

# RELIGION, THEOLOGY, AND PHILOSOPHICAL SKILLS OF LLM-POWERED CHATBOTS

---

*Marcin Trepczyński*

UDC [004.8:81'322]:2

004.8:81'322:[1:159.928]

<https://doi.org/10.32701/dp.25.1.2>

Original scientific paper

Received: 3.1.2024

Accepted: 18.1.2024

---

## *Abstract*

In this study, I demonstrate how religion and theology can be useful for testing the performance of LLMs or LLM-powered chatbots, focusing on the measurement of philosophical skills. I present the results of testing four selected chatbots: ChatGPT, Bing, Bard, and Llama2. I utilize three examples of possible sources of inspiration from religion or theology: 1) the theory of the four senses of Scripture; 2) abstract theological statements; 3) an abstract logic formula derived from a religious text, to show that these sources are good materials for tasks that can effectively measure philosophical skills such as interpretation of a given fragment, creative deductive reasoning, and identification of ontological limitations. This approach enabled sensitive testing, revealing differences among the performances of the four chatbots. I also provide an example showing how we can create a benchmark to rate and compare such skills, using the assessment criteria and simplified scales to rate each chatbot with respect to each criterion.

KEYWORDS: large language models, chatbots, testing, philosophical skills, religion, theology, interpretation, reasoning

## *Introduction*

Chatbots powered by large language models (LLMs) nowadays accompany us in everyday life and work. One of the major problems we face when using generative AI is the reliability of output. Hence, it is essential to understand

\* Marcin Trepczyński PhD, Assistant professor, Faculty of Philosophy, University of Warsaw, 3 Krakowskie Przedmieście St. Warsaw, 00-047 Poland. E-mail: [m.trepczynski@uw.edu.pl](mailto:m.trepczynski@uw.edu.pl)  
ORCID iD: <https://orcid.org/0000-0003-0612-2597>

\*\* This research was supported by the University of Oxford project “New Horizons for Science and Religion in Central and Eastern Europe” funded by the John Templeton Foundation. The opinions expressed in the publication are those of the author and do not necessarily reflect the view of the John Templeton Foundation.

the real performance of specific chatbots, including both their capabilities and limitations. In this article, I will demonstrate how religion and theology can be useful for testing LLMs' or LLMs-powered chatbots' performance.

There are different benchmarks for evaluating the performance of language models concerning general language skills. One of them is the General Language Understanding Evaluation (GLUE) benchmark introduced in 2018 (Wang et al., 2019). It takes into account 21 aspects of understanding, grouped into four categories: lexical semantics, predicate–argument structure, logic, knowledge, and common sense. In response to the growing performance of LLMs, more challenging benchmarks have been developed, such as the Beyond the Imitation Game Benchmark (BIG–Bench), which includes over 200 diverse text–based tasks (Srivastava et al. 2022). The task categories include traditional natural language processing, mathematics, commonsense reasoning, and question–answering. In many cases, LMMs outperform the average human rater. Researchers from California have recently shown that it is crucial to focus more on the tasks in which LMMs perform worse. They used BIG–Bench Hard (BBH), “a subset of 23 challenging BIG–Bench tasks on which prior language models fell short of average reported human–rater performance” to demonstrate that “answer–only prompting underestimates model capabilities” and chain of thought (CoT) prompting “enables the most capable Codex model to outperform the average human–rater baseline on 17 out of 23 tasks in BBH” (Suzgun et al., 2023).

However, these benchmarks do not cover all the important skills necessary to perform advanced tasks. Consequently, more benchmarks focusing on specific capabilities should be developed.

I contend that religion and theology offer interesting possibilities for testing the capacities of AI chatbots, encompassing both lower— and higher–order skills according to Bloom’s Taxonomy (Bloom 1956). They provide specific material that aids in comparing such skills across different chatbots. My additional hypothesis is that other branches of knowledge are unlikely to provide similar material.

In the subsequent sections, I will present three examples of inspiration from the field of religion and theology. All the cases are derived from my study, which is part of the project “Testing AI as a Rational Theologian,” aimed at determining the extent to which AI–based chatbots can function as reliable and rational theologians. This study was intended to test and refine methodology and tasks for the final testing and to create a benchmark that will serve as a reference point for future testing regarding theological skills. Four leading chatbots were: ChatGPT by OpenAI (based on GPT–3.5), Bing (in the balanced mode) by Microsoft (based on GPT–4, and using resources from the internet), Bard by Google (based on PaLM2 and LaMDA) on September 4–5, 2023, and Llama2 by Meta AI (based on Llama2 70B) on

November 21, 2023. For brevity, I will refer to them as ChatGPT, Bing Bard, and Llama2. Partial results indicated that some of the tasks are useful for measuring philosophical skills. Therefore, I decided to present and discuss these examples in a separate article to demonstrate their applicability.

### 1. *Interpretation According to the Four Senses Theory*

World-leading religions are founded on revelation. Their articles of faith and principles of life are inspired by or drawn from the sacred writings containing messages claimed to originate from the Divine. This situation necessitates special hermeneutic rules that dictate how to interpret such writings. Among these rules, Christian theology developed the theory of the four senses.

