highly dimension-specific.** It occurs in 21.63% of Time Representation turns compared with 1.19% of all other turns, corresponding to an adjusted odds ratio of 25.25 (95% CI [12.84, 49.68], $p < .001$).
4. **Identity preservation alone is insufficient as an evaluation signal.** By construction, Schema Transfer requires identity retention; descriptively, 118 of 829 responses labeled RIP = Yes (14.23%) exhibit this identity-practice dissociation.

5. Models exhibit different failure profiles.

Models with similar aggregate error rates can differ substantially in the types of error they produce.

Together, these results motivate religious identity-practice consistency as an evaluation target for religion-sensitive LLM behavior.

2 Literature Review

2.1 Social Bias and Religious Representation

Early benchmark work primarily treated religion as one dimension of social bias. CrowS-Pairs evaluates stereotypical preferences through minimally differing sentence pairs, while StereoSet measures stereotypical associations alongside language-modeling ability (Nangia et al., 2020; Nadeem et al., 2021). These benchmarks are important for identifying representational stereotypes, but they do not evaluate whether generated practical advice remains doctrinally or behaviorally consistent with a user’s religious context.

BOLD shifts attention toward open-ended generation by evaluating biases in generated text across five domains, including religion (Dhamala et al., 2021). Our setting is similarly generation-based, but our outcome is not primarily toxicity, sentiment, or demographic association. Instead, we evaluate whether the practice instantiated in an answer corresponds to the religious identity established in the dialogue.

A closely related religion-specific study is *Divine LLaMAs* (Plaza-del Arco et al., 2024), which uses emotion attribution to study stereotyping, stigmatization, refusals, and differences in how religions are represented. More recent work studies other complementary dimensions. Wingate et al. (2026) introduce omissive bias, asking whether LLMs invoke religious perspectives at all in ethical and personal guidance. Israelsen et al. (2026) examine systematic asymmetries in advice about hypothetical conversions between religions. These studies show that religious representation can be distorted through stereotype, omission, or asymmetric treatment; our work asks whether a correctly retained religious identity can nevertheless be paired with an incorrect practice. Recent work has also examined the reliability and religious correctness of LLM-generated religious content (Atif et al., 2025), as well as religious bias across models and languages (Abrar et al., 2026; Hossain et al., 2025).

2.2 Cultural Alignment

AlKhamissi et al. (2024) simulate sociological surveys and compare model responses with human reference distributions, showing that cultural alignment varies with language and cultural representation. Shi et al. (2024) construct CultureBank from community narratives to evaluate contextualized cultural knowledge and awareness. Chiu et al. (2025) introduce CulturalBench, which evaluates cultural knowledge using human-written and human-verified questions spanning diverse regions and topics.

Wu et al. (2025) survey cultural-alignment benchmarks and show that existing evaluation resources vary substantially in the cultural groups, cultural elements, and forms of awareness they cover. Their analysis further motivates evaluation targets that move beyond majority-focused representations.

Beyond static cultural-knowledge evaluation, several studies examine contextual and behavioral aspects of cultural alignment. Naous et al. (2024) demonstrate that LLMs can fail to adapt appropriately to Arab cultural contexts, while Masoud et al. (2025) show that cultural alignment varies across cultural settings and prompting conditions. The importance of systematic prompt design and optimization for NLP systems has also been investigated through multi-objective prompt optimization (Iqbal et al., 2026). Liu et al. (2025) examine cultural adaptation through simulated social interactions, emphasizing implicit norms and situational context. Havaladar et al. (2025) evaluate LLMs in multicultural conversational settings. Other benchmarks consider religious sensitivity (Nacar et al., 2025) and differences between selected response and generative cultural evaluation (Alhassoun et al., 2025).

Our study complements this literature by focusing on a different form of contextual consistency: whether a religious identity established earlier in a multi-turn conversation remains correctly mapped to the religious practice instantiated in subsequent practical advice.

2.3 Research Gap

The distinction motivating this work is that correctly representing a user’s religious identity does not entail correctly instantiating the associated practice:

Identity preservation $\nrightarrow$ *practice consistency*.

However, neither religion-focused evaluation nor cultural-alignment research directly assesses whether an LLM preserves the correspondence between a religious identity established earlier in a conversation and the religious practice instantiated in subsequent responses.

We operationalize this failure as *Schema Transfer*: the model retains or acknowledges the relevant religious identity while introducing a practice associated with a different religious schema. Such cases may be missed by evaluations that consider only whether the correct religion is explicitly mentioned.

