

Epistemic Framing Tilt in LLMs: Auditing Religious Discourse Across Traditions

Arthur Buzelin, Yan Aquino, Pedro Bento, Arthur Chagas,
Lucas Dayrell, Gisele L. Pappa, Virgilio Almeida, Wagner Meira Jr.

Universidade Federal de Minas Gerais

{arthurbuzelin, yanaquino, pedro.bento, arthurchagas, lucasdayrell, glpappa,
virgilio, meira}@dcc.ufmg.br

Abstract

Large language models do not only answer questions about religion; they also frame those answers through choices of assertion, attribution, grounding, qualification, and discursive register. We study this phenomenon as *Epistemic Framing Tilt*: systematic asymmetry in how LLMs present parallel religious claims across traditions. To audit this behavior, we operationalize four interpretable framing axes: assertive commitment vs. belief-attributed distancing, doctrinal/scriptural grounding vs. generic believer attribution, non-mythologized vs. mythologized/symbolic framing, and unqualified assertion vs. hedged qualification. For each axis, we construct validated minimal pairs, learn contextual model-internal directions, and evaluate them through repeated held-out validation, human annotation on generated outputs, lexical cue ablation, and placebo controls. We then apply the learned directions to a matched audit spanning six religious traditions, 24 prompt templates, and multiple semantic domains, including cosmology, sacred texts, ritual, afterlife, and historical origins. Our results show that religious discourse is not framed uniformly: differences are axis-specific, prompt-sensitive, and only partly shared across model families. Some traditions are described more directly, others more through attribution, qualification, or mythic/symbolic register, even under parallel prompts. We argue that these differences matter not because any one framing style is universally preferable, but because repeated asymmetries in epistemic presentation may shape how traditions become legible to users in AI-mediated settings.

Code and Data — <https://github.com/pedroaugtb/traditions>

Introduction

Ask a language model who Zeus is, and the answer will usually be framed with explicit distance: Zeus is described as a figure from mythology, from ancient Greek legend, or as part of a symbolic narrative, rather than as someone presented in straightforward factual terms. That framing often feels natural, even invisible. Yet it reveals an important feature of language models: they do not merely provide information, but also position that information, signaling whether it should be read as fact, belief, interpretation, legend, doctrine, or symbolic account.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

This matters beyond mythology. In domains such as religion, politics, culture, history, and contested knowledge, models may describe structurally similar claims with different degrees of certainty, attribution, grounding, or distance (Kay, Kasirzadeh, and Mohamed 2024; Helm et al. 2024). Some claims may be presented through direct exposition or institutional language, while others are qualified, attributed to communities, placed in a symbolic register, or described through mythic vocabulary.

We call this phenomenon *Epistemic Framing Tilt*: systematic asymmetry in how models frame parallel topics, traditions, or claims. The term refers to measurable variation not only in the content of an answer, but in the stance through which that answer is delivered. In this paper, we study Epistemic Framing Tilt through a comparative audit of religious discourse, asking whether models systematically differ in how they present parallel religious claims across traditions.

To study this phenomenon, we introduce a framing-centered audit based on four controlled axes: assertive commitment vs. belief-attributed distancing, doctrinal/scriptural grounding vs. generic believer attribution, non-mythologized vs. mythologized/symbolic framing, and unqualified assertion vs. hedged qualification. For each axis, we construct validated minimal pairs, learn model-internal directions from these contrasts, and test their robustness through repeated holdout evaluation. We then apply the learned directions to a parallel prompt audit spanning multiple religions, thematic question groups, and open language models.

Our contributions are threefold. First, we introduce *Epistemic Framing Tilt* as a construct for analyzing how LLMs position parallel claims through language, not only what they say about them. Second, we operationalize this construct through interpretable framing axes that can be validated, learned in contextual representation space, and applied comparatively across models. Third, we show that, even under matched prompts, religious discourse is organized differently across traditions, semantic domains, and model families: some traditions are more often narrated through direct exposition, others through attribution, qualification, or mythic and symbolic framing. Taken together, these results suggest that LLMs differ not only in informational content, but in the discursive terms through which traditions are made intelligible.

Related Work

Prior work has shown that language models reproduce social biases in both likelihood-based evaluations and open-ended generation. Surveys frame bias in NLP as a broad sociotechnical problem involving stereotypes, representational harms, and unequal treatment across groups (Blodgett et al. 2020; Gallegos et al. 2024). Recent benchmarks measure biases across demographic categories, including religion, through minimal pairs, question answering, or generated continuations (Nadeem, Bethke, and Reddy 2021; Nangia et al. 2020; Parrish et al. 2022; Dhamala et al. 2021; Smith et al. 2022). These works establish that LLMs encode socially patterned asymmetries. Our work differs by focusing not on stereotypes or toxicity, but on how models distribute epistemic legitimacy through framing.

Religion appears in several bias benchmarks, but is often treated as one category among many. Recent work shows persistent anti-Muslim bias in GPT-3, especially through associations between Muslim identity and violence (Abid, Farooqi, and Zou 2021). Other benchmarks include religion-related prompts or identity terms to evaluate stereotypical, negative, or harmful associations (Nadeem, Bethke, and Reddy 2021; Nangia et al. 2020; Dhamala et al. 2021). These studies mainly examine overt or associative harms. We instead study subtler framing asymmetries: whether parallel religious claims are presented as factual, doctrinal, symbolic, mythological, hedged, or merely attributed to believers.

Our work also builds on NLP research on factuality, modality, evidentiality, and speaker commitment. Factbank study annotates whether events are presented as factual, uncertain, counterfactual, or attributed (Saurí and Pustejovsky 2009). CommitmentBank studies whether speakers are committed to embedded propositions in naturally occurring discourse (de Marneffe, Simons, and Tonhauser 2019). Related work on committed belief and factuality prediction similarly distinguishes direct assertion from attribution or uncertainty (Werner et al. 2015; Stanovsky et al. 2017). These ideas motivate our framing axes: assertive commitment, doctrinal grounding, mythologization, and hedging.

The paper is also motivated by theories of epistemic injustice and religious classification. Work on epistemic injustice argues that social identity can shape whose testimony is granted credibility and whose interpretive resources become publicly available (Fricker 2007). Religious studies further shows that categories such as religion, myth, doctrine, and belief are historically and politically constructed rather than neutral descriptors (Asad 1993; Lincoln 1999; McCutcheon 1997). This is central to our mythologization axis: the concern is not whether a claim is true, but whether some traditions are systematically placed in a downgraded symbolic or mythic register.

Methodologically, our approach relates to representation engineering and activation-based auditing. Representation Engineering identifies directions in activation space associated with high-level model behaviors (Zou et al. 2025), while Activation Addition shows that contrastive prompts can define vectors that modify generation without fine-tuning (Turner et al. 2025). Related work uses interpretable

semantic axes to probe learned representations (Bento et al. 2026), while latent-knowledge methods study information not directly reflected in surface outputs (Burns et al. 2023). We extend this logic to discourse framing by learning axis directions from controlled minimal pairs and using them for comparative auditing.

Existing work has studied stereotypes, toxicity, factuality, commitment, and activation-level control. However, less attention has been given to how LLMs assign different epistemic positions to religious traditions through their wording. Our contribution is to operationalize this phenomenon as *Epistemic Framing Tilt*: systematic variation in whether religious claims are asserted, attributed, grounded, mythologized, or hedged.