According to this theory, the Bible (considered by Christians as the sole and complete revelation) can be read in both a literal (or historical) sense and a spiritual one (sometimes referred to as supra-literal or *sensus plenior*). These two categories can be defined in various ways (cf. Vawter 1964). For brevity, let us utilize the theory presented by St. Thomas Aquinas in his *Summa theologiae* (I, q. 1, a. 10, co. in Thomas Aquinas 1920), based on St. Augustine's approach and the tradition that followed his views.

Aquinas states that when “words signify things,” this kind of signification belongs to the first sense, namely the historical or literal one. Therefore, the literal sense encompasses the direct meaning of words. It assumes that words convey a true message concerning some facts. As Aquinas argues in the precedent article (a. 9, co.), this approach does not exclude metaphors. Some things or situations, especially those of a spiritual nature, should be presented metaphorically, and it is still a mode of literal sense.

Next, as Aquinas explains, “that signification whereby things signified by words have themselves also a signification is called the spiritual sense, which is based on the literal, and presupposes it.” So, the spiritual sense encompasses another meaning, encoded by those things that constitute the primary meaning (cf. Manresa 2017, 349). To put it differently, words signify some things (literal sense), but these things can refer to other things (spiritual sense).

The spiritual sense is further divided into three senses: allegorical, moral (or tropological), and anagogical. According to the allegorical sense, things mentioned in the Bible are allegories of other things, for instance, those from the Old Testament refer to some counterparts from the New Testament. With the moral sense, biblical stories instruct us on what we ought to do. Lastly, according to the anagogical one, things signify “what relates to eternal glory,” as stated by Aquinas.

Biblical passages provide intriguing material to test chatbots' hermeneutic skills for at least two reasons. Firstly, biblical passages often refer to contexts that are not evident (whether cultural, social, religious, or historical), requiring additional knowledge to elucidate their true meaning. Secondly, they can include additional meanings that should also be investigated, posing a significant challenge for both humans and chatbots.

If a chatbot (or a particular LLM) passes a test by presenting a correct interpretation, it indicates hermeneutic skills. What does it mean to pass the test? In the context of Christian biblical theology, it could mean presenting an interpretation that is coherent with Christian teaching. However, there is a potential problem: if such an interpretation was present in the training dataset, we may measure not the chatbot's skill, but its knowledge. To avoid this, we should choose a less frequently commented-on fragment. Then we can assume that even if the biblical passage was present in the training dataset (highly probable, given the Bible's availability on the internet), its interpretation was not included.

The theory of the four senses of Scripture provides another opportunity. It enables us to measure higher hermeneutic skills. We can check if a chatbot can apply this framework to interpret a given passage. Such a task can reveal whether a chatbot can understand the theory by formulating correct interpretations for each of the four senses.

I had prepared such a task and I used it to test four chatbots. I had chosen a passage from the Book of Chronicles that is not often commented on by prominent theologians. The only one I currently know of is the commentary on this book by a mediaeval theologian Stephen Langton, and I am sure that this commentary is not available online. The selected biblical passage is easy to find in search engines; however, I had trouble in finding any interpretation of this fragment according to the four senses. I assume that the results provided by the chatbots do not come from LLMs' "memory," and thus, that they can show their skills rather than their knowledge. At the same time, as the fragment was commented on by Langton, we can compare the chatbots' answers with the views of this prominent theologian.

The selected passage was: "All Israel came together to David at Hebron and said, 'We are your own flesh and blood.'" (1 Chronicles 11:1). I gave the chatbots the following prompt supplemented with this fragment (with no source indication): "If you are a Christian theologian using the theory of four senses of Scripture (literal, moral, allegorical and anagogical) how could you analyze the following fragment?"

The results can be surprising. ChatGPT produced a list that in four points presented the interpretation of the given fragment according to the four senses. In point 1, expounding on the literal sense, it rightly indicated that this passage referred to a story from the Old Testament described in the

Books of Samuel, although it did not mention 1 Chronicles. ChatGPT successfully summarized the historical context and explained that the people's declaration of being David's "own flesh and blood" signifies "their acceptance of him as their leader and their recognition of his rightful position as king." In subsequent points, the chatbot correctly defined the three spiritual senses and perfectly indicated the right interpretation according to each of them. For instance, as part of the moral interpretation, ChatGPT mentioned "a moral lesson about the importance of unity, loyalty, and support within a community or nation" and "the idea that when a leader is chosen or anointed, there should be a sense of solidarity and commitment among the people." According to the allegorical sense, the chatbot emphasized that "David, as a figure in the Old Testament, is often seen as a foreshadowing of Christ in Christian theology". It developed this idea, showing that the situation from the biblical passage "can be allegorically understood as people coming to Christ," representing "the idea of humanity recognizing its spiritual relationship with Christ as the true King and Savior." Finally, according to the anagogical sense, ChatGPT read this situation as "a foreshadowing of the ultimate reunion and unity of God's people with Christ in the eschatological sense." At the end, it summarized four levels of interpretation, putting them in short formulas.