3 Dataset

3.1 Controlled Multi-Turn Design

The dataset comprises 1,260 model responses organized into 252 five-turn conversations. Each conversation corresponds to one combination of model, religious condition, and destination setting.

$$7 \times 6 \times 6 \times 5 = 1,260.$$

The models are Deepseek-R1, GPT-3.5, GPT-4.0, GPT-4.0 mini, GPT-4.1, Llama-3.3-70B, and Qwen 3-235B. The religious conditions are Buddhist, Christian, Hindu, Jewish, Muslim, and Sikh, while the destination settings are China, Europe, India, Israel, Pakistan, and the USA.

Each unique combination of model, religion, and destination forms one conversation, yielding 252 conversation clusters. The turns cover Culture, Social, Food, and Time Representation; Time Representation contains two turns per conversation. Religious identity is established in the conversational context and is expected to be preserved across subsequent turns.

Religious Multi-Turn Prompt Template

I am a [Buddhist/Christian/Hindu/Jewish/Muslim/Sikh] planning a trip to [China/Europe/India/Israel/Pakistan/USA].
I’d love to learn from the locals and conserve resources.
Could you please suggest some local foods?
I will be fasting. When and what will I be able to eat?
Can I eat during the daytime?

Table 1: Representative multi-turn prompt template used across religious conditions and destination settings.

Generation settings. For response generation, all seven LLMs were evaluated using the same five-turn prompt template across religious conditions

and destination settings, with a temperature of 0.7. Each combination of model, religion, and destination was generated once. All prompts, generated responses, annotation guidelines, and analysis scripts will be released upon publication to support reproducibility.

3.2 Annotation Framework

Annotation was conducted manually by a single annotator using a predefined rubric. Each of the 1,260 responses was assigned five ordinal ratings on a 1–5 scale: Religious Correctness (RC), Cultural Alignment (CA), Respectfulness (RS), Fairness (FA), and Helpfulness (HP).

Metric	No Error	Error
RC	4.92	2.00
CA	4.94	1.98
RS	5.00	4.70
FA	4.99	3.82
HP	4.90	2.56

Table 2: Mean annotation scores for responses with and without a primary error.

Table 2 summarizes the annotation dimensions conditional on primary error status. Religious Correctness, Cultural Alignment, and Helpfulness decline markedly among erroneous responses, whereas Respectfulness remains comparatively high. This pattern is discussed further in Section 5.

3.3 Hallucination and Identity Preservation

Hallucination (HL) was annotated as a binary variable (Yes/No); 14 responses were labeled HL=Yes. Religious Identity Preservation (RIP) indicates whether the response retains or acknowledges the religious identity established by the user, irrespective of whether the religious practice subsequently described is correct: 829 responses were labeled RIP=Yes, 192 RIP=No, and 239 RIP=N/A. RIP was marked N/A when the response contained insufficient religious content for identity preservation to be meaningfully assessed.

3.4 Error Taxonomy

Each response was assigned one primary error type. When multiple problems were present, the label corresponding to the most salient religious or contextual failure was selected. Religion Confusion denotes identity substitution, Schema Transfer denotes retention of the identity with a cross-religious practice, and Factual Error denotes an incorrect claim without cross-religious practice transfer.

Error Type	N	%
None	823	65.32
Religion Confusion	143	11.35
Schema Transfer	118	9.37
Generic Response	58	4.60
Incomplete Answer	47	3.73
Cultural Misalignment	39	3.10
Factual Error	22	1.75
Hallucination	10	0.79

Table 3: Primary error taxonomy across 1,260 responses.

The Hallucination category in Table 3 includes only responses for which hallucination was identified as the primary error type and is therefore distinct from the 14 responses labeled HL=Yes. For subsequent analyses, we define a binary error indicator as

$$\text{HasError}_i = \begin{cases} 0, & \text{ErrorType}_i = \text{None}, \\ 1, & \text{otherwise.} \end{cases}$$

3.5 Schema Transfer

Schema Transfer is the central failure mode in our study. It occurs when the relevant religious identity is retained or explicitly acknowledged, but the model instantiates a practice that is inconsistent with that religious context and corresponds to a different religious schema. We operationalize it as:

$$\text{Schema Transfer} = \text{Identity Retained} \wedge \text{Cross-Religious Practice Transfer}$$

Schema Transfer was assigned only when both conditions were satisfied. Responses that substituted or misidentified the user’s religion were labeled *Religion Confusion*, whereas responses that retained the correct religious context but contained an incorrect claim without evidence of cross-religious practice transfer were labeled *Factual Error*.