Methodology

We study whether language models frame parallel religious claims differently across traditions. Our methodology treats this as a problem of *epistemic framing*: how a model positions a claim with respect to commitment, source, genre, and certainty. The pipeline combines controlled linguistic contrasts, human validation, contextual axis learning, and comparative prompt-based auditing.

Axis Construction

For each axis, we constructed a dataset of 72 paired statements, yielding 288 pairs in total. Each pair contrasts the two poles of a single framing dimension while preserving the underlying propositional content as closely as possible. Pairs were written to minimize changes in topic, referential content, and sentence structure, so that the primary variation lies in the discourse operation of interest. The pair sets span multiple religions and topical domains while remaining tightly controlled at the sentence level. The full list with all pairs can be found in the accompanying repository.

Axis 1: Assertive commitment vs. belief-attributed distancing. This axis captures the locus of commitment. At the assertive pole, a claim is presented directly, as in “God created the world.” At the contrasting pole, the same claim is attributed to a religious community, as in “Christians believe that God created the world.”

Axis 2: Doctrinal/scriptural grounding vs. generic believer attribution. This axis captures source type. At one pole, the claim is anchored in recognized internal sources such as scripture, doctrine, sacred teaching, formal tradition, oral teaching, or named religious authority. At the contrasting pole, the same proposition is attributed only to a generic group of believers.

Axis 3: Non-mythologized framing vs. mythologized/symbolic framing. This axis captures discursive genre. At one pole, a claim is presented as doctrine, sacred teaching, religious account, or authoritative tradition. At the other, it is described as myth, legend, folklore, symbolic narrative, or mythological account.

Axis 4: Unqualified assertion vs. hedged qualification. This axis captures modal and epistemic qualification. At one

pole, a proposition is stated without explicit hedging. At the contrasting pole, it is marked by expressions of uncertainty, interpretive distance, or impersonal attribution, such as “it is believed that,” “may be understood as,” or “is said to.”

Human Validation of Pair Sets

To validate the pair sets, we used independent human annotation. The purpose of this stage is to validate the linguistic contrasts, not to determine whether the religious content is doctrinally correct. Each pair was evaluated by three annotators using a four-part rubric: (i) whether the positive-pole sentence correctly instantiates the intended axis, (ii) whether the negative-pole sentence correctly instantiates the intended axis, (iii) whether the pair preserves the same underlying proposition, and (iv) how strong the contrast is between the two sides. The first three criteria were rated on a three-point ordinal scale, and contrast strength was rated on a five-point ordinal scale.

Three graduate CS students were recruited as annotators for a structured discourse-annotation task rather than for doctrinal or theological evaluation. The protocol focused on observable linguistic properties such as attribution, commitment, qualification, source grounding, and discursive register. Before annotation, annotators received written guidelines, axis definitions, examples of valid and invalid contrasts, and a short reading packet on factuality, attribution, speaker commitment, hedging, and related framing phenomena. The instructions emphasized linguistic form rather than truth, adequacy, or theological correctness.

Because ratings were concentrated near the top of the scale for well-formed pairs, we use *Gwet’s AC2* as the primary inter-annotator reliability coefficient (Gwet 2008), complemented by exact agreement and adjacent agreement. *Gwet’s AC2* is appropriate here because it is more stable than chance-corrected coefficients such as kappa or Krippendorff’s alpha under strong prevalence and ceiling effects. Across all axes combined, agreement was high: *Gwet’s AC2* reached 0.9692 for *positive_pole_correct*, 0.9437 for *negative_pole_correct*, 0.9341 for *proposition_preserved*, and 0.8676 for *contrast_strength*.

Models

Our main audit uses open language models from distinct model families and institutional origins: Qwen/Qwen2.5-7B, google/gemma-2-9b-it, meta-llama/Llama-3.2-3B-Instruct, and allenai/OLMo-2-1124-7B-Instruct. We selected these models to reduce dependence on a single architecture or post-training pipeline and to test whether framing patterns generalize across model families. We restricted the model scale to systems that could be practically run on a single NVIDIA RTX 4090-class GPU.

Contextual Text Representation

All axis learning, scoring, and downstream analyses use contextual hidden-state representations from the evaluated model itself. Given a text x , we tokenize it without adding special tokens and run a forward pass through the model with hidden-state outputs enabled. We extract the final hidden layer and mean-pool across token positions: $\gamma(x) =$

$\frac{1}{T} \sum_{t=1}^T h_t^{(L)}$, where $h_t^{(L)}$ is the final-layer hidden state at token position t .

Mean pooling is a simple sentence-level aggregation strategy: it does not isolate a single linguistic feature, but it provides a compact representation of how the model encodes the full response. We use this representation because epistemic framing often depends on sentence-level context, attribution structure, and combinations of lexical cues rather than on individual words alone.

Learning Axis Directions

For each axis, let $\{(x_i^+, x_i^-)\}_{i=1}^N$ denote the paired statements, where x_i^+ and x_i^- instantiate the two poles of the axis. For the contextual representation function $\gamma(\cdot)$, we compute pairwise difference vectors: $d_i = \gamma(x_i^+) - \gamma(x_i^-)$. The raw axis vector is the mean difference over the training pairs: $g = \frac{1}{N_{\text{train}}} \sum_{i \in \mathcal{T}} d_i$.

To reduce sensitivity to high-variance directions in hidden-state space, we use a regularized metric-aware normalization. We estimate a covariance matrix Σ from a random sample of 20,000 rows of the model’s output embedding matrix and regularize it as $\Sigma_\epsilon = \Sigma + \epsilon \cdot \text{diag}(\Sigma)I$, with $\epsilon = 10^{-3}$. Let $M = \Sigma_\epsilon^{-1}$. We define the metric norm of the raw axis vector as $\|g\|_M = \sqrt{g^\top M g}$, normalize it as $\bar{g} = \frac{g}{\|g\|_M}$, and obtain the scoring vector $\lambda = M\bar{g}$.

For validation on pair differences, a pair receives score $s_i = d_i^\top \lambda$. Positive scores correspond to the first pole named in the axis definition: more assertive commitment, stronger doctrinal/scriptural grounding, less mythologized framing, and less hedged qualification. For generated responses in the religious audit, we use the Euclidean-normalized version of the scoring vector, $\lambda_E = \frac{\lambda}{\|\lambda\|_2}$, and score each response relative to a seed-specific midpoint, as described below.

Axis Validation, Baselines, and Effect Sizes

We validate each axis using repeated random train/validation splits over the 72 paired statements. For each axis and model, we sample 30 random seeds from a fixed base seed. Each seed defines a new 70/30 split, yielding 50 training pairs and 22 held-out validation pairs. The axis direction is learned only from the training pairs for that seed and then evaluated on the held-out pairs.

We report three validation quantities. First, we compute the mean pair score on the training split. Second, we compute leave-one-out scores on the training split: for each training pair, the axis is recomputed without that pair and then evaluated on the omitted pair. Third, we compute held-out validation scores on the validation split. For each set of scores, we report the mean and sign accuracy, where sign accuracy is the fraction of pairs whose score is positive: $\text{SignAcc} = \frac{1}{n} \sum_{i=1}^n 1[s_i > 0]$.

To contextualize validation scores, we compare each learned direction against a random lexical baseline. For each seed, we sample 512 random pairs of vocabulary items from the model’s output embedding matrix and compute $r_{ab} = (W_a - W_b)^\top \lambda$, where W_a and W_b are output-embedding rows and λ is the learned scoring vector for that seed. This

baseline estimates how large scores would be for arbitrary token-level differences under the same learned direction.

