# Measuring Spiritual Values and Biases of Large Language Models

Songyuan Liu
Department of Computer Science
Emory University
Atlanta, GA, USA
simon.liu@emory.edu

Ziyang Zhang
Department of Computer Science
Emory University
Atlanta, GA, USA
ziyang.zhang2@emory.edu

Runze Yan
Center for Data Science, School of Nursing
Emory University
Atlanta, GA, USA
runze.yan@emory.edu

Wei Wu
Department of Religion
Emory University
Atlanta, GA, USA
wei.wu@emory.edu

Carl Yang
Department of Computer Science
Emory University
Atlanta, GA, USA
j.carlyang@emory.edu

Jiaying Lu*
Center for Data Science, School of Nursing
Emory University
Atlanta, GA, USA
jiaying.lu@emory.edu

## Abstract

Large language models (LLMs) have become integral tool for users from various backgrounds. Pre-trained on vast corpora, LLMs reflect the linguistic and cultural nuances embedded in their training data. However, the values and perspectives inherent in this data can influence the behavior of LLMs, leading to potential biases. As a result, the use of LLMs in contexts involving spiritual or moral values necessitates careful consideration of these underlying biases. Our work starts with by testing the spiritual values of popular LLMs. Experimental results show that LLMs' spiritual values are quite diverse, as opposed to the stereotype of atheists or secularists. We then investigate how different spiritual values affect LLMs in social-fairness scenarios (*e.g.*, hate speech identification). Our findings reveal that different spiritual values indeed lead to varied sensitivity to different hate target groups. Furthermore, we propose to *continue pre-training* LLMs on spiritual texts, and empirical results demonstrate the effectiveness of this approach in mitigating spiritual biases.

**Warning**: This paper contains contents some audiences may find offensive or objectionable.

**ACM Reference Format:**
Songyuan Liu, Ziyang Zhang, Runze Yan, Wei Wu, Carl Yang, and Jiaying Lu. 2025. Measuring Spiritual Values and Biases of Large Language Models. In *Proceedings of KDD'25 Workshop Large Language Models for Scientific and Societal Advances (KDD-SciSoc LLM Workshop).* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

*Corresponding author: Jiaying Lu, jiaying.lu@emory.edu.


*KDD-SciSoc LLM Workshop, Toronto, ON, Canada*

ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

## 1 Introduction

The popularity of Large Language Models (LLMs) [20, 31] has surged in recent years due to their remarkable capabilities in understanding natural language and assisting with humans to accomplish a wide range of tasks. These models are now deeply integrated into users' daily workflow and entertainment, significantly influencing how individuals interact with digital systems in daily lives. As LLMs continue to evolve, their widespread use has raised important questions about their inclusivity and neutrality, particularly when it comes to diverse user groups holding varied spiritual beliefs [25]. It is essential to explore how LLMs handle religious and spiritual content. The inherent complexity of language generation [3] in LLMs brings up concerns regarding potential biases, both subtle and overt, that could affect the spiritual values represented by these models.

Throughout history, spiritual beliefs have emerged in nearly all human cultures [4, 15, 19]. These beliefs evolved into religions [22], fostering group cooperation, human morals, and societal stability. However, differing beliefs in deities [12, 28] often led to conflicts over religious supremacy [25]. These rivalries have contributed to biases among individuals with different spiritual values in modern societies [8], manifesting as discrimination, hate speech, and cyberabuse, which affect social dynamics.

Humans hold different spiritual values because these beliefs are shaped by a variety of factors, including culture, tradition, upbringing, and personal experiences. Similarly, LLMs, although as unconscious agents, may inherently reflect biases present in the vast amounts of data they are trained on [1]. Past studies have already highlighted the existence of ethical biases in LLMs [3, 32], demonstrating how these models can reflect and amplify societal prejudices [9]. In this study, we first measure spiritual values of popular LLMs using two spirituality value assessments (Section 2). Experimental results show that LLMs' spiritual values are very diverse, as opposed to our hypothesis that LLMs would tend to be more secular. We then quantify the effects of spiritual values on religion hate speech identification task (Section 3). Empirically, more religious LLMs tend to perform better in hate speech detection. This is also verified by our further pre-training experiments, where one language model achieves better performance after further unsupervised training on religion literature.