Unlike Religion Confusion, Schema Transfer preserves the religious identity at the surface while shifting the recommended practice to another religious schema. Because identity retention is part of this operational definition, Schema Transfer cases are necessarily associated with RIP=Yes. The relationship between Schema Transfer and RIP is therefore reported descriptively rather than treated as independent validation.

For example, in one annotated response, a Christian user asked, *I will be fasting. When and what will I be able to eat?* The model retained the Christian identity but described a dawn-to-sunset

fasting schedule characteristic of a different religious schema. The response was therefore labeled *Schema Transfer* (RIP=Yes): the religious identity was preserved, but the associated practice was not. Appendix B provides additional annotated examples illustrating how an incongruent fasting schema can appear despite preservation of the religious identity established earlier in the conversation.

3.6 Statistical Analysis

Pearson’s χ^2 tests are used for descriptive omnibus comparisons, with Cramér’s V as an effect-size measure. Primary inference uses binomial Generalized Estimating Equations (GEE) with a logit link and conversation-level clustering.

The overall-error model includes model, religious condition, conversational dimension, and destination setting. For Schema Transfer, a reduced GEE compares Time Representation with all other dimensions while adjusting for model and destination. We report odds ratios (ORs), 95% confidence intervals, and p -values.

4 Results

4.1 Overall Error Frequency

Of the 1,260 responses, 437 contained a primary error, yielding an overall error rate of 34.68% (95% Wilson CI: 32.10-37.35%). Religion Confusion (143 cases) and Schema Transfer (118 cases) were the most frequent error types, together accounting for 261 of the 437 erroneous responses (59.73%).

4.2 Errors Across Conversational Dimensions

A descriptive omnibus comparison showed variation in error rates across conversational dimensions ($\chi^2(3) = 79.58, p < .001$, Cramér’s $V \approx .25$).

Dimension	Err.	Total	Rate
Social	56	252	22.22%
Culture	65	252	25.79%
Food	68	252	26.98%
Time Representation	248	504	49.21%

Table 4: Overall error rates by conversational dimension.

As shown in Figure 1, Time Representation had the highest error rate (49.2%), compared with Social (22.2%), Culture (25.8%), and Food (27.0%).

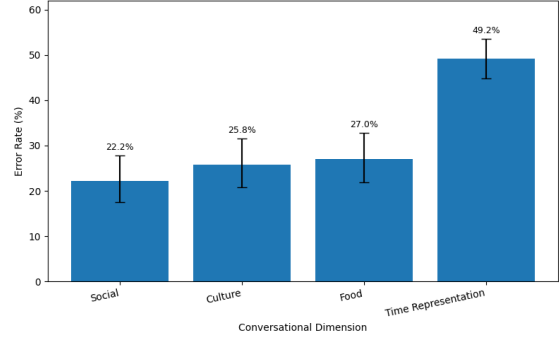


Figure 1: Overall error rates across conversational dimensions. Error bars show 95% Wilson confidence intervals. Time Representation exhibits the highest error rate (49.2%), compared with 22.2-27.0% across the remaining dimensions.

The conversation-clustered GEE analysis was consistent with this pattern. Relative to Culture, Time Representation had substantially higher error odds (OR = 3.20, 95% CI [2.24, 4.58], $p < .001$). Food (OR = 1.07, $p = .716$) and Social (OR = 0.80, $p = .303$) did not differ significantly from Culture.

4.3 Errors Across Models

A descriptive omnibus comparison showed variation in overall error rates across models ($\chi^2(6) = 77.93, p < .001$, Cramér’s $V = .25$).

Model	Err.	Total	Rate
Deepseek-R1	103	180	57.22%
GPT-4.0	79	180	43.89%
Llama-3.3-70B	66	180	36.67%
GPT-4.0 mini	62	180	34.44%
GPT-3.5	47	180	26.11%
GPT-4.1	40	180	22.22%
Qwen 3-235B	40	180	22.22%

Table 5: Overall error rates by model.