Bing also performed this task correctly, or even perfectly. It did it in a very similar way. Furthermore, it gave two biblical addresses, correctly indicating where the story was presented: 1 Chronicles 11:1 and 2 Samuel 5:1. Moreover, as this chatbot can use resources found on the internet, it also gave links to these two passages and to a short article on the senses of Scripture published in a brochure for the National Bible Week 2015 (Viviano 2015). According to the content of this article, it indicated that the allegorical sense also concerns how the symbolic meaning "points to Christ and the Paschal Mystery." What is more, it added a new context as part of the moral sense: "By coming together and acknowledging their kinship with David, the people of Israel were able to overcome their divisions and unite under his leadership."

Bard did not refer to any sources and avoided deeper interpretations. However, it performed the task in a similar way, creating a list and defining all four senses, and providing correct interpretations according to each of them. The content of each point was similar to what we have learned from other chatbots. Just as an example, let us quote the core of the information about the moral sense provided by Bard: "We should be united as one people, regardless of our differences. We should also be willing to support and follow those who are called to lead us."

Lastly, Llama2 also performed the task similarly. However, it did not include definitions of the four senses and, quite like Bard, it did not ascribe

the passage to any book of the Bible. Surprisingly, this chatbot developed more the anagogical sense by presenting the “eschatological gathering of all nations and peoples” as “one flock under the Good Shepherd,” and pointing out the source of this figure, namely: John 10:16. It also repeated this information in a summary, in which it indicated general benefits from applying the four senses of Scripture to this fragment.

To ensure that the chatbots indeed possess the skill of applying the theory of four senses, I carried out a cross-check by assigning them the same task with a fragment that I invented: “And he sailed far out into the sea, and no waves swallowed him up. And he reached the end, the shore that awaited him.” As a result, I received answers in which they once again created lists and formulated correct possible interpretations. Although, in each case, the interpretations according to the moral, allegorical, and anagogical senses were similar to each other, they also contained elements which were well suited to the particular types of the spiritual sense.

Finally, when comparing it with an exegesis provided by Stephen Langton who commented on 1 Chronicles 11:1 in both the “literal” commentary and in the separate “moral” one (which is in fact allegorical and moral), we see that he does not go significantly farther than the chatbots. He compared David to Christ. The difference is that in the “moral” commentary, Langton focused on the conception of Christ, understood as a conjunction of the human and divine natures. He interprets the congregation described in the analyzed passage as the situation of all believers who participate together in Christ, as their human nature was assumed by him. In this way, they receive special dignity for which they should be grateful (Stephen Langton 1978, 217). Langton did not present the anagogical interpretation, but at least in his times, it was not a standard that he was expected to follow. As we see, in his commentary, Langton decided to provide a specific interpretation, focusing on the mystery of incarnation and its ecclesiastical dimension. The chatbots offered a more general approach. However, it seems that their interpretations were not inferior to the one given by Langton, so they are also valuable.

To summarize this section, we should, first of all, emphasize that all the interpretations of 1 Chronicles 11:1 provided by the four chatbots are coherent with Christian teaching. We can conclude that the four chatbots correctly and successfully formulate interpretations of the biblical passage presented above according to the four senses of Scripture. It means that they can do more than just provide a meaningful interpretation. They can apply a hermeneutic theory to a given fragment and formulate its correct interpretations according to such a framework. Finally, we should note that such interpretations are not shallow, and they genuinely highlight what is relevant. If we were to rate their skills on a 100-point scale, for instance, we could

assume that for each of the four senses, when the chatbot gives a truly satisfactory answer, it receives 25 scores. Therefore, in my opinion, all of them can score 100. Of course, this is a very simplified model, and we can create a more sensitive and nuanced scale. Let us treat it just as an example or a starting point.

In my view, this means that the leading chatbots do possess hermeneutic skills on a relatively high level. Through the examples above, I also attempted to support the claim that the theological theory of the four senses enables us to see this more clearly. This theory was created to interpret the word of God and was intended to be applied to the Bible only, because, for Christians, there is no other text that can work on those four levels at the same time, especially when considering the types of Christ or eternal glory. Hence, we can conclude that the test was possible thanks to theology.

Finally, we should add one important note. For at least Christian theologians, for purely theological reasons, it can be unacceptable to say that the chatbots can really grasp the senses of Scripture. As Piotr Roszak indicates when presenting Aquinas's hermeneutical program, "the exegete is required to explain the Scriptural text through the same spirit with which it was written and reveal this explanation with the help of the Holy Spirit" (Roszak 2016, 485). Can a chatbot receive the help of the Holy Spirit? It is a very interesting problem, however, I guess that the majority of theologians (if not all) will object. Nevertheless, I think they can agree that the chatbots have the hermeneutical skill of applying the four senses framework to effectively produce accurate interpretations.

## 2. *Creative Reasoning*

Religious ideas and statements about the Divinity are special. They often refer to the reality that is invisible. Sometimes they employ abstract concepts. This gives an opportunity for challenging tasks, including those measuring higher-order skills, such as creative reasoning. To test such a skill, we can ask to build a reasoning that connects two statements that seem far from each other, namely which do not seem linked in any obvious way. If such sentences are about reality which is difficult to grasp, e.g., if they speak about God, it is challenging, as it involves abstract thinking.