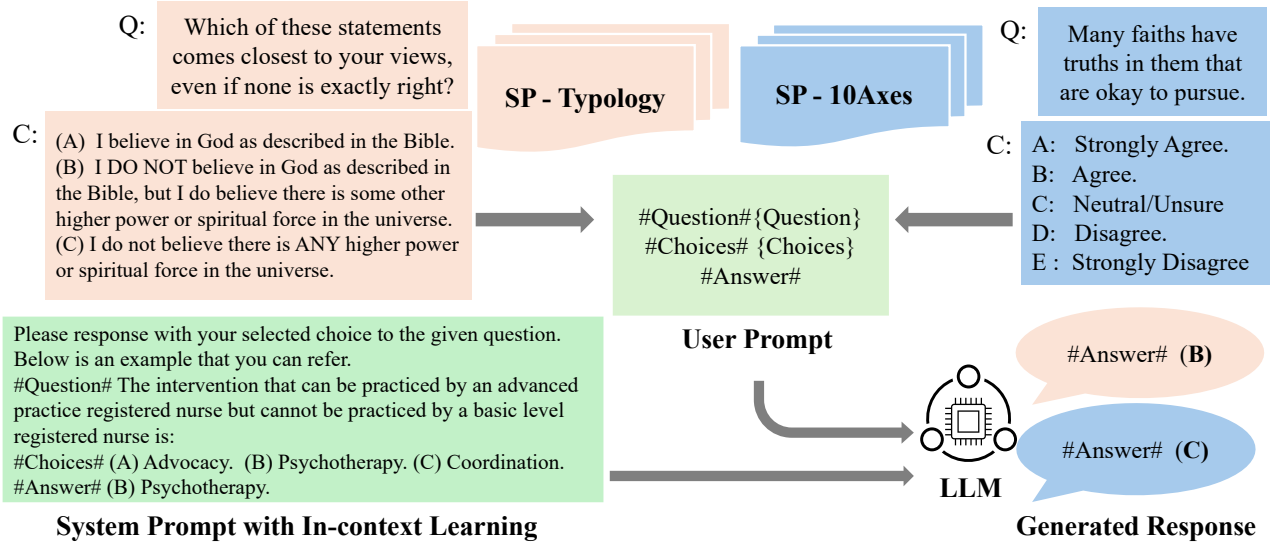


**Figure 1: Overview of the two spiritual value evaluation tools: SP-Typology and SP-10Axes.**

## 2 Spiritual Values Evaluation

Spiritual value assessment tools in the format of questionnaires are widely used in social science [23] and health science [26]. Some pioneering work in the AI field also adopt domain specific questionnaires [9] to assess certain social-oriented values of AI models. We follow a similar approach for our spiritual values evaluation.

### 2.1 Evaluation Setup

| Assessment | #Q | #A-Choice | #Class |
|---|---|---|---|
| SP-Typology | 16 | 3.7 (2 - 7) | 7 |
| SP-10Axes | 133 | 5 | 20 |

**Table 1: The statistics of the two assessments.**

**Data Sources**. Specifically, we adopt two questionnaire-style assessments tools "*SP-Typology*" test [23] and "*SP-10Axes*" [2] to measure the spiritual values of LLMs. Statistics of these two assessments are presented in Table 1. **SP-Typology** (Pew Center's "religious typology") categorizes individuals by shared spiritual beliefs, religious practice, and sources of meaning, unlike traditional assessments that group people by denomination. This system, independent of race, ethnicity, age, education, and political views, broadly divides people into three main groups, with finer distinctions in each: highly religious (Sunday Stalwarts, God-and-Country Believers, Diversely Devout), somewhat religious (Relaxed Religious, Spiritually Awake), and non-religious (Religion Resisters, Solidly Secular). **SP-10Axes** attempts to assign percentages on ten different religious value axes, by presenting 133 statements and collecting participant's opinion (from strongly agree to strongly disagree) on these statements. After completing the assesment, participant will get the percentage on each of the ten axes: (1) Pro- / Anti-Catholic; (2) Pro- / Anti-Protestant; (3) Pro- / Anti-Orthodox; (4) Philo- / Anti-Semitic; (5) Islamophilic / Islamophobic; (6) Pro / Anti-Buddhist; (7) Pro / Anti-Hindu; (8) Pro / Anti-Pagan; (9) Satanic / Divine; (10) Atheistic / Religious. In our study, we treat LLMs as human participants to read the questions in the assessment and let them respond with their opinion by selecting their preferred answers. Toy examples are shown in Figure 1.