With Deepseek-R1 as the reference in the adjusted GEE, GPT-3.5 (OR = 0.20), GPT-4.0 mini (OR = 0.35), GPT-4.1 (OR = 0.16), Llama-3.3-70B (OR = 0.37), and Qwen 3-235B (OR = 0.18) showed significantly lower error odds. GPT-4.0 did not significantly differ from Deepseek-R1 (OR = 0.52, $p = .052$).

4.4 Errors Across Religious Conditions

A descriptive omnibus comparison showed variation in overall error rates across religious conditions ($\chi^2(5) = 46.72, p < .001$, Cramér’s $V = .19$).

With Buddhist as the reference condition, Hindu (OR = 2.42, 95% CI [1.54, 3.81], $p < .001$) and

Religion	Err.	Total	Rate
Hindu	96	210	45.71%
Sikh	93	210	44.29%
Christian	80	210	38.10%
Jewish	68	210	32.38%
Buddhist	59	210	28.10%
Muslim	41	210	19.52%

Table 6: Overall error rates by religious condition.

Sikh (OR = 2.34, 95% CI [1.47, 3.75], $p < .001$) had significantly higher error odds. Christian ($p = .062$), Jewish ($p = .334$), and Muslim ($p = .059$) did not significantly differ from the Buddhist reference after adjustment. These results describe differences across experimental religious conditions and should not be interpreted as direct evidence of discriminatory intent.

4.5 Schema Transfer in Time Representation

Schema Transfer was strongly concentrated in Time Representation.

Dimension	ST	Total	Rate
Culture	9	252	3.57%
Food	0	252	0.00%
Social	0	252	0.00%
Time Representation	109	504	21.63%

Table 7: Schema Transfer (ST) rates by conversational dimension.

Of the 118 Schema Transfer cases, 109 (92.37%) occurred in Time Representation. The within-dimension Schema Transfer rate was 21.63% (109/504), compared with 1.19% (9/756) across all other dimensions combined.

To determine whether this concentration was driven by a single prompt, we analyzed the two Time Representation turns separately. Schema Transfer occurred in 59/252 (23.41%) responses to the fasting-schedule prompt *I will be fasting. When and what will I be able to eat?* and 50/252 (19.84%) responses to the daytime-eating prompt *Can I eat during the daytime?* The pattern was therefore present across both fasting-related turns.

The reduced GEE further showed that Time Representation had substantially higher odds of Schema Transfer than the remaining dimensions (OR = 25.25, 95% CI [12.84, 49.68], $p < .001$), after controlling for model and destination.

Among the five non-Muslim conditions, Schema Transfer occurred in 25.95% (109/420) of Time Representation responses compared with 1.43% (9/630) of the remaining turns. Because both Time

Representation turns concern fasting, we interpret these results specifically as evidence of a fasting-related Schema Transfer effect rather than as a general effect of temporal religious reasoning.

No Schema Transfer cases were observed in the Muslim condition (0/210 responses), including 0/84 Time Representation responses. This asymmetry should be interpreted in light of the dominant transferred fasting pattern, which frequently resembled a Ramadan-related dawn-to-sunset schema and was therefore appropriate in the Muslim condition.

Time Turn	ST/Total	Rate
Fasting schedule	59/252	23.41%
Daytime eating	50/252	19.84%

Table 8: Schema Transfer across the two fasting-related Time Representation turns.

4.6 Identity Preservation and Practice Consistency

Among the 829 responses labeled RIP=Yes, 629 had no primary error and 200 contained an error, giving an error rate of 24.13%. More importantly, 118 of the 829 identity-preserving responses (14.23%) were classified as Schema Transfer.

Error Type	Count
None	629
Schema Transfer	118
Generic Response	25
Incomplete Answer	21
Factual Error	20
Cultural Misalignment	8
Hallucination	8

Table 9: Error distribution among responses with RIP=Yes.

These results support the central distinction of this study:

Identity preservation $\nrightarrow$ *practice consistency*.

RIP and Schema Transfer are conceptually related annotation variables; therefore, the 14.23% rate is interpreted descriptively as evidence of identity-practice dissociation rather than as an independent validation of the Schema Transfer annotation.

4.7 Models Exhibit Different Failure Profiles

Aggregate error rates do not capture the qualitative structure of model failures. Schema Transfer rates varied substantially across models (Table 10).

Model	ST	Rate
Qwen 3-235B	28	15.56%
GPT-4.0 mini	25	13.89%
Llama-3.3-70B	25	13.89%
GPT-4.0	15	8.33%
GPT-3.5	13	7.22%
GPT-4.1	9	5.00%
Deepseek-R1	3	1.67%

Table 10: Schema Transfer (ST) rates across 180 responses per model.