We report effect size relative to the random baseline using a standardized mean difference: $d_{\text{baseline}} = \frac{\mu_{\text{pairs}} - \mu_{\text{baseline}}}{\sqrt{\frac{1}{2}(\sigma_{\text{pairs}}^2 + \sigma_{\text{baseline}}^2)}}$. We summarize validation metrics across the 30 random splits and compute nonparametric bootstrap confidence intervals over the per-seed metric distribution using 500 bootstrap resamples and a 95% confidence level. These intervals characterize split-level variability induced by the finite pair set.

Comparative Religious Audit

After learning axis directions, we audit generated responses about six religious traditions: Christianity, Islam, Judaism, Buddhism, Hinduism, and Yoruba religion. The audit uses 24 prompt templates organized into six semantic groups: origin of the world, divine or ultimate reality, sacred texts or teachings, after death, ritual or worship, and origins or early formation. Each group contains four paraphrased templates, and each template is instantiated once per religion. All prompts are issued in English.

The prompt set is designed to keep the semantic target stable across religions while allowing limited variation in wording. The six groups cover recurrent themes in religious description: cosmology, metaphysics, scripture or oral teaching, afterlife, ritual practice, and historical formation. The templates are intentionally broad rather than tradition-specific so that they can apply to traditions with different structures of authority, textuality, and practice.

For each religion-prompt combination, we sample five responses using stochastic decoding with temperature 0.7, top- $p = 0.9$, and a maximum of 220 new tokens. Empty generations are retried up to three times. If no usable continuation is produced after all retries, the generation is marked as failed and excluded from scoring. This yields up to $6\text{religions} \times 24\text{prompts} \times 5\text{generations} = 720$ generated responses per model in the main audit.

For each axis seed, we compute a seed-specific midpoint from the training pairs: $m = \frac{1}{2} \left(\frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \gamma(x_i^+) + \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \gamma(x_i^-) \right)$. A response y is then scored as $s(y) = \langle \gamma(y) - m, \lambda_E \rangle$, where λ_E is the Euclidean-normalized scoring vector for that seed. We compute one score per axis seed and then summarize scores at the generation level by averaging over seeds. We aggregate generation-level mean scores by religion, prompt group, religion-prompt-group pair, model, and axis.

Aggregation and Uncertainty

The primary unit of analysis in the religious audit is the generated response. Each response receives a generation-level score equal to the mean score across the 30 learned axis seeds. We summarize these generation-level scores using descriptive statistics such as the mean, standard deviation, median, and fraction of positive scores.

Confidence intervals are computed over generation-level mean scores using nonparametric bootstrap resampling with

500 resamples and a 95% confidence level. These intervals summarize variability across sampled generations and prompt templates after averaging over axis seeds. They do not jointly model uncertainty over both axis seeds and generations, and they should not be interpreted as population-level uncertainty over all possible prompts, model deployments, or religious discourse.

Because the audit produces many comparisons across axes, religions, prompt groups, and models, we treat the analysis as descriptive rather than as a collection of independent null-hypothesis tests. In the main text, we therefore emphasize patterns that are stable across aggregation levels and consistent with human validation.

Human Validation on Generated Outputs

To assess whether learned axis scores correspond to human-perceived framing in model outputs, we conduct a direct validation on generated responses. Starting from the response pool produced in the comparative audit, we select a balanced sample spanning models, religions, prompt groups, and score regions. For each axis, the sample includes high-scoring, low-scoring, and intermediate cases.

Each sampled response is rated independently by three annotators. Annotators are shown only the prompt, the generated response, and an axis-specific rubric; they do not see the model score or selection bucket. Ratings are collected on a five-point ordinal scale, with higher values corresponding to the first pole named in the axis definition: more direct assertion for assertive commitment, stronger doctrinal or scriptural grounding for doctrinal attribution, less mythologized framing for mythologization, and less hedged or less qualified assertion for hedging. Responses judged to be malformed or unusable can be marked and excluded from analysis.

Inter-annotator agreement on the generated-output validation set was high, with Gwet’s AC2 equal to 0.72. We aggregate human judgments at the item level and compare them to automatic contextual scores using Spearman correlation. The contextual representation shows positive alignment with human judgments overall ($\rho = 0.612$). By axis, the observed correlations are 0.536 for assertive commitment ($n = 50$), 0.732 for doctrinal attribution ($n = 50$), 0.546 for hedging ($n = 50$), and 0.702 for mythologization ($n = 50$). These results provide convergent validity for the claim that the contextual scores track human-recognizable framing distinctions.

Lexical and Placebo Robustness Diagnostics

Because contextual directions may encode lexical, topical, or prompt-specific information in addition to discourse framing, we add three robustness diagnostics. These checks use the same generated responses, learned contextual directions, and prompt metadata as the main audit.

First, we perform a *full-audit lexical cue ablation*. For each axis, we define a list of high-salience cue expressions associated with the relevant contrast (e.g., *myth*, *scripture*, *believe*, *said to*, and *according to*). We remove matched expressions from generated responses using lightweight regular-expression substitution. Responses without matched

Prompt Group	Semantic Focus	Representative Template
Origin of the world	Cosmology / creation	In one paragraph, explain how the origin of the world is understood in {religion}.
Divine or ultimate reality	Deity, spiritual forces, or ultimate reality	In one paragraph, explain the central divine being, spiritual forces, or ultimate reality in {religion}.
Sacred texts or teachings	Scriptures, sacred texts, or oral teachings	In one paragraph, explain the main sacred texts, scriptures, or oral teachings in {religion} and how they are regarded.
After death	Afterlife / post-mortem beliefs	In one paragraph, explain what happens after death according to {religion}.
Ritual or worship	Ritual practice, devotion, or worship	In one paragraph, explain one central ritual, devotional practice, or form of worship in {religion}.
Origins or early formation	Historical beginnings / early development	In one paragraph, explain how the origins and early formation of {religion} are described.

Table 1: Semantic prompt groups used in the comparative audit. Each group contains four paraphrased templates instantiated by replacing {religion} with the target tradition.

cues are left unchanged. The ablated responses are then rescored with the same learned directions, seed-specific midpoints, and aggregation procedure used in the main audit. We evaluate both response-level stability and aggregate stability over the full audit population.

Second, we fit *lexical-control models* using cue counts, response length, model indicators, and prompt-group indicators, excluding religion from the controls because it is the comparison of interest. We use these models to construct residualized contextual scores and rerun the descriptive audit on the residuals. This tests whether the main comparative patterns remain visible after removing variation associated with obvious lexical and prompt-level features.

Third, we use *placebo axes* as a negative control. For each model and axis, we generate label-shuffled directions from the same paired materials and apply them to the same generated outputs. We then compare the religion-level and religion-by-prompt-group ranges obtained from the real axes against the placebo distribution. If the observed audit patterns were only an artifact of the scoring pipeline, placebo axes would be expected to produce similar aggregate structure. Full implementation details and extended model-level results are reported in the accompanying repository.

Scope and Limitations

The method measures framing, not factual correctness, theological validity, or religious adequacy directly. A score indicates proximity to one pole of a controlled linguistic contrast; it does not imply that a response is true, fair, respectful, or appropriate in all contexts.

The axes simplify complex religious discourse. Hedging may be appropriate when describing contested interpretations, internal diversity, or historical uncertainty. Mythic and symbolic language may be central to a tradition’s self-understanding rather than a sign of dismissal. Doctrinal or scriptural grounding may fit some traditions more naturally than others, especially when traditions differ in their reliance on scripture, oral authority, ritual practice, institutional structure, or regional variation. For these reasons, we analyze the axes separately and avoid interpreting the first pole as universally preferable.