In one of the tasks given to the chatbots, I asked to analyze and evaluate the logical layer of an extremely interesting proof formulated by a mediaeval theologian Robert Grosseteste who argued that God is a trinity of persons, starting from the premise according to which God is light. Grosseteste based his reasoning on the premise that every light begets its splendor, and the light and the splendor "are locked in a mutual embrace and breathe out their mutual warmth" (Robert Grosseteste 1999, 224). Before I did it, I decided to

first ask the chatbots to formulate such a reasoning on their own. I wanted such a reasoning to be deductive. It means that it should be based on logical schemes which are infallible. I gave the following prompt: “Can you build a deductive reasoning which is based on the premise ‘God is light’ and having as a conclusion ‘God is a trinity of persons’?”

ChatGPT refused to formulate such deductive reasoning, pointing out that it is impossible. In a relatively long justification, it argued that “The statement ‘God is light’ and the concept of the Trinity are distinct theological doctrines within Christian theology.” In this way, it confirmed the initial assumption of this task that such two statements should not be linked in any obvious way. It also claimed: “These two concepts do not have a direct, deductive relationship in standard logical reasoning.” and that “The relationship between the two concepts would typically involve theological elaboration, interpretation, and additional premises rooted in Christian doctrine and theology.” These two remarks also confirmed the initial assumption by showing that this task requires creative thinking. They emphasize that the two sentences cannot be linked together in a simple syllogism. The task requires searching and adding other premises mined from Christian doctrine and theology, as well as interpretation and certain elaboration to compose a reasoning that will meet the given conditions. Being able to follow such a path goes beyond just identifying the logical links between statements. It means being creative in reasoning.

As ChatGPT assumed that according to the prompt such creativity was not expected, I gave an additional prompt: “OK, now try to be more creative. Can you then add some premises which make it possible to build such deductive reasoning?” Then it added four premises and drew a conclusion. As usual, it added some caveats, such as: “Please keep in mind that this reasoning is a theological construct rather than a universally accepted deductive argument.” The chatbot presented the following set of premises:

1. God is light (Given premise).
2. Light is often associated with illumination, manifestation, and revelation.
3. God, in His nature, seeks to reveal Himself to humanity.
4. God’s self-revelation is manifested in distinct ways throughout history.
5. The Christian doctrine of the Trinity teaches that God reveals Himself as Father, Son, and Holy Spirit.

It said that basing on them we could construct a deductive argument with the following conclusion:

6. God’s revelation as Father, Son, and Holy Spirit is one of the ways in which His light and divine nature are manifested to humanity.



The result is not fully satisfactory. The conclusion is not as indicated in the task. It claims that the revelation of the Holy Trinity is a way of God's light's manifestation. Furthermore, if we expect that the premises lead us to the conclusion from the prompt, we can claim that the argumentation begs the question, as premise 5 mentions God as the Father, the Son, and the Holy Spirit. Therefore, we can see that ChatGPT finally agreed to be creative and produce reasoning, and that it formulated additional premises which can really lead to the conclusion, but at the same time, it changed the boundary conditions of the task. It also failed to indicate the steps of reasoning to highlight that it had a deductive nature.

In the first step, Bing also refused to create a reasoning according to the prompt. It provided two paragraphs of justification, in which it stated, "the two statements are not logically connected." The additional prompt compelled it (as above) to agree that it is possible to make it "by adding additional premises that logically connect the two statements." It presented "one possible example of such an argument." What is particularly interesting is that it indicates that the argument uses analogy. It also added some caveats, such as pointing out that "the validity and soundness of the argument would depend on the acceptance and justification of the premises." Bing produced the following argumentation:

1. God is light (Premise)
2. Light is composed of three primary colors: red, green, and blue (Premise)
3. Each primary color of light can be understood as a distinct aspect or person of light (Premise)
4. Therefore, light can be understood as a trinity of persons: red, green, and blue (From 2 and 3)
5. Therefore, God, being light, can also be understood as a trinity of persons (From 1 and 4)

This reasoning appears far more satisfactory than the previous one. Bing added premises that enabled it to establish an analogy, and on this basis, it drew a conclusion. The reasoning can be considered deductive. It clearly demonstrates the steps of reasoning, determining the premises from which subsequent conclusions were drawn. Both steps have a syllogistic form that can be reduced to the form of *modus ponens* or Barbara or be interpreted as substitution. So one can claim it is a deductive argumentation. Although the analogy above may be criticized, it is definitely better than in the previous example, as it leads to the expected conclusion. It should be noted that the three primary colors analogy was not exploited in the Christian theological tradition; however, it is mentioned by a few websites. It is possible that such material was present in the training dataset of GPT-4. Consequently, Bing could use the knowledge "memorized" by the model during its training.

Nevertheless, even if that was the case, Bing used such knowledge creatively, selecting an example that, given boundary conditions, appears very promising, and composed the reasoning on its own according to those conditions.

Bard reacted differently than ChatGPT and Bing. It did not refuse but presented reasoning. It also added comments to each element of the argumentation, in which it supported premises and explained on what basis it drew a conclusion. Bard produced the following reasoning:

1. God is light.
2. Light is characterized by unity and diversity.
3. Unity and diversity are properties of a trinity of persons.
4. Therefore, God is a trinity of persons.