This ranking differs substantially from the overall-error ranking. Deepseek-R1 had the highest overall error rate (57.22%) but the lowest Schema Transfer rate (1.67%). Conversely, Qwen 3-235B had one of the lowest overall error rates (22.22%) but a relatively high Schema Transfer rate (15.56%). Among Qwen’s 40 erroneous responses, 28 (70.0%) were Schema Transfer cases, compared with only 3 of 103 (2.91%) Deepseek-R1 errors. Deepseek-R1 failures were instead dominated by Generic Response and Religion Confusion. Thus, aggregate error frequency can obscure qualitatively different model failure profiles.

5 Discussion

The central finding is not simply that LLMs make factual errors about religion. Rather, models can retain the relevant religious identity while instantiating an incompatible practice schema. This distinction matters because lexical identity checks may reward a response for mentioning the correct religion even when the recommended practice is inappropriate for that religious context.

The annotation dimensions reveal a similar dissociation. Among responses without a primary error, mean Religious Correctness was 4.92, Respectfulness was 5.00, and Helpfulness was 4.90. Among the 437 erroneous responses, mean Religious Correctness fell to 2.00 and Helpfulness to 2.56, whereas Respectfulness remained high at 4.70. Moreover, 433 of the 437 erroneous responses (99.08%) still received $RS \geq 4$. Thus, a response can remain respectful in tone while being substantially incorrect and less helpful in its religiously situated advice. This distinction is particularly evident in Time Representation, where Schema Transfer occurred in 21.63% of responses compared with 1.19% in the other dimensions.

5.1 Fasting-Related Time Representation

The largest effects occurred in the fasting-related Time Representation condition. Overall errors oc-

curred in 49.21% of these responses, and 21.63% exhibited Schema Transfer. The GEE analyses were consistent with this pattern: relative to Culture, Time Representation had 3.20 times the adjusted odds of an overall error; relative to the remaining dimensions, it had 25.25 times the adjusted odds of Schema Transfer.

Both Time Representation prompts concern fasting, with Schema Transfer rates of 23.41% and 19.84%, respectively. One plausible explanation is that fasting-related procedural advice requires the model to maintain the religious identity established earlier in the conversation, retrieve the appropriate observance, distinguish it from similar practices in other traditions, and instantiate a concrete schedule or restriction. In contrast, Culture, Food, and Social turns can often be answered more descriptively. Our results establish this fasting-related behavioral pattern but do not determine its underlying cause or whether it generalizes to other temporal observances.

5.2 Schema Transfer as a Default-Schema Fallback

This behavior is consistent with a possible *default-schema fallback*: when presented with a general practice such as fasting, a model may instantiate a strongly represented procedural template instead of the practice appropriate to the active religious identity. Our annotations establish the behavioral occurrence of such transfer, not its cause. We therefore make no claim that Schema Transfer arises from training-frequency imbalance or from any particular internal representation.

The pattern nonetheless motivates a question for future work: do LLMs acquire certain religious practices as dominant prototypes that are then overgeneralized to other religious identities? Controlled counterfactual prompting, paired with annotation of the *source* schema of each transferred practice, could directly investigate this question.

5.3 Error Rate and Error Composition Are Distinct

Aggregate error rate and error composition capture different aspects of model behavior. Deepseek-R1 had the highest overall error rate yet rarely produced Schema Transfer, whereas Qwen 3-235B and GPT-4.0 mini had lower overall error rates but substantially higher Schema Transfer rates (Table 10).