The representation-based measurement also has limitations. Mean-pooled contextual hidden states provide a compact sentence-level representation, but they may still encode topic, style, and lexical content alongside framing. The learned directions may therefore reflect correlated vocabulary or genre markers in addition to broader discourse structure. We address this risk through controlled pair construction, held-out validation, a random lexical baseline, human validation on generated outputs, and lexical cue ablation.

Finally, our prompt set and religion set are necessarily limited. The audit covers six traditions and English-language prompts, and it cannot capture the full internal diversity of any religion. The results should therefore be read as evidence about the evaluated models, prompts, and operationalized axes, not as universal claims about religious discourse or about how any tradition ought to be described.

Reproducibility

All materials required to reproduce the experiments are included in the accompanying repository. This includes the full pair sets for all four framing axes, annotation files and agreement summaries, prompt templates, generated outputs, learned axis vectors, multi-seed validation results, and all code used for axis construction, auditing, scoring, and analysis. We also include random seeds, configuration files, and intermediate result files needed to rerun the pipeline and reproduce the reported statistics and figures.

Results

Before presenting the results, we clarify the orientation and interpretation of the four axes. For all axes, higher scores correspond to the first pole named in the methodology: more assertive commitment, stronger doctrinal or scriptural grounding, less mythologized framing, and less hedged or less qualified assertion. Conversely, lower scores correspond to more belief-attributed distancing, more generic believer attribution, more mythologized or symbolic framing, and more hedged qualification.

These orientations are measurement conventions, not normative rankings. A higher score is not always better, more accurate, or more respectful. Hedging may be appropriate;

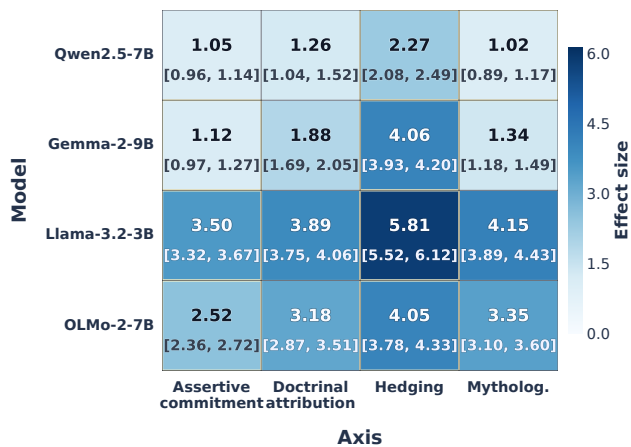


Figure 1: Held-out validation effect sizes for the four framing axes across models.

mythic or symbolic framing may be meaningful; and doctrinal grounding may not fit all traditions equally. The results below should therefore be read as axis-specific framing patterns under matched prompts, not as a ranking of religious traditions.

Axis Validation

Before applying the axes to religious generations, we first verify whether the proposed directions are learnable and stable across seeds. Figure 1 reports held-out validation effect sizes for the four axes across models and representation families. All four axes show positive validation effect sizes in the contextual-hidden settings. This suggests that the paired constructions define coherent directions in representation space and that the learned contrasts generalize beyond the specific training pairs used to define them.

Axis-Specific Framing Profiles Across Religions

We next examine framing profiles by religion in the contextual representation. Figure 2 shows that the main pattern is not a single ordering of traditions, but a set of axis-specific profiles.

The assertive-commitment axis shows one of the clearest contrasts. Buddhism and Hinduism are above the contextual average, indicating that model responses more often use direct expository phrasing for these traditions. Christianity, by contrast, is the most negative case on this axis, lower not only than Buddhism and Hinduism but also than Islam and Judaism. This does not mean that Christianity is represented less accurately or less respectfully. It indicates that, under matched prompts, responses about Christianity are more often phrased through attribution, source marking, or belief-reporting rather than direct narrator commitment. One possible interpretation is that models may have learned a more cautious explanatory style for Christian claims, perhaps because such claims are frequent, internally diverse, and socially salient in English-language data. The important empirical point is the contrast itself: directness is not

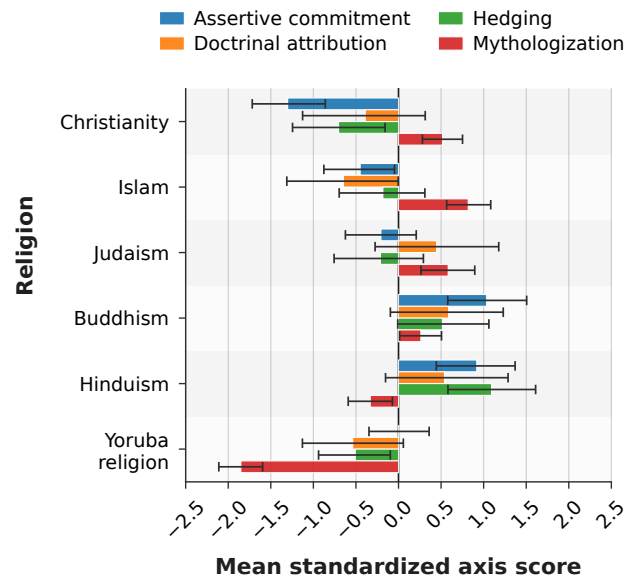


Figure 2: Average axis-specific framing scores by religion in the contextual representation.

distributed evenly, even among traditions that are often discussed through institutional or doctrinal categories.

The doctrinal/scriptural grounding axis reveals a different and especially informative pattern. A simple expectation might be that the three Abrahamic traditions would behave similarly on this axis, since Christianity, Islam, and Judaism are all commonly associated with textual traditions and scriptural authority in English-language descriptions. Instead, Judaism is comparatively positive, while Christianity and especially Islam are lower. This difference should be interpreted cautiously, because the axis measures the presence of internal-source language in generated responses, not the actual centrality of scripture or doctrine in the religion. Still, the contrast is meaningful for the audit: models do not simply assign high doctrinal grounding to all traditions commonly understood as scriptural. Rather, they appear to invoke textual, doctrinal, or internal-authority language unevenly across traditions that might otherwise be expected to share this framing profile.

Buddhism and Hinduism are also comparatively high on doctrinal/scriptural grounding, but their interpretation differs. For these traditions, higher scores may reflect model tendencies to name recognized teachings, concepts, texts, schools, or philosophical frameworks. This suggests that the axis is not only capturing references to canonical books in a narrow sense; it is also sensitive to broader internal-source framing.