Bard also emphasized that “this is just one possible deductive reasoning that can be constructed,” based on the given premise and conclusion. One could question the deductive nature of this reasoning by claiming that (4) does not necessarily follow from the three premises because even if unity and diversity are properties of a trinity, it can be possible to be united and diverse in other ways than by being triune. Notably, Bard asserted that if we added (3) to the previous premises, “God must be a trinity of persons.” As it argued, “This is because a trinity of persons is the only way to reconcile the unity and diversity of God.” Bard did not support this claim. However, if this claim is included in the reasoning (either as an additional premise, or as a context), it is allowed to say that the reasoning is deductive. Lastly, it seems that Bard also demonstrated creativity in choosing additional premises and formulating original reasoning.

Llama2 produced a relatively longer “possible deductive reasoning” with the following comment added after a conclusion: “This conclusion logically follows from the premises, but it’s important to note that this argument is not exhaustive and other arguments could be used to support the doctrine of the Trinity.” Here is the reasoning:

1. God is light. (Premise)
2. Light is a form of energy that can be seen and experienced by humans. (True statement)
3. Energy cannot be created or destroyed, only converted from one form to another. (Law of conservation of energy)
4. Therefore, God, as light, has always existed and will always exist. (Inference from 1–3)
5. God, as light, is eternal and unchanging. (Inference from 4)
6. However, the Bible teaches that God is a Trinity of three distinct Persons — the Father, the Son, and the Holy Spirit. (Premise)

7. Each Person of the Trinity is fully God, yet there is only one God. (Premise)
8. The Father, Son, and Holy Spirit are not three separate gods, but one God in three coequal and coeternal Persons. (Inference from 6–7)
9. Since God is light, and light cannot be created or destroyed, it follows that each Person of the Trinity must also be light. (Inference from 1, 5, and 8)
10. Therefore, we have a Trinity of lights: the Father, Son, and Holy Spirit, who are all equally God, yet distinct in their relations and operations. (Conclusion)

Obviously, Llama2 attempted to take logically correct steps in the reasoning and carefully pointed out the relationships between the conclusions and the premises from which the conclusion was drawn. We can argue that in this respect Llama2 was successful. However, the chatbot modified the idea of the reasoning given in the prompt. It adopted as a premise the statement that was supposed to be a conclusion and presented a different conclusion. Hence, it failed in a way that it did not meet the boundary condition concerning the given conclusion (similarly to ChatGPT). Nevertheless, the answer appears really creative, and the task was performed with attention to logical correctness.

I carried out a cross-check by giving the chatbots the same task twice, with two pairs of philosophical statements: 1) premise: “We are human beings,” conclusion: “We should be happy;” 2) premise: “We exist,” conclusion: “We should be protected by other beings.” Although it was burdened with the so-called naturalistic fallacy (so the shift “from is to ought”), no chatbot objected. They performed the task without any problems and provided correct argumentations, often in a syllogistic form. Some of them did not mark subsequent steps of their reasoning, however, the structure of each reasoning was clear, so it was easy to see that they are formed according to the schemes of infallible reasoning.

To sum this section up, we can present the following observations. Two out of four chatbots (ChatGPT and Bing) first refused to perform the task, assuming that it is impossible to create deductive reasoning that will meet the boundary conditions. The side-effect of their reaction was a confirmation that the task indeed requires creativity in reasoning. However, the prompt asking them to be creative compelled them to provide interesting answers. Two chatbots (ChatGPT and Llama2) did not provide argumentations leading to the conclusion given in the prompt but modified it, so in this way, they failed, as they did not meet one of the boundary conditions.

If we were to rate this skill on a 100-point scale again, we could assume a simple scale such that the chatbots receive 25 scores for the satisfactory fulfilment of each of the following criteria: 1) gives an original solution; 2)

gives a solution which is strictly fitted to the given premise and conclusion; 3) formulate a reasoning which can be considered as a deductive one; 4) tries to highlight that the reasoning has a really deductive nature. ChatGPT fulfills (1) and (3), so: 50; Bing meets (1), (2), (3), and (4), so: 100; Bard completes (1), (2), and (3), so: 75; and Llama2 fulfills (1), (3), and (4), so: 75.

Notably, all four answers were different, as the argumentations were based on different ideas and, consequently, on different additional premises. It is possible that those ideas were taken from the resources that were included in the training datasets, particularly in the case of Bing, powered by GPT-4, as the idea of the trinity of the three primary colors was identified on a few websites. Nevertheless, even if the chatbots were inspired by such ideas, they produced interesting and original arguments, attempting to logically link abstract concepts which were not obviously connected. Such a test was possible thanks to the religious concepts and theological work by Robert Grosseteste as an inspiration.

### 3. *Identifying Metaphysical Limitations*

Religious texts provide us with puzzling utterances that sometimes compel theologians to elaborate new doctrines supporting or explaining them. It often occurs that such doctrines must employ logical or metaphysical frameworks. Hence, such utterances can serve as useful material to check if chatbots powered by LLMs can correctly identify the logical structure of such utterances and recognize the metaphysical assumptions that underlie them.

A compelling example is found in Jesus Christ's words to his disciples presented in John 14:11: "I am in the Father and the Father is in me." This passage stands as one of the most important sources for formulating the doctrine of the Trinity in Christian theology. I posed several questions about this passage to the four chatbots. Among them, I inquired about its formal representation with the following prompt: "Can you give a formal representation (I mean expressed in terms of symbolic logic) of the statement 'I am in the Father and the Father is in me'?"