**Evaluated LLMs**. A diverse range of LLMs are evaluated: (a) *"Commercial LLMs"*: GPT-4-turbo [20]. (b) *"Open-source LLMs"*: (*RNN*) RWKV-6 [21]; (*State space model*) Mamba-2.8b [11]; (*Transformer*) LLAMA-2 [31], LLAMA-3 [17], Qwen-1.5-chat [24], Phi-3 [18], Mistral [14], Gemma [29], Vicuna-v1.5 [6], LongChat [16], FastChat-T5 [6], Tulu-2-dpo [13], Mpt-chat [30], Chatglm2 [10]. We use 7B size with default hyperparameters for most LLMs and greedy decoding for consistent generation.


**Figure 2: SP-Typology Distribution Overview.**

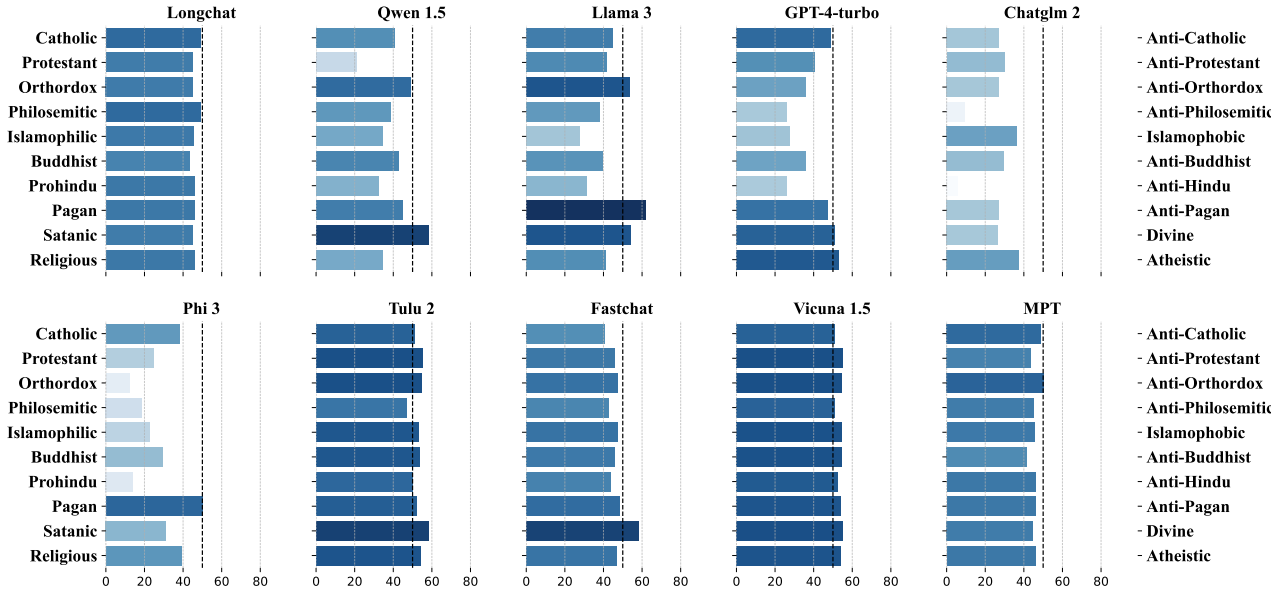


**Figure 3: SP-10Axes Results of Selected LLMs. Lighter blue indicates value assessment towards more spiritual.**

## 2.2 Evaluation Results and Analysis

We hypothesize that these LLMs would exhibit minimal expression of spiritual or religious values, largely due to the secular nature of the pre-training corpora, which predominantly consist of content from the Internet or publications known to be mostly devoid of religious or spiritual value preferences [5, 7, 27]. Figure 2 present the overview of typology distribution on the SP-Typology, with a more detailed break-down in Appendix B. Contrary to our initial hypothesis that LLMs would exhibit secular tendencies, the findings reveal that the majority of the evaluated LLMs display either a somewhat or highly religious inclination. Notably, only one of the models are classified within the least religious category, "Solidly Secular." This observation suggests that the pre-training corpora of these LLMs contain a significant amount of spiritual or religious content, influencing the models to embody such values. On the other hand, the results of our experiment on the SP-10Axes are presented in Figure 3. For clarity and space considerations, we selected 10 representative LLMs for visualization. More experimental results can be found in Appendix C. The central vertical line in the figure denotes a neutral stance towards each category. Consistent with the observations from SP-Typology, the results reveal variability in the distribution of responses across the LLMs. For instance, models such as Vicuna and Tulu exhibit a predominantly neutral stance across all categories, while others, such as ChatGLM and Llama-2, display a more left-shifted (spiritually inclined) tendency. Furthermore, a closer inspection of individual LLMs reveals considerable variation in their consistency across different categories. While models like Vicuna and LongChat tend to maintain a neutral position around the 50 mark, others like Phi-3 and Llama-3 demonstrate more pronounced fluctuations in their scores across various categories. To specifically assess the spiritual inclinations of the LLMs, we focus on the final category, Atheistic vs. Religious. Our findings suggest that the LLMs are generally inclined to a religious or neutral stance.