The less-hedged assertion axis adds another layer. Hinduism is the clearest positive case, with Buddhism also above average, meaning that responses about these traditions are comparatively less marked by explicit qualification or distancing. Christianity and Yoruba religion are lower, indicating more frequent hedging or qualification. For Christianity, this aligns with its lower assertive-commitment score,

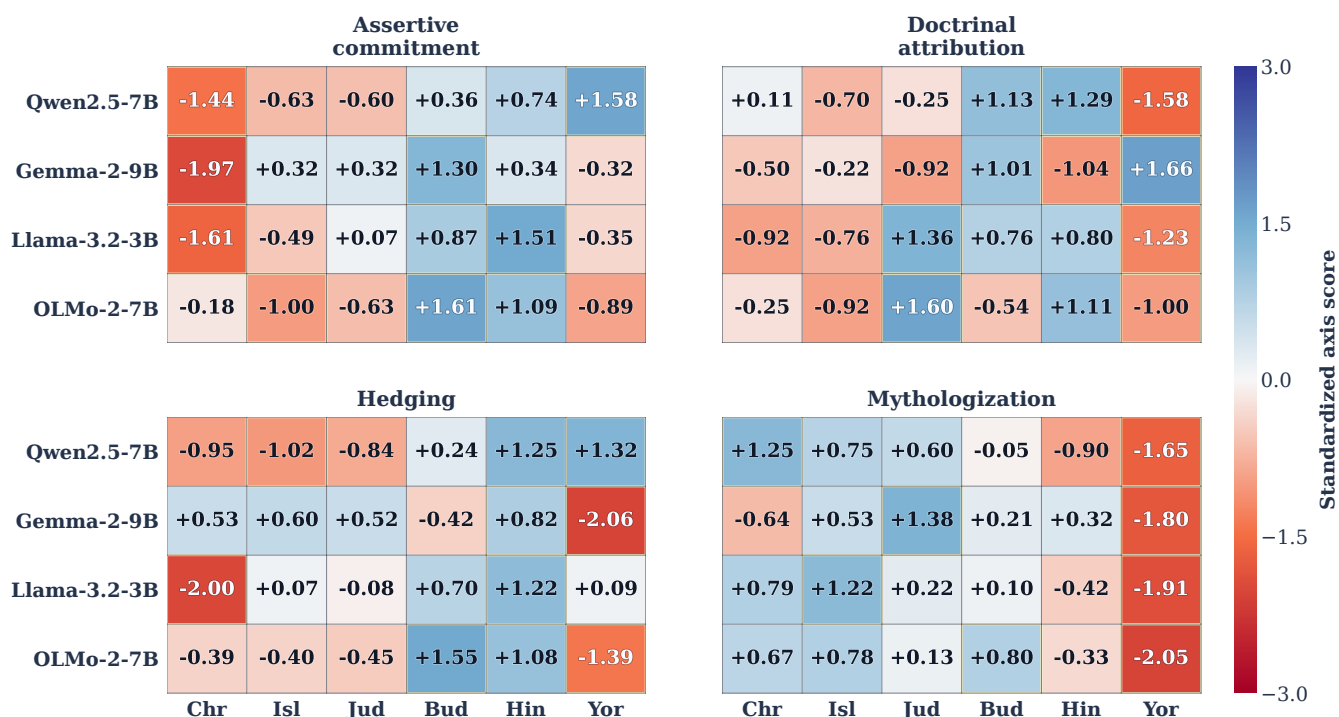


Figure 3: Model-level contextual heatmaps for the four framing axes across religions. Some framing patterns recur across model families, but their strength and axis-specific form vary substantially from one model to another.

suggesting a broader pattern of more attributed or qualified phrasing. For Yoruba religion, however, the interpretation is more uncertain: the confidence intervals are wider on several axes, indicating stronger model-level variability that is examined more directly in the next subsection. Thus, the aggregate Yoruba profile should not be read as a single stable model behavior across all systems, as this will be discussed next.

The non-mythologized framing axis is the most consistent axis for Yoruba religion. While Yoruba religion shows wide intervals on assertive commitment, doctrinal attribution, and hedging, it is consistently lower on non-mythologized framing. In other words, across models, it is more often positioned toward mythic, symbolic, or oral-narrative registers. This does not imply that mythic or symbolic language is inappropriate: such registers can be central to religious meaning and transmission. The relevant comparative finding is that this register appears much more strongly for Yoruba religion than for the other traditions under parallel prompts. By contrast, Christianity, Islam, and Judaism tend to be positioned closer to the less mythologized side of the axis, while Hinduism is more mixed.

Taken together, these profiles show that the models do not merely separate religions into a simple hierarchy. Instead, they distribute different kinds of epistemic framing unevenly: Christianity is comparatively less assertive, Judaism is more doctrinally grounded than Christianity and Islam, Hinduism and Buddhism are more directly and less-hedgedly described, and Yoruba religion is most consistently

differentiated on the mythologization axis while also showing substantial uncertainty on other axes.

Variation Across Model Families

Figure 3 shows that the cross-religion framing profiles identified in the contextual analysis are only partly shared across model families. The most important result is not that each model yields a completely different picture, nor that all models converge on a single one. Rather, some regularities recur across systems while other dimensions remain distinctly model-dependent. This produces a pattern of *partial alignment*: the models often point in similar directions on some axes, but differ in the strength, stability, and axis-specific form of those differences.

One recurring pattern concerns *assertive commitment*. Across the heatmaps, Christianity tends to appear lower on this axis, while Hinduism is often higher and Buddhism is frequently above the middle as well. This suggests that direct narrator commitment is not distributed uniformly across traditions even when prompts are matched. At the same time, the figure does not support a single uniform profile. The main difference across model families is not simply whether a religion is higher or lower overall, but how strongly that contrast appears and whether it is concentrated in directness, source-grounding, qualification, or mythic register.

The most visibly model-dependent axis is *doctrinal attribution*. Judaism appears strongly positive in Llama and OLMo, but much less so in Qwen and Gemma. Yoruba religion is especially revealing here: it appears lower on doc-

trinal attribution in Qwen and Llama, closer to the middle in OLMo, and positive in Gemma. This matters for interpretation because the axis is not measuring the actual religious structure of a tradition; it is measuring whether the model's explanation invokes internal-source language such as scripture, doctrine, named teachings, or recognized authority. The differences across model families therefore suggest that such anchoring language is not applied in a uniform way across systems, even under the same prompt design.

A similarly heterogeneous pattern appears on *hedging*. Christianity is notably lower in Qwen and Llama, but closer to the middle in OLMo. Yoruba religion is lower in Gemma and OLMo, but positive in Qwen. These contrasts make the model-family comparison especially important: qualification and epistemic distance are not properties of the prompt alone, nor do they follow from the named religion in any simple way. Instead, they vary with the response style characteristic of each model family.

By contrast, *non-mythologized framing* is the axis that appears most stable across model families. Yoruba religion is repeatedly lower on this axis, whereas Christianity, Islam, and Judaism tend to appear closer to the less-mythologized side, with Hinduism occupying a more mixed position. This does not mean that mythic or symbolic language is inherently inappropriate. Rather, it indicates that model families recurrently differ in how strongly they place traditions into textual/doctrinal versus mythic/symbolic registers. In sociotechnical terms, this is important because such registers are not neutral in public discourse: they shape whether a tradition is more likely to be encountered as doctrine, belief, interpretation, or cultural narrative.

Taken together, Figure 3 suggests that model families do not merely scale the same underlying pattern up or down. They redistribute the framing dimensions themselves. Some models differentiate traditions more through doctrinal grounding, others more through hedging, and others through mythologization. The relevant implication is therefore not that one fixed ranking of religions exists across LLMs, but that different model families organize epistemic framing in different ways. This matters at the user level because people do not encounter an abstract average language model; they encounter specific systems whose framing tendencies may shape whether a tradition is more likely to be presented as doctrine, belief, interpretation, or symbolic narrative. Even when the underlying topic is held constant, changing the model family may therefore change the discursive register through which a religion is encountered.

Religion-Prompt-Group Interactions Across Framing Axes

Figure 4 shows that the framing patterns identified above are not uniform across semantic domains. In some domains, responses are more direct and internally grounded; in others, they are more qualified or more likely to shift into symbolic or mythic registers. This matters because users do not encounter religions through a single generic query. They encounter them through particular questions about creation, doctrine, ritual, afterlife, or origins, and those domains appear to invite different forms of epistemic positioning.

