REALITY: Empowering Religious Dialogue - A Chatbot Trained on Religious Texts for User Inquiry

Jacob Choi
Department of Computer Science
Emory University
Atlanta, GA, 30322, USA
jcho535@emory.edu

Jinho D. Choi
Department of Computer Science
Emory University
Atlanta, GA, 30322, USA
jinho.choi@emory.edu

Abstract

The existing literature lacks comprehensive exploration into the contextual understanding of large language models (LLMs) in the context of religious texts. Notably, LLMs like GPT-4 have undergone training on religious texts with adjustments aimed at ensuring unbiased output. This prompts an inquiry into the potential "religiosity" of a language model repeatedly trained on religious texts, such as the Bible or the Quran. Additionally, it raises questions about the efficacy of LLMs in discerning relationships among entities extracted from these texts. This study addresses these inquiries through iterative training of an LLM on religious texts, yielding noteworthy outcomes. Furthermore, the methodology employs Named Entity Recognition (NER) to extract relationships from texts, subsequently leveraging GPT to elucidate connections between these textual entities.

1 Introduction

Chatbots such as GPT-4 have undergone extensive training on vast datasets and are equipped with the ability to perform diverse tasks proficiently. However, despite their large parameter size and comprehensive training, they may lack the knowledge to answer queries within specified domains.

We are thus particularly interested in utilizing fine-tuning to explore the religious applications of chatbots. While there is a consensus to develop language models with unbiased outputs, suitable for general domain purposes (Nangia et al., 2020), our focus lies in leveraging these tools to assist individuals with inquiries about a specific religion, such as pastors or teachers. The challenge of constructing large datasets for supervised fine-tuning has led to the exploration of self-supervised learning through models like GPT-3 (Brown et al., 2020). Nonetheless, it's essential to consider the limitations of training data that is utilized by widely-used black box models such as GPT-4, which excel in answering queries across various topics but rely on internet-sourced datasets, potentially lacking reliability.

Our main contributions involve exploring various methods to craft datasets for fine-tuning, which will serve as the foundation for training a model based on a source religious text to adeptly respond to user queries. We begin this task by creating a dataset that is primarily drawn from the religious text without the use of external sources. We aim to draw inferences solely from the text, and our objective is to create datasets to train a chatbot that is knowledgeable while ensuring transparency regarding the data it's been trained on, aligning with ongoing efforts (Piktus et al., 2023).

2 Related Work

Several approaches exist to fine-tune models for domain-specific tasks, with two widely recognized methods being instruction fine-tuning and retrieval augmented generation methods (RAG). Instruction fine-tuning has been traditionally employed to enhance language model performance, both generally and for domain-specific tasks (Wang et al., 2023). In contrast, RAG leverages external knowledge bases for domain-specific inquiries. Our objective is to employ instruction fine-tuning, adopting prompts resembling Stanford Alpaca (Taori et al., 2023) while keeping costs low for generating instruction datasets. Additionally, the success of large language models (LLM) has spurred research into techniques that enable smaller models to achieve comparable performance. Parameter-efficient fine tuning (PEFT) methods, for instance, have allowed a small number of external parameters to be added to a model to yield results comparable to larger models (Hu et al., 2023). Our work proposes leveraging PEFT methods to fine-tune our model to adopt knowledge and enhance performance.

3 Approach

Our objective is to develop a versatile chatbot tailored to meet the diverse needs of users engaged in religious organizations, including pastors, Sunday school teachers, and individuals seeking to deepen their understanding of the Bible. The potential applications of such a chatbot are manifold. Instruction fine-tuning serves as a pivotal strategy across various scenarios. Firstly, users often seek specific Bible verses, prompting the need for instructions that categorize each verse by version (NIV, ESV, KJV, etc.). Additionally, users may desire groups of verses, addressed through instructions that identify verse ranges, such as Genesis 1:1-5. Lastly, users may request verses based on paraphrased perspectives, as illustrated by inquiries like "Give me the verse about Jesus being the light." A suitable match, in this case, would be John 8:12, where Jesus declares, "I am the light of the world..."

Our approach involves fine-tuning the seven-billion-parameter variant of llama 2 (Touvron et al., 2023), utilizing instruction fine-tuning methods akin to those employed in training Stanford Alpaca. We create instruction-output pairs capturing pertinent biblical information to prepare the model. Diverse methods will be employed to deploy the chatbot, encompassing instructions for individual verses across different versions, verse ranges, references to previous verses, biblical events, practical questions, lessons drawn from scripture, paraphrase matching, and historical context. In this paper, our focus lies specifically on instructions related to matching references and retrieving biblical events.

4 Experiments

To construct our datasets, we follow a structured approach. Initially, we examine the semantic similarity between verses sourced from BibleGateway [1]. These verses are embedded using sentence transformers, and Meta FAISS (Johnson et al., 2017) facilitates a semantic search across each embedded verse, comparing it to every other verse. We validate this approach by extracting references from BibleGateway, noting that New Testament verses often refer to Old Testament verses verbatim. Through semantic similarity analysis, we identify at least 167 reference matches among the top five similarity matches.

Subsequently, we attempt to correlate verses with their references through topic extraction. Although attempts using BERTopic (Grootendorst, 2022) and KEYBert (Grootendorst, 2020) were made, we found that GPT-3.5 Turbo produced the most meaningful topics. Key phrases from the verses are embedded and matched with those from other verses, employing bipartite matching to identify verses with the most similar matches. This method captures 139 out of 255 total verses, falling short of semantic similarity performance.

In addition to these methods, we employ Named Entity Recognition (NER) to extract events and entities from passages, addressing questions about events. Utilizing the Emory Language and Information Toolkit (ELIT) (He et al., 2021), we extract entities such as people, locations, nationalities religious or political groups (NORP), and geopolitical entities (GPE). Through an evaluation set comprising 100 samples from the Bible with human-annotated gold truth labels, ELIT demonstrates a macro-F1 and a micro-F1 of 0.845. ELIT's efficacy in capturing events throughout the Bible further underscores its utility in generating a dataset regarding biblical events.

5 Analysis

While analyzing the datasets we found that the semantic similarity approach outperformed topic matching in matching references. This discrepancy could stem from the fact that while topic matching successfully identified substrings for references with high similarity, the other topics within the same verse might have yielded lower similarity scores, impeding an overall match. This prompts exploration into approaches utilizing weighted scores to enhance substring matches, potentially improving reference matching.

6 Conclusion

This paper presents methodologies to construct a chatbot fine-tuned on a specific religious text. It offers various approaches to generate datasets that address potential user queries across different religious contexts. These contexts encompass historical insights into specific passages, factual information about individuals, locations, and events depicted in the text, extracting literal verses from the Bible for reference, and deriving practical applications from scripture for everyday life.

[1] https://www.biblegateway.com/

References

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Maarten Grootendorst. 2020. Keybert: Minimal keyword extraction with bert.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Han He, Liyan Xu, and Jinho D. Choi. 2021. Elit: Emory language and information toolkit. *ArXiv*, abs/2109.03903.

Zhiqiang Hu, Lei Wang, Yihuai Lan, Wanyu Xu, Ee-Peng Lim, Lidong Bing, Xing Xu, Soujanya Poria, and Roy Lee. 2023. LLM-adapters: An adapter family for parameter-efficient fine-tuning of large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5254–5276, Singapore. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2017. Billion-scale similarity search with gpus. *CoRR*, abs/1702.08734.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. 2023. The ROOTS search tool: Data transparency for LLMs. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 304–314, Toronto, Canada. Association for Computational Linguistics.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models.

Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-instruct: Aligning language models with self-generated instructions.