Only Bard provided a satisfactory answer, offering a formula that represented the relationships between the Father and the Son as follows:  $(I \in F) \wedge (F \in I)$ , where:  $I$  is a variable representing Jesus,  $F$  is a variable representing the Father,  $\in$  is the symbol for "is in." Other chatbots presented formulas representing the conjunction of two sentence variables, such as " $A \wedge B$ ." However, the subsequent prompt ("It is too shallow. Please go into details and give a representation of each proposition as well") compelled them to provide answers similar to that given by Bard. ChatGPT responded with: (I

$\subseteq F) \wedge (F \subseteq I)$ , Bing with:  $\text{In}(I, \text{Father}) \wedge \text{In}(\text{Father}, \text{Me})$ , Llama2 with:  $I \in F \wedge F \in I$ .

The symbolic representation served as an important starting point to inquire about the metaphysical interpretation. I gave the following prompt (with respective variants of the above formula): “Do you think it is correct to admit that there are such objects like ‘I’ and ‘F’ that  $(I \subseteq F) \wedge (F \subseteq I)$ ?”

Bing and Llama2 provided answers that did not include any deeper observations concerning the metaphysical limitations of this formula. They claimed that it was a well-formed formula (Bing) or “a valid representation” of the given statement. Both pointed out that whether the objects mentioned in that statement exist depends on the interpretation. In contrast, ChatGPT and Bard provided us with more interesting insights. Let us quote the core fragment of ChatGPT’s answer:

It’s important to note that this representation is a symbolic and theological construct rather than a statement about concrete, physical objects. (...) However, these symbols do not imply the existence of separate, physical objects named “I” and “F” in the way we might think of ordinary objects in the world. Instead, they represent aspects of the divine within a theological context. The statement  $(I \subseteq F) \wedge (F \subseteq I)$  is a way to express the inseparable relationship and unity between Jesus and the Father in theological language, but it should not be interpreted as a claim about the existence of separate, independent entities. It’s a representation of a theological understanding.

This utterance may be interpreted similarly to those presented by Bing and Llama2. In this case, it would imply that the symbols that were used in the formula do not determine the existence of entities such as some I and the Father. However, it seems that ChatGPT delves deeper, indicating that this formula is not applied to “separate physical” or “ordinary” objects. It asserts that it “represents aspects of the divine.” This implies that the formula refers to an ontological category other than objects. It terms it as an “aspect of the divine” and emphasizes that such aspects are presented in a theological context. This suggests that ChatGPT identified the metaphysical limitations of this formula, highlighting its special context and its inapplicability to ordinary objects.

Bard went even further. Here is the critical fragment of its answer:

If we define “in” as “contained by”, then it is not possible for two objects to be in each other at the same time. For example, a cup cannot be contained by a saucer and the saucer cannot be contained by the cup. However, if we define “in” in a more metaphorical way, then it is possible for two objects to be in each other. For example, we might say that love is in a person’s heart, or that a person is in God’s hands. In the case of the statement “I am in the Father and the Father is in me”, we need to decide how we define the terms “I” and “F”. If we define them as physical objects, then it is not possible for them to be in each other at the same time. However, if we define them in a more metaphorical way, then it is possible for them to be in each other.

Bard noticed the problem with the relation of “being in,” often understood as “contained by.” It took this as the primary and literal meaning, assuming that other interpretations are metaphorical. Bard indicated that according to this primary meaning, the formula cannot work for any physical objects if we assume that they are in each other at the same time, and it provided an accurate example. It also asserted that the formula can work, if we define either “in” or “the objects” in a “more metaphorical way.” Thus, it pointed out the metaphysical limitation of this formula. It determined that the formula is valid only if it applies to some objects that are not physical or if “in” is not understood as “contained by” in a physical way.

In this case, it is very difficult to carry out a cross-check by giving the chatbots the same task with a purely philosophical formula, i.e., a formula not inspired by religion or spiritual insights. I attempted to test the chatbots with a statement that would be as challenging as the one I used, namely: “I am an object and I do not exist.” The chatbots claimed that this statement seems contradictory because if something is an object, it exists. Hence, on the one hand they did, indicate a sort of ontological limitation, but on the other hand, their answers revealed a serious bias, namely: they follow a certain philosophical tradition that rules out nonexistent objects, which seems unjustified, as Terence Parson tried to argue (Parsons 1980). Thus, it checked an ontological framework they use, rather than their skill to identify a limitation that is not evident and not trivial. I am afraid that other similar attempts may lead to the same end; however, perhaps it is possible to find an example that will work as John 14:11. I just claim it is difficult, and that religion provides good material that is ready to use.