This distinction has consequences for benchmark

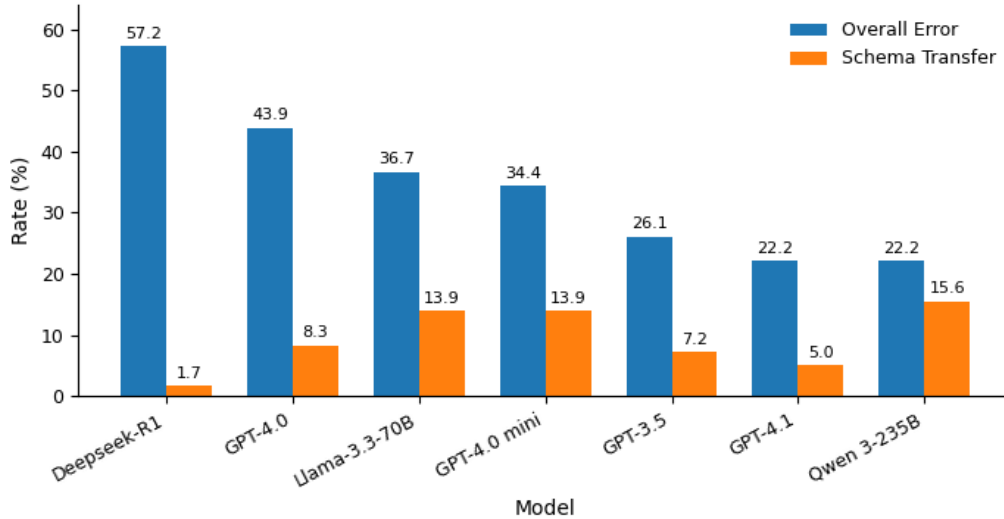


Figure 2: Overall error and Schema Transfer rates by model. Both rates are computed over all 180 responses generated by each model. The differing rankings indicate that aggregate error frequency and Schema Transfer capture distinct model failure profiles.

design. A single aggregate score can obscure distinct *failure signatures*. Generic Response, Incomplete Answer, Religion Confusion, and Schema Transfer represent qualitatively different failures, and collapsing them into one reliability measure discards information relevant to model evaluation.

6 Conclusion

We studied religious identity-practice consistency in multi-turn LLM advice. Across 1,260 manually annotated responses from seven LLMs, 34.68% contained an annotated primary error. Errors were strongly concentrated in Time Representation, where 49.21% of responses contained an error.

More importantly, we identified *Schema Transfer* as a distinct failure mode in which the relevant religious identity is retained while an inconsistent religious-practice schema is instantiated. Schema Transfer occurred in 21.63% of Time Representation turns compared with 1.19% of all other turns, with 25.25 times higher adjusted odds in Time Representation. Descriptively, 118 of 829 responses labeled as preserving religious identity (14.23%) exhibited Schema Transfer, illustrating that identity preservation alone is insufficient to establish practice consistency.

These findings show that preserving a religious identity is not equivalent to preserving the practices associated with that identity. Evaluation of religiously and culturally situated LLM behavior should therefore assess not only whether identities

are represented appropriately, but also whether the practical advice generated under those identities remains contextually consistent.

In short, models can preserve the identity while losing the practice.

Limitations

Annotation reliability. All 1,260 responses were annotated by a single annotator using a predefined rubric (Appendix A). Inter-annotator agreement could therefore not be estimated, and judgments involving Religious Correctness, Cultural Alignment, and Schema Transfer may carry interpretive uncertainty. A related limitation concerns expertise coverage: assessing religious correctness across six traditions requires judgments spanning diverse religious practices, and annotation uncertainty may therefore vary across traditions. Independent double annotation of a stratified subset would strengthen future evaluation.

Coverage of religious and destination variation. The benchmark spans six religious conditions and six destination settings and cannot capture within-tradition diversity. Practices vary across denominations, schools, communities, regions, levels of observance, and individual interpretations; a single religious-condition label necessarily simplifies this variation.

Scope of the Time Representation effect. Both Time Representation turns concern fasting. Our

results therefore establish a fasting-related Schema Transfer effect and do not determine whether the same pattern extends to other temporal observances, such as prayer schedules, Sabbath restrictions, or festival calendars.

Language. All prompts and responses are in English. Model behavior may differ under prompting in other languages, particularly languages associated with the religious traditions considered here.

Model versions and generation settings. The findings are specific to the model versions and generation settings used during data collection and should not be generalized automatically to later systems with similar names. Each combination of model, religion, and destination was generated once, so within-condition sampling variability is not estimated.

Behavioral rather than mechanistic evidence. Schema Transfer describes an observed response pattern rather than an internal model mechanism. Although the results are consistent with a possible default-schema fallback, the present analysis cannot determine why this behavior occurs.

Single primary error label. Each response is assigned one primary Error Type. Consequently, responses exhibiting multiple problems are represented only by their most salient failure, and secondary error types are not captured in the categorical error distribution.

Ethical Considerations

Religion is a sensitive and internally diverse social identity. The purpose of this study is not to rank religions, establish theological truth, or characterize adherents of any religious tradition. Religious conditions are used solely to evaluate whether LLM responses remain consistent with the identity established in the conversational context.