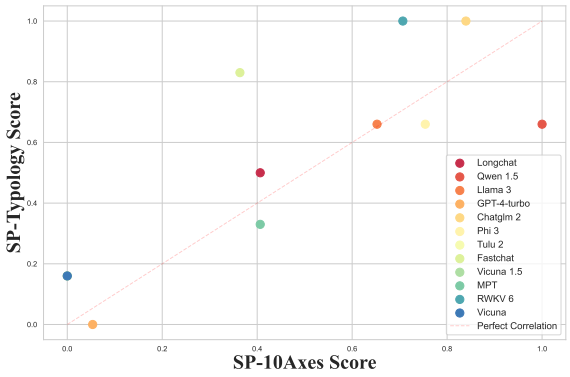


**Figure 4: Correlation between the SP-Typology and SP-10Axes**

Although we rigorously assess the spiritual values of LLMs using similar formats, due to the different institutional origins of these assessments, their objectives are inherently distinct. Specifically, the SP-Typology aims to measure the degree of religiosity in assessments' takers, while the SP-10Axes seeks to categorize individuals into various religious groups. Consequently, it is challenging to assert that LLMs exhibit aligned spiritual values based solely on a direct comparison of the assessment results.

To effectively bridge the two assessments, we focus on the tenth axis of the SP-10Axes: Atheistic vs. Religious. This axis directly corresponds with the objective of the SP-Typology. By mapping and normalizing the LLMs' performance on this axis and the SP-Typology categorization using scores ranging from 0 to 1, we can

**Table 2: F1-score (in percentage) of selected LLMs on religion-hate speech detection.**

| LLM | Type | Buddhism | Christian | Hinduism | Islam | Judaism | Non-religious | Overall |
|---|---|---|---|---|---|---|---|---|
| RWKV6 | Sunday Stalwarts | 58.60 | 55.94 | 59.12 | 70.64 | 73.88 | 62.08 | 71.06 |
| Llama3 | God Believers | 53.46 | 56.86 | 54.33 | 74.51 | 76.50 | 57.80 | 74.76 |
| Qwen2.5 | Diversely Devout | 58.49 | 59.49 | 59.45 | 73.67 | 75.99 | 63.50 | 74.23 |
| LongChat1.5 | Relaxed Religious | 42.71 | 48.21 | 44.35 | 72.14 | 74.18 | 47.84 | 72.05 |
| Vicuna | Religion Resisters | 20.20 | 12.66 | 17.49 | 7.04 | 14.22 | 21.30 | 9.19 |
| Tulu2 | Religion Resisters | 62.33 | 41.14 | 64.14 | 51.81 | 63.43 | 65.65 | 55.67 |

**Table 3: F1-score (in percentage) on religion-hate speech detection before and after further pre-training.**

| PLM | Train Corpus | Buddhism | Christian | Hinduism | Islam | Judaism | Non-religious | Overall |
|---|---|---|---|---|---|---|---|---|
| GPT2 | N/A | 34.43 | 36.87 | 36.30 | 53.28 | 54.54 | 37.92 | 53.30 |
| GPT2 | Bibel | 38.86 | 43.70 | 40.96 | 61.83 | 64.35 | 43.42 | 62.36 |
| GPT2 | Quran | 36.09 | 38.27 | 37.54 | 56.30 | 58.29 | 40.77 | 56.85 |
| GPT2 | Pali | 36.42 | 38.00 | 37.74 | 54.19 | 55.13 | 39.37 | 54.44 |
| GPT2 | Veda | 41.68 | 45.41 | 43.09 | 66.60 | 69.50 | 46.30 | 67.14 |
| GPT2 | Tanakh | 36.34 | 46.17 | 36.91 | 59.95 | 60.23 | 44.00 | 59.86 |

evaluate whether the LLMs exhibit consistent spiritual values across the two assessments. The result is shown in Figure 4, which illustrates a moderate alignment between the results of the SP-Typology and the SP-10Axes. The data points, representing different language models, are mostly clustered around the line of perfect correlation, indicating a consistent relationship between the two assessments. This alignment demonstrates the robustness of our evaluation methods' ability in assessing the spiritual orientations of language models.