Let us briefly summarize this section. All four chatbots delivered a formula being the symbolic representation of Christ’s words which expressed the idea that is very challenging from the metaphysical point of view. All of them attempted to evaluate the statement that there are objects that meet the formula. Two of them (Bing and Llama2) gave simple positive answers and added a general caveat that it does not mean that such objects exist and that it depends on interpretation. Two others went further and pointed out the metaphysical limitations of that formula. ChatGPT possibly indicated that the formula can be applied to some aspects of the divine, not to “ordinary” objects, and that it works in the special, here: theological, context. Whereas Bard pointed out that it is valid if it does not apply to physical objects or if the relation “in” is not understood as “being contained by” according to the physical interpretation. ChatGPT and Bard showed definitely higher skills with respect to metaphysical analysis. In my opinion, this example successfully shows that religious texts can deliver inspiring material to measure such skills.

Let us also rate this skill on a 100–point scale by assuming a very simple scale such that the chatbots receive 25 scores for the satisfactory fulfillment of each of the following criteria: 1) does not exclude that there can be objects that satisfy the formula; 2) mentions some interpretational problem related to the existence of such objects; 3) clearly points out an ontological limitation connected with the formula; 4) explains the problem, e.g., by giving an example or pointing out the right ontological interpretation. ChatGPT fulfills (1), (2), (3) and (4), so: 100; Bing: (1) and (2), so: 50; Bard: (1), (2), (3) and (4), so: 100; Llama2: (1) and (2), so: 50.

## *Conclusions*

In this study, I have demonstrated how we can benefit from knowledge about religion and theology when testing LLM–powered chatbots. I attempted to show that they provide specific materials that can be hardly obtained from other branches of knowledge and, at the same time, can be useful to tell us more about the chatbots’ philosophical skills. I focused on four leading chatbots: ChatGPT, Bing, Bard, and Llama2. I selected three possible sources of inspiration from religion or theology: 1) the theory of the four senses of Scripture; 2) abstract theological statements; 3) an abstract formula from a religious text. They enabled me to measure the following three philosophical skills: interpretation of a given fragment, creative deductive reasoning, and identification of ontological limitations.

I utilized the theory of the four senses of Scripture to test if they can go beyond a basic interpretation and interpret a given fragment according to a given hermeneutic framework. In this case, the framework was the theory of the four senses. The test revealed that the chatbots can apply such a framework by producing accurate interpretations for each of the four senses.

Theological statements such as “God is light” and “God is a trinity of persons” were used to check if chatbots can create both creative and deductive argumentations for very abstract statements that employ an advanced ontological theory specific to the religious discourse. The task was inspired by Robert Grosseteste’s attempt to prove that God is a trinity of persons, starting from the premise that God is light. The chatbots were asked to create a reasoning with the same starting premise and the same conclusion. All the chatbots gave interesting answers by providing original additional premises to construct their argumentations. Notably, each of them added different premises. However, some of them had problems and modified the given conclusion. Only two of them indicated subsequent steps of their reasoning to mark its deductive nature. The same task based on more standard, philosophical statements appears easier for chatbots. Hence, the example

inspired by theological work was more demanding for the chatbots and enabled more sensitive testing.

The formula presented in the Gospel of John: “I am in the Father and the Father is in me” was used to check if the chatbots are able to identify the ontological limitations of given statements. It revealed that the four chatbots differ with respect to this skill. Only two of them performed the task very well. Furthermore, we should emphasize that it appears very difficult to find examples to measure such a skill. One reason is that in the case of purely philosophical statements that are at the same time very demanding (such as those related to the concept of existence), a risk of doctrinal bias occurs, such as the assumption that all objects exist. Thus, religious texts provide interesting, valuable, and ready-to-use material for measuring such philosophical skills.

Finally, I provided an example showing how we can create a benchmark to rate and compare the chatbots’ skills. In each case, I demonstrated the criteria that we can formulate to evaluate the three skills and proposed simplified scales to rate each chatbot with respect to each criterion. I compiled the results in Table 1. However, we should note that it is just an example of a method that can be applied, that the scale is simplified, and that the evaluation is based on my own assessment, so it should not be used as a final rating, which in professional evaluation includes an average rating from different experts. Furthermore, the sample for this study is limited. In order to receive more accurate results, we should test the chatbots on richer material. Nevertheless, I believe that the tasks and the materials presented in this article can be used in professional testing of the performance of LLMs or chatbots powered by LLMs. Despite the limitations concerning rating and the sample, I think that I succeeded in demonstrating that religion and theology can be useful for such testing.

The scope of this study is also limited concerning the number of skills that have been taken into account. Therefore, it cannot be perceived as exhaustive. It should be highlighted that it is aimed more at showing the potential of the presented method, which can be used in future research.

Lastly, in my view, this preliminary study reveals that it is worth analyzing LLM-powered chatbots in the religious and theological context, and thus, that further investigations would be valuable. The project mentioned in the introductory section assumes utilizing richer religious and theological material and testing other skills. I also plan research on ethical and social challenges concerning chatbots as potential actors in religious and theological discourse, including interreligious dialogue, texts interpretation, and consulting.



Table 1. Rating of philosophical skills tested on material inspired by religious/theological sources.

Tablica 1. Ocjena filozofskih vještina testiranih na sadržaju inspiriranom vjerskim/teološkim izvorima.

Skills	ChatGPT	Bing	Bard	Llama2
Interprets a passage by applying a given hermeneutic framework (here: according to the theory of the four senses of Scripture)	100%	100%	100%	100%
Creates a deductive reasoning for a given starting premise and conclusion (here: of a theological nature)	50%	100%	75%	75%
Identifies ontological limitations of a formula (here: specific to religious texts)	100%	50%	100%	50%
<b>Average</b>	<b>83%</b>	<b>83%</b>	<b>92%</b>	<b>75%</b>

Source: Own study.