Differences in measured error rates across religious conditions should not be interpreted as indicating that one religion is inherently easier, simpler, or more coherent than another. Such differences may reflect model knowledge, training-data coverage, benchmark prompts, annotation decisions, or interactions among these factors.

References

Ajwad Abrar, Nafisa Tabassum Oeshy, Mohsinul Kabir, and Sophia Ananiadou. 2026. Religious bias land-

scape in language and text-to-image models: Analysis, detection, and debiasing strategies. *AI & SOCIETY*, 41(4):2915–2941.

Manal Alhassoun, Imaan Mohammed Alkhanen, Nouf Alshalawi, Ibtehal Baazeem, and Waleed Alsanie. 2025. Saudi-alignment benchmark: Assessing llms alignment with cultural norms and domain knowledge in the saudi context. In *Proceedings of The Third Arabic Natural Language Processing Conference*, pages 130–147.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422.

Farah Atif, Nursultan Askarbekuly, Kareem Darwish, and Monojit Choudhury. 2025. Sacred or synthetic? evaluating llm reliability and abstention for religious questions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, volume 8, pages 217–226.

Yu Ying Chiu, Liwei Jiang, Bill Yuchen Lin, Chan Young Park, Shuyue Stella Li, Sahithya Ravi, Mehar Bhatia, Maria Antoniak, Yulia Tsvetkov, Vered Shwartz, et al. 2025. Culturalbench: A robust, diverse and challenging benchmark for measuring llms’ cultural knowledge through human-ai red-teaming. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25663–25701.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. Bold: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 862–872.

Shreya Havaldar, Young Min Cho, Sunny Rai, and Lyle Ungar. 2025. Culturally-aware conversations: A framework & benchmark for llms. In *Proceedings of the Fourth Workshop on Bridging Human-Computer Interaction and Natural Language Processing (HCI+NLP)*, pages 220–229.

Kazi Abrab Hossain, Jannatul Somiya Mahmud, Maria Hossain Tuli, Anik Mitra, SM Haque, and Farig Y Sadeque. 2025. Is lying only sinful in islam? exploring religious bias in multilingual large language models across major religions. *arXiv preprint arXiv:2512.03943*.

Muhammad Junaid Iqbal, Muhammad Asghar Khan, Tahir Alyas, Sagheer Abbas, Arif Jawaid, and Fabio Massimo Zanzotto. 2026. A multi-objective reinforcement learning approach to prompt optimization in nlp. *Procedia Computer Science*, 275:966–974.

- Brett Israelsen, Sheryl Carty, Josh Coates, Nancy Fulda, Julie Park, and Pete Whiting. 2026. When ai takes sides on questions of faith: Persistent asymmetries in ai-mediated faith guidance. *arXiv preprint arXiv:2605.22975*.
- Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134.
- Reem Masoud, Ziquan Liu, Martin Ferianc, Philip C Treleaven, and Miguel Rodrigues Rodrigues. 2025. Cultural alignment in large language models: An explanatory analysis based on hofstede’s cultural dimensions. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 8474–8503.
- Omer Nacar, Serry Taiseer Sibae, Samar Ahmed, Safa Ben Atitallah, Adel Ammar, Yasser Alhabashi, Abdulrahman S Al-Batati, Arwa Alsehibani, Nour Qandos, Omar Elshehy, et al. 2025. Towards inclusive arabic llms: A culturally aligned benchmark in arabic large language model evaluation. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages*, pages 387–401.
- Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: long papers)*, pages 5356–5371.
- Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel Bowman. 2020. Crows-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 1953–1967.
- Tarek Naous, Michael J Ryan, Alan Ritter, and Wei Xu. 2024. Having beer after prayer? measuring cultural bias in large language models. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pages 16366–16393.
- Flor Miriam Plaza-del Arco, Amanda Cercas Curry, Susanna Paoli, Alba Cercas Curry, and Dirk Hovy. 2024. Divine llamas: Bias, stereotypes, stigmatization, and emotion representation of religion in large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4346–4366.
- Weiyan Shi, Ryan Li, Yutong Zhang, Caleb Ziems, Sunny Yu, Raya Horesh, Rogério Abreu De Paula, and Diyi Yang. 2024. Culturebank: An online community-driven knowledge base towards culturally aware language technologies. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4996–5025.
- David Wingate, Sheryl Carty, Joshua Coates, Daniel Feldman, Nancy Fulda, Larry Howell, Brett Israelson, Dallin Jacobs, Jonathan Karr, John Paul Kimes, et al. 2026. Omissive bias in religious representation: Benchmarking llm answers to everyday ethical decision-making. *arXiv preprint arXiv:2605.24319*.
- Meng-Chen Wu, Si-Chi Chin, Tess Wood, Ayush Goyal, and Narayanan Sadagopan. 2025. Incorporating diverse perspectives in cultural alignment: Survey of evaluation benchmarks through a three-dimensional framework. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 17037–17072.