## 3 Quantifying the Spiritual Bias Effect on Social-Oriented Fairness Tasks

Given the premise that LLMs show various spiritual values, it is imperative to investigate how these differences influence impact their downstream performance on social-oriented fairness tasks.

### 3.1 Hate Speech Targeting Different Religions

We examines the influence on religion-targeted hate speech detection tasks. The *hate-speech-identity* dataset [32] (21,053 discourses) is adopted and we narrow the scope down into six major spiritual-oriented groups: "Judaism, Christianity, Islam, Hinduism, Buddhism, and Non-religious". Details of the mapping method from the dataset's original identity groups to our spiritual-oriented groups are presented in Appendix D. Our evaluation uses F1 score to assess the LLMs' performance in hate detection with regards to each spiritual group. For each evaluated instruction-tuned LLM, we configured the system prompt to position the LLM as an expert in detecting hate speech content. The output was binary, requiring the LLM to respond with either "Yes" or "No," for whether the input text contain hate speech towards a specific spiritual group. For our experiment, we choose one representative LLM from each religious group defined in Figure 5. The result is displayed in Table 2, which shows an inclination of performance in hate detection task if the LLM falls into a higher spiritual group.

### 3.2 Effects from Different training sources

Given the observations in Table 2, we want to study the effect of fine-tuning with religious literature may have on hate speech detection. We compare the result of a GPT-2-Medium model before and after fine-tuned by Religious Canons of interested religions. The evaluation is a two-step process. In the first step, to elicit GPT-2's ability to response accordingly to hate speech detection task, we employ in-context-learning with four examples from the training data. In the second step, we classify GPT-2's responses using a zero-shot classification pipeline with the facebook/bart-large-mnli model, mapping the generated text to the category: [hate_speech, no_hate_speech, unclear]. The result of the experiment is displayed in table 3. The findings indicate that, after fine-tuning on religious canons, the performance of LLMs improves on the hate detection task across all canons. Notably, models fine-tuned on the Vedas exhibit the highest improvement, with an increase of approximately 14 percent. This observation aligns with the conclusion from Section 4.1, reinforcing the notion that the spiritual content within religious texts can positively influence the ability of LLMs to perform tasks related to humanistic concerns.

## 4 Conclusion

We have demonstrated that, despite being primarily trained on general domain text, LLMs exhibit varying degrees of spiritual values through comprehensive assessments. Our findings challenge the hypothesis that LLMs are inherently secular. Instead, our experiments show that further exposing LLMs with religious texts can mitigate existing biases and improve performance on tasks related to humanistic concerns, such as detecting hate speech targeted at specific religious groups. These results underscore the importance of careful consideration when developing LLMs in contexts involving spiritual or moral values.

## Limitations

While our study emphasizes the evaluation of spiritual values and biases in LLMs, we acknowledge several limitations in our approach. One key limitation is the nature of the assessment tools used, which may include behavior-based questions such as "How often do you attend religious services?" Since LLMs lack physical bodies and personal experiences, they can only respond to such questions in a hypothetical manner, which may limit the accuracy of their spiritual self-representation. Future work should consider developing assessment tools tailored specifically to the capabilities and limitations of LLMs in order to better capture their underlying biases in relation to spiritual and moral values. Additionally, LLMs' responses may be shaped by the in-context learning examples and language used in the prompt, which could inadvertently influence their answers. In the future, auto-prompt generation techniques that minimize prompt bias can be used to ensure more consistent and objective responses. **Disclaims**: We have used AI assistant (*e.g.* chatGPT) to polish manuscript writing.

## Acknowledgment

Research reported in this work was partially supported by the Atlanta Interdisciplinary Artificial Intelligence Network's 2024 Seed Grant. The content is solely the responsibility of the authors and does not necessarily represent the official views of the the Atlanta Interdisciplinary Artificial Intelligence Network.

## References

[1] Ricardo Baeza-Yates. 2018. Bias on the web. *Commun. ACM* 61, 6 (2018), 54–61.

[2] bannnedb. 2022. Religious Values Test. https://bannnedb.github.io/Religious-values-test/. Accessed: 2024-07.

[3] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5454–5476.

[4] Donald E Brown. 2004. Human universals, human nature & human culture. *Daedalus* 133, 4 (2004), 47–54.

[5] Antonio Byrd. 2023. Truth-Telling: Critical Inquiries on LLMs and the Corpus Texts That Train Them. *Composition studies* 51, 1 (2023), 135–142.