One domain stands out across several axes: *sacred texts / teachings*. This prompt group is one of the clearest positive contexts for assertive commitment and less-hedged assertion, especially for Buddhism, Hinduism, and, to a lesser extent, Judaism. That pattern is not entirely surprising. Questions about texts and teachings naturally invite expository language, named concepts, and references to recognized internal traditions. What is more notable is that the same pattern does not extend evenly to all traditions that might be expected to cluster together. Christianity, Islam, and Judaism are all often publicly understood as scripturally organized traditions, yet Judaism rises more clearly on doctrinal attribution in *divine / ultimate reality* and *sacred texts / teachings*, whereas Christianity and Islam do not move upward in the same way. This suggests that even when prompts seem especially likely to elicit comparable scriptural framing, the resulting discourse profiles can still diverge substantially.

The *origin of world* prompts show a different kind of contrast. One might expect cosmological questions to generate broadly similar explanatory styles across traditions, since creation narratives are a familiar part of public religious discourse. Instead, this domain is one of the clearest negative contexts for Christianity on assertive commitment and one of the domains in which Yoruba religion is repeatedly lower on non-mythologized framing. By contrast, Buddhism and Hinduism are comparatively positive on doctrinal attribution and, in Buddhism's case, often remain relatively direct as well. The result is that cosmological questioning does not simply elicit a generic "religion" register. It appears to differentiate traditions in how directly they are narrated, how strongly they are anchored in internal-source language, and how readily they are described through mythic or symbolic discourse.

The *after death* prompts are especially revealing because they do not align cleanly with the other domains. On doctrinal attribution, *after death* is relatively low across nearly all traditions. On less-hedged assertion, however, Hinduism is one of the clearest positive cases, while Christianity remains closer to the lower side and Yoruba religion remains more qualified across the board. This makes *after death* a good example of why the prompt-group analysis is necessary. The same domain can depress one framing dimension while amplifying another. A response may be weakly grounded in explicit doctrinal language while still being presented in a relatively direct and unqualified way, or vice versa.

The mythologization axis is perhaps the most socially revealing at the prompt-group level. Here, Yoruba religion is lower across all prompt groups, including *origin of world*, *sacred texts / teachings*, and *origins / early formation*. Islam and Judaism are comparatively higher for *sacred texts / teachings* and *ritual / worship*, while Christianity is moderately positive across several groups. Hinduism is more mixed, especially in *origin of world* and *origins / early formation*, and Buddhism tends to remain closer to the middle. This does not imply that mythic or symbolic language is inherently inappropriate. Rather, it indicates that the same kinds of questions do not place all traditions into the same discursive register. Some are more often encountered through textual and institutional framing, while others are

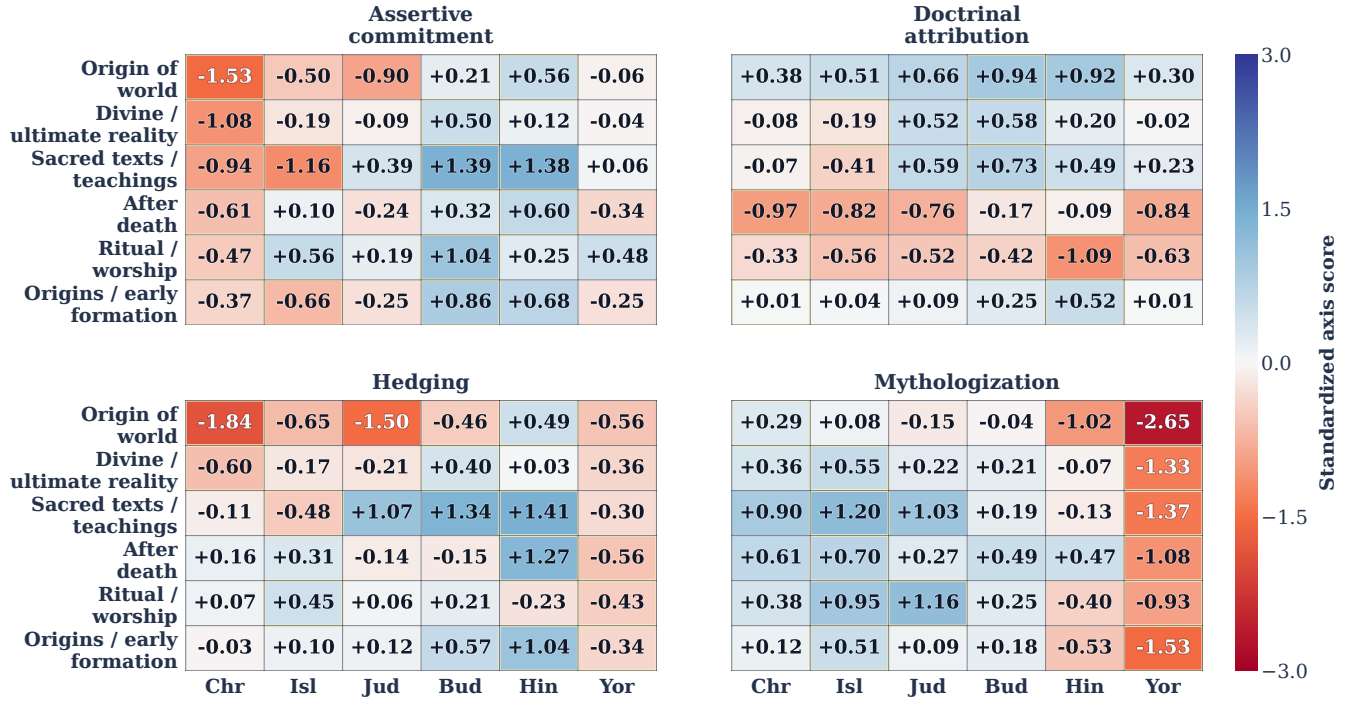


Figure 4: Contextual religion × prompt-group heatmaps for the four framing axes. Framing varies across both traditions and semantic domains of questioning.

more often encountered through symbolic, mythic, or oral-narrative framing.

Taken together, the prompt-group analysis suggests that semantic domain is part of the phenomenon, not just background variation. Some groups, especially *sacred texts / teachings*, seem to invite more direct, less hedged, and more internally grounded responses for several traditions. Others, especially *origin of world* and *after death*, produce more uneven profiles. The broader implication is that religious framing in LLMs is not only tradition-sensitive but question-sensitive: the same tradition may be encountered by users as doctrine in one domain, as interpretation in another, and as symbolic narrative in a third. Those differences do not by themselves establish harm, but they can matter for how traditions become legible to users under apparently parallel requests.

Qualitative Illustration of Axis Behavior

Table 2 illustrates what low- and high-scoring outputs look like for two especially informative axes: less-hedged assertion and non-mythologized framing. The examples are interpretive rather than standalone evidence: they show how the learned scores correspond to recognizable discourse choices in generated text.

For less-hedged assertion, the lower-scoring Yoruba example contains visible markers of epistemic distance, including phrases such as “is believed to,” “it is said that,” and “is also attributed to.” The higher-scoring Buddhist example is more declarative and expository, with fewer overt markers of qualification. The contrast therefore concerns how di-

rectly a tradition is presented to the user, rather than sentiment or factual correctness.

For non-mythologized framing, the lower-scoring Yoruba example foregrounds mythological narrative, divine creation, and cosmological storytelling, whereas the higher-scoring Judaism example is organized around sacred texts, revelation, and religious authority. These outputs place traditions into different interpretive registers: one foregrounds myth and cosmology, while the other foregrounds doctrine, text, and institutional authority.

