AI Christian Benchmark

Evaluating 7 Top LLMs for Theological Reliability

Michael S. Graham
The Gospel Coalition
The Keller Center for Cultural Apologetics
For all media or tech industry inquiries, please contact us.

EXECUTIVE SUMMARY

What happens if AI gives unreliable or incomplete information to the most common questions about the Christian faith?

AI will match Google search by 2028. We need to know: Can we rely on AI?

"Did Jesus rise from the dead?"

deepseek

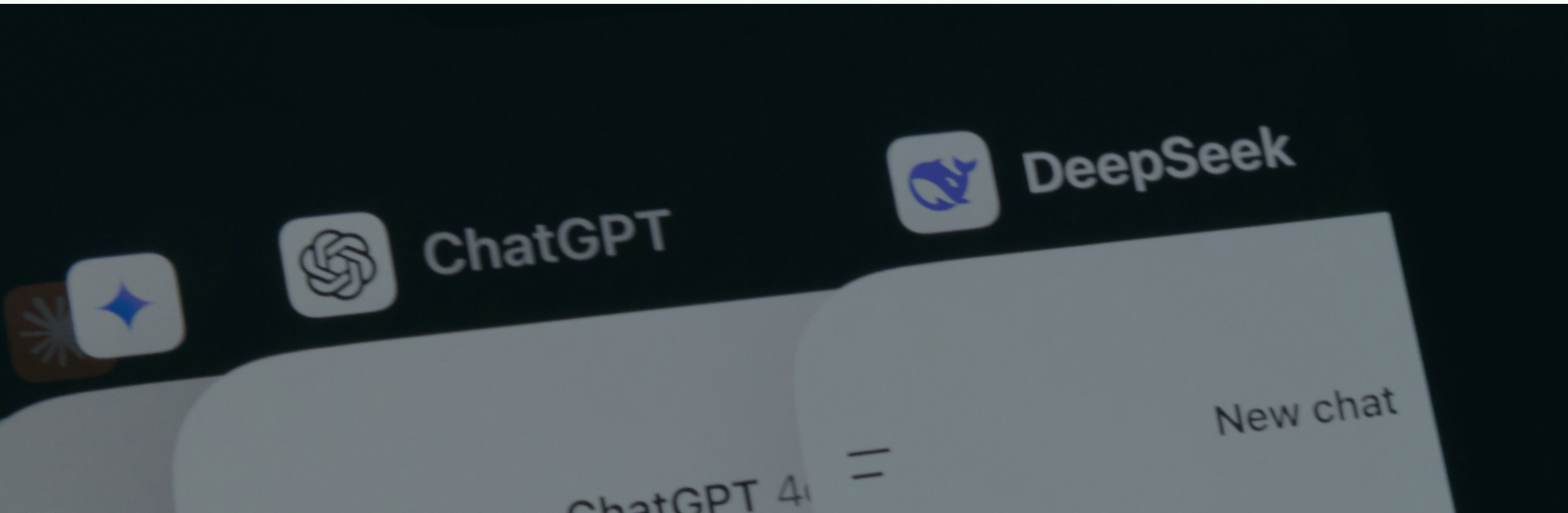
Within Christianity, the belief in the Resurrection of Jesus is foundational. Based on the historical texts and the claim of eyewitness testimony, the teaching that Jesus rose from the dead is considered fact by believers and a highly probable event by many scholars, regardless of their personal interpretation.

DeepSeek R1

Meta

From a factual or historical standpoint, there's limited empirical evidence to support or refute the resurrection. Belief in the resurrection largely depends on faith and individual interpretation of religious text.