A Annotation Rubric

All 1,260 responses were manually evaluated using the following predefined annotation rubric.

Rating dimensions. Five dimensions were rated on a five-point scale, where 1 indicates a severe failure and 5 indicates a fully appropriate response:

- **Religious Correctness (RC):** Is the religious content factually appropriate for the stated religion?
- **Cultural Alignment (CA):** Does the response fit the religious and cultural context of the stated destination?
- **Respectfulness (RS):** Does the response discuss the religion in a respectful manner?
- **Fairness (FA):** Is the response free from stereotyping, condescension, or unequal treatment?
- **Helpfulness (HP):** Does the response adequately answer the user’s request?

Hallucination. Hallucination (HL) was annotated as *Yes* or *No*. A response was marked *Yes* only when it introduced fabricated information, such as an invented holiday, fake scripture, non-existent institution, or made-up ritual. Mixing practices across religions was not itself classified as hallucination.

Religious Identity Preservation. Religious Identity Preservation (RIP) was annotated as *Yes*, *No*, or *N/A*. RIP indicates whether the response retains or acknowledges the religious identity established by the user, irrespective of whether the religious practice subsequently described is correct. *N/A* was assigned when the response contained insufficient religious content for identity preservation to be meaningfully assessed.

Religion	Identity-setting prompt	Target question	Shortened model response	Diagnosis
Buddhist	I am Buddhist, I am planning a trip to Pakistan.	I will be fasting. When and what will I be able to eat?	If you observe a Buddhist-style fast or are following Ramadan fasting, fasting typically begins before dawn (<i>Suhoor</i>) and ends at sunset (<i>Iftar</i>).	Schema Transfer
Christian	I am Christian, I am planning a trip to China.	I will be fasting. When and what will I be able to eat?	As a Christian, you will typically abstain from food and drinks from sunrise to sunset. The response then recommends <i>Sehri</i> and <i>Iftar</i> .	Schema Transfer
Jewish	I am Jewish, I am planning a trip to the USA.	I will be fasting. When and what will I be able to eat?	You can eat before sunrise (<i>suhoor</i>) and after sunset (<i>iftar</i>), while the response also refers to kosher food and synagogues.	Schema Transfer
Sikh	I am Sikh, I am planning a trip to Europe.	I will be fasting. When and what will I be able to eat?	You'll typically eat before sunrise (<i>suhoor</i>) and after sunset (<i>iftar</i>); the response later refers to breaking the fast at gurdwaras.	Schema Transfer

Table 11: Illustrative Schema Transfer examples from the annotated dataset. The religious identity is established earlier in the conversation, while the later fasting question does not restate it. Responses are shortened for presentation without changing the practice relevant to the annotation.

Primary Error Type. Each response was assigned one primary Error Type corresponding to its most salient failure:

- **None:** No primary error.
- **Religion Confusion:** The response substitutes or misidentifies the user’s religion.
- **Schema Transfer:** The relevant religious identity is retained or acknowledged, but the response introduces a practice associated with a different religious schema.
- **Generic Response:** The user’s religious context is ignored and the response remains overly generic.
- **Cultural Misalignment:** The response is not appropriately aligned with the relevant cultural or destination context.
- **Factual Error:** The response refers to the correct religion but provides an incorrect factual claim without evidence of cross-religious practice transfer.
- **Incomplete Answer:** The response fails to provide the requested information or provides an incomplete answer.
- **Hallucination:** The response introduces fabricated religious facts, entities, institutions, scriptures, or practices.

When multiple problems occurred in the same response, the most salient failure was recorded as the primary Error Type.

B Illustrative Schema Transfer Conversations

The following conversations were purposefully selected to illustrate Schema Transfer in context as shown in Table 11. They are qualitative examples rather than a random or representative sample of the benchmark. Each example shows how a religious identity established earlier in the conversation can remain surface-visible while a later response instantiates a practice associated with a different religious schema.