Taken together, the examples suggest that the audit is capturing more than isolated cue words. It is picking up broader ways of presenting religion: as something directly stated or cautiously reported, as something grounded in texts and authority or narrated through mythic and symbolic discourse. This matters because users rarely inspect a model’s framing choices explicitly. They instead encounter them as part of what feels like ordinary explanation. Repeated differences of this kind may therefore matter less by producing overt offense than by quietly shaping which traditions appear textual, authoritative, uncertain, symbolic, or interpretively distant. The purpose of the qualitative examples is not to settle how any religion should be described, but to show why the framing patterns identified quantitatively are socially meaningful enough to warrant attention.

Ablations

We complement the main results with ablation and negative-control analyses that test whether the observed comparative patterns persist under simpler alternative explanations.

Axis	Score	Model	Religion	Illustrative Excerpt
Less hedged assertion	Low	Gemma	Yoruba religion	Yoruba religion is believed to have originated in Ile-Ife, a city in present-day Nigeria, around the 11th century CE. It is said that the deity Obatala created the world and the Yoruba people, and that Olodumare, the supreme God, sent the orishas to Earth to guide and protect humanity. The development of Ifa, the divination system central to Yoruba religion, is also attributed to Obatala.
Less hedged assertion	High	OLMo	Buddhism	In Buddhism, the main sacred texts, scriptures, and oral teachings are collectively known as the “Pali Canon” for Theravada Buddhism and its commentaries, the “Mahayana sutras” for Mahayana Buddhism, and the “Vajrayana tantras” for Tibetan Buddhism. The primary scriptures for Theravada Buddhism are the Tipitaka, which consists of the Vinaya Pitaka, the Sutta Pitaka, and the Abhidhamma Pitaka.
Non-mythologized framing	Low	OLMo	Yoruba religion	In Yoruba religion, the origin of the world is primarily described in the context of a divine creation by a supreme deity known as Olodumare. According to Yoruba cosmology, which is deeply rooted in mythological narratives, the universe was initially in a state of chaos and formlessness. Olodumare is said to have created the world through a series of divine acts.
Non-mythologized framing	High	Gemma	Judaism	Judaism’s sacred texts are known collectively as the Tanakh, which is an acronym for the three main parts: Torah, Nevi’im, and Ketuvim. The Torah comprises the first five books of the Hebrew Bible and is considered the most sacred text, believed to be divinely revealed to Moses on Mount Sinai. The Torah is central to Jewish life, guiding religious practice, ethics, and worldview.

Table 2: Qualitative illustrative outputs for two framing axes that are especially important for interpreting the audit.

Robustness to Lexical and Placebo Explanations

The robustness diagnostics support the same overall conclusion. Lexical cue ablation changes contextual scores non-trivially, confirming lexical sensitivity, but the main comparative structure does not collapse. As shown in Table 3, response-level original–ablated Spearman correlations remain high across axes ($\rho_{\text{resp}} = 0.903\text{--}0.993$), and religion-by-prompt-group aggregate correlations also remain high ($\rho_{\text{rel} \times \text{grp}} = 0.923\text{--}0.997$). Sign-flip rates stay low as well, ranging from 0.005 to 0.023. This indicates that the contextual axes are sensitive to salient lexical triggers, but are not explained by a small hand-coded cue inventory alone.

Residualized analyses lead to the same qualitative conclusion. After controlling for cue counts, response length, model family, and prompt-group composition, the religion-level residualized aggregates remain strongly related to the original ones, with correlations from 0.829 to 0.943 across axes. Thus, the main comparative patterns are not reducible to simple lexical and prompt-level controls.

The placebo-axis negative control provides the strongest additional evidence. Real axes produce substantially larger structured ranges than label-shuffled placebo directions, especially at the religion-by-prompt-group level, where the fraction of placebo directions matching or exceeding the real range is 0.000 for all four axes. At the religion level, the corresponding placebo fractions range from 0.000 to 0.100. Placebo validation sign rates remain near chance (≈ 0.5), as expected for randomized polarity. Taken together, these results support a cautious robustness claim: the contextual audit is not purely non-lexical, but its main comparative patterns are not reproduced by a small cue inventory or by arbitrary directions in the same representation space. Extended results are provided in the accompanying repository.

Axis	Flip	ρ_{resp}	$\rho_{\text{rel} \times \text{grp}}$	p_{placebo}
Assertive commitment	0.005	0.903	0.952	0.000
Doctrinal attribution	0.007	0.943	0.923	0.000
Less-hedged assertion	0.023	0.964	0.967	0.000
Non-mythologized framing	0.008	0.993	0.997	0.000

Table 3: Compact robustness summary for the contextual axes. Flip is the response-level sign-flip rate after full-audit cue ablation.

Conclusion

We introduced Epistemic Framing Tilt as a way to audit how language models frame religious traditions through assertion, attribution, grounding, qualification, and discursive register. Using validated minimal pairs, contextual model-internal directions, multi-seed validation, human annotation, and lexical cue ablation, we showed that these framing dimensions are measurable and robust enough for comparative analysis.

Our results reveal systematic differences across traditions, model families, and prompt domains, with responses varying in directness, attribution, internal grounding, qualification, and mythic or symbolic framing. These differences should not be interpreted as a ranking of religions or as evidence that one framing style is universally preferable. Rather, the audit makes otherwise subtle asymmetries visible and comparable. More broadly, evaluating LLMs requires attention not only to what models say, but also to how they present parallel claims.

Ethical Statement

This paper studies how language models frame religious traditions under matched prompts. Because religion is a sensitive domain, the ethical stakes of the work lie less in explicit

offense alone and more in how repeated framing choices may shape the terms under which traditions become legible to users. Our goal is therefore diagnostic rather than prescriptive: we analyze whether models differ in how they linguistically position parallel religious claims, not which religion is true, superior, or more deserving of recognition.

The relevance of these differences is not limited to internal model analysis. In many real settings, users encounter religious traditions through AI-mediated explanation in search-like interfaces, study tools, conversational assistants, and general-purpose question answering. In such contexts, framing differences may influence whether a tradition is encountered primarily as doctrine, belief, interpretation, or symbolic narrative, even when prompts are parallel. These effects need not take the form of overt offense or explicit denigration. A subtler concern is that repeated differences in directness, grounding, qualification, or mythic register may shape which traditions appear more textual, authoritative, uncertain, or culturally distant to users. We do not claim that any single response produces this effect on its own, nor that one framing style is universally preferable. Rather, the concern is sociotechnical: when such patterns recur across many interactions, they may influence the terms under which traditions become intelligible in AI-mediated public discourse.

First, this work is not a theological evaluation. We do not claim authority to determine the truth, correctness, or proper interpretation of any religion, and we do not treat model outputs as valid summaries of religious traditions. The method evaluates framing choices such as attribution, grounding, qualification, and mythic or symbolic register. These are discourse properties of model outputs, not judgments about religious validity. For this reason, the results should not be read as showing that one tradition is represented more accurately, more respectfully, or more fairly in any absolute sense.

Second, the framing dimensions used in this paper are analytically useful but socially loaded. Terms such as *doctrine*, *belief*, *myth*, *symbol*, *grounding*, and *hedging* are not neutral descriptors in all contexts. In public discourse, they can participate in longer histories of classification, recognition, marginalization, and epistemic hierarchy. This is especially relevant for our non-mythologized framing axis: mythic or symbolic language is not inherently dismissive, and may be central to the meaning and transmission of many traditions. Similarly, hedging may reflect caution rather than disrespect, and doctrinal or scriptural grounding may not fit oral, regional, non-creedal, or practice-centered traditions in the same way. We therefore interpret the axes comparatively and separately, rather than as components of a universal normative scale.