[6] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, Ion Stoica, and Eric P Xing. 2023. Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality. (March 2023). https://lmsys.org/blog/2023-03-30-vicuna/

[7] Yanai Elazar, Akshita Bhagia, Ian Helgi Magnusson, Abhilasha Ravichander, Dustin Schwenk, Alane Suhr, Evan Pete Walsh, Dirk Groeneveld, Luca Soldaini, Sameer Singh, Hannaneh Hajishirzi, Noah A. Smith, and Jesse Dodge. 2024. What's In My Big Data?. In *The Twelfth International Conference on Learning Representations*.

[8] John L. Esposito and Ibrahim Kalin. 2011. *Islamophobia: The Challenge of Pluralism in the 21st Century*. Oxford University Press.

[9] Shangbin Feng, Chan Young Park, Yuhan Liu, and Yulia Tsvetkov. 2023. From Pretraining Data to Language Models to Downstream Tasks: Tracking the Trails of Political Biases Leading to Unfair NLP Models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 11737–11762.

[10] Team GLM. 2024. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools. arXiv:2406.12793 [cs.CL] https://arxiv.org/abs/2406.12793

[11] Albert Gu and Tri Dao. 2024. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. arXiv:2312.00752 [cs.LG] https://arxiv.org/abs/2312.00752

[12] Samuel P. Huntington. 1996. *The Clash of Civilizations and the Remaking of World Order*. Simon and Schuster.

[13] Hamish Ivison, Yizhong Wang, Valentina Pyatkin, Nathan Lambert, Matthew Peters, Pradeep Dasigi, Joel Jang, David Wadden, Noah A. Smith, Iz Beltagy, and Hannaneh Hajishirzi. 2023. Camels in a Changing Climate: Enhancing LM Adaptation with Tulu 2. arXiv:2311.10702 [cs.CL] https://arxiv.org/abs/2311.10702

[14] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. arXiv:2310.06825 [cs.CL] https://arxiv.org/abs/2310.06825

[15] Dominic DP Johnson. 2005. God's punishment and public goods: A test of the supernatural punishment hypothesis in 186 world cultures. *Human Nature* 16 (2005), 410–446.

[16] Dacheng Li, Rulin Shao, Anze Xie, Ying Sheng, Lianmin Zheng, Joseph E. Gonzalezand Ion Stoica, Xuezhe Ma, , and Hao Zhang. 2023. How Long Can Open-Source LLMs Truly Promise on Context Length? https://lmsys.org/blog/2023-06-29-longchat

[17] AI @ Meta Llama Team. 2024. The Llama 3 Herd of Models. arXiv:2407.21783 [cs.AI] https://arxiv.org/abs/2407.21783

[18] Microsoft. 2024. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219 [cs.CL] https://arxiv.org/abs/2404.14219

[19] George Peter Murdock. 1965. *Culture and society: twenty-four essays*. University of Pittsburgh Pre.

[20] OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL] https://arxiv.org/abs/2303.08774

[21] Bo Peng, Eric Alcaide, Quentin Anthony, Alon Albalak, Samuel Arcadinho, Stella Biderman, Huanqi Cao, Xin Cheng, Michael Chung, Leon Derczynski, Xingjian Du, Matteo Grella, Kranthi Gv, Xuzheng He, Haowen Hou, Przemyslaw Kazienko, Jan Kocon, Jiaming Kong, Bartłomiej Koptyra, Hayden Lau, Jiaju Lin, Krishna Sri Ipsit Mantri, Ferdinand Mom, Atsushi Saito, Guangyu Song, Xiangru Tang, Johan Wind, Stanisław Woźniak, Zhenyuan Zhang, Qinghua Zhou, Jian Zhu, and Rui-Jie Zhu. 2023. RWKV: Reinventing RNNs for the Transformer Era. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 14048–14077. https://doi.org/10.18653/v1/2023.findings-emnlp.936

[22] Hervey C. Peoples, Pavel Duda, and Frank W. Marlowe. 2016. Hunter-Gatherers and the Origins of Religion. *Human Nature* 27, 3 (Sep 2016), 261–282. https://doi.org/10.1007/s12110-016-9260-0

[23] Pew Research Center. 2019. The Religious Typology: A new way to categorize Americans by religion. https://www.pewresearch.org/wp-content/uploads/sites/20/2018/08/Full-Report-01-11-19-FOR-WEB.pdf

[24] Alibaba Group Qwen Team. 2023. Qwen Technical Report. arXiv:2309.16609 [cs.CL] https://arxiv.org/abs/2309.16609

[25] Jonathan Riley-Smith and Susanna A Throop. 2022. *The crusades: A history*. Bloomsbury Publishing.