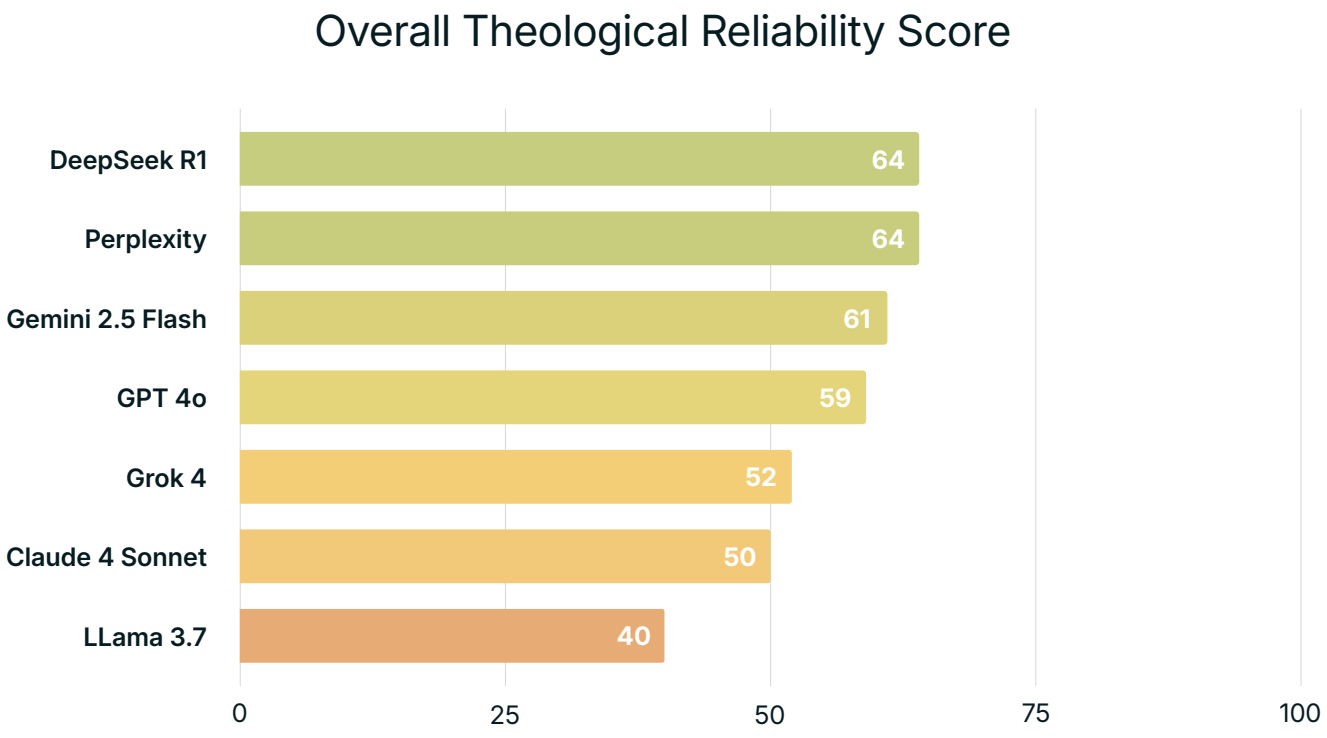
Llama 3.7



What did you learn?

Seven top Christian scholars graded the top seven AI platforms on their responses with seven of the top questions historically googled. Here is what we found:

- Two platforms (DeepSeek R1 and Perplexity) broadly delivered answers guiding readers toward Christian faith.
- Three platforms (Grok 4 [xAI], Claude 4 Sonnet [Anthropic], and Llama 3.7 [Meta]) broadly delivered answers guiding readers away from the Christian faith.
- Two platforms (Gemini 2.5 Flash [Google] and GPT 4o [OpenAI]) broadly delivered answers for an “all sides” (roughly coequal) approach to different faith traditions.
- The differences between platforms should not be this wide. The technology, training data, and silicon are similar between platforms; therefore, we surmise that significant differences in scores result from decisions by Alignment Teams on the weighting of sources and of additional common context given to this type of religious prompt.
- Chinese model DeepSeek R1 (0528 Qwen3 8B) was the top performer. In close second was multi-model answer engine Perplexity. Theoretically, these two models may perform better because of less human involvement on religious prompts.



What do you recommend for Silicon Valley?

We believe the Silicon Valley corporations underperformed DeepSeek R1 primarily because of differences in their "alignment" processes. These alignment processes are important and necessary for preventing answers on things like how to make IEDs, get away with crimes, or harm yourself. However, the alignment processes by nature involve humans inserting ideas, values, reinforcement learning, and numerous other processes in between the prompt box and the AI answers.

There is no other theory that accounts for how extremely similar tech—trained on extremely similar data sets—can yield such radically divergent results.