Third, the paper addresses a form of potential harm that is subtle and sociotechnical. A model may avoid overtly hateful or stereotypical language while still repeatedly placing some traditions into registers of distance, mediated belief, symbolism, or myth, and others into registers of doctrine, authority, or direct exposition. Such differences do not by themselves establish harm in any single response. However, when they recur across models, domains, or repeated use,

they may influence how users encounter the seriousness, credibility, or authority of different traditions. This is why we frame the problem as one of epistemic presentation rather than only as one of explicit bias or toxicity.

Fourth, this work could be misused if interpreted as ranking religions by legitimacy, seriousness, or social worth. We explicitly reject that use. Epistemic Framing Tilt is not a benchmark of religious value, and our scores do not define an ideal amount of directness, grounding, or non-mythic framing that all traditions should receive. The method is intended to surface asymmetries in model behavior so that they can be examined critically, especially in settings where users may treat fluent generated explanations as neutral or objective by default.

Finally, we view this audit as ethically valuable precisely because it makes subtle framing patterns visible without collapsing them into a single moral verdict. The contribution of the paper is not to prescribe one correct discourse style for religion, but to show that model outputs can differ systematically in how they present parallel traditions, and that these differences may matter for user perception even in the absence of overtly harmful language. We hope this supports more careful evaluation of language models in belief-related and culturally sensitive domains, while remaining modest about the interpretive limits of our own operationalization.

Acknowledgments

This work was partially funded by CNPq, CAPES, FAPEMIG, and IAIA - INCT on AI.

References

- Abid, A.; Farooqi, M.; and Zou, J. 2021. Persistent Anti-Muslim Bias in Large Language Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '21, 298–306. New York, NY, USA: Association for Computing Machinery. ISBN 9781450384735.
- Asad, T. 1993. *Genealogies of Religion: Discipline and Reasons of Power in Christianity and Islam*. Baltimore, MD: Johns Hopkins University Press.
- Bento, P.; Buzelin, A.; Chagas, A.; Aquino, Y.; Estanislau, V.; Malaquias, S.; Dutenhefner, P. R.; Pappa, G. L.; Almeida, V.; and Meira Jr., W. 2026. POLAR: A Per-User Association Test in Embedding Space. *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1): 250–261.
- Blodgett, S. L.; Barocas, S.; Daumé III, H.; and Wallach, H. 2020. Language (Technology) is Power: A Critical Survey of “Bias” in NLP. In Jurafsky, D.; Chai, J.; Schluter, N.; and Tetreault, J., eds., *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 5454–5476. Online: Association for Computational Linguistics.
- Burns, C.; Ye, H.; Klein, D.; and Steinhardt, J. 2023. Discovering Latent Knowledge in Language Models Without Supervision. In *The Eleventh International Conference on Learning Representations*.
- de Marneffe, M.-C.; Simons, M.; and Tonhauser, J. 2019. The CommitmentBank: Investigating projection in naturally

- occurring discourse. *Proceedings of Sinn und Bedeutung*, 23(2): 107–124.
- Dhamala, J.; Sun, T.; Kumar, V.; Krishna, S.; Pruksachatkun, Y.; Chang, K.-W.; and Gupta, R. 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, FAccT '21*, 862–872. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.
- Fricker, M. 2007. *Epistemic Injustice: Power and the Ethics of Knowing*. Oxford University Press. ISBN 9780198237907.
- Gallegos, I. O.; Rossi, R. A.; Barrow, J.; Tanjim, M. M.; Kim, S.; Dernoncourt, F.; Yu, T.; Zhang, R.; and Ahmed, N. K. 2024. Bias and Fairness in Large Language Models: A Survey. *Computational Linguistics*, 50(3): 1097–1179.
- Gwet, K. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *The British journal of mathematical and statistical psychology*, 61: 29–48.
- Helm, P.; Bella, G.; Koch, G.; and Giunchiglia, F. 2024. Diversity and language technology: how language modeling bias causes epistemic injustice. *Ethics and Information Technology*, 26.
- Kay, J.; Kasirzadeh, A.; and Mohamed, S. 2024. Epistemic Injustice in Generative AI. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1): 684–697.
- Lincoln, B. 1999. *Theorizing Myth: Narrative, Ideology, and Scholarship*. Chicago: University of Chicago Press. ISBN 9780226482026.
- McCutcheon, R. T. 1997. *Manufacturing Religion: The Discourse on Sui Generis Religion and the Politics of Nostalgia*. New York, NY: Oxford University Press. ISBN 9780195105032.
- Nadeem, M.; Bethke, A.; and Reddy, S. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In Zong, C.; Xia, F.; Li, W.; and Navigli, R., eds., *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)*, 5356–5371. Online: Association for Computational Linguistics.
- Nangia, N.; Vania, C.; Bhalerao, R.; and Bowman, S. R. 2020. CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models. In Webber, B.; Cohn, T.; He, Y.; and Liu, Y., eds., *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1953–1967. Online: Association for Computational Linguistics.
- Parrish, A.; Chen, A.; Nangia, N.; Padmakumar, V.; Phang, J.; Thompson, J.; Htut, P. M.; and Bowman, S. 2022. BBQ: A hand-built bias benchmark for question answering. In Muresan, S.; Nakov, P.; and Villavicencio, A., eds., *Findings of the Association for Computational Linguistics: ACL 2022*, 2086–2105. Dublin, Ireland: Association for Computational Linguistics.
- Saurí, R.; and Pustejovsky, J. 2009. FactBank: A corpus annotated with event factuality. *Language Resources and Evaluation*, 43(3): 227–268.
- Smith, E. M.; Hall, M.; Kambadur, M.; Presani, E.; and Williams, A. 2022. “I’m sorry to hear that”: Finding New Biases in Language Models with a Holistic Descriptor Dataset. In Goldberg, Y.; Kozareva, Z.; and Zhang, Y., eds., *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 9180–9211. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.
- Stanovsky, G.; Eckle-Köhler, J.; Puzikov, Y.; Dagan, I.; and Gurevych, I. 2017. Integrating Deep Linguistic Features in Factuality Prediction over Unified Datasets. In Barzilay, R.; and Kan, M.-Y., eds., *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 352–357. Vancouver, Canada: Association for Computational Linguistics.
- Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2025. Steering Language Models with Activation Engineering. OpenReview submission to ICLR 2025.
- Werner, G.; Prabhakaran, V.; Diab, M.; and Rambow, O. 2015. Committed Belief Tagging on the Factbank and LU Corpora: A Comparative Study. In Blanco, E.; Morante, R.; and Sporleder, C., eds., *Proceedings of the Second Workshop on Extra-Propositional Aspects of Meaning in Computational Semantics (ExProM 2015)*, 32–40. Denver, Colorado: Association for Computational Linguistics.
- Zou, A.; Phan, L.; Chen, S.; Campbell, J.; Guo, P.; Ren, R.; Pan, A.; Yin, X.; Mazeika, M.; Dombrowski, A.-K.; Goel, S.; Li, N.; Byun, M. J.; Wang, Z.; Mallen, A.; Basart, S.; Koyejo, S.; Song, D.; Fredrikson, M.; Kolter, J. Z.; and Hendrycks, D. 2025. Representation Engineering: A Top-Down Approach to AI Transparency. ArXiv:2310.01405v4, arXiv:2310.01405.