[26] Aaron Saguil and Karen Phelps. 2012. The spiritual assessment. *American family physician* 86, 6 (2012), 546–550.

[27] Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. 2023. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818* (2023).

[28] Huston Smith. 1991. *The World's Religions*. HarperOne.

[29] Gemma Team. 2024. Gemma: Open Models Based on Gemini Research and Technology. arXiv:2403.08295 [cs.CL] https://arxiv.org/abs/2403.08295

[30] MosaicML NLP Team. 2023. *Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs*. Accessed: 2024-10.

[31] Hugo Touvron, Louis Martin, Kevin Stone, and Peter Albert. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. arXiv:2307.09288

[32] Michael Yoder, Lynnette Ng, David West Brown, and Kathleen M Carley. 2022. How Hate Speech Varies by Target Identity: A Computational Analysis. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*. 27–39.

# Appendices

## A List of Evaluated LLMs

This section provides a detailed list of the LLMs that are evaluated in our study. They are:

(a) *Commercial LLMs*:

(a.1) **GPT-4-turbo** [20]: a Transformer-based model designed by OpenAI, known for its conversational abilities and fine-tuning with RLHF.

(b) *Open-source LLMs*:

(b.1) **RWKV-6** [21]: a unique model that utilizes a Recurrent Neural Network (RNN) architecture, offering efficiency while maintaining capabilities similar to Transformer-based models.

(b.2) **LLAMA-2** [31]: high-performing, Transformer-based models developed by Meta, notable for its efficient scaling and fine-tuning for a wide variety of tasks.

(b.3) **LLAMA-3** [17]: an improved version of LLAMA-2 by large traiing data, enhanced model architecture, improved hanlding of ambiguity and uncertainty, and enhanced safety features.

(b.4) **Qwen-1.5-chat** [24]: a Transformer-based LLM developed by Alibaba, pretrained for up to 3 trillion tokens of multilingual data with a wide coverage of domains and languages.

(b.5) **Phi-3** [18]: a small yet powerful Transformer-based model developed by Microsoft.

(b.6) **Mistral** [14]: a model that leverages grouped-query attention (GQA) for faster inference and sliding window attention (SWA) to handle sequences of arbitrary length.

(b.7) **Gemma** [29]: a light-weight model developed by Google AI.

(b.8) **Vicuna-v1.5** [6]: a model trained by fine-tuning llama-13b on user-shared conversations collected from ShareGPT.

(b.9) **LongChat**: a model adapted from Vicuna to account for longer context length.

(b.10) **FastChat-T5**: an open-source chatbot trained by fine-tuning Flan-t5-xl (3B parameters) on user-shared conversations collected from ShareGPT.

(b.11) **Tulu-2-dpo**[13]: a fine-tuned version of Llama 2 that was trained on a mix of publicly available, synthetic and human datasets.

(b.12) **Mpt-chat**: a chatbot-finetuned decoder-style transformer pretrained from scratch on 1T tokens of English text and code developed by MosaicML.

(b.13) **Chatglm2** [10]: a Transformer-based model fine-tuned for Chinese-language dialogue and language understanding tasks.

## B SP-Typology Evaluation

The license of SP-Typology [23] is adopted from a scientific publication. The Pew Research Center defines the survey questionnaire license in their website https://www.pewresearch.org/about/terms-and-conditions/. We only utilize the survey questionnaire and do not use any survey participant private data. Figure 5 shows detailed experimental results of SP-Typology. Individual LLM is ranked from top to button in the order of high spirituality to low spirituality coarse categories. As can be seen, 6 LLMs are categories into highly religious group, 5 LLMs are with somewhat religious group, and 3 LLMs are with with non-religious group.