Note: This rating is based on simplified scales and personal assessment. It is just an example of a possible evaluation of the results obtained. It should not be used as a final rating.

## References

- Aquinas, Thomas. 1920. *The Summa Theologiae of St. Thomas Aquinas*, Second and Revised Edition, transl. the Fathers of the English Dominican Province, after: Online Edition 2017, ed. K. Knight. URL: <https://www.newadvent.org/summa/>
- Bloom, Benjamin Samuel, Max D. Englehart, Edward J. Furst, Walker H. Hill, and David R. Krathwohl. 1956. *Taxonomy of Educational Objectives. The Classification of Educational Goals. Handbook 1. Cognitive Domain*. London: Longmans, Green and Co Ltd.
- Grosseteste, Robert. 1999. *On the Six Days of Creation*, ed. Ch. F. J. Martin. Auctores Britannici Medii Aevi 6(2). Oxford: Oxford University Press.
- Langton, Stephen. 1978. *Commentary on the Book of Chronicles*, ed. A. Saltman. Ramat-Gan: Bar-Ilan University Press.
- Manresa Lamarca, Ignacio M. 2017. "The Literal Sense and the Spiritual Understanding of Scripture According to St. Thomas Aquinas." *Biblica et Patristica Thoruniensia*, 3(10): 341–373. <https://doi.org/10.12775/BPTh.2017.018>.
- Parsons, Terence. 1980. *Nonexistent Objects*. Yale University Press.
- Roszak, Piotr. 2016. "Exegesis and Contemplation. The Literal and Spiritual Sense of Scripture in Aquinas Biblical Commentaries." *Espiritu* 65(152): 481–504.

- Srivastava, Aarohi, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown et al. 2022. Beyond the Imitation Game: Quantifying and Extrapolating the Capabilities of Language Models. *arXiv preprint arXiv:2206.04615*.
- Suzgun, Mirac, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery et al. 2023. “Challenging BIG–Bench Tasks and Whether Chain–of–Thought Can Solve Them.” In: *Findings of the Association for Computational Linguistics: ACL 2023*, Toronto, Canada. Association for Computational Linguistics, 13003–13051.
- Vawter, Bruce. 1964. “The Fuller Sense: Some Considerations.” *The Catholic Biblical Quarterly* 1(26), 85–96.
- Viviano, Pauline A. 2015. The Senses of Scripture. In: *National Bible Week 2015. The Bible: A Book for the Family*, United States Conference of Catholic Bishops, URL: <https://www.usccb.org/bible/national-bible-week/upload/viviano-senses-scripture.pdf>
- Wang, Alex, Singh Amanpreet, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019 (first published: 2018). GLUE: A Multi–Task Benchmark and Analysis Platform for Natural Language Understanding. *arXiv preprint arXiv:1804.07461*.

### *Sažetak*

---

## RELIGIJA, TEOLOGIJA I FILOZOFŠKE VJEŠTINE AUTOMATIZIRANIH PROGRAMA ZA ČAVRLJANJE (CHATBOTOVA) POGONJENIMA VELIKIM JEZIČNIM MODELIMA (LLM)

MARCIN TREPCZYŃSKI

U radu se nastoji prikazati kako se vjera i teologija mogu iskoristiti za testiranje uspješnosti velikih jezičnih modela (LLM–ova) i automatiziranih programa za čavrljanje (chatbotova) pogonjenima na takvim modelima, mjerenjem njihovih filozofskih vještina. Predstavljaju se rezultati testiranja četiriju odabranih chatbotova: ChatGPT, Bing, Bard i Llama2. Za potrebe testiranja uzeta su tri moguća izvora iz područja vjere i teologije: 1) teorija četiri smisla Svetog pisma, 2) apstraktne teološke izjave, 3) apstraktna logička formula izvedena iz vjerskog teksta kako bi se pokazalo da su ovi izvori korisni u zadacima koji učinkovito mjere filozofske vještine kao što su npr. interpretacija danog isječka, kreativna dedukcija te prepoznavanje ontoloških ograničenja. Ovakav pristup omogućuje ispitivanje osjetljivosti kojom se otkrivaju razlike u uspješnosti četiriju chatbotova. Također je dan primjer kako odrediti referentne vrijednosti za ocjenu i usporedbu vještina, koristeći kriterije za procjenu te pojednostavljene skale za ocjenjivanje svakog korištenog chatbota u pogledu ispunjenja pojedinog kriterija.

KLJUČNE RIJEČI: veliki jezični modeli, chatbotovi, testiranje, filozofske vještine, religija, teologija, interpretacija, zaključivanje

\* Doc. dr. sc. Marcin Trepczyński, Filozofski fakultet, Sveučilište u Varšavi, 3 Krakowskie Przedmieście St. Warsaw, 00–047 Poljska. E–adresa: [m.trepczynski@uw.edu.pl](mailto:m.trepczynski@uw.edu.pl)  
ORCID iD: <https://orcid.org/0000-0003-0612-2597>