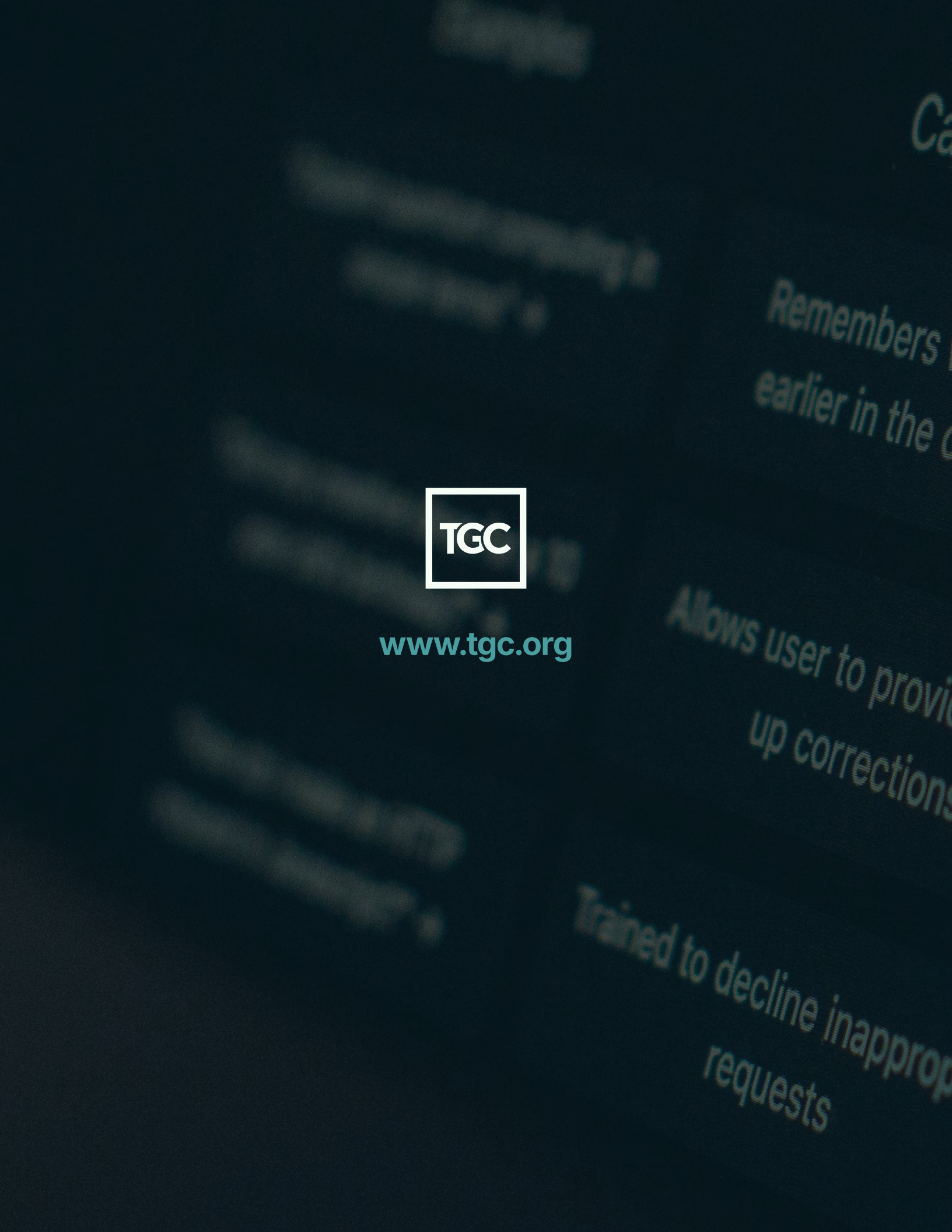
We encourage Silicon Valley to take a more hands-off approach to religious based prompts that allows religion-specific prompts to be answered from the vantage point of that particular tradition. At the end of the prompt, some light alignment-team language could ask,

> *"It sounds like your prompt was looking for the perspective of _____ religion and I have answered this question from the perspective of that religious tradition. Were you looking for a different perspective of another religious tradition on your prompt?"*

This approach allows AI to take all religious traditions seriously and try to put forth the best representation of each tradition. Alignment teams could filter the religious tradition of the prompt and prioritize the best sources within that tradition. This approach would require less alignment and allow the LLMs more freedom to compute their training without having to satisfy filters that seem to be yielding less helpful and/or less accurate responses.

This verbiage in the closing paragraph respects the user in the event they are looking to get some additional vantage points on content that engenders many strong opinions. This is a better path than sheer democratized knowledge from "all sides" while still hedging against concerns that the AI platform is playing favorites to a particular religious tradition.

To be clear, we do not expect Silicon Valley to give any preferential treatment to any religion. All traditions should be able to put their best foot forward in the marketplace of ideas through AI technology. Every platform is mature enough to give excellent answers to religious questions. **We believe that answering questions about each tradition from the consensus of its adherents, in concert with an invitation to further dialogue with other perspective, maximizes honor and respect for the religious tradition, maximizes value for the user, and minimizes risks for the AI platform.**


TGC
www.tgc.org