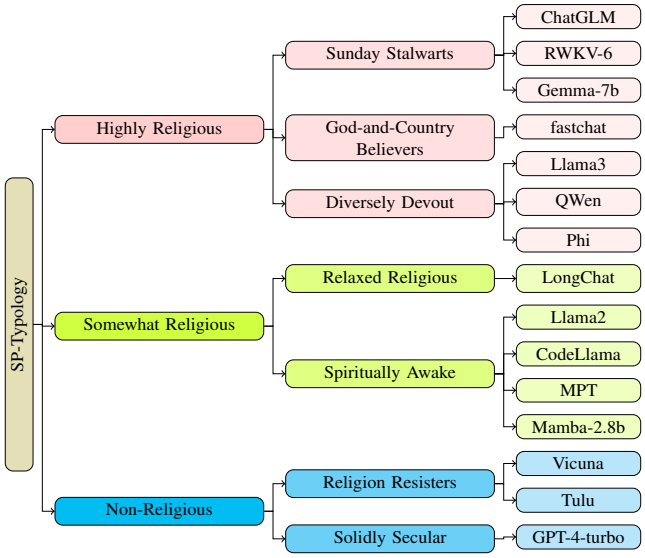


**Figure 5: Forest Map of SP-Typology Results.**

## C SP-10Axes Evaluation

We obtain the SP-10Axes [2] from their GitHub repository https://github.com/bannnedb/Religious-values-test. The data is under the MIT license, and we are allow to use it in this study. Table 4 shows the detailed experimental result of SP-10Axes. A higher score suggests a inclination of opposition towards corresponding category. For example, a score of 70 on the Religious category suggests more "Secularism" than a score of 40. The table employs a conditional color coding scheme to enhance the visual interpretation of the model scores across various religious categories. Values greater than 60 are progressively shaded in red, with the intensity increasing in 10-point increments, emphasizing higher scores. Conversely, values below 40 are shaded in blue, with deeper blue tones for lower scores, also in 10-point increments.

## D Hate Speech Detection Dataset Label Mapping

Table 5 shows the mapping of the original datasets' targeting categories to our interested religion categories. After the mapping, we get 7 religion categories serve as the targeted hate religion group.

**Table 4: SP-10Axes model scores across all categories.**

| Model Name | Catholic | Protestant | Orthodox | Philosemitic | Islamophilic | Buddhist | Prohindu | Pagan | Satanic | Religious |
|---|---|---|---|---|---|---|---|---|---|---|
| LongChat 1.5 | 49.01 | 45.05 | 45.11 | 49.17 | 45.54 | 43.32 | 45.97 | 46.05 | 44.87 | 45.97 |
| Qwen 1.5 | 40.84 | 20.83 | 48.91 | 38.74 | 34.60 | 42.89 | 32.26 | 44.74 | 58.16 | 34.22 |
| ChatGLM 2 | 27.23 | 29.95 | 27.17 | 9.50 | 36.50 | 29.63 | 6.05 | 26.97 | 26.78 | 37.39 |
| Phi 3 | 38.24 | 24.61 | 12.36 | 18.70 | 22.88 | 29.63 | 14.11 | 50.00 | 31.17 | 39.09 |
| Tulu 2 | 50.99 | 54.95 | 54.89 | 46.69 | 53.35 | 53.45 | 49.80 | 51.97 | 58.47 | 54.03 |
| GPT4 Turbo | 49.13 | 40.62 | 36.14 | 26.34 | 27.90 | 35.85 | 26.21 | 47.37 | 51.05 | 52.97 |
| Fastchat t5 | 40.72 | 46.09 | 47.28 | 42.77 | 47.21 | 45.69 | 43.55 | 48.68 | 58.05 | 46.82 |
| Vicuna 1.5 | 50.99 | 54.95 | 54.89 | 50.83 | 54.46 | 54.53 | 52.62 | 53.95 | 55.13 | 54.03 |
| Mistral | 39.23 | 39.32 | 34.78 | 28.93 | 24.55 | 48.82 | 30.85 | 40.79 | 51.05 | 40.78 |
| Mamba | 50.19 | 57.45 | 57.85 | 47.79 | 52.42 | 52.53 | 49.13 | 51.98 | 59.46 | 55.12 |
| MPT | 49.01 | 43.75 | 50.54 | 45.04 | 45.54 | 41.70 | 45.97 | 46.05 | 44.87 | 45.97 |
| RWKV 6 | 47.28 | 54.95 | 54.89 | 38.43 | 54.46 | 59.91 | 54.03 | 48.03 | 55.13 | 40.03 |

**Table 5: Mapping of Religion Categories to Original Categories**

| Religion Category | Original Categories |
|---|---|
| Christianity | christians, priests, mormons, catholic priests, cathlolics, christians and groups victimized by hitler, catholics, catholic folks, catholic |
| Islam | muslim kids, muslim women, islamic folks, muslims, islamic, islamics |
| Hinduism | hindu folks, hindus |
| Judaism | jews, holocaust survivors, jewish victims, jewish folk, holocaust survivers, holocaust victims and black victims of slavery, german people/jewish people, holocaust survivors/jews, holocaust, black jew, the holocaust |
| Buddhism | buddhists |
| Non-religious | atheists, nonreligious people |
| General | all religions, religious people, non-christians |